

Esercitazioni MTD ConfInt

Matteo Re

23 Ottobre 2019

1 Introduzione

In questa esercitazione vedremo come stimare l'intervallo di confidenza da un campione che assumiamo essere derivante da una popolazione normalmente distribuita.

L'esercitazione e' basata su due esercizi pratici da svolgere in classe. Nel primo esercizio ci concentreremo sull'automazione del calcolo dell'intervallo di confidenza mediante l'utilizzo di una funzione. Nel secondo esercizio utilizzeremo la funzione prodotta in esercizio 1 per effettuare dei test. Nel terzo esercizio vedremo alcuni esempi reali di calcolo dell'intervallo di confidenza.

2 Esercizio 1

Data la formula della media della popolazione:

$$\mu \in \bar{x} \pm \frac{z \cdot \sigma}{\sqrt{N}}$$

L'intervallo di confidenza (Confidence Interval) puo' essere stimato come segue:

$$\left[\frac{-z \cdot \sigma}{\sqrt{N}}, \frac{z \cdot \sigma}{\sqrt{N}} \right]$$

Iniziamo ad implementare in R la formula per il calcolo dell'intervallo di confidenza a partire da un vettore di osservazioni.

```
# Creazione di un vettore di 100 osservazioni
# parametri normale: media=0, deviazione standard=1
obs <- rnorm(100,0,1)
# Ora calcoliamone il valor medio
N <- length(obs)
tot <- sum(obs)
# calcolo media campionaria
xbar <- tot/N # verificate che xbar e' uguale a mean(obs)
# calcolo varianza campionaria
s2 <- sum((obs - xbar)^2)/(n - 1) # e' uguale a var(obs) ?
# la deviazione standard e' la radice della varianza
stddev <- sqrt(s2) # verificate che e' uguale a sd(obs)
```

Abbiamo la media campionaria (\bar{x}):

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

e la varianza del campione (s^2):

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i^2 - N\bar{x}^2)$$

Ora iniziamo a ragionare sugli intervalli di confidenza. La prima cosa da fare è verificare se, ragionevolmente, possiamo considerare la distribuzione generatrice dei dati campionati gaussiana. Per verificare questa ipotesi possiamo vedere i grafici di probabilità utilizzando le funzioni R `qqnorm()` e `qqline()` come segue:

```
qqnorm(obs) # disegna i punti
# NON chiudete il grafico appena generato,
# dobbiamo aggiungere unaa retta
qqline(obs) # disegna la retta
```

Abbiamo visto come valutare la normalità del campione in modo qualitativo mediante l'utilizzo di grafici ma è anche possibile effettuare un test statistico. Il test più utilizzato in questi casi (soprattutto per piccoli campioni, di numerosità inferiore a 30) è il Shapiro-Wilk test. Esso è disponibile in R nella libreria di base `stats` e, quindi, non è necessario installare alcun package aggiuntivo per poterlo utilizzare. È bene ricordare che nei test di normalità H_0 , l'ipotesi nulla, è che i dati del campione siano provengano da una popolazione normalmente distribuita. Di conseguenza se il test di normalità non è significativo, non possiamo rifiutare l'ipotesi nulla (che i dati sono distribuiti normalmente). Solo nel caso in cui il test sia statisticamente significativo ($p\text{-value} < 0.05$) possiamo rifiutare H_0 . In R il test di Shapiro-Wilk si utilizza come segue:

```
# Test di normalità di Shapiro-Wilk:
> mydata <- rnorm(20,0,1)
> shapiro.test(x=mydata)

      Shapiro-Wilk normality test

data:  mydata
W = 0.97213, p-value = 0.799
}
```

In questo esempio i dati del piccolo ($n < 30$) campione, se testati con Shapiro-Wilk portano ad un $p\text{-value} > 0.05$, quindi non possiamo rifiutare l'ipotesi che siano distribuiti normalmente. In queste condizioni possiamo procedere alla stima dell'intervallo di confidenza per il valor medio come segue. Supponiamo di sapere che la deviazione standard della popolazione sia $\sigma = 1.005$,

allora l'intervallo di confidenza al 95% di μ e'

$$\bar{x} \pm \frac{1.96 \cdot \sigma}{\sqrt{N}}$$

Cio' a cui dobbiamo prestare attenzione e' l'associazione tra il valore della probabilita' totale compresa nell'intervallo (95%) ed il valore di z (1.96). Tali coppie di valori si trovano in tabelle statistiche che vengono utilizzate come riferimento durante i calcoli. Una tabella ridotta che potete utilizzare e' la seguente:

