ESERCITAZIONE 11– DOCENTE: MAURO PICCIONI, TUTOR: HLAFO ALFIE MIMUN

May 26, 2020

1. ESERCIZI

Ex. 1: Per stimare μ, la quantità di nicotina media contenuta in un nuovo marchio di sigarette, vengono scelte 44 sigarette a caso e viene determinata la quantità di nicotina in esse contenuta.

(1a) Se viene trovata una quantità di nicotina media pari a 1.74 mg, qual è l'intervallo di confidenza per la media μ a livello 0.95?

(1b) Trovare il più grande valore di α tale che 1.8 sia nell'intervallo di confidenza per la media μ di livello $1-\alpha$ (ovvero trovare il p-value di $\mu_0 = 1.8$).

(1c) Quanto grande deve essere il campione affinché la lunghezza dell'intervallo di confidenza per la media μ a livello 0.95 sia minore o uguale a 0.3 mg?

Si assuma che sia noto dall'esperienza passata che la deviazione standard della nicotina contenuta in una sigaretta sia pari a 0.7 mg.

Ex. 2: Dato il seguente set di 20 dati

$$16, 0, 0, 2, 3, 6, 8, 2, 5, 0, 12, 10, 5, 7, 2, 3, 8, 17, 9, 1$$

si calcolino

(2a) l'intervallo di confidenza per la media di livello 0.95.

(2b) l'intervallo di confidenza per la media di livello 0.99.

N.B.: si noti che a differenza dell'esercizio precedente la deviazione standard non è nota e dunque dovrà essere stimata con la deviazione standard campionaria.

Ex. 3 Ad un di un campione di 100 studenti universitari viene chiesto se sono fumatori o meno. 82 hanno affermato di non esserlo. Basandosi su tali informazioni, costruire un intervallo di confidenza di livello 0.99 per il parametro p definito come la proporzione di studenti universitari che non sono fumatori.

Ex. 4: Se 3 monete vengono lanciate insieme e tale tipo di lancio viene ripetuto 200 volte con i seguenti esiti

- per 20 lanci in nessuna delle tre monete è comparsa una testa;
- per 63 lanci in una sola moneta è comparsa testa;
- per 84 lanci in due sole monete è comparsa testa;
- per 35 lanci in tutte e tre le monete è comparsa testa.

Si usi il test χ^2 per discutere se le monete siano o meno truccate.

Ex. 5: Date le seguenti coppie di dati per le variabili (X, Y)

$$(1,4),\ (2,7),\ (3,8),\ (5,12)\,.$$

Si stimi la retta di regressione (dove $X = x$ è la variabile data in input, mentre Y è la risposta all'input x) e l'errore commesso con l'approssimazione lineare. Si stimi il valore atteso di Y condizionata a $X = 4$ e si determini un intervallo di confidenza per tale stima, con livello di confidenza 0.95.

2. SOLUZIONI

Ex. 1:(1a) Indichiamo con X_i la variabile aleatoria che denota la quantità di nicotina contenuta in una sigaretta. Dunque le $\{X_i\}_i$ sono i.i.d.. Assumiamo che $X_i \sim \mathcal{N}(\mu, \sigma^2)$ con $\sigma = 0.7$ (in realtà essendo grande la taglia del campione, cioè $n = 44$, tale assunzione non è necessaria in virtù del teorema del limite centrale e del fatto che al posto della varianza della popolazione si potrebbe mettere la varianza campionaria). Definiamo

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

Sappiamo che l'intervallo di confidenza per la media μ di livello $1-\alpha$ è dato da

$$\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]\,,$$

dove $z_\alpha = \Phi^{-1}(1-\alpha)$ e Φ è la CDF di $Z \sim \mathcal{N}(0,1)$. Nel nostro caso $\bar{X}_n = 1.74$ mg, $n = 44$, $\sigma = 0.7$ mg ed

$$1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow z_{\alpha/2} = z_{0.025} = 1.96\,.$$

Dunque l'intervallo diventa

$$\left[1.74 - \frac{0.7\cdot 1.96}{44}, 1.74 + \frac{0.7\cdot 1.96}{44}\right] = [1.533, 1.947]\,.$$

(1b) Dobbiamo imporre

$$1.8 = \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \Rightarrow z_{\alpha/2} = \sqrt{n}\cdot\frac{1.8-\bar{X}_n}{\sigma} = \sqrt{44}\cdot\frac{0.06}{0.7} \approx 0.57\,,$$

da cui

$$1-\alpha/2 = \Phi(z_{\alpha/2}) \approx \Phi(0.57) \approx 0.72 \Rightarrow \alpha \approx 0.56\,.$$

(1c) La lunghezza dell'intervallo di confidenza per la media μ di livello $1-\alpha$ è data da

$$2\frac{\sigma}{\sqrt{n}} z_{\alpha/2}\,.$$

Nel nostro caso $\sigma = 0.7$ e $z_{\alpha/2} = 1.96$. Vogliamo che l'intervallo sia lungo al più 0.3. Dunque imponiamo

$$2\frac{\sigma}{\sqrt{n}} z_{\alpha/2} = 2\frac{0.7}{\sqrt{n}}\cdot 1.96 \le 0.3 \Rightarrow n \ge 83.7\,.$$

Poiché n è intero, ciò ci dice che n deve essere almeno 84.

