

**Esercizi di Statistica della 4<sup>a</sup> settimana (Corso di Laurea in Biotecnologie, Università degli Studi di Padova).**

**Esercizio 1.** Il cosiddetto "test del DNA" non fa altro che misurare la lunghezza di  $K$  geni, senza controllare le basi azotate che li compongono. Per ognuno di tali geni, la probabilità che due dati individui presentino una lunghezza uguale viene assunta come pari a  $1/10$ . Un'altra ipotesi che viene comunemente fatta è che le lunghezze di geni diversi siano indipendenti l'una dall'altra.

Supponiamo di misurare la lunghezza di  $K = 7$  geni da un campione di DNA trovato su una "scena del crimine".

1. Calcolare la probabilità che un dato individuo abbia la lunghezza dei suoi geni uguali a quella del campione.
2. Supponendo di avere un database di  $n = 101905$  individui, calcolare la probabilità che almeno uno di questi individui abbia le lunghezze dei suoi geni uguali a quelle del campione incriminato.
3. Calcolare la probabilità che almeno due di questi individui abbiano le lunghezze dei loro geni uguali a quelle del campione incriminato.
4. Supponendo di aver trovato un individuo con le lunghezze dei geni uguali a quelle del campione incriminato, calcolare la probabilità che ce ne sia almeno un altro.

**Esercizio 2.** Sia  $X \sim B(30; 1/3)$  e supponiamo di voler calcolare le probabilità  $\mathbb{P}\{X \leq 15\}$  e  $\mathbb{P}\{X \geq 16\}$ .

1. Dire se si può utilizzare l'approssimazione normale, motivando la risposta.
2. Calcolare  $\mathbb{P}\{X \leq 15\}$ .
3. Calcolare  $\mathbb{P}\{X \geq 16\}$  e confrontare il risultato con quanto ottenuto al punto 2.
4. Come si fa a far venire i due risultati sopra in modo che la loro somma dia esattamente 1?

**Esercizio 3.** Si supponga che il peso (in tonnellate) di un autoveicolo si distribuisca come una variabile aleatoria di media 3 e deviazione standard 0.3. Nel seguito, supporremo di poter applicare l'approssimazione normale.

1. Se consideriamo  $n$  autoveicoli con " $n$  grande", con che variabile aleatoria possiamo approssimare il peso totale?
2. Supponiamo che la portata della campata di un ponte sia 400 tonnellate, prima di riportare danni strutturali. Se il numero massimo di veicoli che ci possono transitare contemporaneamente è uguale a 100, qual è la probabilità che si possa danneggiare?
3. Rispondere alla stessa domanda supponendo che la portata della campata sia una variabile aleatoria gaussiana di media 400 e di deviazione standard 40.

4. Supponiamo di voler controllare se l'assunzione iniziale (media = 3, dev. standard = 0.3) era corretta. Pesando 10 autoveicoli, otteniamo i seguenti valori (in tonnellate):

2.53 1.91 3.01 6.12 3.42 2.95 3.24 2.89 4.52 3.07

Quali sono la media e la varianza stimate da questo campione?

**Esercizio 4.** Supponiamo di voler studiare la relazione tra abuso di analgesici e livello di creatinina nel sangue. In particolare, consideriamo 15 persone che lavorano in una fabbrica e sono conosciuti per “abuso di analgesici” (cioè più di 10 pillole al giorno) e misuriamo il loro livello di creatinina, con i seguenti risultati:

0.9, 1.1, 1.6, 2.0, 0.8, 0.7, 1.4, 1.2, 1.5, 0.8, 1.0, 1.1, 1.4, 2.2, 1.4

1. Stimare media, deviazione standard ed errore standard della media del livello di creatinina in base ai dati sopra.
2. Supponendo di sapere che la vera deviazione standard è uguale a quella stimata, calcolare l'intervallo di confidenza al 95% della media.
3. Supponendo sempre di conoscere la vera deviazione standard, quanto grande dovremo prendere  $n$  in modo che l'intervallo di confidenza sia largo meno di 0.1?
4. Supponiamo ora di sapere che nella popolazione “normale” il livello di creatinina sia 1.0. Dire, in base all'intervallo di confidenza, se si ritiene plausibile che il livello medio sia 1.0 anche in questo campione.

