

Esame di Statistica del 17 luglio 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. Calcolare la probabilità che una persona sia nata il 30 aprile:

1. sapendo che è nata in un anno bisestile;
2. sapendo che è nata in un anno non bisestile;
3. non avendo questa informazione (e supponendo per semplicità che gli anni bisestili siano $1/4$ dei totali).
4. Come cambiano le risposte sopra se la data, invece del 30 aprile, è il 29 febbraio?

Supponiamo che due persone si sposino. Calcolare la probabilità che gli sposi:

5. siano entrambi nati il 30 aprile;
6. siano nati nello stesso giorno (che non sia il 29 febbraio).

In un anno, nella città di New York, si celebrano 80000 matrimoni. Dai una stima della probabilità che per una almeno delle coppie entrambi gli sposi:

7. siano nati il 30 aprile;
8. siano nati nello stesso giorno (che non sia il 29 febbraio).

Esercizio 2. Per verificare se il fare jogging porti ad un abbassamento della frequenza cardiaca a riposo, 8 volontari hanno iniziato un programma di 1 mese di jogging. I dati prima e dopo il mese sono i seguenti:

individuo	1	2	3	4	5	6	7	8
prima	74	86	98	102	78	84	79	70
dopo	70	85	90	110	71	80	69	74

1. Che test bisogna effettuare per vedere se il jogging ha effetto sulla frequenza cardiaca?
2. Effettuare il test col metodo degli intervalli di confidenza, utilizzando $\alpha = 5\%$.
3. Effettuare il test col metodo del valore critico, utilizzando $\alpha = 5\%$. Riportare poi limitazioni al valore P .

Esercizio 3. Un sondaggio elettorale su un quotidiano riporta il candidato A in vantaggio sul candidato B con il 53% contro il 47% delle preferenze, con un margine di errore di $\pm 4\%$. Il giornale conclude dicendo che, siccome il margine di errore è inferiore allo scarto tra i due (6%), la vittoria del candidato A è ormai certa. Vogliamo dimostrare che il quotidiano ha torto.

1. Supponendo che il “margine di errore” sia la metà dell’ampiezza dell’intervallo di confidenza (che quindi ha estremi 0.53 ± 0.04), determinare la taglia del campione su cui è stato fatto il sondaggio nel caso in cui l’intervallo di confidenza sia al 95% e nel caso in cui sia al 99%.
2. La nostra supposizione è che in realtà l’elettorato sia ancora fortemente indeciso sui due candidati. Che test bisogna effettuare per stabilirlo?
3. Effettuare il test supponendo che l’intervallo di confidenza fosse al 95%. Riportare il valore P .
4. Che conclusione se ne trae sul risultato delle elezioni?
5. È possibile pervenire ad un risultato analogo anche senza effettuare il test?

Esercizio 4. Studiamo il livello di corrosione di una certa sostanza metallica esposta ad un’atmosfera di ossigeno puro a 500 gradi Celsius. L’aumento relativo di massa della sostanza viene usato come indicatore della quantità di ossigeno che ha reagito. I dati raccolti sono i seguenti:

Ore di esposizione (x)	1	2	2.5	3	3.5	4
Incremento percentuale (y)	0.02	0.03	0.035	0.042	0.05	0.054

1. Trova la relazione lineare che interpola meglio i dati.
2. La relazione trovata è significativa? Riporta limitazioni al valore P .
3. Fornisci una previsione dell’intervallo di confidenza al 95% dell’incremento di massa dopo 3.2 ore di esposizione.

Soluzioni

Esercizio 1. Innanzitutto chiamiamo X_i il giorno di nascita dell' i -esima persona, e B l'evento "anno bisestile". Supponiamo, per simmetria, che la data di nascita di una persona abbia legge uniforme su tutti i giorni dell'anno.

1. Chiamiamo X_1 la data di nascita di questa prima persona. Poichè un anno bisestile ha 366 giorni,

$$\mathbb{P}(\{X_1 = 30/4\} \mid B) = \frac{1}{366}$$

2. Poichè un anno non bisestile ha 365 giorni,

$$\mathbb{P}(\{X_1 = 30/4\} \mid B^c) = \frac{1}{365}$$

3. Applicando la formula della probabilità totale con $\mathbb{P}(B) = 1/4$, abbiamo

$$\mathbb{P}\{X_1 = 30/4\} = \mathbb{P}(\{X_1 = 30/4\} \mid B)\mathbb{P}(B) + \mathbb{P}(\{X_1 = 30/4\} \mid B^c)\mathbb{P}(B^c) = \frac{1}{366} \cdot \frac{1}{4} + \frac{1}{365} \cdot \frac{3}{4} = 0.002738$$

