Esame di Statistica del 22 giugno 2006 (Corso di Laurea Triennale in Biotecnologie, Università degli Studi di Padova).

Cognome Nome Matricola

Es. 1	Es. 2	Es. 3	Es. 4	Somma	Voto finale

Attenzione: si consegnano SOLO i fogli di questo fascicolo.

Esercizio 1. In alcune specie di topi, il colore nero è una caratteristica genetica dominante rispetto al colore marrone. Supponiamo che un topo nero (che chiameremo "topo 1") con due genitori neri abbia un fratello marrone.

1. Dimostrare che il genotipo dei genitori è per forza Aa per entrambi, dove $A =$ gene del colore nero, $a =$ gene del colore marrone.
2. Qual è la probabilità che il topo 1 abbia genotipo AA? E genotipo Aa? E genotipo aa? (ricordarsi di utilizzare il fatto che il topo è nero!)

Supponiamo ora di prendere il topo 1 e di incrociarlo con un altro topo di genotipo aa.

3. Se il topo 1 ha genotipo Aa e la coppia partorisce 5 topini, qual è la probabilità di ottenerli tutti e 5 neri? Supporre che i genotipi dei topini siano indipendenti tra di loro.
4. Rispondere alla stessa domanda nel caso in cui il topo 1 abbia genotipo AA.
5. Supponiamo che la coppia abbia effettivamente partorito 5 topini neri; qual è la probabilità che il topo 1 abbia genotipo AA?

Esercizio 2. È stato condotto uno studio per correlare l'uso di contraccettivi orali col livello di colesterolo. Il livello di colesterolo tra 66 utilizzatrici di contraccettivi orali era di 201 ± 37 (media $\pm$ deviazione standard), mentre il livello di colesterolo tra 97 non utilizzatrici di contraccettivi orali era di 193 ± 37 (media $\pm$ deviazione standard).

1. Effettuare un test per stabilire se c'è una differenza significativa nel livello medio di colesterolo tra i due gruppi. Riportare limitazioni al valore P.
2. Calcolare l'intervallo di confidenza al 95% per la differenza di livello di colesterolo tra i gruppi.
3. Supporre che, invece dei risultati già calcolati, il test del punto 1. avesse dato un valore $P = 0.03$ e che l'intervallo di confidenza fosse venuto $(-0.6; 7.3)$. Questi risultati si contraddicono l'un l'altro? Perchè o perchè no?
4. Supponiamo ora di voler costruire un test per stabilire se la "vera" differenza tra le medie è almeno di 10, supponendo che la deviazione standard sia di 37, e che livello e potenza debbano essere rispettivamente $\alpha = 0.01$ e $1 - \beta = 0.99$. Qual è il numero minimo di donne che serve avere nei due gruppi?

Esercizio 3. L'incidenza di melanoma maligno in donne dai 35 ai 59 anni è circa di 15 nuovi casi ogni 100.000 donne all'anno. Viene pianificato uno studio per seguire 10.000 donne con esposizione eccessiva al sole.

1. Dimostrare che l'incidenza di melanoma maligno su 4 anni in donne che all'inizio avevano 35 anni è $p \simeq 6 \cdot 10^{-4}$.
2. Calcolare il numero atteso di melanomi maligni su 10.000 donne.
3. Supponiamo che invece siano osservati 9 nuovi casi. Che test bisogna effettuare per stabilire se questa osservazione è in contrasto con quanto ci si aspettava o no?
4. Eseguire il test del punto 3. con un livello $\alpha = 0.05$.

Esercizio 4. Vogliamo testare fino a che punto l'ipertensione è un fenomeno genetico. A questo scopo, sono state esaminate 20 famiglie, prendendo la pressione arteriosa di madre (y), padre (x) e primogenito (t), con i seguenti risultati:

$$\sum_{i=1}^{20} x_i = 2980, \qquad \sum_{i=1}^{20} y_i = 2620, \qquad \sum_{i=1}^{20} t_i = 2030$$

$$\sum_{i=1}^{20} x_i^2 = 451350, \qquad \sum_{i=1}^{20} y_i^2 = 351350, \qquad \sum_{i=1}^{20} t_i^2 = 210850$$

$$\sum_{i=1}^{20} x_i y_i = 390825, \qquad \sum_{i=1}^{20} x_i t_i = 305700, \qquad \sum_{i=1}^{20} y_i t_i = 269550$$

1. Trovare la retta di regressione della pressione del figlio (var. dipendente) rispetto a quella del padre (var. indipendente).
2. Eseguire un test per vedere se in effetti c'è una relazione lineare. Riportare limitazioni al valore P.
3. Qual è la pressione media attesa del figlio se la pressione del padre è 130? E se è 150? E se è 170?
4. Trovare gli errori standard delle stime del punto 3. Perchè non sono gli stessi?

