Esercizi Unità 3

Analisi dei dati 2024/25

Cristiano Varin

1. In una recente ispezione in un ospedale, è stato misurato il rumore in decibel in 74 corridoi e stanze di ricovero. La media delle misure ottenute è stata 61.3 decibel e lo scarto quadratico medio 7.8 decibel. Si calcoli un intervallo di confidenza al 90% per la media dei decibel a cui sono esposti i ricoverati dell'ospedale considerato. (Tratto dal materiale didattico di Guido Masarotto, Università di Padova).

   Soluzione. Vista la numerosità campionaria, possiamo calcolare un intervallo di confidenza per il rumore medio con livello approssimato del 90% usando la statistica Z, ovvero

   $$\bar{X} \pm z_{0.05} \frac{S}{\sqrt{74}}.$$

   Questo intervallo avrà livello di confidenza esattamente pari al 90% se le osservazioni sono normalmente distribuite. Nel caso specifico abbiamo $\bar{x} = 61.3$ e $s = 7.8$ decibel. Siccome $z_{0.05} = 1.64$, l'intervallo di confidenza è $[59.81, 62.79]$ decibel.

   Calcoli con **R**:

```
z <- qnorm(.95)
z

## [1] 1.644854

61.3 + c(-1, 1) * z * 7.8 / sqrt(74)

## [1] 59.80856 62.79144
```

2. (Continuazione dell'esercizio precedente) È stato misurato il rumore in decibel in 85 corridoi e stanze di ricovero di un secondo ospedale. La media delle misure è stata 59.7 decibel con uno scarto quadratico medio di 8.1 decibel. Si calcoli un intervallo di confidenza al 90% per la differenza del rumore medio nei due ospedali. Si discuta il risultato ottenuto.

   Soluzione. In questo caso siamo interessati al parametro

   $$\theta = \mu_x - \mu_y,$$

   dove μ_x indica il rumore medio nel primo ospedale mentre μ_y nel secondo. Uno stimatore non distorto di questa differenza è $\hat{\theta} = \bar{X} - \bar{Y}$. La stima di θ con i dati osservati è $\hat{\theta} =$

$\bar{x} - \bar{y} = 61.3 - 59.7 = 1.6$ decibel. Viste le due numerosità campionarie, possiamo calcolare un intervallo di confidenza per θ con livello approssimato del 90% usando la statistica Z, ovvero

$$\hat{\theta} \pm z_{0.05}\, \widehat{\mathrm{SE}}(\hat{\theta}).$$

L'errore standard stimato di $\hat{\theta}$ è

$$\widehat{\mathrm{SE}}(\hat{\theta}) = \sqrt{\frac{7.8^2}{74} + \frac{8.1^2}{85}} = 1.26.$$

L'intervallo di confidenza per θ al 90% è, quindi,

$$1.6 \pm 1.64\ (1.26) = [-0.47, 3.67] \text{ decibel.}$$

Siccome l'intervallo contiene sia valori negativi che positivi, allora non possiamo escludere che il livello di rumore medio nei due ospedali sia lo stesso ad un livello di confidenza del 90%.

Calcoli con **R**:

```
z <- qnorm(.95)
z

## [1] 1.644854

(61.3 - 59.7) + c(-1, 1) * z * sqrt(7.8 ^ 2 / 74 + 8.1 ^ 2 / 85)

## [1] -0.4767178  3.6767178
```

3. Viene condotto un esperimento per valutare quanto sia efficiente un algoritmo nel risolvere un dato problema. L'esperimento è condotto dieci volte e l'efficienza dell'algoritmo viene valutata utilizzando il tempo, misurato in secondi CPU, necessario per risolvere il problema. I tempi osservati nell'esperimento sono

$$9, 9, 6, 10, 3, 9, 4, 5, 3, 7.$$

   (a) Assumendo che i tempi formino un campione casuale semplice da una distribuzione normale di media e varianza ignote, si costruisca un intervallo di confidenza di livello 99% per il tempo necessario a risolvere il problema.