**Esercizio 5.** L'ente che gestisce un tratto di autostrada conserva sale a sufficienza per eliminare un tratto di 80 pollici di neve. Supponiamo che la quantità di neve che cade al giorno sia una variabile aleatoria di media 1.5 pollici e deviazione standard di 0.3 pollici.

1. Trova la probabilità approssimata che il sale a disposizione basti per 50 giorni.
2. Quali sono le ipotesi fatte per rispondere al punto 1.? Possono ritenersi giustificate?
3. Supponiamo che nei primi 10 giorni del periodo siano caduti in totale 20 pollici di neve. Supponendo che la deviazione standard sia sempre pari a 0.3, calcolare l'intervallo di confidenza al 99% per la media.
4. Possiamo ancora supporre che la media sia 1.5? I calcoli del punto 1. sono quindi ancora giustificati? (non eseguire nuovi calcoli)

Soluzioni su <http://www.math.unipd.it/~vargiolu/Statistica/>

## Soluzioni

### Esercizio 1.

1. Definiamo gli eventi

$$\begin{aligned} A_j &:= \{ \text{la lunghezza del } j\text{-esimo gene dell'individuo è uguale a quella del campione} \}, \quad j = 1, \\ B &:= \{ \text{le lunghezze di tutti i geni dell'individuo sono uguali a quelle del campione} \} \end{aligned}$$

Allora l'evento  $B$  ha probabilità pari a

$$\mathbb{P}(B) = \mathbb{P}\left(\bigcap_{j=1}^K A_j\right) = \prod_{j=1}^K \mathbb{P}(A_j) = \frac{1}{10^K} = 10^{-7}$$

dove abbiamo usato l'indipendenza delle lunghezze dei geni, e quindi degli eventi  $(A_j)_j$ .

2. Definiamo, per ogni individuo  $i = 1, \dots, n$ , l'evento  $B_i := \{ \text{le lunghezze di tutti i geni dell}'i\text{-esimo individuo sono uguali a quelle del campione} \}$ . Allora  $\mathbb{P}(B_i) = 10^{-7}$  dal punto 1. Definiamo poi le variabili aleatorie di Bernoulli  $X_i := \mathbf{1}_{B_i}$ , che si possono supporre indipendenti tra di loro, e la variabile aleatoria  $S_n := \sum_{i=1}^n X_i$ , che per l'indipendenza delle  $X_i$  ha legge  $B(n, p)$ , con  $p := 10^{-7}$ . Poichè nel nostro caso  $n = 101905 > 100$  e  $p = 10^{-7} < 0.01$ , possiamo utilizzare l'approssimazione di Poisson, e quindi  $S_n \approx Po(\lambda)$ , con  $\lambda = np = 0.01019$ . Allora bisogna calcolare

$$\mathbb{P}\{S_n \geq 1\} = 1 - \mathbb{P}\{S_n < 1\} = 1 - \mathbb{P}\{S_n = 0\} \simeq 1 - e^{-\lambda} \simeq 0.01019$$

3. Bisogna calcolare

$$\mathbb{P}\{S_n \geq 2\} = 1 - \mathbb{P}\{S_n = 0\} - \mathbb{P}\{S_n = 1\} \simeq 1 - e^{-\lambda} - \lambda e^{-\lambda} = 0.00005192$$

4. Bisogna calcolare

$$\mathbb{P}\{S_n \geq 2 \mid S_n \geq 1\} = \frac{\mathbb{P}(\{S_n \geq 2\} \cap \{S_n \geq 1\})}{\mathbb{P}\{S_n \geq 1\}} = \frac{\mathbb{P}\{S_n \geq 2\}}{\mathbb{P}\{S_n \geq 1\}} \simeq 0.00509$$

### Esercizio 2.

1. Calcoliamo  $np(1-p) = 30 \cdot \frac{1}{3} \cdot \frac{2}{3} = 6.66 > 5$ , quindi si può usare l'approssimazione normale.