Ex. 2: Ricordiamo che se $x_1, \dots, x_n$ sono un set di n dati, allora

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad S_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X}_n)^2 .$$

Per il set di dati fornito dall'esercizio si ha ($n = 20$)

$$\bar{X}_{20} = 5.8 , \qquad S_{20} = 5.085 .$$

Ricordiamo che, nel caso in cui la deviazione standard sia ignota, l'intervallo di confidenza per la media μ di livello $1-\alpha$ è dato da

$$\left[\bar{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1,\alpha/2}\right] ,$$

dove $t_{n,\alpha}$ è definito come il valore per cui $\mathbb{P}(T_n > t_{n,\alpha}) = \alpha$ e dove T_n è una variabile con distribuzione t-Student con n gradi di libertà. Si ricorda che per calcolare la probabilità $\mathbb{P}(T_n > t_{n,\alpha}) = \alpha$ bisogna utilizzare le tavole della distribuzione t-Student con n gradi di libertà.

(2a) Dobbiamo porre

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow t_{n-1,\alpha/2} = t_{19,0.025} = 2.093 ,$$

e dunque l'intervallo di confidenza per la media μ di livello 0.95 è dato da

$$\left[5.8 - \frac{5.085}{\sqrt{20}} \cdot 2.093, 5.8 + \frac{5.085}{\sqrt{20}} \cdot 2.093\right] = [3.42, 8.18] .$$

(2b) Dobbiamo porre

$$1 - \alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow t_{n-1,\alpha/2} = t_{19,0.005} = 2.861 ,$$

e dunque l'intervallo di confidenza per la media μ di livello 0.99 è dato da

$$\left[5.8 - \frac{5.085}{\sqrt{20}} \cdot 2.861, 5.8 + \frac{5.085}{\sqrt{20}} \cdot 2.861\right] = [2.55, 9.05] .$$

Ex. 3: Posso assumere che le variabili $X_1, \dots, X_{100}$ definite come

$$X_i = \begin{cases} 1, & \text{se l'}i\text{-esimo studente non fuma;} \\ 0, & \text{altrimenti;} \end{cases}$$

siano i.i.d. con distribuzione Bernoulliana di parametro p. Dobbiamo trovare un intervallo di confidenza per p.

Chiamiamo $\hat{p}$ una stima per p. Un'opzione per il valore di $\hat{p}$ è data da 0.82 (per i dati forniti nel testo dell'esercizio).

Notiamo che $\hat{p}(1-\hat{p})$ è una stima consistente della varianza della popolazione per cui applicando il teorema del limite centrale e il lemma di Slutzky si ha che

$$\mathbb{P}\left(\frac{\sum_{i=1}^{100} X_i - 100p}{10\sqrt{p(1-p)}} \cdot \sqrt{\frac{p(1-p)}{\hat{p}(1-\hat{p})}} \le t\right) \approx \mathbb{P}(Z \le t) ,$$

dove $Z \sim \mathcal{N}(0,1)$. Poichè $X \sim \mathrm{Ber}(0.82)$, si ha

$$\mathbb{E}[X] = 0.82, \qquad \mathrm{Var}(X) = 0.82 \cdot (1 - 0.82) \approx 0.1476 \Rightarrow \sigma = \sqrt{0.1476} = 0.3842 .$$

Notiamo che

$$1-\alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow z_{\alpha/2} = z_{0.005} = 2.576\,.$$

Dunque un intervallo di confidenza per p a livello 0.99 è dato da

$$\left[0.82 - \frac{0.3842}{\sqrt{100}} \cdot 2.576, 0.82 + \frac{0.3842}{\sqrt{100}} \cdot 2.576\right] \approx [0.721, 0.919]\,.$$

Ex. 4: Se le tre monete non fossero truccate, sapremmo che

- la probabilità di ottenere in un lancio 0 teste (sulle 3 monete) è

$$p_0 = \binom{3}{0}(0.5)^0(0.5)^{3-0} = 0.5^3$$

  e dunque il numero medio di lanci in cui si ottengono 0 teste (sulle 3 monete) è pari a

$$p_0 \cdot 200 = 0.5^3 \cdot 200 = 25\,.$$

- la probabilità di ottenere in un lancio 1 testa (sulle 3 monete) è

$$p_1 = \binom{3}{1}(0.5)^1(0.5)^{2-0} = 3 \cdot 0.5^3$$

  e dunque il numero medio di lanci in cui si ottengono 1 testa (sulle 3 monete) è pari a

$$p_1 \cdot 200 = 3 \cdot 0.5^3 \cdot 200 = 75\,.$$

- la probabilità di ottenere in un lancio 2 teste (sulle 3 monete) è

$$p_2 = \binom{3}{2}(0.5)^2(0.5)^{1-0} = 3 \cdot 0.5^3$$

  e dunque il numero medio di lanci in cui si ottengono 2 teste (sulle 3 monete) è pari a

