

# R per l'analisi statistica multivariata

Unità M: proprietà degli stimatori

**Tommaso Rigon**

**Università Milano-Bicocca**



## Argomenti affrontati

- Distribuzione di uno stimatore
- Distorsione, varianza ed errore quadratico medio
- Esercizi **R** associati: [https://tommasorigon.github.io/introR/exe/es\\_4.html](https://tommasorigon.github.io/introR/exe/es_4.html)

# Inferenza statistica parametrica (recap)

- In questa unità discuteremo di **stimatori** e delle loro proprietà.
- Ricordiamo che, nel caso assolutamente continuo, un **modello statistico** è una collezione di funzioni di densità

$$\mathcal{F} = \{f(\cdot; \theta) : \theta \in \Theta\},$$

indicizzata da un vettori di parametri  $\theta \in \Theta$ , dove  $\Theta \subseteq \mathbb{R}^p$  è lo **spazio parametrico**.

- Assumiamo inoltre che i dati  $y_1, \dots, y_n$  siano realizzazioni iid di variabili aleatorie  $Y_1, \dots, Y_n$  con legge  $f(y; \theta)$ , per un qualche **ignoto** valore del parametro  $\theta$ .
- **Obiettivo**. Il nostro scopo è stimare l'ignoto valore di  $\theta$  nel miglior modo possibile usando i **dati** osservati  $y = (y_1, \dots, y_n)$ .

# Stime e stimatori (recap)

- Uno **stimatore**  $\hat{\theta}(Y)$  è una qualsiasi funzione delle variabili aleatorie  $Y = (Y_1, \dots, Y_n)$  che “si avvicina” al vero valore di  $\theta$ . Lo stimatore è una **variabile aleatoria**.
- Una **stima**  $\hat{\theta} = \hat{\theta}(y)$  è una qualsiasi funzione dei dati  $y = (y_1, \dots, y_n)$  che “si avvicina” al vero valore di  $\theta$ . La stima è la realizzazione di  $\hat{\theta}(Y)$ , perciò è un **numero**.
- Abbiamo bisogno di **criterio** per stabilire se uno stimatore “funziona” o meno. Il sostegno logico e filosofico proviene dal seguente principio (**frequentista**).

## Il principio del campionamento ripetuto

- Immaginiamo che sia possibile, almeno ipoteticamente, **ripetere l'esperimento** varie volte, ottenendo ogni volta un nuovo campione  $y$  e quindi una nuova stima  $\hat{\theta}$ .
- Di conseguenza, lo stimatore  $\hat{\theta}(Y)$  è una variabile aleatoria, per la quale possiamo parlare di distribuzione, valore atteso e così via.

# Il principio del campionamento ripetuto (recap)

- Se accettiamo il principio del campionamento ripetuto, valuteremo le bontà della singola stima  $\hat{\theta}$  sulla base delle **proprietà dello stimatore**  $\hat{\theta}(Y)$ .
- In altri termini, ci chiediamo come si comporterebbero le varie stime  $\hat{\theta}$  se potessimo osservare tanti campioni, non solo quello che abbiamo a disposizione.
- Ci aspettiamo che **mediamente** la distribuzione di  $\hat{\theta}(Y)$  sia **concentrata** attorno al vero ed ignoto valore  $\theta$ . Ovviamente, questo non è assicurato campione per campione.
- Una proprietà tipicamente richiesta è che all'aumentare della dimensione del campione  $n$ , la distribuzione di  $\hat{\theta}(Y)$  sia **concentrata** attorno a  $\theta$ .

# Distorsione, varianza ed errore (recap)

- Una prima semplice aspettativa rispetto allo stimatore è che mediamente esso sia corretto o **non distorto**, ovvero

$$\mathbb{E}\{\hat{\theta}(Y)\} = \theta, \quad \theta \in \Theta.$$

- La **distorsione** è infatti definita come la differenza semplice

$$\text{BIAS}\{\hat{\theta}(Y)\} := \mathbb{E}\{\hat{\theta}(Y)\} - \theta, \quad \theta \in \Theta.$$

Se uno stimatore è non distorto allora ovviamente  $\text{BIAS}\{\hat{\theta}(Y)\} = 0$ .

