Statistica Applicata
Corso di Laurea in Scienze Naturali
a. a. 2018/2019

prof. Federico Plazzi

28 Giugno 2019

Nome: _____

Cognome: _____

Matricola: _____

Alcune indicazioni:

- La prova è costituita 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

La teballa 1 mostra le altezze in centimetri e i risultati in trentesimi di 60 studenti del secondo anno di Scienze Naturali nell'esame di Zoologia Sistematica. Gli studenti sono divisi per sesso e in base alla prima lettera del cognome (A-L o M-Z).

Tabella 1: Altezze degli studenti (cm) e relativi risultati in Zoologia Sistematica.

Maschi				Femmine			
A-L		M-Z		A-L		M-Z	
Altezza	Voto	Altezza	Voto	Altezza	Voto	Altezza	Voto
187	23	171	26	178	29	174	28
174	29	189	28	181	29	162	28
175	24	188	28	167	26	170	29
176	28	175	26	172	28	169	27
181	23	183	22	177	28	166	29
182	26	177	26	166	27	169	29
191	22	181	26	164	26	173	27
178	24	182	27	162	27	175	29
177	26	185	24	166	29	166	28
187	25	182	27	166	28	178	29
176	29	184	24	163	25	166	27
180	25	179	25	178	27	173	27
181	28	177	28	165	30L	169	26
171	24	190	25	171	29	168	29
181	26	177	24	177	28	167	25

2 Esercizi

2.1 Statistiche di base

Calcola media, devianza, varianza e deviazione standard dell'altezza delle femmine del gruppo A-L.

Media	*Devianza*	*Varianza*	*Deviazione standard*
$170,2$	$582,4$	$38,83$	$6,23$

2.2 Distribuzione dei dati

La figura 1 mostra il Q-Q plot relativo a tutte e 60 le altezze e a tutti e 60 i voti. Di seguito sono mostrati i risultati del test di Shapiro e Wilk condotti sugli stessi due insiemi di dati. Cosa possiamo concludere sulla distribuzione dei dati? Motiva la tua risposta.

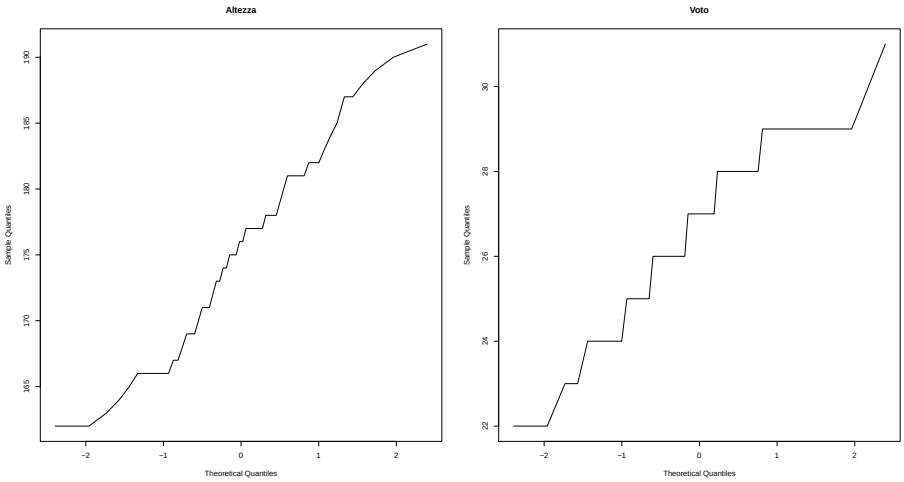

Figura 1: Q-Q plot di altezze (*a sinistra*) e voti (*a destra*).

- Test di Shapiro e Wilk eseguito sulle altezze:

```
Shapiro-Wilk normality test
data:  altezza
W = 0.97179, p-value = 0.1785
```

- Test di Shapiro e Wilk eseguito sui voti:

```
Shapiro-Wilk normality test
data:  voto
W = 0.94385, p-value = 0.008086
```

L'ipotesi nulla del test di Shapiro e Wilk è la normalità della distribuzione dei dati. Il p-value del test eseguito sulle altezze non è significativo ($0,1785$*), per cui non possiamo rifiutare l'ipotesi nulla di normalità: i punti del relativo Q-Q plot, coerentemente, si dispongono approssimativamente lungo una linea. Al contrario, i punti del Q-Q plot relativo ai voti non si dispongono bene lungo una linea e il p-value del test di Shapiro e Wilk è significativo (*$0,0081$*), per cui possiamo rifiutare l'ipotesi nulla di normalità per questi dati: è possibile che ciò sia legato al fatto al fatto che i voti sono prossimi a un loro valore massimo (*30*).*

2.3 Altezze ed esiti a confronto

Per verificare se ci siano differenze significative in altezza e in rendimento tra maschi e femmine, vengono eseguiti diversi test statistici tra i 30 maschi e le 30 femmine. I risultati sono mostrati di seguito; qual è il test corretto? Perché? Cosa possiamo concludere?

