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

prof. Federico Plazzi

16 Aprile 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

Sono ormai largamente usate diverse tecniche di sequenziamento massivo del DNA o del RNA che consentono la produzione di enormi quantità di dati in un tempo molto breve: più o meno tutte sfruttano l'idea di spezzare la grande quantità (in termini di paia di basi) di DNA o di RNA presente in una cellula in piccoli frammenti, di lunghezza variabile a seconda delle tecnologie. Questi frammenti vengono poi sequenziati su larga scala e si ottengono le cosiddette *read*, brevi sequenze che devono poi essere rielaborate per ottenere il DNA o il RNA di partenza.

Nella tabella seguente sono forniti alcuni dati riguardanti 20 *read* tratte da due esperimenti di sequenziamento diversi: viene mostrata la lunghezza della *read* e una stima della qualità (cioé dell'affidabilità della sequenza) in una scala che varia a seconda della tecnologia utlizzata, ma sempre fatta in modo che ai numeri più alti corrispondano qualità più alte.

Tabella 1: Dati di 20 *read* per due esperimenti di sequenziamento massivo.

	Esperimento A		Esperimento B	
read	Lunghezza (pb)	Qualità media	Lunghezza (pb)	Qualità media
1	513	40	506	43
2	499	36	497	42
3	495	34	492	38
4	496	38	503	45
5	496	38	502	40
6	499	38	498	42
7	502	37	497	39
8	504	33	507	47
9	511	44	505	40
10	502	37	492	40
11	498	38	499	42
12	495	31	504	44
13	514	37	507	46
14	501	39	496	41
15	489	43	507	42
16	516	40	501	42
17	485	36	497	43
18	523	35	496	45
19	513	38	507	44
20	493	36	507	43

2 Esercizi

2.1 Statistiche di base

Calcolare media, varianza e deviazione standard della qualità media delle *read* dell'esperimento A mostrate in tabella 1.

Media	*Varianza*	*Deviazione standard*
$37,4$	$8,84$	$2,973214$

2.2 Distribuzione dei dati

La distribuzione della qualità media delle *read* viene indagata con la tecnica del Q-Q plot, mostrato in figura 1. La tabella 2 mostra i parametri di un modello di correlazione lineare calcolato sui punti del Q-Q plot. Che cosa si può concludere sulla distribuzione dei dati?

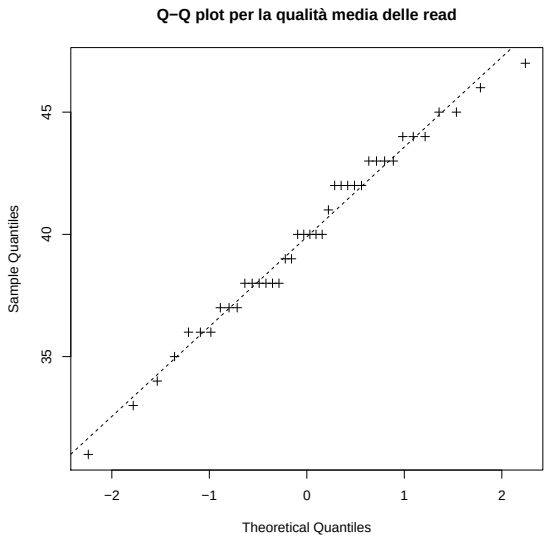

Figura 1: Q-Q plot delle qualità medie delle *read* di entrambi gli esperimenti.

Tabella 2: Regressione lineare tra i punti del Q-Q plot in figura 1.

	Stima	p-value	r	R^2
Intercetta	$39,9$			
Pendenza	$3,6737$	0	$0,9918662$	$0,9837986$

I punti del Q-Q plot approssimano molto bene una linea retta ($R^2 = 0,98$, $p = 0$). Questo è il risultato che si vede attraverso un Q-Q plot quando la distibuzione dei dati è normale.

2.3 Esperimenti a confronto

Per verificare se l'esperimento A e l'esperimento B abbiano restituito dati con qualità medie comparabili, si procede con diversi statistici. Qual è il test corretto? Perché? Possiamo dire che uno dei due esperimenti ha dato qualità medie più alte? Se sì, quale?

1. Test t a campioni indipendenti:

   ```
   Two Sample t-test

   t = -5.8083, df = 35.668, p-value = 1.294e-06
   alternative hypothesis:  true difference in means is not equal
   to 0
   ```

2. Test t a campioni appaiati:

   ```
   Paired t-test

   t = -5.1087, df = 19, p-value = 6.24e-05
   alternative hypothesis:  true difference in means is not equal
   to 0
   ```

3. Test di Mann e Whitney

   ```
   Wilcoxon rank sum test with continuity correction

   W = 38.5, p-value = 1.212e-05
   alternative hypothesis:  true location shift is not equal to 0
   ```

4. Test di Wilcoxon

   ```
   Wilcoxon signed rank test with continuity correction

   V = 10.5, p-value = 0.000436
   alternative hypothesis:  true location shift is not equal to 0
   ```