- Un requisito un po' meno stringente è che lo stimatore sia **asintoticamente non distorto**, ovvero

$$\lim_{n \rightarrow \infty} \mathbb{E}\{\hat{\theta}(Y)\} = \theta, \quad \theta \in \Theta,$$

dove  $n$  è la dimensione campionaria.

# Distorsione, varianza ed errore (recap)

- La non-distorsione (asintotica) è una proprietà auspicabile, ma spesso meno importante dello **errore** o **scarto quadratico medio**.
- Lo scarto quadratico medio (**mean squared error**) misura la distanza media tra stimatore e vero valore del parametro, ovvero

$$\text{MSE}\{\hat{\theta}(Y)\} = \mathbb{E} \left\{ [\hat{\theta}(Y) - \theta]^2 \right\}, \quad \theta \in \Theta.$$

- Esercizio - proprietà. Dimostrare che vale la seguente scomposizione:

$$\text{MSE}\{\hat{\theta}(Y)\} = \text{var} \{ \hat{\theta}(Y) \} + \text{BIAS} \{ \hat{\theta}(Y) \}^2, \quad \theta \in \Theta.$$

- Nota. Se uno stimatore è non-distorto, allora il suo scarto quadratico medio coincide con la varianza dello stimatore.

# Consistenza (recap)

- Uno stimatore si dice **consistente in media quadratica** se

$$\lim_{n \rightarrow \infty} \text{MSE}\{\theta(Y)\} = 0, \quad \theta \in \Theta.$$

oppure equivalentemente se

$$\lim_{n \rightarrow \infty} \mathbb{E}\{\hat{\theta}(Y)\} = \theta, \quad \lim_{n \rightarrow \infty} \text{var}\{\hat{\theta}(Y)\} = 0, \quad \theta \in \Theta.$$

- La consistenza in media quadratica implica la **convergenza in probabilità**, per cui scriveremo che

$$\hat{\theta}(Y) \xrightarrow{P} \theta, \quad n \rightarrow \infty, \quad \theta \in \Theta.$$

- **Esercizio.** Lo studente è invitato a rivedersi la definizione di convergenza in probabilità, le sue proprietà e la sua relazione con la consistenza in media quadratica (in  $L^2$ ).
- **Nota linguistica.** Il termine “consistente” deriva da un’errata traduzione del termine inglese *consistent*. Purtroppo, l’uso del termine è ormai troppo consolidato per porvi rimedio e non resta che subirlo. Lo stesso può dirsi del termine “stima puntuale”.



- Stabiliti i criteri per valutare la bontà di uno stimatore, rimane da capire come utilizzarli in pratica.
- Nei corsi di Statistica II vengono presentati modelli e stimatori per i quali è possibile calcolare l' $MSE$  analiticamente. Questo capita di rado nelle **applicazioni reali**.
- Fortunatamente il metodo **Monte Carlo** che abbiamo visto nell'**unità I** può venire in aiuto in assenza di risultati analitici.
- Ad esempio, lo scarto quadratico medio è per definizione un valore atteso, che possiamo quindi approssimare tramite **integrazione Monte Carlo**.

# Modello gaussiano con varianza nota

- Sia  $y = (y_1, \dots, y_n)$  un campione iid da una variabile casuale normale con **media ignota**  $\mu$  e **varianza nota** e pari  $\sigma^2 = 16$ , ovvero  $Y \sim N(\mu, 16)$ .
- Il parametro  $\mu$  è **ignoto** e siamo interessati a stimarlo.
- Una stima naturale per  $\mu$ , che oltretutto coincide con la SMV, è la media aritmetica

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i.$$

- La distribuzione (esatta!) dello stimatore  $\hat{\mu}(Y)$  è nota ed è pari

$$\hat{\mu}(Y) \sim N\left(\mu, \frac{16}{n}\right).$$

- **Esercizio.** Si dimostri che  $\text{MSE}\{\hat{\mu}(Y)\} = 16/n$ . Se ne deduca che lo stimatore è consistente.

