UNIVERSITÀ DEGLI STUDI DI BERGAMO

Dipartimento di Ingegneria Gestionale, dell'Informazione e della Produzione

IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Corso di Laurea Magistrale in INGEGNERIA INFORMATICA

Lezione 2: Richiami di statistica

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di Bergamo

Syllabus

Parte I: sistemi statici

1. **Richiami di statistica**
2. **Teoria della stima**
   - 2.1 Proprietà degli stimatori
3. **Stima a minimi quadrati**
   - 3.1 Stima di modelli lineari
   - 3.2 Algoritmo del gradient descent
4. **Stima a massima verosimiglianza**
   - 4.1 Proprietà della stima
   - 4.2 Stima di modelli lineari
5. **Regressione logistica**
   - 5.1 Stima di un modello di regressione logistica
6. **Fondamenti di machine learning**
   - 6.1 Bias-Variance tradeoff
   - 6.2 Overfitting
   - 6.3 Regolarizzazione
   - 6.4 Validazione
7. **Cenni di stima Bayesiana**
   - 7.1 Probabilità congiunte, marginali e condizionate
   - 7.2 Connessione con Filtro di Kalman

IMAD

Parte I: sistemi statici

Stima parametrica $\hat{\theta}$

- **θ deterministico**
  - ***NO assunzioni su ddp dei dati***
    - ✓ Stima parametri popolazione
    - ✓ Stima modello lineare: minimi quadrati
  - ***SI assunzioni su ddp dei dati***
    - ✓ Stima massima verosimiglianza parametri popolazione
    - ✓ Stima modello lineare: massima verosimiglianza
    - ✓ Regressione logistica
- **θ variabile casuale**
  - ***SI assunzioni su ddp dei dati***
    - ✓ Stima Bayesiana

Machine learning

Parte II: sistemi dinamici

Stima parametrica $\hat{\theta}$

- **θ deterministico**
  - ***NO assunzioni su ddp dei dati***
    - ✓ Modelli lineari di pss
    - ✓ Predizione
    - ✓ Identificazione
    - ✓ Persistente eccitazione
    - ✓ Analisi asintotica metodi PEM
    - ✓ Analisi incertezza stima (numero dati finito)
    - ✓ Valutazione del modello

Outline

1. Definizione e proprietà delle variabili casuali: caso scalare
2. Definizione e proprietà delle variabili casuali: caso multivariabile
3. Stima e stimatori
4. Proprietà degli stimatori

Outline

1. **Definizione e proprietà delle variabili casuali: caso scalare**
2. Definizione e proprietà delle variabili casuali: caso multivariabile
3. Stima e stimatori
4. Proprietà degli stimatori

Variabili casuali (random variables)

Intuizione: una **variabile casuale** v è una variabile definita a partire dall'**esito** s di un **esperimento casuale**

Esempio: l'esperimento è il lancio di una moneta. A seconda se l'esito è $s = \text{testa}$ o $s = \text{croce}$, la variabile v assume un valore diverso

$$v = \begin{cases} 1 & s = \text{testa} \\ 0 & s = \text{croce} \end{cases}$$

- Indichiamo una variabile casuale (v.c.) come $v(s)$
- Il valore assunto da una v.c. v a seguito di un particolare esito $\bar{s}$ è $v(\bar{s})$

Problema: dato che v può assumere diversi valori (a seconda del valore assunto da s), come posso descriverli?

⇩

Assegno una probabilità che ogni esito accada. Questo influisce sulla probabilità che v assuma i valori che può assumere *(distribuzione di probabilità)*

Variabili casuali (random variables)

Caso 1) v assume valori DISCRETI (v è una variabile casuale discreta)

- **Funzione di probabilità di massa (pmf)** $p(x) = P(v = x)$

  Associa ad ogni valore x di v una probabilità

Indichiamo con x_i i valori di v. Se v può assumere m diversi valori, allora $\sum_{i=1}^{m} p(x_i) = 1$