   (b) Quanto dovrebbe essere grande il campione per far sì che il margine d'errore dell'intervallo di confidenza non sia più grande di 0.5 secondi CPU? Si risponda utilizzando la stima della varianza ottenuta con i dati del campione a disposizione.

Soluzione

   (a) Vista l'assunzione di normalità, calcoliamo un intervallo di confidenza per la media basato sulla statistica T, ovvero

$$\bar{X} \pm t_{0.005} \frac{S}{\sqrt{10}},$$

dove $t_{0.005}$ è il quantile di posizione 0.995 della distribuzione T con 9 gradi di libertà. Con il campione osservato abbiamo $\bar{x} = 6.5$ e $s = 2.68$ e $t_{0.005} = 3.25$. Con questi valori si ottiene l'intervallo $[3.75, 9.25]$.

Calcoli con **R**:

```
x <- c(9,9, 6,10, 3, 9, 4, 5, 3, 7)
mean(x)

## [1] 6.5

sd(x)

## [1] 2.677063

t <- qt(.995, df = 9)
t

## [1] 3.249836

mean(x) + c(-1, 1) * t * sd(x) / sqrt(10)

## [1] 3.748814 9.251186
```

(b) Per assicurare il margine d'errore richiesto, la numerosità campionaria dovrebbe soddisfare la diseguaglianza

$$n \geq \left(\frac{z_{0.005}\,\sigma}{0.5}\right)^2 = \left(\frac{2.58}{0.5}\right)^2 \sigma^2 = 26.63\,\sigma^2.$$

Non conosciamo σ^2 ma il testo dell'esercizio ci suggerisce di sostituirlo con la sua stima calcolata con i dati a disposizione, ovvero con $s^2 = 7.17$. La diseguaglianza diventa

$$n \geq 26.63\,(7.17) = 190.94,$$

ovvero il campione dovrebbe avere almeno 191 osservazioni.

4. L'azienda *The Machine Learning Company* afferma che il suo software per il riconoscimento facciale ha una accuratezza del 97%. Prima di acquistare il software, si decide di effettuare un esperimento su un campione casuale di 500 fotografie trovando che il software individua correttamente i volti in 479 fotografie.

   (a) Si calcoli un intervallo di confidenza con livello approssimato del 95% per il grado di accuratezza del software. L'intervallo permette di affermare che il livello di accuratezza dichiarato dall'azienda sia corretto?

   (b) Un'altra azienda produce un software concorrente che individua correttamente i volti in 488 fotografie tratte da un altro campione casuale di 500 fotografie. Questo secondo campione di fotografie ha lo stesso grado di difficoltà di quello precedente. Il risultato ottenuto nell'esperimento casuale ci permette di affermare che vi sia una qualche differenza nell'accuratezza nell'individuare i volti da parte dei due software?

Soluzione

(a) Sia $X_1, \ldots, X_{500}$ il campione casuale semplice utilizzato per valutare l'accuratezza del software. Le variabili X_i assumono il valore 1 se il volto nella fotografia viene correttamente individuato e 0 se, invece, non viene correttamente individuato. Si tratta, quindi, di un campione casuale semplice da una distribuzione Bernoulliana di parametro p che rappresenta la probabilità di corretta classificazione del volto.
Uno stimatore non distorto di p è la media campionaria $\hat{p} = \bar{X}$. Vista la numerosità campionaria, il "Teorema del limite centrale" ci assicura che $\hat{p}$ ha distribuzione approssimativamente normale. Quindi, l'intervallo di confidenza basato sulla statistica Z con livello approssimativamente pari al 95% è

$$\hat{p} \pm z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{500}} = \bar{X} \pm z_{0.025}\sqrt{\frac{\bar{X}(1-\bar{X})}{500}}.$$

Abbiamo che $z_{0.025} = 1.96$ e la media campionaria vale $\bar{x} = 479/500 = 0.96$, quindi l'intervallo di confidenza approssimato è $[0.94, 0.98]$. Siccome l'intervallo contiene il valore 0.97 non possiamo escludere che il livello di accuratezza dichiarato dall'azienda sia corretto.

