

Inferenza Statistica

Note in R

V. Gioia (e R. Pappadà, N. Torelli,)

11/12/2024

Contents

Stima intervallare	1
Intervallo di confidenza per la media di una normale con varianza nota	1
Intervallo di confidenza per la media di una normale con varianza ignota	4
Esercizio 3 - Esercitazione 8	5

Stima intervallare

Un intervallo di confidenza per un parametro è costruito in modo tale che, se si immaginasse di ripetere più volte l'estrazione del campione dalla medesima popolazione, esso conterrebbe molto spesso il valore che vogliamo stimare. Gli estremi dell'intervallo sono infatti funzione dei dati campionari, per cui sono variabili aleatorie al variare del campione. In molti casi siamo in grado di costruire tali intervalli casuali in modo tale che la probabilità che essi coprano il vero valore ignoto del parametro sia elevata e pari a un valore fissato, detto livello di confidenza.

Si noti tuttavia che dopo aver costruito tale intervallo esso o conterrà il vero valore o non lo conterrà (e questo probabilmente non lo sapremo mai). Ma possiamo affermare che se costruiamo l'intervallo di confidenza (con livello $1 - \alpha = 0.95$) per un parametro e replichiamo l'esperimento un numero elevato di volte, circa il 95% degli intervalli di confidenza osservati dovrebbe contenere il vero valore del parametro. Non è corretto infatti affermare che la probabilità che il parametro ignoto sia contenuto nell'intervallo è pari a 0.95 (cosa però possibile in una impostazione di inferenza Bayesiana). Il valore 0.95 rappresenta una proprietà dell'intervallo casuale ottenuto che è, come già sottolineato, la seguente: se io estraessi moltissime volte il campione otterrei intervalli sempre diversi, ma che contengono il valore vero 95 volte su 100.

Intervallo di confidenza per la media di una normale con varianza nota

Quale primo esempio consideriamo l'intervallo di confidenza per la media di una variabile Y che nella popolazione è distribuita come una gaussiana di varianza nota σ^2 . Fissato il livello $1 - \alpha$, la realizzazione intervallo di confidenza per μ al livello $(1 - \alpha)100\%$ che ha ampiezza più piccola possibile ha estremi

$$IC_{\mu}^{1-\alpha} = (\bar{y} - z_{1-\alpha/2}\sigma/\sqrt{n}, \bar{y} + z_{1-\alpha/2}\sigma/\sqrt{n})$$

dove $z_{1-\alpha/2}$ è il quantile di ordine $1 - \alpha/2$ della normale standard.

Se volessimo verificare empiricamente il comportamento dell'intervallo così definito, potremmo utilizzare ancora una volta una simulazione Monte Carlo e vedere cosa accade generando un gran numero di campioni casuali.

Ad esempio, usiamo R per ricavare i limiti dell'intervallo per μ da diversi campioni di dimensione $n = 10$ estratti da una $\mathcal{N}(\mu = 5, \sigma^2 = 4)$. Quindi verifichiamo che la proporzione di intervalli che contiene il valore

