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

prof. Federico Plazzi

30 Luglio 2018

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

Si sa che il genoma mitocondriale di una certa specie di nematodi è sottoposto a complicati vincoli selettivi. Per indagare, si scelgono 16 regioni a caso del genoma per 4 tipologie: regioni che codificano per proteine ("PCGs"), regioni che codificano per rRNA ("rRNAs"), regioni che codificano per tRNA ("tRNAs") e regioni non assegnate, prive di funzione nota ("URs"). Per ognuna di queste regioni del genoma mitocondriale si calcola la diversità nucleotidica su un ampio numero di individui (tabella 1): la diversità nucleotidica indica la variabilità di una certa zona del genoma e va da 0 (tutti gli individui hanno la stessa sequenza in una certa regione) a 1 (ogni individuo ha una propria variante in una certa regione).

Tabella 1: Diversità nucleotidica di 4 tipi di regioni del genoma mitocondriale.

PCGs	rRNAs	tRNAs	URs
0.4644	0.1523	0.4022	0.5540
0.4535	0.3979	0.2162	0.6928
0.4920	0.3414	0.4106	0.8221
0.4926	0.3215	0.3315	0.5673
0.4462	0.3550	0.3219	0.8985
0.4230	0.2318	0.3220	0.5851
0.5257	0.3189	0.3229	0.6232
0.5243	0.1922	0.2229	0.7150
0.4291	0.3977	0.2541	0.6643
0.4193	0.2772	0.1212	0.7714
0.5170	0.0000	0.0001	0.8250
0.4853	0.0000	0.1181	0.7736
0.4893	0.2238	0.2481	0.8886
0.1356	0.0425	0.0000	0.4034
0.3603	0.3552	0.0000	0.5173
0.0000	0.3860	0.0000	0.0000

2 Esercizi

2.1 Statistiche di base

Calcolare media, varianza e deviazione standard della variabilità nucleotidica delle regioni che codificano per una proteina (PCG).

Media	*Varianza*	*Deviazione standard*
$0,4161$	$0,0197$	$0,1404$

2.2 Distribuzione dei dati

Prima di ulteriori analisi, è necessario indagare la distribuzione dei dati. Viene eseguito il test di Shapiro e Wilk su tutti i 64 valori di diversità nucleotidica messi insieme e questo è il risultato:

```
Shapiro-Wilk test

data:  pi.overall
W = 0.96445, p-value = 0.06213
```

Cosa si può concludere?

Dato che il valore di P ($0,06213$*) è superiore a* $0,05$ *non è possibile rifiutare l'ipotesi nulla di normalità.*

2.3 Differenze in diversità nucleotidica tra le regioni

L'ipotesi preliminare è che le regioni senza funzione (UR) nota siano più libere di variare delle altre, perché sottoposte a vincoli selettivi più rilassati. D'altro canto, tra gli altri tre tipi di regioni, quelle che codificano per una proteina (PCG) dovrebbero poter essere più variabili delle altre, perché grazie alla degenerazione del codice sono possibili mutazioni sinonime e quindi silenti.
Si scartano le regioni codificanti per i tRNA perché sono molto poche nel genoma e si dicide di testare questa ipotesi preliminare con un test t a campioni appaiati e non appaiati, confrontando prima UR e PCG e poi PCG e regioni che codificano per gli rRNA.

Tabella 2: Risultati del test t a coppie.

	URs-PCGs	PCGs-rRNAs
Campioni appaiati	$7,2062$	$3,1779$
Campioni non appaiati	$3,44$	$3,3285$

La tabella 2 mostra i valori ottenuti per la statistica t. Discuti l'uso dei test t in questa situazione. È corretto? Quale? Perché? Cosa si potrebbe concludere?

L'uso del test t *in questa situazione non è consigliabile, perché il test non è in grado di correggere per campioni multipli: qui ci sono tre campioni (se non quattro, considerando le regioni che codificano per i tRNA che sono state momentaneamente escluse). L'ANOVA, introdotta nell'esercizio seguente, è sicuramente un approccio preferibile.*
Ad ogni modo, se si vogliono indagare i dati in modo preliminare con confronti a coppie, l'esercizio precedente conferma la distribuzione normale della variabile e l'approccio corretto è quello a campioni non appaiati (non c'è ragione di confrontare il primo dato del primo campione con il primo dato del secondo e

così via). Consultando le tabelle, si vede che il valore soglia di t *per 15 gradi di libertà (*$16-1$*) al 5% è* $2,131$*, per cui entrambi i confronti a coppie evidenziano valori di* t *significativi (*$3,44$ *e* $3,3285$*) e l'ipotesi iniziale risulta verificata in via preliminare.*

2.4 One-Way ANOVA

Nella tabella 3 viene impostata una One-Way ANOVA per indagare la diversità nucleotidica nei 4 tipi di regioni considerati.

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

	D	σ^2	g.l.	F	p-value
tra	$1,8926$	$0,63088$	3		
entro	$1,6629$	$0,02772$	60	$22,763$	$< 0,001$

1. Per quale motivo l'ANOVA è un approccio valido 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?

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 60 per la varianza entro gruppi (perché ci sono 16 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, perché il valore soglia è di* $2,76$*: il valore esatto sarebbe* $5,813 \times 10^{-10}$*.*
Possiamo quindi concludere che esiste una differenza in diversità nucleotidica tra le quattro regioni genomiche. Il passo successivo potrebbe essere il test di Tukey per verificare se alcuni gruppi si discostano significativamente dagli altri.