La variabile è a distribuzione normale in base ai risultati dell'esercizio precedente; i campioni non sono appaiati, perché le 20 read *dell'esperimento A vanno confrontate con le 20* read *dell'esperimento B nella loro globalità, non una per una. Il test corretto è quindi il test t a campioni indipendenti. Il p-value altamente significativo ($1,294 \times 10^{-6}$) indica una certa differenza tra le medie dei due campioni: visto che la qualità media delle* read *del campione A è $37,4$ e quella di quelle del campione B è $42,4$, possiamo concludere che l'esperimento B ha restituito dati di qualità più elevata.*

2.4 Qualità e lunghezza delle *read*

Bisogna capire a questo punto se la tecnologia utilizzata porta ad aumentare la qualità media di certe *read* a scapito di altre, per esempio quelle più corte o quelle più lunghe. La figura 2 mostra lunghezza e qualità media delle *read*; la tabella 3 mostra i parametri dei relativi modelli di correlazione lineare. Cosa possiamo concludere? C'è effettivamente un legame tra lunghezza e qualità media? Se sì, c'è in entrambi gli esperimenti?

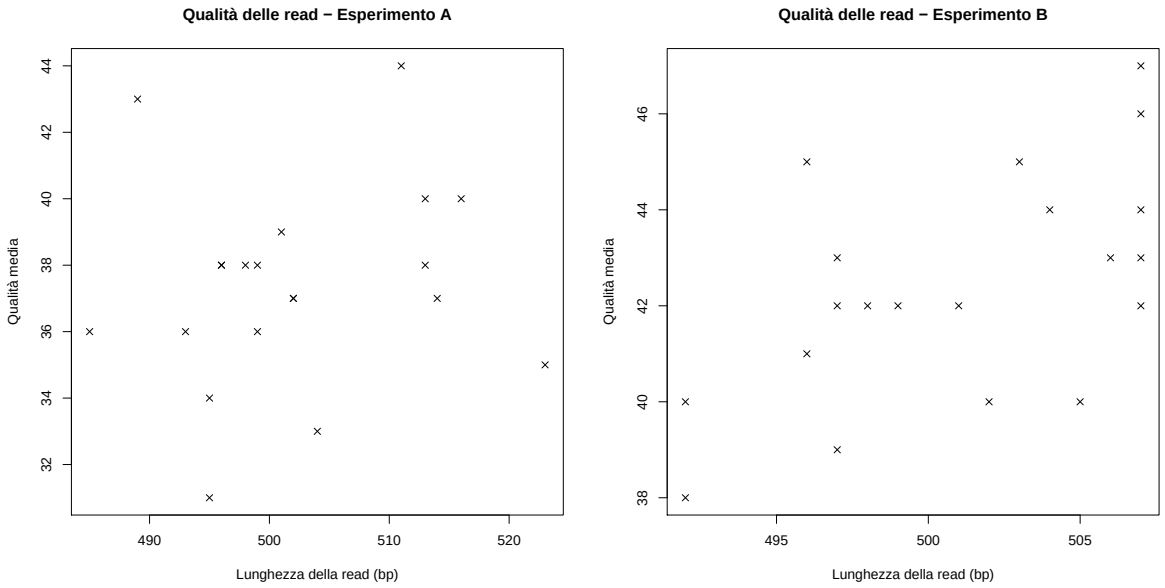

Figura 2: Plot di lunghezza e qualità media delle *read*.

Tabella 3: Modelli di correlazione lineari relativi alla figura 2

	Esperimento A	Esperimento B
Intercetta	$12,876$	$-86,7949$
Pendenza	$0,0488$	$0,2579$
r	$0,158$	$0,5678$
R^2	$0,025$	$0,3223$
p-value	$0,5058$	$0,009$

La correlazione non è significativa per l'esperimento A (p-value$=0,5058$*), ma lo è per l'esperimento B (p-value*$=0,009$*), quantunque non molto forte (*$R^2 = 0,3223$*). Si tratta di una correlazione positiva (*$r > 0$*), come è visibile anche nella figura 2: le* read *con le lunghezze maggiori hanno tendenzialmente qualità medie più alte.*

2.5 Numero di *read* di qualità media elevata

La soglia minima per indicare una certa *read* come di qualità media "elevata" viene di solito fissata a 40. Esiste un modo per capire se il numero di *read* di qualità media elevata è lo stesso in entrambi gli esperimenti? Se sì, quale? Che cosa porterebbe a concludere?

Il test corretto è il χ^2, perché si tratta di classificare, per ogni esperimento, le read *secondo la qualità media elevata o non elevata. Risulta una tabella 2×2: ci sono 4* read *di qualità media elevata nell'esperimento A e 18 nell'esperimento B. Il valore di χ^2 risultante è $17,071$ con un grado di libertà, che è molto significativo (il valore esatto del p-value sarebbe $3,601 \times 10^{-5}$). Possiamo quindi concludere che l'esperimento B ha dato significativamente più* read *di qualità media elevata dell'esperimento A.*