Esempio: Esperimento «lancio di un dado»

$$
\begin{array}{lll}
m = 6 & x_1 = 1 & p(x_1) = P(v = x_1) = P(v = 1) = 1/6 \\
 & x_2 = 2 & p(x_2) = P(v = x_2) = P(v = 2) = 1/6 \\
 & \vdots & \vdots \\
 & x_6 = 6 & p(x_6) = P(v = x_6) = P(v = 6) = 1/6
\end{array}
$$

Variabili casuali (random variables)

Caso 2) v assume valori CONTINUI (v è una variabile casuale continua)

- **Funzione di densità di probabilità (pdf)** $f_v(x)$

In questo caso, dire $P(v = x)$ **non ha senso**. Infatti, dato che v può assumere **infiniti valori**, la probabilità che v assuma esattamente un valore specifico è praticamente zero!

$$\Rightarrow \quad P(v = x) = 0$$

Intuizione: se la variabile v (continua) assumesse valori tutti equiprobabili (come nel caso del dado), la probabilità che v assuma una valore specifico sarebbe $\frac{1}{\infty} = 0$

Esempio: sia v l'altezza di un uomo adulto. Non ha senso chiedersi la probabilità che un uomo sia alto **esattamente** $1{,}74254154787951217953 87$ metri

Variabili casuali (random variables)

La pdf $f_v\ (x)$ definisce la probabilità che v appartenga ad un **intervallo di valori** $[a, b]$

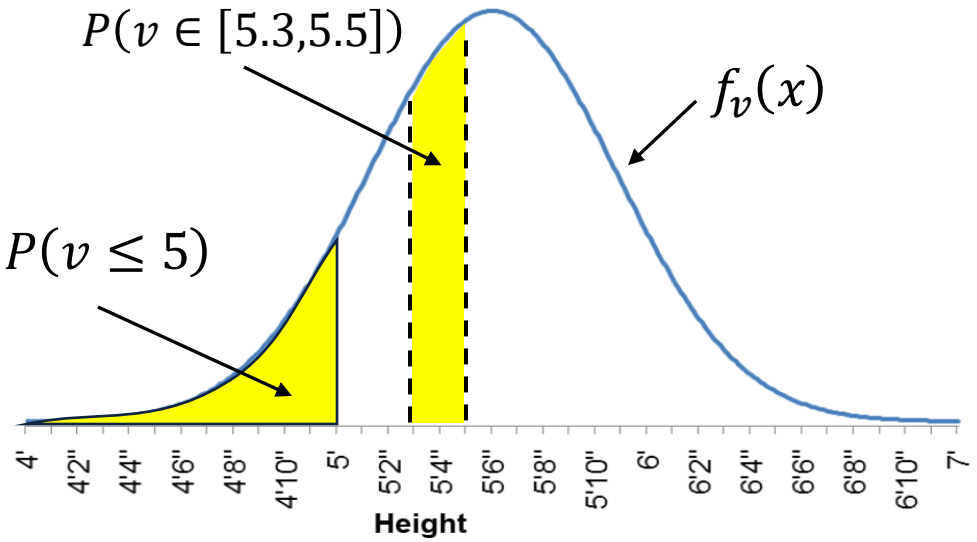
$$P(v \in [a,b]) = \int_a^b f_v(x)\, dx$$

- $f_v(x) \geq 0$
- $\int_{-\infty}^{+\infty} f_v(x) = 1$

- **Funzione di densità cumulata (cdf)** $F_v(z)$

$$F_v(z) = \int_{-\infty}^{z} f_v(x)\, dx = P(v \leq z)$$

$F_v(5) = P(v \leq 5)$



Valore atteso

Il **valore atteso** di una variabile casuale v è:

$$\mathbb{E}_s[v] = \int_{-\infty}^{+\infty} x \cdot f_v(x)\, dx$$

Somma pesata dei valori x che v può assumere. I pesi sono la probabilità di osservare il valore x

Il valore atteso gode della proprietà di **linearità:**

$$\mathbb{E}_s[\alpha \cdot v_1 + \beta \cdot v_2 + \gamma] = \alpha \cdot \mathbb{E}_s[v_1] + \beta \cdot \mathbb{E}_s[v_2] + \gamma \qquad \forall\, \alpha, \beta, \gamma \in \mathbb{R}$$