Calcoli con **R**:

```
p.hat <- 479 / 500
z <- qnorm(0.975)
z

## [1] 1.959964

p.hat + c(-1, 1) * z * sqrt(p.hat * (1 - p.hat) / 500)

## [1] 0.9404179 0.9755821
```

(b) Per rispondere alla domanda, costruiamo un intervallo di confidenza per la differenza delle probabilità di individuare correttamente un volto con i due software, ovvero un intervallo per il parametro

$$\theta = p_1 - p_2,$$

dove p_1 indica la probabilità di corretta classificazione del software prodotto da *The Machine Learning Company* e p_2 invece la probabilità di corretta classificazione del software prodotto dall'azienda concorrente. Uno stimatore non distorto di θ è $\hat{\theta} = \hat{p}_1 - \hat{p}_2$, dove $\hat{p}_1 = \bar{X}$ e $\hat{p}_2 = \bar{Y}$ sono le proporzioni campionarie di corretta classificazione per *The Machine Learning Company* (X) e l'azienda concorrente (Y). La stima di θ con i dati osservati è $\hat{\theta} = 479/500 - 488/500 = -0.018$. Viste le numerosità dei due campioni, possiamo invocare il "Teorema del limite centrale" e costruire l'intervallo confidenza basato sulla statistica Z con livello di confidenza approssimativamente pari al 95%,

$$\hat{p}_1 - \hat{p}_2 \pm z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{500} + \frac{\hat{p}_2(1-\hat{p}_2)}{500}}.$$

Con i dati osservati otteniamo l'intervallo $[-0.04, 0.004]$. Siccome l'intervallo di confidenza contiene sia valori positivi che negativi non possiamo dire, al livello di confidenza del 95%, che vi sia una significativa differenza fra il grado di accuratezza dei

sue software.

Calcoli con **R**:

```
p1.hat <- 479 / 500
p2.hat <- 488 / 500
z <- qnorm(0.975)
z

## [1] 1.959964

diff.p <- p1.hat - p2.hat
diff.p

## [1] -0.018

se.diff <- sqrt(p1.hat * (1 - p1.hat) / 500 + p2.hat * (1 - p2.hat)
/ 500)
diff.p + c(-1, 1) * z * se.diff

## [1] -0.040115482  0.004115482
```

5. Tradizionalmente il numero medio di richieste che arrivano ad un dato server è pari a 4.2 richieste al secondo. Osservando i dati, sembra plausibile che le richieste seguano una distribuzione di Poisson. Se in un campione casuale di 100 secondi si osservano complessivamente 440 richieste, possiamo ancora credere che il numero medio di richieste sia pari a 4.2?

   Soluzione. Dobbiamo calcolare un intervallo di confidenza per la media di una variabile casuale di Poisson $\lambda = \mathrm{E}(X)$. Lo stimatore $\hat{\lambda} = \bar{X}$ è uno stimatore non distorto di λ. Inoltre, il "Teorema del limite centrale" afferma che $\hat{\lambda}$ ha distribuzione approssimativamente normale, per cui possiamo considerare l'intervallo di confidenza approssimato basato sulla statistica Z,

$$\hat{\lambda} \pm z_{\alpha/2}\widehat{\mathrm{SE}}(\hat{\lambda}).$$

   Siccome la varianza di una variabile di Poisson è λ, allora l'errore standard di $\hat{\lambda} = \bar{X}$ è pari a $\sqrt{\lambda/n}$ e l'intervallo di confidenza approssimato è

$$\bar{X} \pm z_{\alpha/2}\sqrt{\frac{\bar{X}}{100}}.$$

   La media campionaria con i dati osservati vale $\bar{x} = 440/100 = 4.4$. Se consideriamo un intervallo con un livello di confidenza del 95% (questa è una scelta arbitraria visto che l'esercizio non specifica il livello di confidenza da considerare), allora il corrispondente intervallo di confidenza è $[3.99, 4.81]$. Siccome questo intervallo contiene il valore 4.2, non possiamo escludere con un grado di confidenza del 95% che il numero medio di richieste sia rimasto invariato.