incognito, ovvero valutiamo il **grado di copertura effettivo**, è circa pari al livello di confidenza fissato (nel nostro caso 0.95), cioè il **grado di copertura nominale**.

```
set.seed(13)
R <- 1000 #numero di replicazioni
n <- 10   #dimensione del campione

# Parametri
mu <- 5
sig2 <- 4

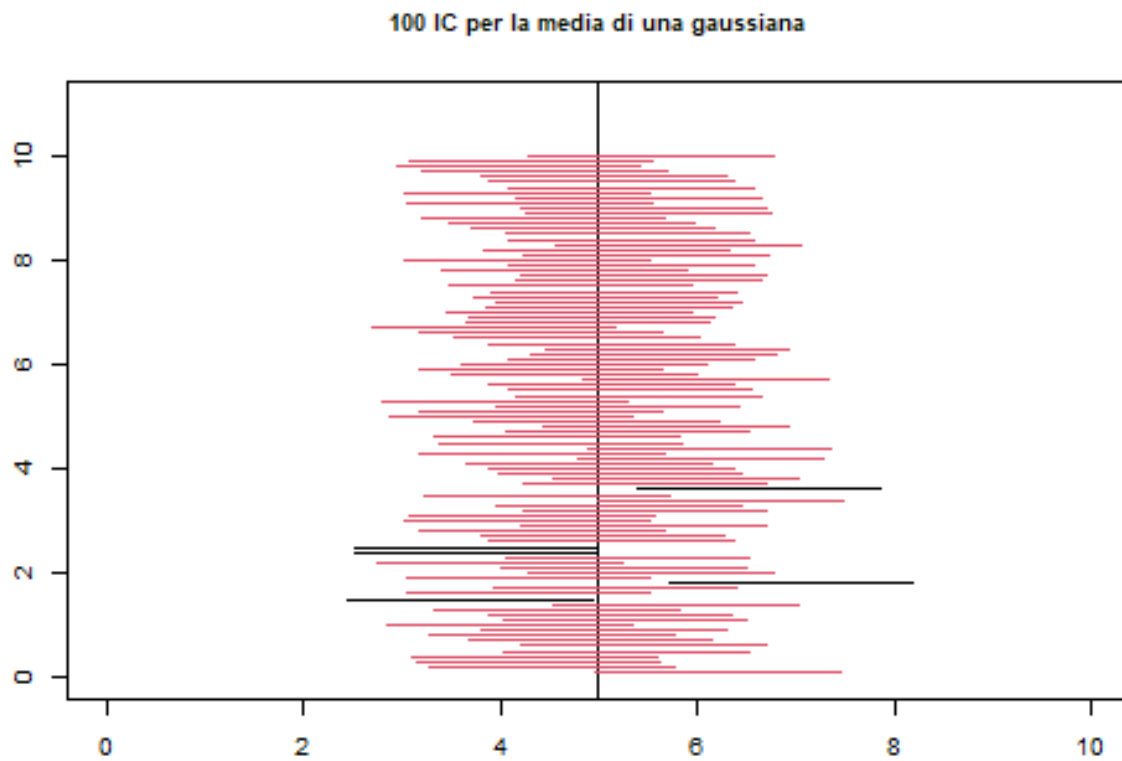
alpha <- 0.05 # livello di confidenza 1- alpha = 0.95

inf <- sup <- ecop <- vector(mode = "numeric", length = R)
for (i in 1 : R){
  y <- rnorm(n, mu, sqrt(sig2))
  inf[i] <- mean(y) - qnorm(1 - alpha/2) * sqrt(sig2/n)
  sup[i] <- mean(y) + qnorm(1 - alpha/2) * sqrt(sig2/n)
  ecop[i] <- (mu > inf[i] & mu < sup[i]) #valori 1(=TRUE) e 0(=FALSE)
}
mean(ecop)
```

```
## [1] 0.955
```

Successivamente rappresentiamo gli intervalli ottenuti in un grafico, distinguendo quelli che non contengono il valore vero del parametro ($\mu = 5$ nel nostro caso) con un colore diverso.

```
# Rappresentiamo i primi 100 IC
plot(0, 0, xlim = c(0, 10), ylim = c(0, 11), type = "n", xlab = "", ylab = "",
     main = "100 IC per la media di una gaussiana", cex.main = 0.9)
abline(v = mu)
d <- 0
for(i in 1 : 100){
  d <- d + 0.1 # per distanziare le barre orizzontali
  lines(seq(inf[i], sup[i], length = 100), rep(d, 100), col= (ecop[i] + 1))
}
```



Esercizio Si valuti il grado di copertura di un intervallo di confidenza per la proporzione p di una popolazione Bernoulliana, ponendo $n < 25$. Nei casi visti, esso dovrebbe corrispondere al livello di confidenza da noi fissato. Cosa accade nel caso in cui utilizziamo l'approssimazione Gaussiana per costruire l'intervallo di confidenza? Utilizzare uno studio Monte Carlo per verificare le proprietà dell'intervallo costruito.