Nota: l'operatore valore atteso $\mathbb{E}_s[v]$ considera **tutti i possibili esiti** s della variabile casuale v. Di seguito, per semplicità, **renderemo implicita la dipendenza** da s, esplicitandola quando necessario

Varianza

La **varianza** di una variabile casuale v è:

$$\mathrm{Var}[v] = \int_{-\infty}^{+\infty} (x - \mathbb{E}[v])^2 \cdot f_v(x)\, dx$$

- Quanto i valori x si discostano dalla loro media
- Se varianza piccola, v assume valori x molto vicini fra loro

Osservazioni

- $\mathrm{Var}[v] \geq 0$. Se $\mathrm{Var}[v] = 0$, la variabile v è deterministica (assume sempre un solo valore)
- Deviazione standard: $\sigma[v] = \sqrt{\mathrm{Var}[v]}$
- $\mathrm{Var}[v] = \mathbb{E}[(v - \mathbb{E}[v])^2] = \mathbb{E}[v^2 - 2\mathbb{E}[v]v + \mathbb{E}[v]^2] = \mathbb{E}[v^2] - 2\mathbb{E}[\mathbb{E}[v]v] + \mathbb{E}[\mathbb{E}[v]^2]$

  $= \mathbb{E}[v^2] - 2\mathbb{E}[v] \cdot \mathbb{E}[v] + \mathbb{E}[v]^2$ $= \mathbb{E}[v^2] - \mathbb{E}[v]^2$
- $\mathrm{Var}[\alpha \cdot v + \beta] = \alpha^2 \cdot \mathrm{Var}[v] \quad \forall\, \alpha, \beta \in \mathbb{R}$

Correlazione

Date due variabili casuali v_1 e v_2, si definisce il **coefficiente di correlazione** come:

$$\rho[v_1, v_2] = \frac{\mathbb{E}[(v_1 - \mathbb{E}[v_1]) \cdot (v_2 - \mathbb{E}[v_2])]}{\sigma[v_1] \cdot \sigma[v_2]}$$

- ρ indica il grado di **dipendenza lineare** tra v_1 e v_2. Infatti, se $v_2 = \alpha v_1 + \beta$, si ha $\rho = 1$
- Se $\rho = 0$, le due variabili si dicono **scorrelate**

Covarianza

Date due variabili casuali v_1 e v_2, si definisce la **covarianza** come:

$$\mathrm{Cov}[v_1, v_2] = \mathbb{E}[(v_1 - \mathbb{E}[v_1]) \cdot (v_2 - \mathbb{E}[v_2])]$$

E quindi

$$\rho[v_1, v_2] = \frac{\mathrm{Cov}[v_1, v_2]}{\sigma[v_1] \cdot \sigma[v_2]}$$

- Le variabili casuali v_1 e v_2 sono **scorrelate** se $\mathrm{Cov}[v_1, v_2] = 0$

Outline

1. Definizione e proprietà delle variabili casuali: caso scalare

2. **Definizione e proprietà delle variabili casuali: caso multivariabile**

3. Stima e stimatori

4. Proprietà degli stimatori

Variabili casuali, caso multivariabile

Le precedenti definizioni si possono estendere al caso di **vettore** di variabili casuali

$$\underset{d \times 1}{\boldsymbol{v}} = [v_1, v_2, \dots, v_d]^\top \in \mathbb{R}^{d \times 1}$$

Assumiamo che v sia una v.c. continua

- **Funzione di densità cumulata (cdf)** $F_{\boldsymbol{v}}(z_1, z_2, \dots, z_d)$

$$F_{\boldsymbol{v}}(z_1, z_2, \dots, z_d) = P(v_1 \leq z_1, v_2 \leq z_2, \dots, v_d \leq z_d)$$

$$= \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} \cdots \int_{-\infty}^{z_d} f_{v_1, v_2, \dots, v_d}(x_1, x_2, \dots, x_d)\, dx_1\, dx_2 \, \dots dx_d$$

Pdf congiunta

Variabili casuali, caso multivariabile

- Il **valore atteso** è un vettore colonna di d componenti

$$\underset{d \times 1}{\mathbb{E}[\boldsymbol{v}]} = \left[\mathbb{E}[v_1], \mathbb{E}[v_2], \ldots, \mathbb{E}[v_d]\right]^\top \in \mathbb{R}^{d \times 1}$$