Soluzioni

Esercizio 1.

1. Entrambi i genitori sono neri, quindi devono avere un gene di tipo A. Inoltre, per generare un figlio marrone (e quindi di genotipo aa), devono anche avere entrambi un gene di tipo a.

2. Per le leggi di Mendel, in assenza di altre informazioni il genotipo del topo 1 sara uno tra i 4 seguenti $\{AA, Aa, aA, aa\}$ con uguale probabilità. Sappiamo però che il topo ha il pelo nero, e quindi non può avere genotipo aa. Allora il genotipo del topo potrà essere uno tra i 3 seguenti $\{AA, Aa, aA\}$. Considerando Aa e aA come lo stesso genotipo, si ha che $\mathbb{P}\{T1 = AA\} = 1/3$, $\mathbb{P}\{T1 = Aa\} = 2/3$, e $\mathbb{P}\{T1 = aa\} = 0$ (dove chiamiamo $T1$ il genotipo relativo al pelo del topo 1). Si può arrivare allo stesso risultato anche condizionando la probabilià iniziale (legge uniforme su 4 genotipi) all'evento "il topo 1 ha pelo nero".

3. Chiamiamo B_i l'evento "i-esimo topino nero". Siccome dobbiamo condizionare la probabilità all'evento $\{T1 = Aa\}$, si ha
$$\mathbb{P}(B_i|\{T1 = Aa\}) = 1/2 \quad \forall i = 1, \ldots, 5$$
difatti i possibili genotipi dell'i-esimo topino sono $\{Aa, Aa, aa, aa\}$, ognuno con probabilità 1/4. Dato che i B_i sono indipendenti (condizionatamente ai genitori), si ha che
$$\mathbb{P}\left(\bigcap_{i=1}^{5} B_i \middle| \{T1 = AA\}\right) = \prod_{i=1}^{5} \mathbb{P}(B_i|\{T1 = Aa\}) = \left(\frac{1}{2}\right)^5$$

4. Stavolta dobbiamo condizionare la probabilità all'evento $\{T1 = AA\}$, e si ha
$$\mathbb{P}(B_i|\{T1 = AA\}) = 1 \quad \forall i = 1, \ldots, 5$$
difatti i possibili genotipi dell'i-esimo topino sono $\{Aa, Aa, Aa, Aa\}$, ognuno con probabilità 1/4. Dato che i B_i sono indipendenti (condizionatamente ai genitori), si ha che
$$\mathbb{P}\left(\bigcap_{i=1}^{5} B_i \middle| \{T1 = Aa\}\right) = \prod_{i=1}^{5} \mathbb{P}(B_i|\{T1 = Aa\}) = 1$$

5. Dobbiamo calcolare
$$\mathbb{P}\left(\{T1 = AA\} \middle| \bigcap_{i=1}^{5} B_i\right) = \frac{\mathbb{P}\left(\bigcap_{i=1}^{5} B_i \middle| \{T1 = AA\}\right)\mathbb{P}\{T1 = AA\}}{\mathbb{P}(\cap_{i=1}^{5} B_i)}$$
dove abbiamo utilizzato la formula di Bayes. Per calcolare la probabilità al denominatore, utilizziamo la formula della probabilità totale:
$$\begin{aligned} \mathbb{P}(\cap_{i=1}^{5} B_i) &= \mathbb{P}(\cap_{i=1}^{5} B_i|\{T1 = AA\})\mathbb{P}\{T1 = AA\} + \mathbb{P}(\cap_{i=1}^{5} B_i|\{T1 = Aa\})\mathbb{P}\{T1 = Aa\} = \\ &= 1 \cdot \frac{1}{3} + \frac{1}{32} \cdot \frac{2}{3} = \frac{17}{48} \end{aligned}$$
Abbiamo quindi
$$\mathbb{P}\left(\{T1 = AA\} \middle| \cap_{i=1}^{5} B_i\right) = \frac{1 \cdot \frac{1}{3}}{\frac{17}{48}} = \frac{16}{17}$$

Esercizio 2.

