# Inferenza Statistica

Note in R

V. Gioia (e R. Pappadà, N. Torelli,)

03/12/2024

## Contents



## Monte Carlo Esercizio 3 - Esercitazione 7

Sia $(y_1, \dots, y_n)$ un campione proveniente da una variabile $Y$ descritta dalla densità $f(y;\theta) = 3y^2/\theta^3$, $0 \leq y \leq \theta$.

Al punto a. si chiedeva lo stimatore di $\theta$ con il metodo dei momenti ($\hat{\theta}_{MM}$). Esso risultava $\hat{\theta}_{MM} = 4\overline{Y}/3$

Al punto b. si chiedeva lo stimatore di massima verosimiglianza di $\theta$ ($\hat{\theta}_{MV}$). Esso risultava $\hat{\theta}_{MV} = Y_{(n)}$

Al punto c. si chiedeva la verifica della non distorsione e della consistenza in media quadratica di $\hat{\theta}_{MM}$. Essendo lo stimatore non distorto, l'errore quadratico medio risultava

$$MSE_{\hat{\theta}_{MM}}(\theta) = \mathbb{V}(\hat{\theta}_{MM}) = \frac{\theta^2}{15n}$$

Dunque lo stimatore risulta consistente in media quadratica, poichè

$$\lim_{n\to\infty} MSE_{\hat{\theta}_{MM}}(\theta) = \lim_{n\to\infty} \mathbb{V}(\hat{\theta}_{MM}) = \lim_{n\to\infty} \frac{\theta^2}{15n} = 0$$

Al punto d. la richiesta era

- Usando il calcolo delle probabilità si ottiene che $\mathbb{E}(\hat{\theta}_{MV}) = 3n\theta/(3n+1)$ e $\mathbb{V}(\hat{\theta}_{MV}) = 3n\theta^2/[(3n+1)^2(3n+2)]$. Per $n = 5$ si dica quale tra $\hat{\theta}_{MV}$ e $\hat{\theta}_{MM}$ sia preferibile.

Quindi sappiamo che lo stimatore di massima verosimiglianza non è corretto e avrà distorsione pari a

$$B_{\hat{\theta}_{MV}}(\theta) = -\frac{\theta}{3n+1}$$

mentre l'erore quadratico medio risulta essere pari a

$$MSE_{\hat{\theta}_{MV}}(\theta) = \mathbb{V}(\hat{\theta}_{MV}) + B_{\hat{\theta}_{MV}}(\theta)^2 = \frac{\theta^2}{136}$$

I risultati relativi allo stimatore di massima verosimiglianza forniti nella traccia discendono dal notare che

$$Y_{(n)} = X_{(n)}\theta$$

dove $X_{(n)} \sim \text{Beta}(3n, 1)$. Questo verrà utilizzato per generare dei dati da $Y$, ovvero $Y = X\theta$ con $X \sim \text{Beta}(3, 1)$. In alternativa, per poter generare da $Y$ si poteva utilizzare il metodo dell'inversione, visto nell'esercitazione 3. Quindi conduciamo una simulazione Monte Carlo per analizzare le proprietà degli stimatori ottenuti. A tal proposito, decidiamo un vero vlore del parametro e consideriamo il caso $n = 5$.

```
theta <- 3
R <- 10000
n <- 5
set.seed(13)
theta_MLE <- theta_MOM <- rep(NA, R)
for(i in 1:R){
  x <- rbeta(n, 3, 1)
  y <- x * theta
  theta_MOM[i] <- 4 * mean(y)/3
  theta_MLE[i] <- max(y)
}


biasMOM <- mean(theta_MOM) - theta # valore teorico 0
biasMOM
```

```
## [1] 0.001706246
```

```
biasMLE <- mean(theta_MLE) - theta
biasMLE
```

```
## [1] -0.1869313
```

```
-theta/(3 * n+1) # Valore teorico bias
```

```
## [1] -0.1875
```

```
varMOM <- var(theta_MOM)
varMLE <- var(theta_MLE)


MSE_MOM <- var(theta_MOM) + biasMOM^2
MSE_MOM
```

```
## [1] 0.1190152
```

```
theta^2/(15 * n) # Valore teorico
```

```
## [1] 0.12
```

```
MSE_MLE <- var(theta_MLE) + biasMLE^2
MSE_MLE
```

```
## [1] 0.0661122
```