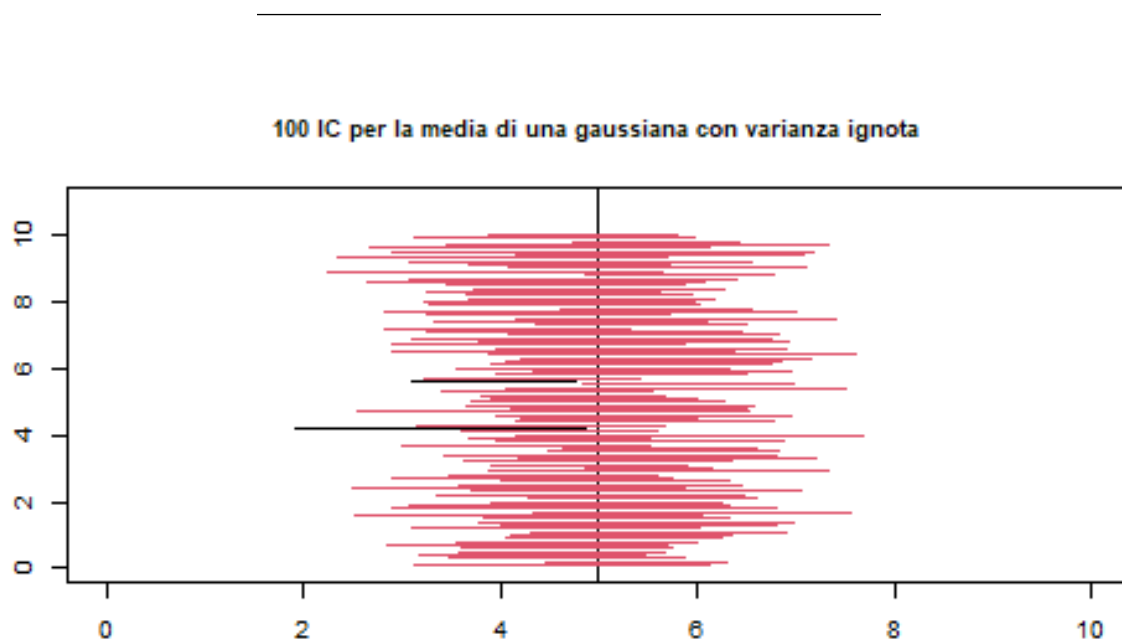
Intervallo di confidenza per la media di una normale con varianza ignota

In questo secondo esempio consideriamo ancora una popolazione Gaussiana ma supponiamo che la varianza sia ignota. Quindi costruiamo un grafico analogo a quello ottenuto poc'anzi e ne valutiamo il grado di copertura empirico mediante studio di simulazione.

```
for (i in 1 : R){  
  y <- rnorm(n, 5, 2)  
  inf[i] <- mean(y) + qt(alpha/2, n - 1) * sqrt(var(y)/n)  
  sup[i] <- mean(y) + qt(1 - alpha/2, n - 1) * sqrt(var(y)/n)  
  ecop[i] <- (mu > inf[i] & mu < sup[i])  
}  
mean(ecop)
```

```
## [1] 0.948
```

```
plot(0, 0, xlim = c(0, 10), ylim = c(0, 11), type = "n", xlab = "", ylab = "",  
     main = "100 IC per la media di una gaussiana con varianza ignota", cex.main = 0.9)  
abline(v = mu)  
d <- 0  
for(i in 1 : 100){  
  d <- d + 0.1 # per distanziare le barre orizzontali  
  lines(seq(inf[i], sup[i], length = 100), rep(d, 100), col=(ecop[i] + 1))  
}
```



Esercizio Utilizzare il metodo Monte Carlo per valutare le proprietà di un intervallo di confidenza per la differenza tra due medie sulla base di diversi campioni di dimensione $n_1 = 20$ e $n_2 = 25$, provenienti da una $\mathcal{N}(14, 12)$ e $\mathcal{N}(11, 18)$, rispettivamente.