# Modello gaussiano con varianza nota

- Un secondo possibile stimatore per  $\mu$  è la **mediana** campionaria  $Me$ .
- La **distribuzione dello stimatore**  $Me(Y)$  è ignota. Di conseguenza, anche le relative proprietà sono ignote.
- La mediana è uno stimatore distorto? Il suo scarto è maggiore o minore di quello della media aritmetica?
- In assenza di risultati analitici, possiamo provare a fornire una **risposta parziale** tramite Monte Carlo.
- In altri termini, indagheremo quale stimatore funziona meglio per degli specifici valori di  $\mu$ , ad esempio  $\mu = 10$  oppure  $\mu = 15$ .

# Modello gaussiano con varianza nota

- Supponiamo di voler investigare il caso  $\mu = 10$ . Supponiamo inoltre che  $n = 20$ .
- Cominciamo simulando un singolo campione  $y_1, \dots, y_n$  da una distribuzione  $N(\mu, 16)$ .

---

```
set.seed(100)
n <- 20 # Numerosità campionaria
mu <- 10 # Media teorica (solitamente ignota)

# Campione  $y_1, \dots, y_n$ 
y <- rnorm(n, mean = mu, sd = sqrt(16))

# Vero valore è  $\mu = 10$ 
mean(x)
# [1] 10.43147
median(x)
# [1] 10.37232
```

---

- In questo caso specifico, la mediana si avvicina di più al vero valore della media ( $\mu = 10$ ). Tuttavia, questo vale per questo **specifico campione**.

# Modello gaussiano con varianza nota

- Coerentemente con quanto discusso nelle slides precedenti, un modo preciso per valutare la bontà dello stimatore si basa sul **campionamento ripetuto**.
- In altri termini, vogliamo confrontare gli **scarti quadratici medi** dei due stimatori

$$\text{MSE}\{\hat{\mu}(Y)\}, \quad \text{MSE}\{\text{Me}(Y)\}.$$

- Nel caso della media aritmetica con  $\sigma^2 = 16$  ed  $n = 20$  i conti analitici implicano che  $\text{MSE}\{\hat{\mu}(Y)\} = 16/20 = 0.8$ . Ma nel caso della mediana?
- Utilizzando il metodo Monte Carlo, ottengo una **stima** dello scarto quadratico medio dello **stimatore** mediana (**!!**), ovvero

$$\widehat{\text{MSE}\{\text{Me}(Y)\}}.$$

- **Esercizio**. Lo studente rilegga questa frase fino a convincersi della sua correttezza.

# Modello gaussiano con varianza nota

- L'approssimazione  $\widehat{\text{MSE}\{\text{Me}(Y)\}}$  si basa sul metodo di integrazione Monte Carlo.
- Si supponga che  $\text{Me}_1, \dots, \text{Me}_R$  siano  $R$  estrazioni casuali della mediana calcolata su un campione iid gaussiano ( $\mu = 10$ ,  $\sigma^2 = 16$ ) di dimensione  $n = 20$ .
- Possiamo ottenere  $\text{Me}_1, \dots, \text{Me}_R$  simulando  $R$  campioni  $Y$  e calcolandone la mediana:

---

```
set.seed(156)
R <- 10^5
# Ottengo R estrazioni della mediana campionaria Me_1, ... Me_R
median_hat <- replicate(R, median(rnorm(n = n, mean = mu, sd = sqrt(16))))
```

---

- L'approssimazione Monte Carlo è quindi pari a

$$\widehat{\text{MSE}\{\text{Me}(Y)\}} = \frac{1}{R} \sum_{r=1}^R (\text{Me}_r - \mu)^2 \approx \mathbb{E}\{(\text{Me}(Y) - \mu)^2\} = \text{MSE}\{\text{Me}(Y)\}.$$

---

```
mean((median_hat - mu)^2) # Stima dello scarto quadratico medio (MSE) della mediana
# [1] 1.172079
```

---

# Modello gaussiano con varianza nota

- La mediana sembra essere **meno efficiente** della media aritmetica, quantomeno se  $\mu = 10, \sigma^2 = 16$  ed  $n = 20$ .
- Nel seguito sono riportati alcuni risultati aggiuntivi, incluse le stime Monte Carlo relative alla media aritmetica.

