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

prof. Federico Plazzi

27 Febbraio 2017

Nome: _____

Cognome: _____

Matricola: _____

Alcune indicazioni:

- La prova è constituita 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

Ad un gruppo di 20 studenti viene somministrato un test di statistica il cui punteggio è un numero (naturale) da 0 a 100. Lo stesso test viene somministrato di nuovo agli stessi studenti alla fine di un corso di statistica inferenziale lungo un mese. I risultati sono mostrati in Tabella 1.

Tabella 1: Risultati del test di statistica.

	Primo test	Secondo test
Studente 1	69	96
Studente 2	63	92
Studente 3	74	94
Studente 4	60	98
Studente 5	53	98
Studente 6	82	85
Studente 7	91	81
Studente 8	95	78
Studente 9	81	96
Studente 10	79	84
Studente 11	88	84
Studente 12	46	95
Studente 13	88	91
Studente 14	79	95
Studente 15	90	97
Studente 16	82	83
Studente 17	78	86
Studente 18	92	91
Studente 19	88	87
Studente 20	74	83

2 Esercizi

2.1 Statistiche di base

Calcolare media, devianza, varianza e deviazione standard dei risultati degli studenti per quanto riguarda il secondo test.

```
> mean(Secondo.test)
89.7
> SS(Secondo.test)
764.2
> variance(Secondo.test)
38.21
> standard.deviation(Secondo.test)
6.181424
```

2.2 Distribuzione dei risultati

Per indagare la distribuzione dei dati, vengono seguite tre diverse strategie: la costruzione di un Q-Q plot, l'osservazione dell'istogramma delle frequenze e il test di Shapiro e Wilk. Per un errore nei comandi forniti al computer, però, una di queste tre analisi resituisce un risultato sbagliato. Quale potrebbe essere l'analisi sbagliata, visto che non si accorda con il risultato delle altre due? Perché? Cosa si può concludere sulla distribuzione dei dati?

- Q-Q Plot

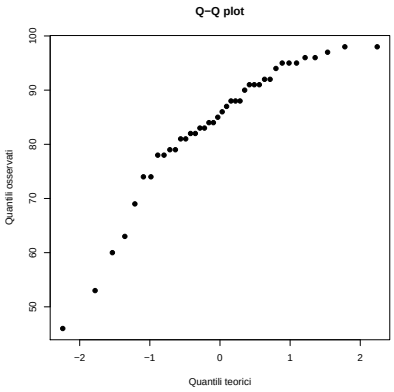

Figura 1: Q-Q Plot dei risultati.

- Istogramma delle frequenze

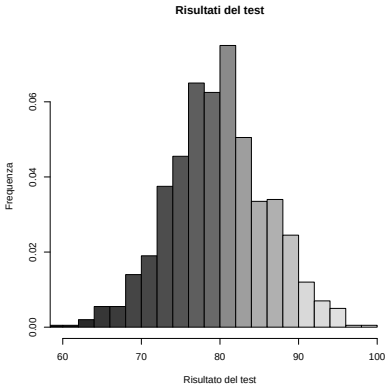

Figura 2: Istogramma delle frequenze dei risultati.

- Test di Shapiro e Wilk

```
Shapiro-Wilk test

data:  Risultati del test
W = 0.88181, p-value = 0.000589
```

Il risultato sbagliato è l'istogramma: i punti del Q-Q Plot non si allineano lungo una retta e il p-value del test di Shapiro e Wilk è inferiore a $0,05$, due risultati che porterebbero ad escludere l'ipotesi nulla di normalità, mentre l'istogramma mostra una distribuzione tendenzialmente normale.

2.3 Miglioramento in seguito al corso

Si vuole verificare se in seguito al corso ci sia stato un effettivo miglioramento degli studenti. A questo proposito, vengono eseguiti quattro diversi test statistici, i cui risultati sono elencati qui di seguito. Qual è l'approccio corretto? Perché? Cosa si può concludere sul miglioramento degli studenti?

- Test t a una coda a campioni indipendenti

```
Two Sample t-test

data:  Primo test and Secondo test
t = -3.634, df = 27.026, p-value = 0.080772
alternative hypothesis:  true difference in means is less than 0
```

- Test t a una coda a campioni appaiati

```
Paired t-test

data:  Primo test and Secondo test
t = -3.0356, df = 19, p-value = 0.003402
alternative hypothesis:  true difference in means is less than 0
```

- Test di Wilcoxon a una coda

```
Wilcoxon signed rank test

data:  Primo test and Secondo test
V = 35, p-value = 0.004719
alternative hypothesis:  true location shift is less than 0
```

- Test di Mann e Whitney a una coda

```
Mann and Whitney rank sum test

data:  Primo.test and Secondo.test
W = 79.5, p-value = 0.095783
alternative hypothesis:  true location shift is less than 0
```

La distribuzione dei dati non è normale e i campioni sono appaiati, per cui il test corretto è il test di Wilcoxon, il cui risultato è significativo ($p < 0,05$): il miglioramento dopo il corso è significativo.

2.4 Miglioramento del singolo studente

Ci si aspetterebbe che ci sia correlazione lineare tra i risultati del singolo studente nel primo e nel secondo test. La tabella 2 mostra i risultati di un test di correlazione lineare: come si possono commentare? Che legame c'è tra il voto di uno studente nel primo test e il voto dello stesso studente nel secondo test?

La correlazione è positiva ($r = 0,56201$) e significativa ($p < 0,05$): uno studente con un voto alto nel primo test tende a prendere un voto alto anche nel secondo e viceversa, anche se non c'è un legame strettissimo tra i due voti ($R^2 = 0,31586$).

2.5 Studenti di altre scuole

A prescindere dalla distribuzione dei risultati degli studenti della scuola in esame, che può essere normale o no, si sa che la distribuzione dei voti di questo test di tutte le scuole della regione è normale, con $\mu = 79$ e $\sigma = 11$. Come si può verificare se i risultati (la seconda volta che il test è stato somministrato) dei 20 studenti di cui alla tabella 1 siano migliori o peggiori di quelli regionali? Nella tabella 3 vengono indicati i risultati di diversi test effettuati per rispondere a questa domanda. Qual è quello corretto? Cosa si può concludere?

Tabella 2: Correlazione tra il voto nel primo test e il voto nel secondo.

	Stima	p-value	r	R^2
Intercetta	76, 22953			
Pendenza	0, 79456	0, 00991	0, 56201	0, 31586

Tabella 3: Test effettuati per confrontare i risultati dei 20 studenti (la seconda volta) con i risultati regionali.

Test	Valore della statistica
Test t a due code	$t_{19} = 7,5452$
One-Way ANOVA	$F_{1,5} = 2,6846$
Test Z a due code	$Z = 0,9727$

Il test corretto è il t di Student, visto che si sa che la popolazione è a distribuzione normale. Consultando la tabella del t di Student per 19 gradi di libertà, si vede che un valore di $7,5452$ è altamente significativo: siccome il t si calcola come $\frac{\overline{X}-\mu_{pop}}{\sigma_{pop}}$, un valore positivo indica che questi 20 studenti sono andati significativamente *meglio* della media regionale.