Esercizio 3 - Esercitazione 8

Le durate in ore di una serie di lampadine sono descritte da una v.a. distribuita secondo una $\text{Gamma}(2, \theta)$, con $\theta > 0$ e incognito. Un campione di n lampadine è selezionato e si dispone delle durate $y_1, \dots, y_n$.

- [a.] Determinare un intervallo di confidenza approssimato per θ , a livello del 90%.
- [b.] Determinare un intervallo di confidenza approssimato per ψ , a livello del 90%, dove ψ è il valore atteso della durata di ore delle lampadine.

Sia Y la variabile aleatoria che descrive la durata delle lampadine, e sappiamo che $Y \sim \text{Gamma}(2, \theta)$. Per una generica v.a. $\text{Gamma}(\alpha, \theta)$ la funzione di densità è data da

$$f_Y(y; \alpha, \theta) = \frac{\theta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\theta y} 1_{\{y>0\}}$$

Per cui nel nostro caso

$$f_Y(y; \theta) = \theta^2 y e^{-\theta y} 1_{\{y>0\}}$$

Per il punto a) siano $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Gamma}(2, \theta)$, per cui la funzione di verosimiglianza è data da

$$\mathcal{L}(\theta) = \prod_{i=1}^n f(y_i; \theta) \propto \prod_{i=1}^n \theta^2 e^{-\theta y_i} = \theta^{2n} e^{-\theta \sum_{i=1}^n y_i}$$

Quindi

$$\ell(\theta) = 2n \log(\theta) - \theta \sum_{i=1}^n y_i$$

da cui la funzione di punteggio

$$s(\theta) = \frac{2n}{\theta} - \sum_{i=1}^n y_i$$

Pertanto la stimatore di massima verosimiglianza, ottenuto dall'equazione di verosimiglianza, è

$$s(\theta; \mathbf{Y}) = 0 \implies \hat{\theta}_{MV} = \frac{2n}{\sum_{i=1}^n Y_i} = \frac{2}{\bar{Y}}$$

e l'informazione osservata è

$$J(\theta) = -\frac{d}{d\theta} s(\theta) = \frac{2n}{\theta^2} = I(\theta)$$

per cui si ha che $\hat{\theta} \sim \mathcal{N}(\theta, \theta^2/2n)$ quindi per ottenere un IC approssimato al 90% per θ abbiamo che

$$IC_{\theta}^{0.90} = (\hat{\theta} - 1.645 \times \sqrt{\hat{\theta}^2/2n}, \hat{\theta} + 1.645 \times \sqrt{\hat{\theta}^2/2n})$$

e una sua realizzazione si ottiene rimpiazzando la stima $\hat{\theta} = 2/\bar{y}$

$$IC_{\theta}^{0.90} = (2/\bar{y} - 1.645 \times \sqrt{2/(n\bar{y}^2)}, 2/\bar{y} + 1.645 \times \sqrt{2/(n\bar{y}^2)})$$

Utilizziamo Monte Carlo e costruiamo una funzione che ci ritorni la probabilità empirica copertura per diversi valori di n

```

set.seed(1)
R <- 10000  #numero di replicazioni
n <- c(2,10,100,1000)  #dimensione del campione

alpha <- 0.1
# Parametri
theta <- 5

sim_ICtheta <- function(R, n, alpha, theta){
  ecop <- rep(NA, length(n))
  bias <- rep(NA, length(n))
  mse <- rep(0, length(n))
  theta_MLE <- rep(0, R)
  var_theta_MLE <- rep(0, R)
  inf <- rep(0, R)
  sup <- rep(0, R)
  ecopi <- rep(0,R)
  for(j in 1: length(ecop)){
    for (i in 1 : R){
      y <- rgamma(n[j], 2, theta)
      theta_MLE[i] <- 2/mean(y)
      var_theta_MLE[i] <- 2/(n[j]*mean(y)^2)
      inf[i] <- theta_MLE[i] - qnorm(1-alpha/2) * sqrt(2/(n[j]*mean(y)^2))
      sup[i] <- theta_MLE[i] + qnorm(1-alpha/2) * sqrt(2/(n[j]*mean(y)^2))
      ecopi[i] <- (theta > inf[i] & theta < sup[i])
    }
    ecop[j] <- mean(ecopi)
    bias[j] <- mean(theta_MLE) - theta
    mse[j] <- mean( var_theta_MLE) + bias[j]^2
  }
  return(list(ecop = ecop, bias = bias, mse = mse))
}

res <- sim_ICtheta(R,n, alpha, theta)
res

## $ecop
## [1] 0.9230 0.9037 0.9027 0.8986
##
## $bias
## [1] 1.629227677 0.278513204 0.028699416 0.002566717
##
## $mse
## [1] 19.56293216 1.54848470 0.12788365 0.01252571