1. Facciamo un test t con ipotesi $H_0 : \mu_O = \mu_C$ e alternativa $H_1 : \mu_O \neq \mu_C$. Abbiamo

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}} = \frac{201 - 193}{\sqrt{\frac{37^2}{66} + \frac{37^2}{97}}} = 1.36$$

I gradi di libertà sono $\nu = 66 + 97 - 2 = 161$. Siccome $t_{0.95}(161) > t_{0.95}(200) = 1.653 > 1.36 = t$, possiamo riportare $P > 0.1$. Questo significa che possiamo accettare H_0.

2. Approssimiamo il quantile che ci serve con $t_{0.975}(200) = 1.972$. Allora l'intervallo di confidenza ha estremi $\bar{X} - \bar{Y} \pm t_{0.975}(200) s_{\bar{X}-\bar{Y}} = 8 \pm 1.972 \cdot 5.904$, e quindi risulta $[-3.64; 19.64]$.

3. Un intervallo di confidenza al 95% che contiene lo zero significa che, facendo il test del punto 1 con livello $\alpha = 5\%$, avremmo accettato H_0. Questo è però in contrasto con $P = 0.03$, che significa che avremmo rifiutato H_0 con qualunque $\alpha > 0.03$ (e quindi in particolare con $\alpha = 5\%$). I risultati forniti quindi si contraddicono l'un l'altro.

4. Il numero minimo di donne da utilizzare in entrambi i campioni è

$$\bar{n} = 2 \left(\frac{\sigma}{\delta}\right)^2 (q_{1-\beta} + q_{1-\alpha/2})^2 = 2 \cdot \left(\frac{37}{10}\right)^2 (q_{0.99} + q_{0.995})^2 = 654.71$$

Bisogna quindi utilizzare almeno 655 donne in ogni campione.

Esercizio 3.

1. Possiamo definire gli eventi

$$\begin{aligned} A_i &:= \{ \text{ ammalarsi di melanoma l'}i\text{-esimo anno } \} \\ A &:= \{ \text{ ammalarsi di melanoma in un periodo di 4 anni } \} \end{aligned}$$

Allora sappiamo che $\mathbb{P}(A_i) = 15/100000 = 1.5 \cdot 10^{-4}$ per ogni anno i. Supponiamo che gli $(A_i)_i$ siano indipendenti tra di loro. Dobbiamo allora calcolare

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}\left(\bigcup_{i=1}^{4} A_i\right) = \mathbb{P}\left(\left(\bigcap_{i=1}^{4} A_i^c\right)^c\right) = 1 - \mathbb{P}\left(\bigcap_{i=1}^{4} A_i^c\right) = 1 - \prod_{i=1}^{4} \mathbb{P}(A_i^c) = \\ &= 1 - 0.99985^4 = 1 - 0.9994001 = 0.0005999 \simeq 6 \cdot 10^{-4} \end{aligned}$$

2. Per ogni donna, possiamo definire una variabile aleatoria di Bernoulli

$$X_i := \begin{cases} 1 & \text{se la } i\text{-esima donna è malata,} \\ 0 & \text{altrimenti} \end{cases}$$

di parametro $p = 6 \cdot 10^{-4}$. Il numero di donne malate su n donne è quindi $S_n = \sum_{i=1}^{n} X_i$; dato che le $(X_i)_i$ si possono considerare indipendenti, abbiamo che $S_n \sim B(n,p)$ per ogni n, e quindi $E[S_{10000}] = 10000 \cdot p = 6$.

3. Dobbiamo fare un test di ipotesi $H_0 : p = 6 \cdot 10^{-4}$ e alternativa $H_1 : p \neq 6 \cdot 10^{-4}$. Dato che $10000 \cdot 6 \cdot 10^{-4} = 6 > 5$, possiamo usare il test Z.

4. Abbiamo che $\hat{p} = \frac{9}{10000} = 9 \cdot 10^{-4}$, e quindi

$$Z = \frac{6 \cdot 10^{-4} - 9 \cdot 10^{-4}}{\sqrt{\frac{6 \cdot 10^{-4}(1 - 6 \cdot 10^{-4})}{10000}}} = -1.23$$

Dobbiamo confrontare $|Z|$ con il quantile $q_{1-\alpha/2} = q_{0.975} = 1.96$. Siccome $|Z| = -1.23 < 1.96$, accettiamo H_0: le osservazioni non sono in contrasto con quanto ci si aspetta.

