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

prof. Federico Plazzi

20 Febbraio 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

La selezione naturale agisce di solito sulle sequenze di DNA in modo da conservare le regioni che svolgono una funzione in qualche modo adattativa, mentre l'accumulo di mutazioni è normalmente più comune in regioni con meno vincoli funzionali. In molti casi, si può cercare di stabilire se una certa regione abbia o meno un ruolo preciso all'interno del genoma proprio misurando la variabilità di quella regione tra diversi individui di una stessa popolazione.
Alcuni biologi molecolari sono interessati ad un certo gene e, per scoprire eventuali regioni connesse alla sua regolazione ed espressione, sequenziano, in un grande numero di individui di una stessa popolazione, le 400 paia di basi immediatamente prima dell'inizio del messaggero di questo gene. Le 400 paia di basi sequenziate vengono poi divise in 4 regioni da 100 nucleotidi ciascuna; ogni regione viene a sua volta divisa in 10 "finestre" consecutive da 10 paia di basi ciascuna e, per ciascuna "finestra", viene misurata la diversità nucleotidica (indicata con "Pi" o con "π") osservata tra tutti gli individui: la diversità nucleotidica indica la frequenza di mutazioni in quella "finestra", in modo tale che un valore pari a 0 indica che tutti gli individui presentano esattamente le stesse 10 paia di basi in quella "finestra", mentre valori sempre più alti indicano una sempre maggior variabilità (e sono quindi indizio di sempre minori vincoli funzionali). I risultati sono mostrati in tabella 1 e in figura 1.

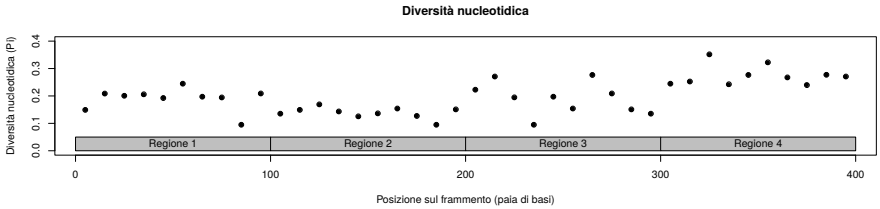

Figura 1: Diversità nucleotidica lungo le 400 paia di basi di interesse.

Tabella 1: Diversità nucleotidica (π) in "finestre" da 10 nucleotidi.

"Finestra"	Regione 1	Regione 2	Regione 3	Regione 4
1	0,1490	0,1353	0,2228	0,2447
2	0,2087	0,1490	0,2709	0,2527
3	0,2006	0,1693	0,1949	0,3519
4	0,2061	0,1435	0,0949	0,2427
5	0,1921	0,1256	0,1973	0,2764
6	0,2447	0,1365	0,1542	0,3227
7	0,1973	0,1542	0,2764	0,2677
8	0,1946	0,1268	0,2091	0,2393
9	0,0949	0,0949	0,1510	0,2772
10	0,2091	0,1510	0,1353	0,2709

2 Esercizi

2.1 Statistiche di base

Calcolare media, devianza, varianza e deviazione standard della diversità nucleotidica nella regione 4.

Media	*Devianza*	*Varianza*	*Deviazione standard*
$0,2746$	$0,0120$	$0,0012$	$0,0346$

2.2 Distribuzione dei dati

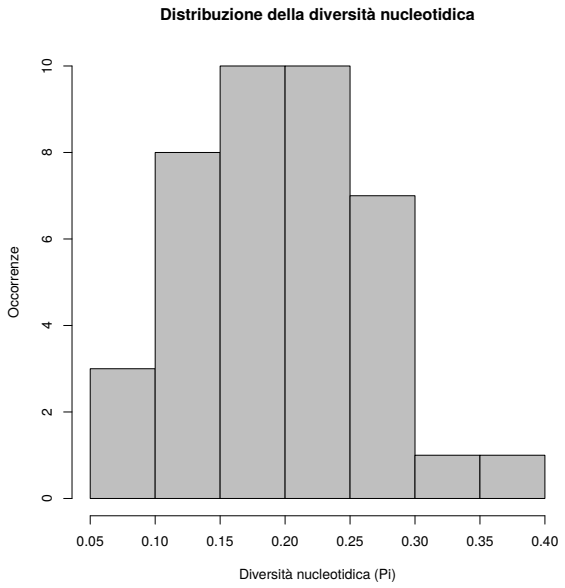

Figura 2: Distribuzione della diversità nucleotidica.

Giudicando dalla precedente figura 1 e dalla figura 2 in questa pagina, che mostra la distribuzione della diversità nucleotidica (π), quale dei seguenti potrebbe essere un verosimile risultato del test di Shapiro e Wilk condotto su tutti e 40 i valori di π? Perché? Che cosa indicherebbe?

- `Shapiro-Wilk test`

  ```
  data:  Pi
  W = 0.96557, p-value = 0.2583
  ```

- `Shapiro-Wilk test`

  ```
  data:  Pi
  W = 0.98951, p-value = 1.26e-15
  ```

- `Shapiro-Wilk test`

  ```
  data:  Pi
  W = 0.48364, p-value = 0.0451
  ```

Il risultato più verosimile è il primo, che ha restituito un p-value non significativo ($0,2583$): la distribuzione appare infatti abbastanza vicino alla normale nelle due figure indicate. Il secondo test, inoltre, indica un p-value molto molto basso a fronte di un W molto vicino a 1, il che è quantomeno insolito, mentre il terzo test, a fronte di un p-value significativo, ma vicinissimo a $0,05$, presenta un W molto lontano da 1, che è improbabile per una distribuzione come quella in figura 2.

2.3 Cambiamento della diversità nucleotidica (π)

Per indagare se i vincoli funzionali aumentino o diminuiscano regolarmente man mano che ci si avvicina al messaggero, si calcola un modello di correlazione lineare tra la posizione sul frammento in questione (figura 1) e π. Cosa possiamo concludere dai dati mostrati in tabella 2?

Tabella 2: Correlazione lineare tra posizione sul DNA e π.

	Stima	p-value	r	R^2
Intercetta	$1,441 \times 10^{-1}$			
Pendenza	$2,713 \times 10^{-4}$	$9,831 \times 10^{-4}$	$0,5013$	$0,2513$

Esiste una correlazione significativa (il p-value è infatti minore di $0,05$) e positiva (il coefficiente r è positivo) tra le due variabili. Tale correlazione, tuttavia, non è fortissima.

2.4 One-Way ANOVA

Per concludere lo studio, bisogna capire esattamente se esistano differenze in diversità nucleotidica tra le quattro regioni mostrate in figura 1. Si calcola perciò la devianza *entro* gruppi e la devianza *tra* gruppi, usando come gruppi le quattro regioni identificate. I risultati sono trascritti in Tabella 3.

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?

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

	D	σ^2	g.l.	F	p-value
tra	$0,095172$	$0,031724$	3		
entro	$0,060925$	$0,001692$	36	$18,745$	$1,713 \times 10^{-7}$

1. *L'ANOVA è indicata perché la variabile è a distribuzione normale; i campioni sono poi tutti della stessa dimensione.*

2. *I gradi di libertà sono 3 per la varianza tra gruppi (perché i gruppi in tutto sono 4) e 36 per la varianza entro gruppi (perché ci sono 40 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: 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.*