```

Per il punto b), richiamiamo che vogliamo un IC (a livello 0.90) per $\psi = \mathbb{E}(Y) = 2/\theta = h(\theta)$. Sappiamo per l'equivarianza dello stimatore di massima verosimiglianza che $\hat{\psi}_{MV} = 2/\hat{\theta}_{MV} = \bar{Y}$. Al fine di determinare la distribuzione asintotica di $\hat{\psi}_{MV}$ possiamo fare ricorso al metodo delta per cui

$$\hat{\psi}_{MV} \sim \mathcal{N}\left(\psi, \mathcal{I}(\theta)^{-1} \left(\frac{d}{d\theta} h(\theta)\right)^2\right) \equiv \mathcal{N}\left(\psi, \frac{\theta^2}{2n} \frac{4}{\theta^4} = \frac{2}{n\theta^2} = \frac{\psi^2}{2n}\right)$$

poichè $\frac{d}{d\theta} h(\theta) = -2/\theta^2$ e a questo punto utilizziamo la stima di ψ per stimare la varianza dello stimatore che sarà quindi $\hat{V}(\hat{\psi}_{MV}) = \bar{Y}^2/2n$, da cui l'intervallo di confidenza approssimato stimato è

$$IC_{\psi}^{0.90} = (\bar{y} - z_{0.95} \times \bar{y}/\sqrt{2n}, \bar{y} + z_{0.95} \times \bar{y}/\sqrt{2n})$$

```
set.seed(1)
R <- 10000 #numero di replicazioni
n <- c(2,10,100,1000) #dimensione del campione
alpha <- 0.1
theta <- 5
sim_ICpsi <- function(R, n, alpha, theta){
  ecop <- rep(NA, length(n))
  bias <- rep(NA, length(n))
  mse <- rep(0, length(n))
  psi_MLE <- rep(0, R)
  var_psi_MLE <- rep(0, R)
  inf <- sup <- rep(0, R)
  ecopi <- rep(0,R)
  for(j in 1: length(ecop)){
    for (i in 1 : R){
      y <- rgamma(n[j], 2, theta)
      theta_MLE <- 2/mean(y)
      psi_MLE[i] <- 2/theta_MLE
      var_psi_MLE[i] <- sqrt(mean(y)^2/(2*n[j]))
      inf[i] <- psi_MLE[i] - qnorm(1-alpha/2) * sqrt(mean(y)^2/(2*n[j]))
      sup[i] <- psi_MLE[i] + qnorm(1-alpha/2) * sqrt(mean(y)^2/(2*n[j]))
      ecopi[i] <- (2/theta > inf[i] & 2/theta < sup[i])
    }
    ecop[j] <- mean(ecopi)
    bias[j] <- mean(psi_MLE) - 2/theta
    mse[j] <- mean(var_psi_MLE) + bias[j]^2
  }
  return(list(ecop = ecop, bias = bias, mse = mse))
}

res <- sim_ICpsi(R,n, alpha, theta)
res
```

```
## $ecop
## [1] 0.8243 0.8847 0.9012 0.9000
##
## $bias
## [1] 3.772958e-03 -1.060480e-03 -3.356993e-04 -4.392987e-06
##
## $mse
## [1] 0.201900714 0.089206713 0.028260646 0.008944174
```