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

prof. Federico Plazzi

19 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

I genomi mitocondriali completi di cinque specie contengono quattro regioni: i tratti che contengono i geni per le proteine, quelli che contengono i geni per gli RNA ribosomali (rRNA), quelli che contengono i geni per gli RNA di trasporto (tRNA) e i tratti rimanenti (Regioni Non Codificanti, RNC). Nella Tabella 1 è indicato il numero di A, C, G e T che si trovano in ciascuna delle quattro regioni, specie per specie. Il contenuto in A+T è la proporzione, regione per regione, di A e T sul totale di basi.

Tabella 1: Composizione nucleotidica di cinque genomi mitocondriali.

Specie	Regione	A	C	G	T	A+T
1	Proteine	2701	2730	3045	2724	0,484
1	rRNA	456	311	367	466	0,576
1	tRNA	199	130	161	310	0,636
1	RNC	850	263	345	942	0.747
2	Proteine	3206	3073	3094	2877	0,497
2	rRNA	459	378	417	496	0,546
2	tRNA	259	167	232	217	0,544
2	RNC	701	425	560	939	0.625
3	Proteine	2886	2994	2940	2730	0,486
3	rRNA	447	380	345	478	0,561
3	tRNA	268	178	147	232	0,606
3	RNC	1088	323	336	728	0,734
4	Proteine	3187	3217	3217	2979	0,489
4	rRNA	500	358	396	546	0,581
4	tRNA	253	162	175	310	0,626
4	RNC	891	583	371	855	0,647
5	Proteine	3262	3243	3114	2806	0,488
5	rRNA	455	429	372	519	0,549
5	tRNA	261	219	183	225	0,547
5	RNC	831	533	487	811	0,617

2 Esercizi

2.1 Statistiche di base

Calcolare media, varianza e deviazione standard tra le cinque specie del numero di A nelle regioni che codificano per proteine e nelle RNC.

	Media	*Varianza*	*Deviazione standard*
Proteine	$3048,40$	$47346,64$	$217,59$
RNC	$872,20$	$15684,56$	$125,24$

2.2 Distribuzione dei risultati

Viene eseguito un test di Shapiro e Wilk sulla colonna A+T e si ottengono i seguenti risultati.

```
Shapiro-Wilk test

data:  A+T
W = 0.92086, p-value = 0.1029
```

La Figura 1 mostra il Q-Q Plot riferito alla stessa variabile. Se si andasse a verificare la correlazione lineare tra i punti del grafico, quale delle tre tabelle seguenti potrebbe esserne il risultato? Perché?

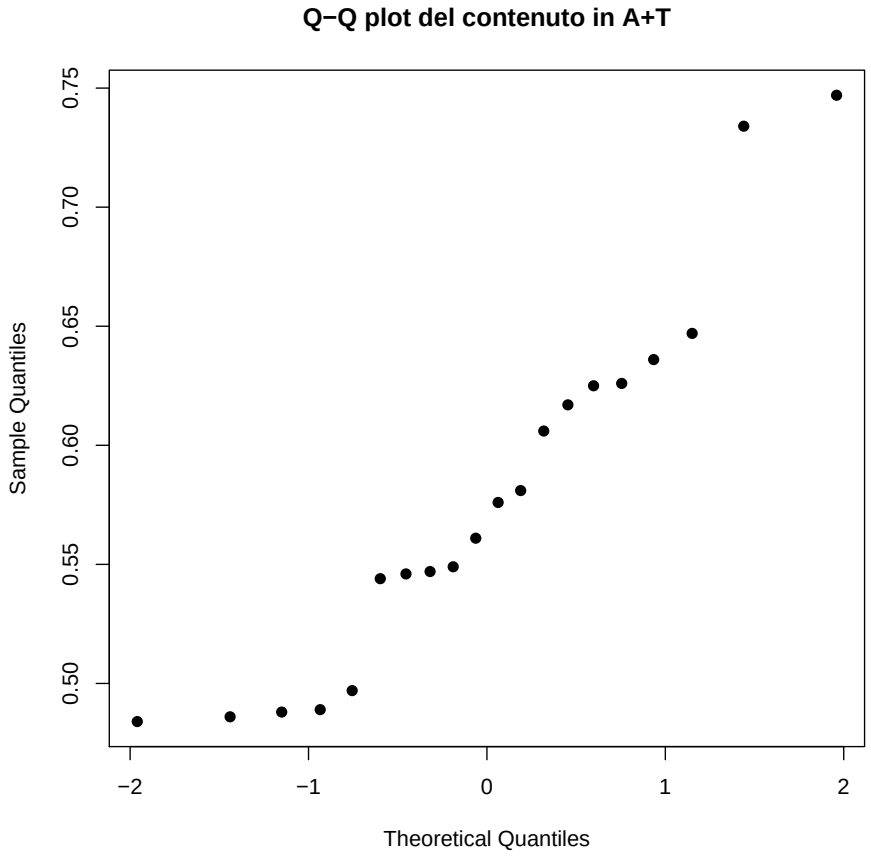

Figura 1: Q-Q Plot della variabile A+T

Tabella 2: Correlazione lineare tra i punti del Q-Q plot in Figura 1.

