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

prof. Federico Plazzi

13 Febbraio 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

Di una certa proteina enzimatica interessa in particolare il *sito attivo*, ossia quella parte che prende parte direttamente alla reazione chimica catalizzata da quell'enzima; si sa che esso comprende gli aminoacidi che vanno da 194 a 208 sulla sequenza aminoacidica. Per compiere uno studio comparativo su questa proteina, vengono analizzate le sequenze da 98 specie e per ognuna di queste specie gli aminoacidi[1] ai siti da 194 a 208 vengono classificati come acidi (D, E), basici (H, K, R), polari (C, G, N, Q, S, T, Y) o idrofobici (A, F, I, L, M, P, V, W). I risultati sono mostrati in Tabella 1.

Tabella 1: Aminoacidi nel sito attivo della proteina su 98 sequenze analizzate, ripartiti per categorie di idropatia.

Sito	Acidi	Basici	Polari	Idrofobici
194	0	0	23	75
195	0	0	13	85
196	0	0	6	92
197	0	97	1	0
198	0	0	0	98
199	0	0	9	89
200	0	0	14	84
201	0	0	1	97
202	0	97	0	1
203	70	0	15	13
204	1	17	76	4
205	0	0	98	0
206	0	0	95	3
207	0	0	98	0
208	0	0	98	0

[1] Le sigle che seguono sfruttano il codice IUPAC ad una lettera per indicare gli aminoacidi: ad esempio, "A" indica l'alanina, "C" la cisteina e così via.

2 Esercizi

2.1 Statistiche di base

Qual è il numero medio di aminoacidi polari per sito? E di aminoacidi idrofobici? Qual è la deviazione standard del numero medio di aminoacidi polari per sito? Qual è la deviazione standard del numero medio di aminoacidi idrofobici per sito? Come si possono commentare questi dati?

```
mean(polari)
36.47
> mean(idrofobici)
42.73
> standard.deviation(polari)
40.74615
> standard.deviation(idrofobici)
43.28736
```

Entrambe le deviazioni standard sono molto grandi rispetto alle media, addirittura più delle medie stesse. Questo indica che il numero medio di aminoacidi polari/idrofobici per sito non ha un grande valore descrittivo.

2.2 Distribuzione di aminoacidi polari ed idrofobici

La presenza di un aminoacido polare o di uno idrofobico ha importanti ripercussioni sulla struttura tridimensionale della proteina, per cui è importante capire se queste categorie di aminoacidi siano distribuite a caso o meno. Di seguito vengono eseguiti tre test per stabilire se la distribuzione degli aminoacidi polari o idrofobici sia significativamente diversa tra siti: un test t a due campioni appaiati, un test del χ^2 e una Two-Way ANOVA. Qual è l'approccio più corretto? Cosa possiamo concludere?

1. Test t a due campioni appaiati

   ```
   Paired t-test

   data:  polari and idrofobici
   t = -0.30739, df = 14, p-value = 0.7631
   alternative hypothesis:  true difference in means is not equal
   to 0
   ```

2. Test del χ^2

   ```
   Pearson's Chi-squared test

   data:  polari and idrofobici
   X-squared = 908.87, df = 14, p-value < 2.2e-16
   ```

Tabella 2: Risultati della Two-Way ANOVA.

	g.l.	D	σ^2	F	p-value
Sito	14	27599	$1971,35$	$1,1425$	$0,3246$
Idropatia	1	0	$0,01$	$0,0000$	$0,9986$
Interazione	14	28284	$2020,25$	$1,1708$	$0,3018$
entro	14	293335	$1725,50$		

Il test corretto è il χ^2 perché si tratta di una variabile qualitativa (categorica). Il test è significativo, per cui possiamo rifiutare l'ipotesi nulla: ci sono delle differenze tra siti nell'idropatia degli aminoacidi.

2.3 Regioni del sito attivo

Si può dire, usando un modello lineare, che c'è correlazione negativa tra l'uso di un aminoacido polare e l'uso di un aminoacido idrofobico? In altre parole, possiamo di dire su questa base che, se in un certo sito ci sono aminoacidi polari in molte sequenze, saranno poche quelle con aminoacidi idrofobici? **L'uso di una correlazione lineare non è molto corretto, perché si tratta di una variabile categorica la cui distribuzione normale, oltretutto, non è nemmeno stata testata.**

Tabella 3: Correlazione tra uso di aminoacidi polari e uso di aminoacidi idrofobici.

	Stima	p-value	r	R^2
Intercetta	$62,5187$			
Pendenza	$-0,6096$	$0,009038$	$-0,6476622$	$0,4195$

2.4 Regione a valle

Per ognuno dei siti viene calcolato un "tasso di idrofobia", ossia il rapporto tra il numero di sequenze che contengono un aminoacido idrofobico ed il numero di sequenze totali (98), a meno di diverse correzioni di natura empirica e biochimica. Questo tasso viene calcolato anche per i tre siti successivi, per cui la tabella 1 si espande come illustrato in basso (Tabella 4). Sapendo che il tasso di idropatia dei 15 siti originali è distribuito in modo normale ($\mu = 0,436$, $\sigma = 0,442$), cosa possiamo dire sui tre siti inseriti successivamente (da 209 a 211)? Hanno un tasso di idropatia significativamente diverso dagli altri 15 o no? **Solamente il terzo è significativamente diverso dagli altri: il test Z, infatti, resituisce un p-value di 0.553081 per il primo, 0.6931277 per il secondo e 0.9608996 per il terzo.**

Tabella 4: Aminoacidi nel sito attivo della proteina, con tasso di idropatia.

Sito	Acidi	Basici	Polari	Idrofobici	Tasso di idropatia
194	0	0	23	75	0,014
195	0	0	13	85	0,501
196	0	0	6	92	0,091
197	0	97	1	0	0,144
198	0	0	0	98	0,644
199	0	0	9	89	0,377
200	0	0	14	84	0,003
201	0	0	1	97	0,767
202	0	97	0	1	0,362
203	70	0	15	13	0,120
204	1	17	76	4	0,027
205	0	0	98	0	0,510
206	0	0	95	3	0,293
207	0	0	98	0	0,338
208	0	0	98	0	0,442
209					0,495
210					0,659
211					1,214