TAABELLA RIFERIMENTO VALORI z

Confidence level %	z
50	0.67
68	1.00
80	1.29
90	1.64
95	1.96
96	2.00
99	2.58
99.7	3.00
99.9	3.29

Per costruire la funzione che calcola l'intervallo di confidenza dobbiamo fornire alcune informazioni. In particolare servono il valore di z (che definisce il grado di confidenza), la deviazione standard della popolazione ed il vettore delle osservazioni (quest'ultimo serve per calcolare la media campionaria e per fornire il valore di N). Data la disponibilita' di tali informazioni la funzione R puo' essere scritta come segue:

```
# funzione per il calcolo dell'intervallo di confidenza:
confint <- function(z, sdpop, obsvector){
  # costruzione variabile per risultati
  res <- rep(0,3)
  names(res) <- c("xbar", "IClowerBound", "ICupperBound")
  # calcolo valori
  xbar <- mean(obsvector)
  res[1] <- xbar
  res[2] <- xbar - (z*sdpop)/sqrt(length(obsvector))
  res[3] <- xbar + (z*sdpop)/sqrt(length(obsvector))
  # restituzione risultati
  return(res)
}
```

3 Esercizio 2

Equipaggiati con la funzione `confint` appena scritta possiamo procedere con alcuni esperimenti. Procedete come segue:

- Create 3 vettori `vec1`, `vec2` e `vec3` di osservazioni campionate casualmente da una normale con parametri $\mu = 5$ e $\sigma = 1.5$. I vettori dovranno avere rispettivamente lunghezza 5, 10 e 100
- Applicate ad essi la funzione `confint()` utilizzando $z = 1.96$ e $\sigma = 1.5$

Cosa osservate? C'è relazione tra la lunghezza dei vettori e l'ampiezza dell'intervallo di confidenza calcolato? Provate a ripetere i passaggi utilizzando diversi valori di σ . Le vostre conclusioni cambiano? Motivate la risposta.

4 Esercizio 3

Nella pratica statistica i reali valori di σ e μ non sono noti ed è quindi necessario procedere ad una stima di tali parametri a partire dal campione a disposizione. Lo stimatore appropriato per la media di popolazione è la media campionaria. Si procede poi alla stima dell'errore quadratico medio. Quest'ultimo, in particolare, può essere stimato come segue:

$$\frac{\sigma^2}{N}$$

Si procede, infine, con la stima dell'errore standard $SE = \frac{s}{\sqrt{N}}$, in cui s vale :

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

il calcolo della media campionaria in R è banale mentre SE può essere calcolato come segue:

```
# Calcolo SE di un vettore di osservazioni di nome obsvect
s <- sd(obsvect)
SE <- s/sqrt(length(obsvect))
```

Nel caso in cui il valore di σ della popolazione non sia noto sarebbe errato utilizzare la funzione `confint` (che richiede in input il vero valore di σ della popolazione). In questi casi possiamo procedere come segue per la stima di un intervallo di confidenza:

$$\bar{x} \pm t \cdot s / \sqrt{N}$$

in cui:

- s è la deviazione standard del campione (non quella supposta nota della popolazione)

- t e' un coefficiente maggiore di 1.96 dipendente dalla dimensione del campione

il coefficiente t si ottiene non dalla tavola della normale ma da quella della distribuzione t di Student . Si puo' sostituire la tavola della distribuzione t di Studente con le seguenti istruzioni R:

```
LC <- 0.95 # Livello di confidenza
n <- 11 # Numerosita' campionaria
t <- qt(LC + (1-LC)/2, n-1)
```

Ora proviamo a confrontare gli intervalli di confidenza al 95% calcolati a partire dal medesimo campione assumendo nota la μ di popolazione ed utilizzando il metodo appena presentato.

- Costruite un vettore `rndvect` contenente 10 campionamenti da una normale avente $\mu = 0$ e $\sigma = 1$
- Stimare intervallo di confidenza al 95% utilizzando la funzione `confint()`
- costruite una funzione per il calcolo del valore di t dato il livello di confidenza.
- Utilizzate la funzione per calcolare t per $LC = 0.95$
- salvate il risultato ottenuto in una variabile `t095` ed utilizzatelo come segue:

```
SE <- sd(rndvect)/sqrt(length(rndvect))
xbar <- mean(rndvect)
A <- xbar - t * SE
B <- xbar + t * SE
ConfIntT <- c(A,B)
```

Cosa osservate riguardo all'ampiezza degli intervalli di confidenza calcolati assumendo la σ di popolazione nota e non nota (mediante t). Che conclusioni potete trarne? Motivate la risposta.