Tabella 2: Risultati di diversi test statistici nel confronto tra maschi e femmine. Tutti i test sono eseguiti a due code.

Test	Altezza	Voto
Test t a campioni appaiati	$P = 1,191 \times 10^{-7}$	$P = 0,8522$
Test t a campioni indipendenti	$P = 2,254 \times 10^{-10}$	$P = 6,798 \times 10^{-2}$
Test di Wilcoxon	$P = 8,699 \times 10^{-2}$	$P = 0,000142$
Test di Mann e Whitney	$P = 2,254 \times 10^{-8}$	$P = 3,024 \times 10^{-5}$

La scelta del test più corretto avviene in funzione della distribuzione della variabile: la variabile "altezza" è distribuita in modo normale, per cui si può usare il test t*, parametrico; la variabile "voto" non è a distribuzione normale, per cui bisogna usare un test non parametrico.*

In entrambi i casi il test è a campioni indipendenti (non c'è ragione di confrontare il primo maschio con la prima femmina, il secondo con la seconda e così via), per cui i test corretti sono il test t a campioni indipendenti per l'altezza ($P = 2,254 \times 10^{-10}$) e il test di Mann e Whitney per il voto ($P = 3,024 \times 10^{-5}$). In entrambi i casi il p-value è inferiore al 5%, per cui rifiutiamo l'ipotesi nulla di appartenenza alla stessa popolazione e diciamo che le differenze sia in altezza sia in voto dei due gruppi sono significative.

2.4 One-Way ANOVA

Cercando di capire se esistano differenze tra i quattro gruppi in termini di rendimento, si confrontano separatamente i voti di maschie e femmine suddivisi ulteriormente in base ai cognomi (A-L o M-Z). L'approccio scelto è l'ANOVA, i cui risultati preliminari sono mostrati in tabella 3.

1. Per quale motivo, nonostante tutto, l'ANOVA è un approccio possibile in questo caso?
2. Completa la tabella 3 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?

Tabella 3: One-Way ANOVA. D, devianza; σ^2, varianza; g.l., gradi di libertà.

	D	σ^2	g.l.	F	p-value
entro	167, 467	2, 9905	56		
tra	73, 133	24, 3778	3	8, 1518	0, 0001359

1. *L'ANOVA è indicata perché, sebbene la variabile non sia a distribuzione normale, i campioni sono poi tutti della stessa dimensione e hanno variabilità comparabili (*omoschedasticità*).*

2. *I gradi di libertà sono 3 per la varianza tra gruppi (perché i gruppi in tutto sono 4) e 56 per la varianza entro gruppi (perché ci sono 60 osservazioni e 4 gruppi). Le varianze si ottengono dividendo le rispettive devianze per i gradi di libertà.*

3. *Il valore di F si ottiene dividendo la varianza* tra *gruppi per la varianza* entro *gruppi. Consultando le tabelle allegate si vede che F è altamente significativo, sia considerando 50 gradi di libertà per la varianza entro sia considerandone 60: il valore esatto è quello tabulato qui sopra. Possiamo quindi concludere che esiste una differenza nella variabilità nucleotidica della quattro regioni considerate.*

4. *Il passo successivo potrebbe essere il test di Tukey per verificare se alcuni gruppi si discostano significativamente dagli altri.*

2.5 Le persone più basse hanno risultati migliori?!

Analizzando la possibile correlazione tra tutti e 60 i voti e tutte e 60 le altezze, si ottiene il seguente modello di correlazione lineare. Stranamente, i dati porterebbero a vedere una correlazione negativa tra l'altezza in centimetri e il voto in Zoologia Sistematica: è vero? Commenta i dati mostrati in tabella 4.

Tabella 4: Modello di correlazione lineare tra altezza e voto in Zoologia.

	Stima	p-value	r	R^2
Intercetta	45, 64805			
Pendenza	−0, 10812	0, 00134	−0, 40472	0, 1638

La correlazione negativa ($r < 0$) è in effetti significativa ($P = 0, 00134$), ancorché molto debole ($R^2 = 0, 1638$). La significatività di una correlazione, tuttavia, non implica necessariamente un rapporto di causa-effetto tra le due variabili in gioco, anche perché la correlazione è un concetto simmetrico: è la maggiore statura che diminuisce le prestazioni in campo zoologico o sono le buone prestazioni in campo zoologico a rimpicciolire la statura?

Di fatto, bisogna individuare una spiegazione logica per questo comportamento, che sta probabilmente nel fatto che le ragazze, che hanno voti tendenzialmente migliori dei ragazzi, hanno anche stature inferiori (vedi esercizio 2): questo probabilmente è alla base di questo modello di correlazione lineare.