4. Poichè il 29 febbraio è presente solo negli anni bisestili,

$$\mathbb{P}(\{X_1 = 29/2\} \mid B) = \frac{1}{366}, \quad \mathbb{P}(\{X_1 = 30/4\} \mid B^c) = 0$$

e quindi

$$\mathbb{P}\{X_1 = 29/2\} = \mathbb{P}(\{X_1 = 29/2\} \mid B)\mathbb{P}(B) + \mathbb{P}(\{X_1 = 29/2\} \mid B^c)\mathbb{P}(B^c) = \frac{1}{366} \cdot \frac{1}{4} + 0 \cdot \frac{3}{4} = 0.000683$$

5. È plausibile che le due date di nascita, che chiamiamo X_1 e X_2 , siano indipendenti. Allora

$$p_1 := \mathbb{P}\{X_1 = X_2 = 30/4\} = \mathbb{P}\{X_1 = 30/4\}\mathbb{P}\{X_2 = 30/4\} = 0.002738^2 = 0.0000075$$

6. Dobbiamo calcolare

$$p_2 := \mathbb{P}\{X_1 = X_2 \neq 29/2\} = \sum_{i=1}^{365} \mathbb{P}\{X_1 = X_2 = i\} = \sum_{i=1}^{365} \mathbb{P}\{X_1 = i\}\mathbb{P}\{X_2 = i\} = \sum_{i=1}^{365} 0.002738^2 = 0.002735$$

dove abbiamo sommato sui giorni dell'anno diversi dal 29/2.

7. Per ogni $k = 1, \dots, n$, con $n = 80000$, definiamo la variabile aleatoria di Bernoulli

$$Y_k := \begin{cases} 1 & \text{se nella } k\text{-esima coppia gli sposi sono del } 30/4 \\ 0 & \text{altrimenti} \end{cases}$$

che ha quindi parametro p_1 . Allora $S_n := \sum_{k=1}^n Y_k$ rappresenta il numero di coppie in cui gli sposi sono nati il 30/4. Dobbiamo calcolare

$$\mathbb{P}\{S_n \geq 1\} = 1 - \mathbb{P}\{S_n = 0\} = 1 - \frac{n}{0} p_1^0 (1 - p_1)^n = 1 - 0.9999925^{80000} = 1 - 0.548992 = 0.451008$$

Poichè $n > 100$ e $p_1 < 0.01$, allo stesso risultato si poteva arrivare usando l'approssimazione di Poisson con $\lambda = np_1 = 80000 \cdot 0.0000075 = 0.5996 \simeq 0.6$.

8. Per ogni $k = 1, \dots, n$, con $n = 80000$, definiamo la variabile aleatoria di Bernoulli

$$Z_k := \begin{cases} 1 & \text{se nella } k\text{-esima coppia gli sposi sono dello stesso giorno} \\ 0 & \text{altrimenti} \end{cases}$$

che ha quindi parametro p_2 . Allora $T_n := \sum_{k=1}^n Z_k$ rappresenta il numero di coppie in cui gli sposi sono nati lo stesso giorno. Dobbiamo calcolare

$$\mathbb{P}\{T_n \geq 1\} = 1 - \mathbb{P}\{T_n = 0\} = 1 - \frac{n}{0} p_2^0 (1 - p_2)^n = 1 - 0.997265^{80000} \simeq 1 - 0 = 1$$

Poichè $n > 100$ e $p_2 < 0.01$, allo stesso risultato si poteva arrivare usando l'approssimazione di Poisson con $\lambda = np_2 = 80000 \cdot 0.002737 = 218.87$ (la probabilità risulta $1 - e^{-218.87}$, che è chiaramente molto vicino a 1).

Esercizio 2.

1. Innanzitutto notiamo che le due serie di dati non sono indipendenti; bisogna quindi fare un test t accoppiato: per fare questo, bisogna considerare le differenze tra le due pressioni di ogni individuo, che risultano essere le seguenti:

$$-4 \quad -1 \quad -8 \quad +8 \quad -7 \quad -4 \quad -10 \quad +4$$

e su questi dati si fa un test t sulla media, con ipotesi $H_0 : \mu = 0$ e alternativa $H_1 : \mu \neq 0$.

2. Risulta $\bar{X} = 2.75$, $s_X = 6.15$, quindi $s_{\bar{X}} = 2.18$. Il quantile che ci serve è $t_{0.975}(7) = 2.364$. Allora l'intervallo di confidenza ha estremi $\bar{X} \pm t_{0.975}(7)s_{\bar{X}}$, e quindi risulta $[-2.40; 7.90]$, che contiene lo 0. Questo significa che, facendo il test del punto 1 con livello $\alpha = 5\%$, accettiamo H_0 : non sembra che l'influenza dello jogging sulla frequenza cardiaca sia significativa.
3. Calcoliamo la statistica t :

$$t = \frac{\bar{X} - 0}{s_{\bar{X}}} = \frac{2.75}{2.18} = 1.26$$

e risulta $|t| < t_{0.975}(7) = 2.364$: anche con questo metodo accettiamo H_0 . Siccome poi $t_{0.95}(7) = 1.894$ è il più basso valore sulle tavole, possiamo riportare $P > 0.1$.

Esercizio 3.