Calcoli con **R**:

```
lambda.hat <- 440 / 100
z <- qnorm(0.975)
z

## [1] 1.959964

lambda.hat + c(-1, 1) * z * sqrt(lambda.hat / 100)

## [1] 3.988874 4.811126
```

6. Riprendiamo i dati relativi ai 23 lanci dello Space Shuttle Challenger precedenti alla sua esplosione. Vi sono stati $n_X = 16$ lanci senza danni agli o-ring e $n_Y = 7$ lanci con il cedimento di almeno un o-ring. La temperatura media nei lanci senza danni agli o-ring è stata $\bar{x} = 72.12$ °F con deviazione standard $s_X = 4.84$ °F, mentre la temperatura media nei lanci con cedimenti degli o-ring è stata $\bar{y} = 63.71$ °F con deviazione standard $s_Y = 8.16$ °F.

   Supponendo che le temperature siano normalmente distribuite:

   (a) Si calcoli un intervallo di confidenza al 95% per la differenza fra la temperatura media μ_X nei lanci in cui nessun o-ring è stato danneggiato e la temperatura media μ_y dei lanci in cui almeno un o-ring ha subito un cedimento strutturale assumendo che le varianze σ_X^2 e σ_Y^2 siano uguali.

   (b) Si ricalcoli l'intervallo senza l'assunzione che le varianze σ_X^2 e σ_Y^2 siano uguali.

   (c) Si interpretino i risultati ottenuti.

   Soluzione

   (a) Sotto le assunzioni di normalità e uguali varianze, possiamo calcolare l'intervallo per la differenza $\mu_X - \mu_Y$ con la seguente statistica T:

   $$(\bar{X} - \bar{Y}) \pm t_{0.025}\, S_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}},$$

   dove $t_{0.025}$ è il quantile di posizione 0.975 della distribuzione T di Student con $7 + 16 - 2 = 21$ gradi di libertà, ovvero $t_{0.025} = 2.08$. La quantità S_P^2 è la varianza 'pooled' pari a

   $$S_P^2 = \frac{1}{7 + 16 - 2}(15S_X^2 + 6S_Y^2).$$

   Con i dati a disposizione otteniamo la stima della varianza 'pooled'

   $$s_P^2 = \frac{1}{21}\{15(4.84^2) + 6(8.16^2)\} = 35.75.$$

   Quindi, l'intervallo di confidenza corrispondente ai dati osservati è

   $$(72.12 - 63.71) \pm 2.08 \sqrt{35.75\left(\frac{1}{16} + \frac{1}{7}\right)} = [2.77, 14.04]\ ^\circ\text{F}.$$

Calcoli con **R**:

```
t <- qt(1 - 0.025, df = 7 + 16 - 2)
t

## [1] 2.079614

(72.12 - 63.71) + c(-1, 1) * t * sqrt(35.75 * (1 / 16 + 1 / 7))

## [1]  2.775237 14.044763
```

(b) Se le varianze non sono uguali allora un intervallo di confidenza approssimato per la differenza $\mu_X - \mu_Y$ è dato dalla "Formula di Satterthwaite":

$$(\bar{X} - \bar{Y}) \pm t_{0.025} \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}},$$

dove $t_{0.025}$ è il quantile di posizione 0.975 della distribuzione T di Student con gradi di libertà

$$\nu = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{S_X^4}{n_X^2(n_X-1)} + \frac{S_Y^4}{n_Y^2(n_Y-1)}}.$$