- La **varianza** è una matrice $d \times d$ **semidefinita positiva** e **simmetrica**

$$\mathrm{Var}[\boldsymbol{v}] = \int_{\mathbb{R}^d} (\boldsymbol{x} - \mathbb{E}[\boldsymbol{v}])\,(\boldsymbol{x} - \mathbb{E}[\boldsymbol{v}])^\top f_{\boldsymbol{v}}(\boldsymbol{x})\, d\boldsymbol{x}$$

$$= \begin{bmatrix} \mathrm{Var}[v_1] & \cdots & \mathrm{Cov}[v_1, v_d] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[v_d, v_1] & \cdots & \mathrm{Var}[v_d] \end{bmatrix}$$

- «simile» al ≥ 0 per numeri reali
- Una matrice M reale e simmetrica è definita positiva se $\boldsymbol{z}^\top M \boldsymbol{z} \geq 0 \; \forall \boldsymbol{z} \in \mathbb{R}$
- Autovalori di M sono ≥ 0

Indipendenza

Due variabili casuali v_1 e v_2 con funzione di probabilità congiunta $f_{v_1,v_2}(x_1, x_2)$ si dicono **indipendenti** se

$$f_{v_1,v_2}(x_1, x_2) = f_{v_1}(x_1) \cdot f_{v_2}(x_2)$$

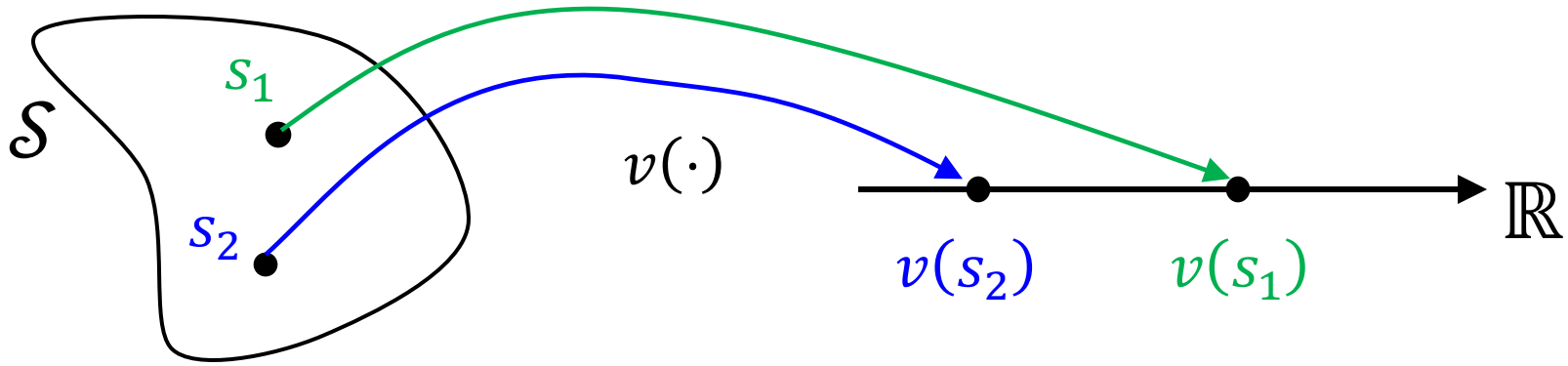
- Se due variabili v_1 e v_2 sono **indipendenti**, allora sono anche scorrelate (non vale il viceversa in quanto potrebbero essere dipendenti in modo **non lineare**)

Variabili casuali: approfondimento

APPROFONDIMENTO

Una definizione più rigorosa di variabile casuale è quella di considerare una v.c. come una **funzione**, che, in funzione di un valore dell'esito s, ritorna un valore della v.c.