2. Usando la correzione di continuità, possiamo calcolare

$$\mathbb{P}\{X \leq 15\} = \mathbb{P}\{X \leq 15.5\} = \mathbb{P} \simeq F_Z \left( \frac{15.5 - 30 \cdot \frac{1}{3}}{\sqrt{30 \cdot \frac{1}{3} \cdot \frac{2}{3}}} \right) = F_Z \left( \frac{5.5}{\sqrt{6.66}} \right) = F_Z(2.13) = 0.9834$$

dove  $Z$  è una generica variabile aleatoria  $N(0, 1)$ .

3. Ancora utilizzando la correzione di continuità, calcoliamo

$$\begin{aligned}\mathbb{P}\{X \geq 16\} &= \mathbb{P}\{X \leq 15.5\} = \mathbb{P}\left\{\frac{X - 30 \cdot \frac{1}{3}}{\sqrt{30 \cdot \frac{1}{3} \cdot \frac{2}{3}}} \geq \frac{15.5 - 30 \cdot \frac{1}{3}}{\sqrt{30 \cdot \frac{1}{3} \cdot \frac{2}{3}}}\right\} = \mathbb{P}\left\{X^* \geq \frac{5.5}{\sqrt{6.66}}\right\} = \\ &= 1 - \mathbb{P}\{X^* < 2.13\} \simeq 1 - F_Z(2.13) = 1 - 0.9834 = 0.0166\end{aligned}$$

$$\text{dove } X^* = \frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}[X]}}.$$

Gli eventi  $\{X \leq 15\}$  e  $\{X \geq 16\}$  sono complementari se  $X$  è una binomiale, e in effetti, poichè abbiamo usato la correzione di continuità, le loro probabilità calcolate con l'approssimazione normale hanno somma  $0.9834 + 0.0166 = 1$ . Se invece non avessimo usato la correzione di continuità ma avessimo approssimato la binomiale con una normale in modo "ingenuo", saremmo incappati nel problema che gli eventi  $\{X \leq 15\}$  e  $\{X \geq 16\}$  non includono  $\{15 < X < 16\}$ , e usando l'approssimazione normale senza correzione di continuità questo evento ha probabilità bassa (circa 0.01) ma positiva.

4. Il modo corretto è usare la correzione di continuità, come è stato fatto.

### Esercizio 3.

1. Chiamiamo  $X_i$  il peso dell' $i$ -esimo veicolo. Allora  $\mathbb{E}[X_i] = 3$ ,  $\text{Var}[X_i] = 0.3^2$ , ed è ragionevole supporre che le  $(X_i)_i$  siano indipendenti. Se consideriamo  $n$  veicoli, allora il loro peso totale sarà uguale a  $S_n := \sum_{i=1}^n X_i$ , e quindi sicuramente  $\mathbb{E}[S_n] = \sum_{i=1}^n \mathbb{E}[X_i] = 3n$ , e  $\text{Var}[S_n] = \sum_{i=1}^n \text{Var}[X_i] = 0.3^2 n$ . Se ora supponiamo di poter utilizzare l'approssimazione normale, allora possiamo approssimare la legge di  $T$  con una normale  $N(3n, 0.09n)$ .

2. Utilizzando l'approssimazione normale, calcoliamo

$$\begin{aligned}\mathbb{P}\{S_{100} > 400\} &= \mathbb{P}\left\{\frac{S_{100} - 300}{0.3\sqrt{100}} > \frac{400 - 300}{3}\right\} = 1 - \mathbb{P}\left\{\frac{S_{100} - 300}{3} \leq 33.3\right\} \simeq \\ &\simeq 1 - F_Z(33.3) < 1 - F_Z(3.49) = 1 - 0.9998 = 0.0002\end{aligned}$$

dove ci siamo dovuti limitare all'ultimo valore in tavola. Disponendo di un calcolatore, possiamo calcolare una stima più precisa:

$$\mathbb{P}\{S_{100} > 400\} \simeq 1 - F_Z(33.3) < 1 - F_Z(8.25) = 1.1102 \cdot 10^{-16}$$