1. Secondo i dati del problema, abbiamo $0.04 = s_{\hat{p}}q_{1-\alpha/2}$, quindi $s_{\hat{p}} = 0.04/q_{1-\alpha/2}$. Si ha anche che $s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$, e quindi

$$n = \frac{\hat{p}(1-\hat{p})}{s_{\hat{p}}^2} = \frac{\hat{p}(1-\hat{p})q_{1-\alpha/2}^2}{0.04^2} = \frac{0.53 \cdot 0.47}{0.04^2} q_{1-\alpha/2}^2$$

e n dipende quindi da α . Se $\alpha = 5\%$, allora $n = 598.09 \simeq 598$; se invece $\alpha = 1\%$, allora $n = 1028.3 \simeq 1028$.

2. Dobbiamo fare un test sulla proporzione di ipotesi $H_0 : p = 1/2$ e alternativa $H_1 : p \neq 1/2$. Bisogna verificare che $np(1-p) > 5$, che è equivalente a $n/4 > 5$, cioè $n > 20$, che è verificato sia per $\alpha = 5\%$ che per $\alpha = 1\%$. Possiamo quindi usare il test Z .
3. Se $\alpha = 0.05$ abbiamo che

$$Z = \frac{0.53 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{598}}} = 1.47$$

Abbiamo che $F_Z(1.47) = 0.92922 = 1 - P/2$, quindi $P = 0.14156$. Con un valore P così alto, accettiamo l'ipotesi H_0 .

4. Avendo accettato l'ipotesi $H_0 : p = 1/2$, sembra che l'elettorato sia ancora fortemente indeciso tra i due candidati, contrariamente a quanto affermato dal giornale.
5. Sì: siccome il valore $1/2 = 0.5$ è contenuto nell'intervallo di confidenza al 95% (ricordiamo che era $[0.49; 0.57]$), possiamo accettare l'ipotesi $H_0 : p = 1/2$ senza fare conti.

Esercizio 4. Facendo i calcoli si ottiene

$$\sum_{i=1}^6 X_i = 16, \quad \sum_{i=1}^6 Y_i = 0.231, \quad \sum_{i=1}^6 X_i Y_i = 0.685$$

$$\sum_{i=1}^6 X_i^2 = 48.5, \quad \sum_{i=1}^6 Y_i^2 = 0.010$$

Partiamo calcolando le quantità:

$$\begin{aligned} \bar{X} &= 2.6667 \quad (0,5 \text{ punti}), & \bar{Y} &= 0.0385 \quad (0,5 \text{ punti}), \\ s_X^2 &= 1.1666 \quad (0,5 \text{ punti}), \\ s_Y^2 &= 0.000162 \quad (0,5 \text{ punti}), \\ s_{XY} &= 0.0137 \quad (0,5 \text{ punti}) \end{aligned}$$

1. Calcoliamo i coefficienti della retta di regressione:

$$\begin{aligned} b_1 &= \frac{s_{XY}}{s_X^2} = 0.0117 \\ b_0 &= \bar{Y} - b_1 \bar{X} = 0.0071 \end{aligned}$$

La retta di regressione è quindi $y = 0.0117x + 0.0071$.

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

$$\begin{aligned} s_{Y|X} &= \sqrt{\frac{n-1}{n-2}(s_T^2 - b_1^2 s_X^2)} = 0.0013 \\ s_{b_1} &= \frac{s_{T|X}}{\sqrt{(n-1) \cdot s_X^2}} = 0.00055 \end{aligned}$$

Abbiamo allora

$$t = \frac{b_1 - 0}{s_{b_1}} = 21.26$$

Bisogna confrontare $|t|$ con una legge di Student a $\nu = 9 - 2 = 7$ gradi di libertà. Il più alto valore in tavola è $t_{0.999}(4) = 7.173 < |t|$, quindi possiamo riportare $P < 0.002$. Con un valore P così basso, possiamo rifiutare H_0 in favore di H_1 , concludendo che la relazione lineare sembra essere significativa.

3. Supponendo che la relazione lineare sia valida, la percentuale attesa per 3.2 ore di esposizione è $y = b_1 x + b_0 = 0.0447$, con un errore standard della stima di

$$s := s_{Y|X} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{(n-1)s_X^2}} = 0.00147$$

e $t_{0.975}(4) = 2.776$, quindi l'intervallo di confidenza cercato ha estremi $y \pm t_{1-\alpha/2}(\nu)s = 0.0447 \pm 2.776 \cdot 0.00147$, ed è quindi $[0.0406; 0.0488]$

Esame di Statistica del 17 luglio 2006 (Corso di Laurea in Biotecnologie, Università degli Studi di Padova) (docente: Tiziano Vargiolu)

Hanno superato la prova:

D'Alessandro Stefano	17.5
Isoli Fabio	17

Visione compiti corretti, registrazione voto e/o orali: giovedì 20 luglio ore 10.00 oppure ore 14.30 aula LuF1 via Luzzatti (complesso Paolotti).

Verrà data precedenza alla registrazione voti a chi accetta il voto dello scritto e ha il bonus di + 3.