---

```
set.seed(156)
```

```
R <- 10^5
```

```
mu_hat <- replicate(R, mean(rnorm(n = n, mean = mu, sd = sqrt(16))))
```

```
median_hat <- replicate(R, median(rnorm(n = n, mean = mu, sd = sqrt(16))))
```

```
mean(mu_hat) - mu # Distorsione dello stimatore; valore teorico: 0
```

```
# [1] -0.004089293
```

```
mean((mu_hat - mu)^2) # Scarto quadratico medio dello stimatore; valore teorico: 0.8
```

```
# [1] 0.8008738
```

```
mean(median_hat) - mu # Distorsione dello stimatore; valore teorico: ??
```

```
# [1] -0.001820327
```

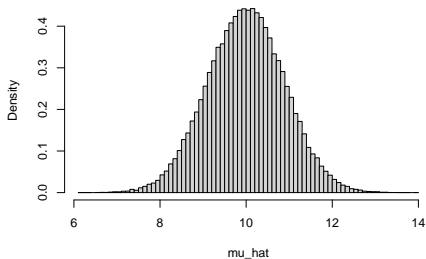
```
mean((median_hat - mu)^2) # Scarto quadratico medio dello stimatore; valore teorico: ??
```

```
# [1] 1.172248
```

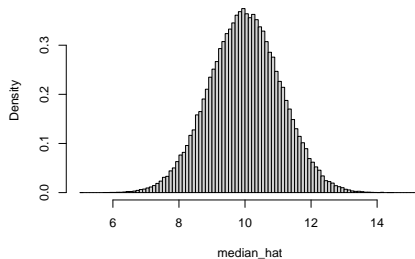
---

# Distribuzione degli stimatori

Histogram of  $\mu_{\text{hat}}$



Histogram of  $\text{median}_{\text{hat}}$



---

```
par(mfrow = c(1, 2))  
hist(mu_hat, breaks = 100, freq = F)  
hist(median_hat, breaks = 100, freq = F)
```

---



# Commenti conclusivi

- Le approssimazioni coinvolte in questa ultima discussione sono di due **differenti tipologie**.
- Da un lato abbiamo la **variabilità** di  $\hat{\theta}$ , che è legata ai dati  $y_1, \dots, y_n$  e alla loro numerosità campionaria  $n$ .
- Dall'altro abbiamo la **variabilità Monte Carlo** di  $\widehat{\text{MSE}\{\theta(Y)\}}$ , che è invece legata alle repliche Monte Carlo e al numero di simulazioni  $R$ .
- Questi due concetti sono ben distinti e non vanno confusi tra loro.
- Inoltre, mentre aumentare il numero di simulazioni  $R$  è sempre possibile (basta aspettare più tempo), non sempre disponiamo di dati aggiuntivi.

# Esercizio riassuntivo (da fare a casa)

- Si supponga che  $Y_1, \dots, Y_n$  sono variabili aleatorie iid distribuite come un normale di media nota  $\mathbb{E}(Y_1) = 0$ , con  $n = 20$ .
- La varianza  $\sigma^2$  è **ignota** e siamo interessati a stimarla.
- Si calcolino tramite simulazione la **distorsione** e l'**errore quadratico** dei seguenti stimatori della varianza, quando  $\sigma^2 = 16$ :

$$\begin{aligned} S_1^2 &= \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2, & S_2^2 &= \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2, \\ S_3^2 &= \frac{1}{n} \sum_{i=1}^n Y_i^2, & S_4^2 &= \frac{1}{n+1} \sum_{i=1}^n (Y_i - \bar{Y})^2. \end{aligned}$$