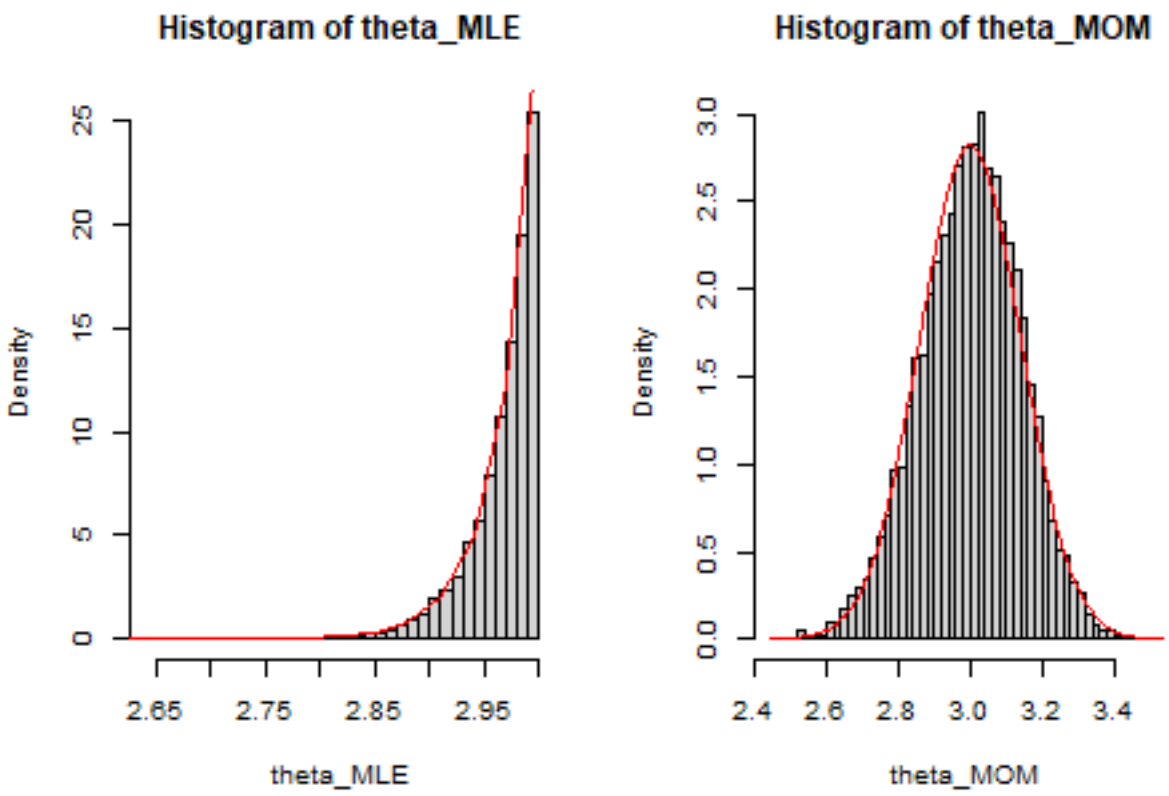
```
2 * theta^2/((3 * n + 1) * (3 * n + 2)) # Valore teorico
```

```
## [1] 0.06617647
```

Aumentiamo un pò il nostro $n$ per la seguente verifica distributiva.

```
theta <- 3
R <- 10000
n <- 30
set.seed(13)
theta_MLE <- theta_MOM <- rep(NA, R)
for(i in 1:R){
  x <- rbeta(n, 3, 1)
  y <- x * theta
  theta_MOM[i] <- 4 * mean(y)/3
  theta_MLE[i] <- max(y)
}

par(mfrow = c(1, 2))
hist(theta_MLE, breaks = 50, freq =FALSE)
curve(dbeta(x/(theta), 3 * n, 1)/theta, from = 0, to = theta - 1e-14, col = "red", add = TRUE)
hist(theta_MOM, breaks = 50, freq = FALSE)
curve(dnorm(x, theta, sqrt(theta^2/(15 * n))), add = TRUE, col = "red")
```

---



# 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 $\sigma^2$. Fissato il livello $1-\alpha$, la realizzazione intervallo di confidenza per $\mu$ 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 quatile 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 $\mu$ 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**.