Con i dati a disposizione troviamo

$$\nu = \frac{\left(\frac{4.84^2}{16} + \frac{8.16^2}{7}\right)^2}{\frac{4.84^4}{16^2(16-1)} + \frac{8.16^4}{7^2(7-1)}} = 7.91.$$

Possiamo calcolare il quantile $t_{0.025}$ della distribuzione T di Student con 7.91 gradi di libertà usando **R** trovando $t_{0.025} = 2.31$. Il corrispondente intervallo di confidenza è

$$(72.12 - 63.71) \pm 2.31 \sqrt{\frac{4.84^2}{16} + \frac{8.16^2}{7}} = [0.76, 16.06]\ ^\circ\text{F}.$$

Calcoli con **R**:

```
nu <- (4.84 ^ 2 / 16 + 8.16 ^ 2 / 7) ^ 2 /
      (4.84 ^ 4 / (16 ^ 2 * 15) + 8.16 ^ 4 / (7 ^ 2 * 6))
nu

## [1] 7.914159

t <- qt(1 - 0.025, df = nu)
t

## [1] 2.310365

(72.12 - 63.71) + c(-1, 1) * t * sqrt(4.84 ^ 2 / 16 + 8.16 ^ 2 / 7)

## [1]  0.7556351 16.0643649
```

(c) I due intervalli di confidenza calcolati ai punti precedenti contengono solamente valori positivi. Quindi, concludiamo che la differenza fra le temperature medie nei lanci senza e con danni agli o-ring è statisticamente rilevante ad un livello di confidenza del 95%.

7. Si risolvano i seguenti esercizi del libro di testo Baron (2014): 9.7 (a), 9.9 (a), 9.10 (a), 9.12 (a), 9.16 (a), 9.17, 9.18 (a).

   Soluzione. Le soluzioni degli esercizi sono disponibili nel documento "Soluzioni esercizi libro di testo Baron (2014)" pubblicato nella pagina Moodle del corso.

8. Si consideri l'esercizio 3 della seconda unità. Si calcoli un intervallo di confidenza per il parametro θ con un livello di confidenza del 99%.

   Soluzione. Sappiamo dalla soluzione dell'esercizio 3 della seconda unità che lo stimatore di massima verosimiglianza di θ è

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log X_i} - 1$$

   e che la derivata seconda della log-verosimiglianza è

$$\ell''(\theta) = -\frac{n}{(\theta+1)^2}.$$

   Quindi, possiamo stimare l'errore standard di $\hat{\theta}$ al crescere della dimensione campionaria con la quantità

$$\widehat{\mathrm{SE}}(\hat{\theta}) \approx \frac{\hat{\theta}+1}{\sqrt{n}}.$$

   Di conseguenza un intervallo di confidenza per θ di livello approssimativamente 99% è

$$\hat{\theta} \pm z_{0.005} \frac{\hat{\theta}+1}{\sqrt{n}}.$$

   Con i dati osservati otteniamo $\hat{\theta} = 100/32.71 - 1 = 2.06$, $\widehat{\mathrm{SE}}(\hat{\theta}) \approx (2.06+1)/\sqrt{100} = 0.31$, $z_{0.005} = 2.58$, per cui l'intervallo di confidenza approssimato è

$$2.06 \pm z_{0.005}(0.31) = [1.27, 2.84].$$

   Calcoli con **R**:

```
theta.hat <- 100 / 32.71 - 1
theta.hat

## [1] 2.057169

se.theta <- (theta.hat + 1) / sqrt(100)
se.theta
```

```
## [1] 0.3057169

z <- qnorm(0.995)
z

## [1] 2.575829

theta.hat + c(-1, 1) * z * se.theta

## [1] 1.269694 2.844644
```

9. Si consideri l'esercizio 5 della seconda unità. Si calcoli un intervallo di confidenza per il parametro θ con livello i confidenza del 90%.