# Schema della soluzione

```
set.seed(520)
R <- 10^5; n <- 20
mu <- 0; sigma2 <- 16

# Definisco le funzioni che calcolano gli stimatori
var1 <- function(x) mean(x^2) - mean(x)^2
var2 <- function(x) var(x) # Coincide con la definizione di R
var3 <- function(x) mean(x^2)
var4 <- function(x) (length(x) - 1) / (length(x) + 1) * var(x)

# Esecuzione della simulazione
S2_1 <- replicate(R, var1(rnorm(n = n, mean = mu, sd = sqrt(sigma2))))
S2_2 <- replicate(R, var2(rnorm(n = n, mean = mu, sd = sqrt(sigma2))))
S2_3 <- replicate(R, var3(rnorm(n = n, mean = mu, sd = sqrt(sigma2))))
S2_4 <- replicate(R, var4(rnorm(n = n, mean = mu, sd = sqrt(sigma2))))

# Distorsioni (approssimate)
round(mean(S2_1 - sigma2), 2)
round(mean(S2_2 - sigma2), 2)
round(mean(S2_3 - sigma2), 2)
round(mean(S2_4 - sigma2), 2)

# Errore quadratico medio (approssimato)
mean((S2_1 - sigma2)^2)
mean((S2_2 - sigma2)^2)
mean((S2_3 - sigma2)^2)
mean((S2_4 - sigma2)^2)
```

- Sia  $y_1, \dots, y_n$  un campione iid da una distribuzione uniforme in  $(0, \theta)$ , dove  $\theta > 0$  è un parametro ignoto. La stima di massima verosimiglianza in questo caso è pari a

$$\hat{\theta}_n = \max\{X_1, \dots, X_n\}.$$

- Vogliamo verificare tramite simulazione se lo stimatore è **consistente**, ovvero se

$$\hat{\theta}(Y) \xrightarrow{P} \theta.$$

- In pratica, ciò che possiamo fare è simulare alcuni valori di  $\hat{\theta}$  per valori di  $n$  crescenti e controllare se questi si avvicinano sempre più a  $\theta$ .

# Consistenza II

- Supponiamo che il **vero valore** del parametro sia  $\theta = 40$ .

---

```
theta0 <- 40

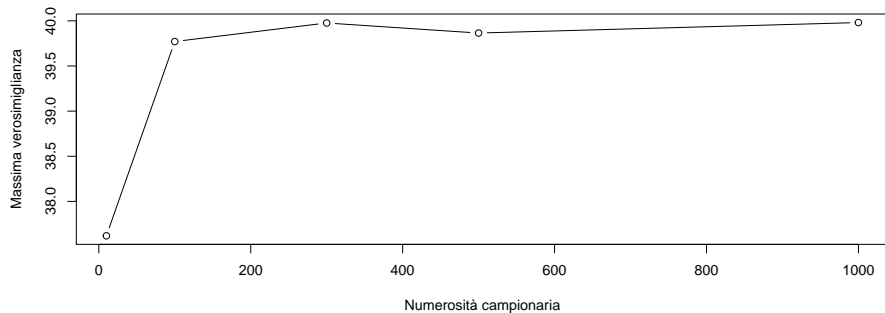
# Numerosità campionarie
nn <- c(10, 100, 300, 500, 1000)

# Stime di massima verosimiglianza
set.seed(123)
theta_hat <- c(
  max(runif(nn[1], min = 0, max = theta0)),
  max(runif(nn[2], min = 0, max = theta0)),
  max(runif(nn[3], min = 0, max = theta0)),
  max(runif(nn[4], min = 0, max = theta0)),
  max(runif(nn[5], min = 0, max = theta0))
)
theta_hat
# [1] 37.61869 39.77079 39.97618 39.86469 39.98096
```

---

- All'aumentare di  $n$ , lo stimatore **tende** a diventare sempre più preciso. Per valori di  $n$  ancora maggiori di 1000, la precisione aumenta ulteriormente.

# Consistenza III



---

```
plot(nn, theta_hat,  
     type = "b",  
     xlab = "Numerosità campionaria",  
     ylab = "Massima verosimiglianza"  
)
```

---