3. Definiamo  $X \sim N(400, 40^2)$  indipendente da  $S_n$ . Allora la legge di  $S_n - X$  si può approssimare con una legge  $N(3n - 400, 0.09n + 40^2)$ . Dobbiamo ora calcolare

$$\begin{aligned}\mathbb{P}\{S_{100} > X\} &= \mathbb{P}\{S_{100} - X > 0\} = \mathbb{P}\left\{\frac{S_{100} - X - (300 - 400)}{\sqrt{9 + 1600}} > \frac{400 - 300}{\sqrt{9 + 1600}}\right\} = \\ &\simeq 1 - F_Z\left(\frac{100}{40.1}\right) = 1 - F_Z(2.49) = 1 - 0.99361 = 0.00639\end{aligned}$$

4. Calcoliamo media e deviazione standard usando i corrispondenti stimatori:

$$\bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i = 3.07, \quad s_X = \sqrt{\frac{1}{9} \left( \sum_{i=1}^{10} X_i^2 - 10 \cdot \bar{X}^2 \right)} = 0.66$$

**Esercizio 4.**

1. Le stime richieste sono:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = 1.27, \quad s_X = \sqrt{\frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n \bar{X}^2 \right)} = 0.435, \quad s_{\bar{X}} = \frac{s_X}{\sqrt{n}} = 0.112$$

2. Supponendo che  $\sigma = 0.435$ , l'intervallo di confidenza per la media ha estremi

$$\bar{X} \pm q_{0.05/2} \frac{\sigma}{\sqrt{n}} = 1.27 \pm 1.96 \cdot 0.112 = 1.27 \pm 0.22$$

e quindi è uguale a  $[1.05; 1.49]$ .

3. L'intervallo di confidenza è largo

$$\bar{X} + q_{0.05/2} \frac{\sigma}{\sqrt{n}} - \left( \bar{X} - q_{0.05/2} \frac{\sigma}{\sqrt{n}} \right) = 2q_{0.05/2} \frac{\sigma}{\sqrt{n}}$$

Imponendo che la larghezza sia minore di 0.1, otteniamo

$$2q_{0.05/2} \frac{\sigma}{\sqrt{n}} < 0.1 \implies n \geq \left( 2q_{0.05/2} \frac{\sigma}{0.1} \right)^2 = \left( 2 \cdot 1.96 \cdot \frac{0.435}{0.1} \right)^2 = 290.77$$

e quindi per avere un intervallo di confidenza ampio meno di 0.1 bisogna avere  $n \geq 291$ .

4. Siccome 1.0 non appartiene all'intervallo di confidenza, a cui la vera media appartiene con probabilità 0.95, riteniamo poco plausibile che la media di questo campione sia 1.0.

**Esercizio 5.** Se chiamiamo  $X_i$  la variabile aleatoria che designa la quantità di neve per l' $i$ -esima giornata,  $i = 1, \dots, 50$ , allora abbiamo che  $\mathbb{E}[X_i] = 1.5$ ,  $\text{Var}[X_i] = 0.3^2$ .

1. Bisogna calcolare

$$\mathbb{P} \left\{ \sum_{i=1}^{50} X_i \leq 80 \right\} = \mathbb{P} \left\{ \frac{\sum_{i=1}^{50} X_i - 1.5 \cdot 50}{0.3\sqrt{50}} \leq \frac{80 - 75}{2.1213} \right\}$$

Supponendo di poter applicare l'approssimazione normale (vedi punto 2.), abbiamo

$$\mathbb{P} \left\{ \sum_{i=1}^{50} X_i \leq 80 \right\} \simeq \Phi(2.35) = 0.99061$$

2. Supponendo che le variabili aleatorie  $(X_i)_i$  siano i.i.d., il testo ci dice che hanno varianza finita: poichè il loro numero è pari a 50, può essere sufficiente a giustificare l'approssimazione normale.

3. In questo caso abbiamo

$$\bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i = \frac{20}{10} = 2$$

Supponendo che  $s_X = 0.3$ , l'intervallo di confidenza ha estremi

$$\bar{X} \pm \frac{\sigma}{\sqrt{10}} q_{0.01/2} = 2 \pm \frac{0.3}{\sqrt{10}} \cdot 2.57 = 2 \pm 0.308$$

e quindi è (1.692; 2.308).

4. Siccome 2 non appartiene all'intervallo di confidenza, a cui la vera media appartiene con probabilità 0.99, riteniamo poco plausibile che la media di questo campione sia 2.