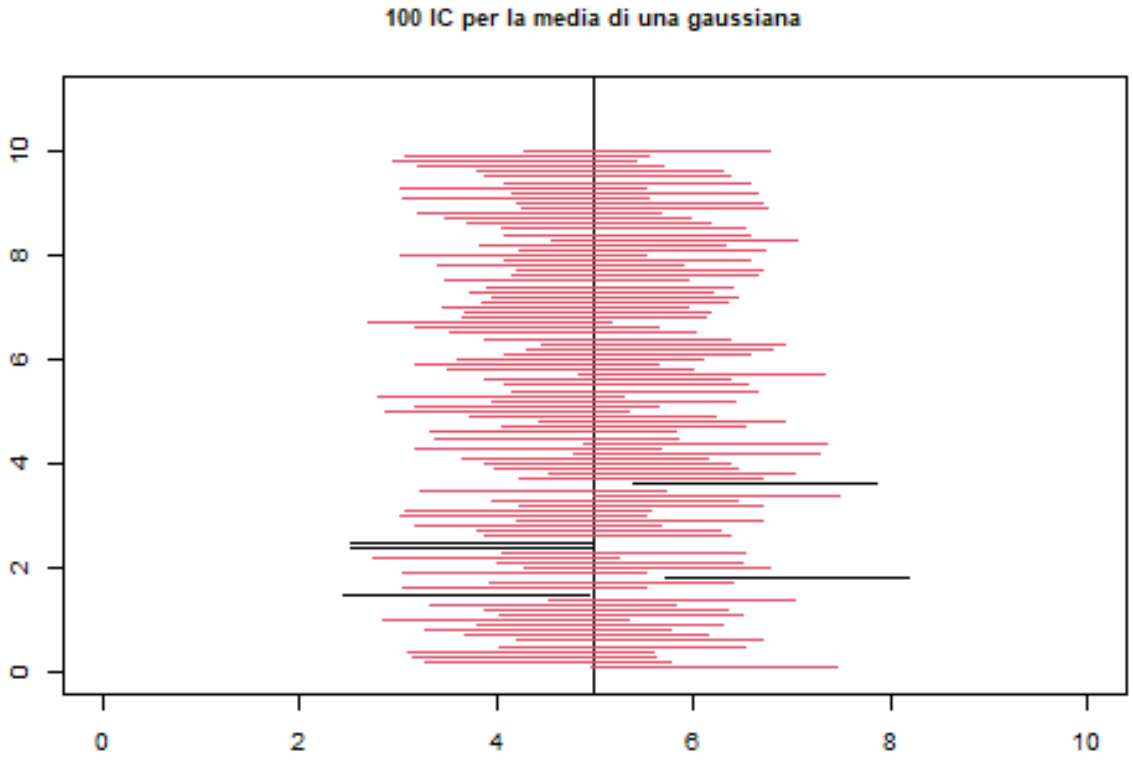
```
set.seed(13)
R <- 1000   #numero di replicazioni
n <- 10    #dimensione del campione

# 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

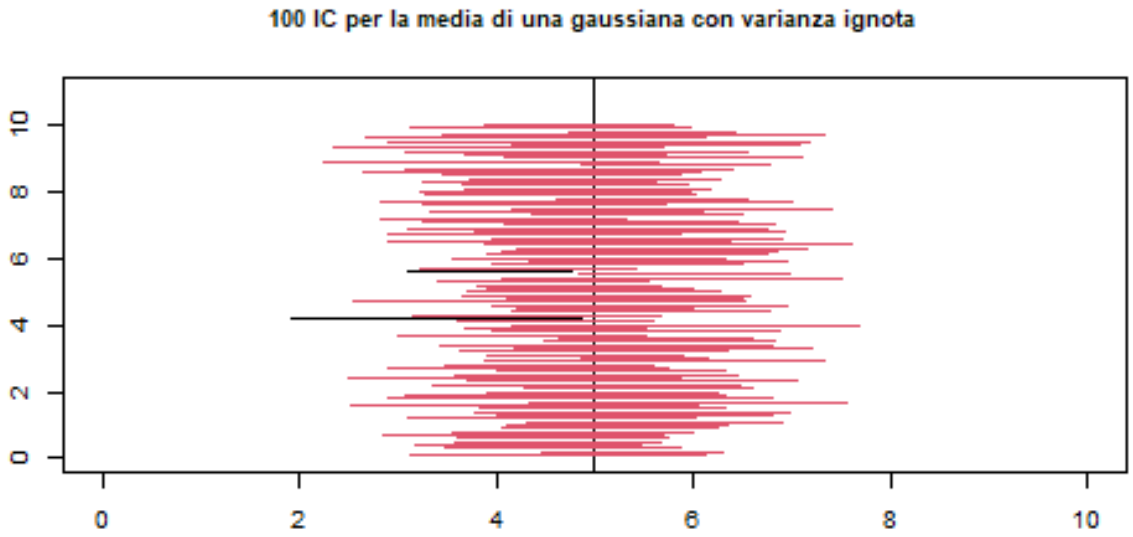
In questo secondo esempio consideriamo ancora una popolazione Gaussiana ma supponiamo che la varianza sia ignota. Quindi costruiamo un grafico analogo a quello ottenuto poc'anzi e ne valutiamo il grado di copertura empirico mediante studio di simulazione.

```
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.