$$p_2 \cdot 200 = 3 \cdot 0.5^3 \cdot 200 = 75\,.$$

- la probabilità di ottenere in un lancio 3 teste (sulle 3 monete) è

$$p_3 = \binom{3}{3}(0.5)^3(0.5)^{3-3} = 0.5^3$$

  e dunque il numero medio di lanci in cui si ottengono 3 teste (sulle 3 monete) è pari a

$$p_3 \cdot 200 = 0.5^3 \cdot 200 = 25\,.$$

Calcoliamo χ^2

$$\chi^2 = \frac{(20-25)^2}{25} + \frac{(63-75)^2}{75} + \frac{(84-75)^2}{75} + \frac{(20-25)^2}{25} = 8\,.$$

Sono state fatte 4 osservazioni (ovvero 4 possibili valori del numero di teste in 3 lanci) e dunque i gradi di libertà sono 3. Usando le tabelle per la distribuzione χ^2 con 3 gradi di libertà si ha che

- $\chi^2 > 7.815$ con probabilità 0.05, ovvero il p-value associato a 7.815 è 0.05;

– $\chi^2 > 9.348$ con probabilità 0.025, ovvero il p-value associato a 9.348 è 0.025.

Dunque il p-value associato a 8 è nell'intervallo $(0.025, 0.05)$

Ex. 5: Il modello della regressione standard assegna la legge della risposta Y, condizionata alla covariata $X = x$, uguale a $\mathcal{N}(\alpha + \beta x, \sigma^2)$.

Supponiamo di avere un set di dati

$$(x_1, Y_1), \dots, (x_n, Y_n)$$

e definiamo

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \,,$$

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \,,$$

allora la retta di regressione stimata è data da

$$Y = \hat{\alpha} + \hat{\beta} x \,,$$

dove

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \,,$$

è una stima per β e

$$\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n \,,$$

è una stima per α.

Dunque (essendo $n = 4$)

$$\bar{x}_n = \frac{11}{4} = 2.75 \,, \qquad \bar{Y}_n = 7.75 \,,$$

$$\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \approx 8.75 \,. \tag{1}$$

$$\sum_{i=1}^{n}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n) = 16.75 \,.$$

Quindi

$$\hat{\beta} = \frac{16.75}{8.75} \approx 1.91 \,.$$

$$\hat{\alpha} = 7.75 - 1.91 \cdot 2.75 = 2.5 \,.$$

Dunque la retta di regressione stimata è

$$Y = 2.5 + 1.91 \cdot x \,.$$

Si veda la figura sotto per una rappresentazione grafica della retta e dei dati.

Per ciò detto in precedenza, il modello di regressione standard assegna la legge di Y, condizionata a $X = x$, uguale a $\mathcal{N}(2.5 + 1.91x, \sigma^2)$, dove σ^2 è stimato da

$$\frac{\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n-2} \approx 0.34 \,. \tag{2}$$

Dunque

$$\mathbb{E}[Y \mid X = 4] = 2.5 + 1.91 \cdot 4 = 10.14 \,.$$

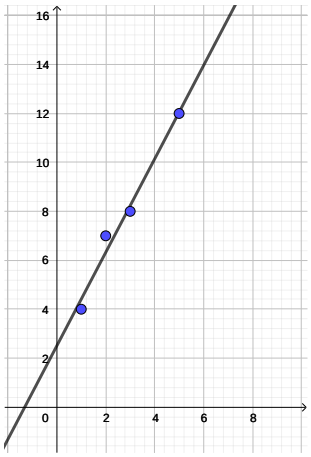

FIGURE 1. I pallini blu corrispondono ai dati forniti mentre in nero è disegnata la retta di regressione calcolata nell'esercizio.

Dobbiamo ora determinare un intervallo di confidenza per il valore atteso di Y condizionata a $X = 4$, con livello di confidenza 0.95. Sappiamo che se il livello fosse $1 - \alpha$, tale intervallo sarebbe (si ricordi che $n = 4$)

$$\left[10.14 - t_{\alpha,n-2} \cdot \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(4-\bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right)}, 10.14 + t_{\alpha,n-2} \cdot \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(4-\bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right)}\right],$$

dove σ^2 è stata stimata in (2) con 0.34. Nel nostro caso $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$, $n = 4$ e $t_{0.05,2} = 4.303$. Dunque l'intervallo diventa (si ricordi (1))

$$\left[10.14 - 4.303 \cdot \sqrt{0.34 \left(\frac{1}{4} + \frac{(4-2.75)^2}{8.75}\right)}, 10.14 + 4.303 \cdot \sqrt{0.34 \left(\frac{1}{4} + \frac{(4-2.75)^2}{8.75}\right)}\right] =$$

$$= [10.14 - 1.643, 10.14 + 1.643] = [8.497, 11.783] .$$