	Stima	p-value	r	R^2
Intercetta	$0,58$			
Pendenza	$0,07$	$9,92 \times 10^{-12}$	$0,96$	$0,93$

Tabella 3: Correlazione lineare tra i punti del Q-Q plot in Figura 1.

	Stima	p-value	r	R^2
Intercetta	$0,56$			
Pendenza	$-0,1$	$5,51 \times 10^{-5}$	$-0,91$	$0,83$

Tabella 4: Correlazione lineare tra i punti del Q-Q plot in Figura 1.

	Stima	p-value	r	R^2
Intercetta	$0,59$			
Pendenza	$0,02$	$6,14 \times 10^{-1}$	$0,03$	$0,02$

La tabella corretta è la prima: il test di Shapiro e Wilk indica che la variabile ha una distribuzione normale (p-value $> 0,05$) e quindi ci aspettiamo che i punti del Q-Q Plot si dispongano lungo una retta a pendenza positiva, come infatti si può vedere in Figura 1: la seconda tabella indica però una correlazione negativa ($r < 0$) e la terza un'assenza di correlazione (p-value $> 0,05$). Nella terza tabella, tra l'altro, R^2 non è il quadrato di r.

2.3 Composizione nucleotidica

Gli studiosi sono interessati a capire se la composizione nucleotidica (numero di A, C, G e T) nelle regioni che codificano per le proteine sia diversa nella specie 2 rispetto alla specie 3. Per farlo, vengono usati tre diversi test: un test t a campioni appaiati, un test di Wilcoxon e un test del χ^2, i cui risultati sono mostrati di seguito. Qual è l'approccio corretto? Perché? Cosa possiamo concludere in base al risultato?

1. Test t a campioni appaiati:

```
Paired t-test

data:  Proteine (Specie 2) and Proteine (Specie 3)
t = 31.641, df = 3, p-value = 6.937e-05
```

2. Test di Wilcoxon a campioni appaiati

```
Wilcoxon signed rank test

data:  Proteine (Specie 2) and Proteine (Specie 3)
V = 10, p-value = 0.125
```

3. Test del χ^2:

```
Chi-squared test

data:  Proteine (Specie 2) and Proteine (Specie 3)
X-squared = 5.0381, df = 3, p-value = 0.169
```

L'approccio corretto è il test del χ^2: contare quante basi, nelle regioni date, sono A, quante sono C, quante G e quante T equivale infatti ad usare quattro variabili qualitative e chiedersi quante basi sono da registrare sotto ciascuna di esse. Il p-value del test del χ^2 è maggiore di $0,05$, per cui il test non è significativo: la specie 2 e la specie 3 non hanno una composizione nucleotidica significativamente differente nelle regioni che codificano per le proteine.

2.4 One-Way ANOVA

Esistono differenze significative tra i valori di A+T calcolati nelle quattro diverse regioni genomiche (proteine, rRNA, tRNA e RNC)? A questo scopo, si calcola la devianza *entro* gruppi e la devianza *tra* gruppi, usando come gruppi i quattro tipi di regioni genomiche. I risultati sono trascritti in Tabella 5.

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

	D	σ^2	g.l.	F	p-value
tra	$0,0879674$	$0,02932247$	3		
entro	$0,0240048$	$0,0015003$	16	$19,5444$	$< 0,001$

1. Per quale motivo l'ANOVA è un approccio valido in questo caso?
2. Completa la tabella 5 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 3 per la varianza tra gruppi (perché i gruppi in tutto sono 4) e 16 per la varianza entro gruppi (perché ci sono 20 osservazioni e 4 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. Consultando le tabelle allegate si vede che F è altamente significativo: il valore esatto sarebbe* $1,34 \times 10^{-5}$.
Possiamo quindi concludere che esiste una differenza nel contenuto in A+T tra le quattro regioni genomiche. Il passo successivo potrebbe essere il test di Tukey per verificare se alcuni gruppi si discostano significativamente dagli altri.