

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

prof. Federico Plazzi

12 Febbraio 2019

Nome: _____

Cognome: _____

Matricola: _____

Alcune indicazioni:

- La prova è costituita 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

Il trascrittoma di una data cellula è dato dall'insieme degli RNA che vengono trascritti in un certo momento dal DNA. Questi RNA possono essere di diversi tipi: RNA messaggeri che verranno tradotti in proteine, RNA ribosomali, RNA di trasporto, piccoli RNA non codificanti, ... In particolare, esistono diversi tipi di piccoli RNA non codificanti, a seconda della loro lunghezza, del modo in cui vengono prodotti a partire dal DNA e del meccanismo con il quale regolano altri processi cellulari, come per esempio la traduzione dei messaggeri. La tabella 1 mostra, per ogni tipo di piccolo RNA non codificante, quante molecole di una certa lunghezza sono state trovate nel trascrittoma cellulare nel corso di un esperimento.

Tabella 1: Caratteristiche di alcuni dei piccoli RNA non codificanti.

Tipo di RNA	Lunghezza (pb)	Numero di molecole
miRNA	18	1
miRNA	19	1
miRNA	20	3
miRNA	21	5
miRNA	22	9
miRNA	23	7
miRNA	24	3
piRNA	23	1
piRNA	25	1
piRNA	26	2
piRNA	28	6
piRNA	29	10
piRNA	30	5
siRNA	23	4
siRNA	24	5

2 Esercizi

2.1 Statistiche di base

Calcolare media, devianza, varianza e deviazione standard della lunghezza dei nove siRNA registrati in tabella 1.

<i>Media</i>	<i>Devianza</i>	<i>Varianza</i>	<i>Deviazione standard</i>
23,55556	2,222222	0,2469136	0,496904

2.2 Distribuzione dei dati

La prima cosa da capire è la distribuzione delle lunghezze. Di seguito è mostrato il risultato del test di Shapiro e Wilk condotto sulle lunghezze delle 90 molecole mostrate in tabella 1: quale delle due figure che seguono potrebbe corrispondere al relativo Q-Q plot?

Shapiro-Wilk normality test

data: sncRNA.length
W = 0.98734, p-value = 0.538

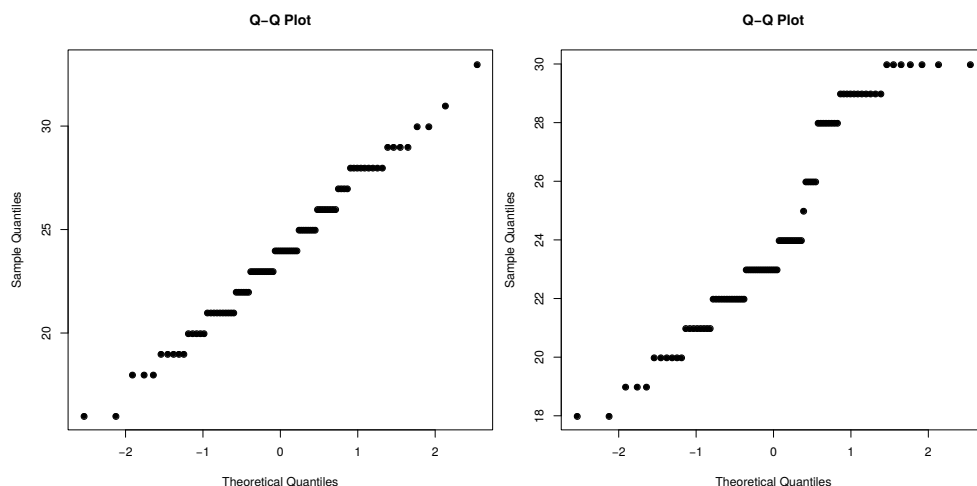


Figura 1: Due possibili Q-Q plot delle lunghezze dei piccoli RNA non codificanti mostrati in tabella 1.

Le lunghezze si distribuiscono in modo normale: il p-value del test di Shapiro e Wilk (0,538) non è significativo, per cui non si può rifiutare l'ipotesi nulla di normalità. I punti del Q-Q plot dovrebbero quindi approssimare una linea, per cui la figura corrispondente è la prima delle due.

2.3 Tipi di piccoli RNA non codificanti

È noto che i miRNA, nella specie in esame e nei tessuti analizzati fino a questo esperimento, dovrebbero essere in numero circa triplo di piRNA e siRNA messi insieme. L'uso di quale test statistico può permettere di verificare questa ipotesi con i dati a disposizione? Qual è il risultato?

L'attribuzione di una molecola a una delle tre categorie di piccoli RNA non codificanti è una variabile qualitativa e, come tale, può essere indagata attraverso il test del χ^2 . Le frequenze attese si calcolano in base ai rapporti attesi tra le tre specie di RNA: se i miRNA sono attesi essere il triplo di piRNA e siRNA messi insieme, le rispettive frequenze attese saranno 0,75, 0,125 e 0,125 (in questo modo, infatti, $(0,125 + 0,125) \cdot 3 = 0,75$). Il valore di χ^2 risulta essere 44,45, che con due $(3 - 1)$ gradi di libertà è significativo (il p-value esatto sarebbe di $2,228 \times 10^{-10}$). Si può quindi concludere che in questo particolare trascrittoma i diversi tipi di piccoli RNA non codificanti non rispettano le proporzioni attese.

2.4 One-Way ANOVA

Per concludere lo studio, bisogna capire esattamente se esistano differenze in lunghezza tra i tre tipi di piccoli RNA non codificanti mostrati in tabella 1. Si calcola perciò la devianza *entro* gruppi e la devianza *tra* gruppi, usando come gruppi i tre tipi di RNA. I risultati sono trascritti in Tabella 2.

1. Per quale motivo l'ANOVA è un approccio valido in questo caso?
2. Completa la tabella 2 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 2: One-Way ANOVA. D, devianza; σ^2 , varianza; g.l., gradi di libertà.

	D	σ^2	g.l.	F	p-value
<i>tra</i>	578,52	289,259	2		
<i>entro</i>	131,80	2,197	60	131,68	$< 2,2 \times 10^{-16}$

1. *L'ANOVA è indicata perché la variabile è a distribuzione normale; sebbene la variabile non sia continua (il numero di paia di basi è per forza naturale), almeno due gruppi sono di dimensione comparabile tra loro; per maggiore sicurezza, bisognerebbe testare l'omoschedasticità, ossia la comparabilità delle varianze.*
2. *I gradi di libertà sono 2 per la varianza tra gruppi (perché i gruppi in tutto sono 3) e 60 per la varianza entro gruppi (perché ci sono 63 osservazioni e 3 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. Possiamo quindi concludere che esiste una differenza in lunghezza tra i tre tipi di piccoli RNA non codificanti considerati.*
4. *Il passo successivo potrebbe essere il test di Tukey per verificare se un gruppo si discosta significativamente dagli altri due.*