Esercizio 4. Partiamo calcolando le quantità:

$$\begin{aligned}
\bar{X} &= 149 \quad (0,5 \text{ punti}), \qquad \bar{Y} = 101.5 \quad (0,5 \text{ punti}),\\
s_X^2 &= 385.78 \quad (0,5 \text{ punti}),\\
s_T^2 &= 252.89 \quad (0,5 \text{ punti}),\\
s_{XT} &= 170.00 \quad (0,5 \text{ punti})
\end{aligned}$$

1. Calcoliamo i coefficienti della retta di regressione:

$$\begin{aligned}
b_1 &= \frac{s_{XY}}{s_X^2} = 0.441\\
b_0 &= \bar{Y} - b_1\bar{X} = 35.842
\end{aligned}$$

   La retta di regressione è quindi $y = 0.441x + 35.842$.

2. Bisogna eseguire un test di ipotesi $H_0 : \beta_1 = 0$ e alternativa $H_1 : \beta_1 \neq 0$. Bisogna allora calcolare:

$$\begin{aligned}
s_{T|X} &= \sqrt{\frac{n-1}{n-2}(s_T^2 - b_1^2 s_X^2)} = 13.707\\
s_{b_1} &= \frac{s_{T|X}}{\sqrt{(n-1)\cdot s_X^2}} = 0.160
\end{aligned}$$

   Abbiamo allora

$$t = \frac{b_1 - 0}{s_{b_1}} = 2.75$$

   Bisogna confrontare $|t|$ con una legge di Student a $\nu = 20 - 2 = 18$ gradi di libertà. Abbiamo $t_{0.99}(18) = 2.55 < |t| < 2.87 = t_{0.995}(18)$, quindi possiamo riportare $0.01 < P < 0.02$. Con una P così bassa, possiamo rifiutare H_0 e accettare H_1, concludendo che in effetti c'è una relazione lineare significativa.

3. La pressione attesa del figlio con un padre di pressione uguale a x è $y = b_1x + b_0$. Per le 3 quantità domandate si ha quindi:

$$\begin{aligned}
y_1 &= 0.441\cdot 130 + 35.842 = 93.13\\
y_2 &= 0.441\cdot 150 + 35.842 = 101.94\\
y_3 &= 0.441\cdot 170 + 35.842 = 110.75
\end{aligned}$$

4. La formula generale per l'errore standard della media della y è

$$s := s_{Y|X}\sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{(n-1)s_x^2}}$$

   che quindi dipende da x: in particolare, più x si allontana dalla media (in questo caso $\bar{X} = 149$), più l'errore standard sarà alto. Per i 3 casi abbiamo

$$\begin{aligned}
s_1 &= 13.707\sqrt{\frac{1}{20} + \frac{(130-149)^2}{19\cdot 385.79}} = 4.32\\
s_2 &= 13.707\sqrt{\frac{1}{20} + \frac{(150-149)^2}{19\cdot 385.79}} = 3.07\\
s_3 &= 13.707\sqrt{\frac{1}{20} + \frac{(170-149)^2}{19\cdot 385.79}} = 4.55
\end{aligned}$$

Esame di Statistica del 22 giugno 2006 (Corso di Laurea in Biotecnologie, Universitá degli Studi di Padova) (docente: Tiziano Vargiolu)

Hanno superato la prova:

Basso Stefania	20.5 + 3
Bolognese Paolo	25.5 + 3
Bortolaso Rossella	16.5 + 3
Castagnaro Silvia	26.5 + 3
Cocchetto Alessandro	19.5
Colizzi Enrico Sandro	26.5 + 3
D'Agnolo Lorenzo	18 + 3
De Rossi Matteo	17 + 3
Dinaro Ylenia Maria	17 + 3
Franzoso Mauro	26 + 3
Galozzi Paola	25 + 3
Giobbe Giovanni Giuseppe	18.5 + 3
Giulitti Stefano	21.5
Giurgola Laura	17 + 3
Lanciai Federico	19
Martewicz Sebastian	24.5 + 3
Mazzocato Alberto	20.5
Pegoraro Valentina	14.5 + 3
Pellegrini Alice	19 + 3
Perin Angela	16.5 + 3
Prando Claudio	18
Prevato Marua	19 + 3
Procopio Maria	19.5 + 3
Ranchio Alessandro	24
Santini Gaia Cecilia	19.5 + 3
Stradiotto Damiano	24
Usai Carla	23 + 3
Vannozzi Giulia	15.5 + 3
Vono Maria	22 + 3
Zampieri Davide	23

Visione compiti corretti, registrazione voto e/o orali: martedì 27 giugno ore 11.00 aula 1B/50 Torre Archimede.
Verrà data precedenza alla registrazione voti a chi accetta il voto dello scritto e ha il bonus di + 3.