   Soluzione. Sappiamo dalla soluzione dell'esercizio 5 della seconda unità che la stima di massima verosimiglianza di θ è $\hat{\theta} = 0.5$ e che la derivata seconda della log-verosimiglianza è

   $$\ell''(\theta) = -\frac{5}{\theta^2} - \frac{5}{(1-\theta)^2}.$$

   Quindi, la stima dell'errore standard di $\hat{\theta}$ è

   $$\widehat{\mathrm{SE}}(\hat{\theta}) \approx \frac{1}{\sqrt{-\ell''(\hat{\theta})}} = \frac{1}{\sqrt{40}} = 0.16.$$

   Di conseguenza un intervallo di confidenza per θ di livello approssimativamente 90% è

   $$0.5 \pm z_{0.05}\, 0.16 = [0.24, 0.76].$$

   Calcoli con **R**:

```
se.theta <- 0.5 / sqrt(2 * 5)
se.theta

## [1] 0.1581139

z <- qnorm(0.95)
z

## [1] 1.644854

0.5 + c(-1, 1) * z * se.theta

## [1] 0.2399258 0.7600742
```

10. Si considerino gli esercizi 6 e 7 della seconda unità. Si calcolino degli intervalli di confidenza per i parametri σ^2, σ e $\log \sigma^2$ con un livello di confidenza del 95%.

Soluzione. Sappiamo dalla soluzione dell'esercizio 6 della seconda unità che lo stimatore di massima verosimiglianza di σ^2 è

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

con un errore standard approssimato pari a

$$\widehat{\mathrm{SE}}(\hat{\sigma}^2) = \hat{\sigma}^2 \sqrt{\frac{2}{n}}.$$

Possiamo usare queste due quantità per calcolare gli intervalli di confidenza richiesti. Come spiegato alla fine dell'unità 3 conviene costruire l'intervallo di confidenza per $\psi = \log \sigma^2$ e poi ricavare gli intervalli per σ^2 e σ tramite trasformazioni. Lo stimatore di massima verosimiglianza per $\psi = \log \sigma^2$ è semplicemente $\hat{\psi} = \log \hat{\sigma}^2$ mentre lo standard error approssimato è pari a

$$\begin{aligned}
\widehat{\mathrm{SE}}(\hat{\psi}) &\approx \left. \frac{d \log(\sigma^2)}{d\sigma^2} \right|_{\hat{\sigma}^2} \widehat{\mathrm{SE}}(\hat{\sigma}^2) \\
&= \frac{1}{\hat{\sigma}^2} \hat{\sigma}^2 \sqrt{\frac{2}{n}} \\
&= \sqrt{\frac{2}{n}}.
\end{aligned}$$

Quindi, un intervallo di confidenza di livello approssimativamente 95% per $\psi = \log \sigma^2$ è

$$\log \hat{\sigma}^2 \pm z_{0.025} \sqrt{\frac{2}{n}}.$$

Con i dati osservati abbiamo $\log \hat{\sigma}^2 = \log(2.75) = 1.01$ e $n = 60$ per cui l'intervallo di confidenza approssimato è $1.01 \pm z_{0.025}(0.18) = [0.654, 1.369]$. Il corrispondente intervallo di confidenza approssimato per σ^2 è $[e^{0.654}, e^{1.369}] = [1.92, 3.93]$, mentre quello per σ è $[\sqrt{1.92}, \sqrt{3.93}] = [1.39, 1.98]$.

Calcoli con **R**:

```
sigma2.hat <- 2.75
n <- 60
z <- qnorm(0.975)
z


## [1] 1.959964


## intervallo di confidenza per log(sigma2)
ci.log <- log(sigma2.hat) + c(-1, 1) * z * sqrt(2 / n)
ci.log
```

```
## [1] 0.6537621 1.3694397

## intervallo di confidenza per sigma2
exp(ci.log)

## [1] 1.922761 3.933146

## intervallo di confidenza per sigma
sqrt(exp(ci.log))

## [1] 1.386637 1.983216
```