Definizione: una variabile casuale (scalare, reale) è una funzione definita sull'insieme degli esiti $\mathcal{S}$, che, ad ogni esito s_i, restituisce un numero reale $v(\cdot): \mathcal{S} \rightarrow \mathbb{R}$



Probabilità: approfondimento

APPROFONDIMENTO

Una definizione più rigorosa di probabilità include la definizione di un **insieme degli eventi**, ovvero di **combinazioni di esiti**

La probabilità è assegnata ad ogni **singolo evento**, e non all'esito (nel caso in cui gli eventi siano i singoli esiti, si ritorna alla nostra definizione intuitiva basata sugli esiti)

Esempio: Lancio di un dado. Supponiamo di essere interessati alla probabilità che esca un numero pari o un numero dispari

Definisco l'insieme degli eventi $\mathcal{P} = \{\{1,3,5\},\{2,4,6\}\}$, i cui elementi (eventi) sono $\{1,3,5\}$ e $\{2,4,6\}$, ed assegno una probabilità ad ognuno di essi:

$$P(\{1,3,5\}) = 1/2 \qquad P(\{2,4,6\}) = 1/2$$

Outline

1. Definizione e proprietà delle variabili casuali: caso scalare
2. Definizione e proprietà delle variabili casuali: caso multivariabile
3. **Stima e stimatori**
4. Proprietà degli stimatori

Teoria della stima

Per **gestire l'incertezza** presente nei dati (e.g. rumore di misura) **interpretiamo** i dati come variabili casuali. I **dati osservati** saranno i valori assunti dalle variabili casuali

In questo corso ci concentreremo sul problema della **stima parametrica**. Vogliamo quindi stimare il vettore di parametri $\boldsymbol{\theta}^0$ che ha generato i **dati** $\mathcal{D} = \{y(1), \dots, y(N)\}$

Esempio: Lancio di una moneta. Osserviamo $N = 8$ dati $\mathcal{D} = \{1,0,0,1,1,1,0,1\}$. In questo caso, il parametro di interesse θ^0 è la probabilità che esca testa

Quindi, i dati $\mathcal{D}$ dipendono sia dall'esito s, sia dai parametri $\boldsymbol{\theta}^0$ ⇨ $\mathcal{D}(s, \boldsymbol{\theta}^0)$

I **dati osservati** dipendono da uno specifico esito $\bar{s}$ ⇨ $\mathcal{D} = \mathcal{D}(\bar{s}, \boldsymbol{\theta}^0)$

Teoria della stima

Uno **stimatore** è una **funzione** $T\left(\mathcal{D}(s, \boldsymbol{\theta}^0)\right)$ dei dati (ovvero, una funzione di variabili casuali)

La **stima** è il risultato di uno stimatore su una specifica realizzazione dei dati $\mathcal{D}(\bar{s}, \boldsymbol{\theta}^0)$

$$\widehat{\boldsymbol{\theta}} = T\left(\mathcal{D}(\bar{s}, \boldsymbol{\theta}^0)\right)$$

Osservazione

Poiché il risultato di $T(\quad)$ dipende dall'esito s (dal quale dipendono i dati), allora **lo stimatore è una variabile casuale** che dipende da s

Teoria della stima

Esempio: Supponiamo di voler **stimare l'altezza media** degli studenti e delle studentesse che seguono il corso di IMAD

Supponiamo di poter misurare solo $N = 10$ persone (se misurassimo tutti, non sarebbe più una stima, ma avremmo il valore vero del parametro «altezza media», cioè θ^0)

- **esito** s_1: primi 10 studenti «estratti» ⇨ $T\left(\mathcal{D}(s_1, \theta^0)\right) = \hat{\theta}_{(s_1)}$
- **esito** s_2: altri 10 studenti «estratti» ⇨ $T\left(\mathcal{D}(s_2, \theta^0)\right) = \hat{\theta}_{(s_2)} \neq \hat{\theta}_{(s_1)}$

La stima $\hat{\theta}_{(s)}$ fornita da $T(\quad)$ dipende da s. Quindi, lo **stimatore è una variabile casuale**

⇩

«Ha senso» parlare di **distribuzione di probabilità**,
valore atteso e **varianza** dello stimatore

Outline

1. Definizione e proprietà delle variabili casuali: caso scalare
2. Definizione e proprietà delle variabili casuali: caso multivariabile
3. Stima e stimatori
4. **Proprietà degli stimatori**

Proprietà di uno stimatore

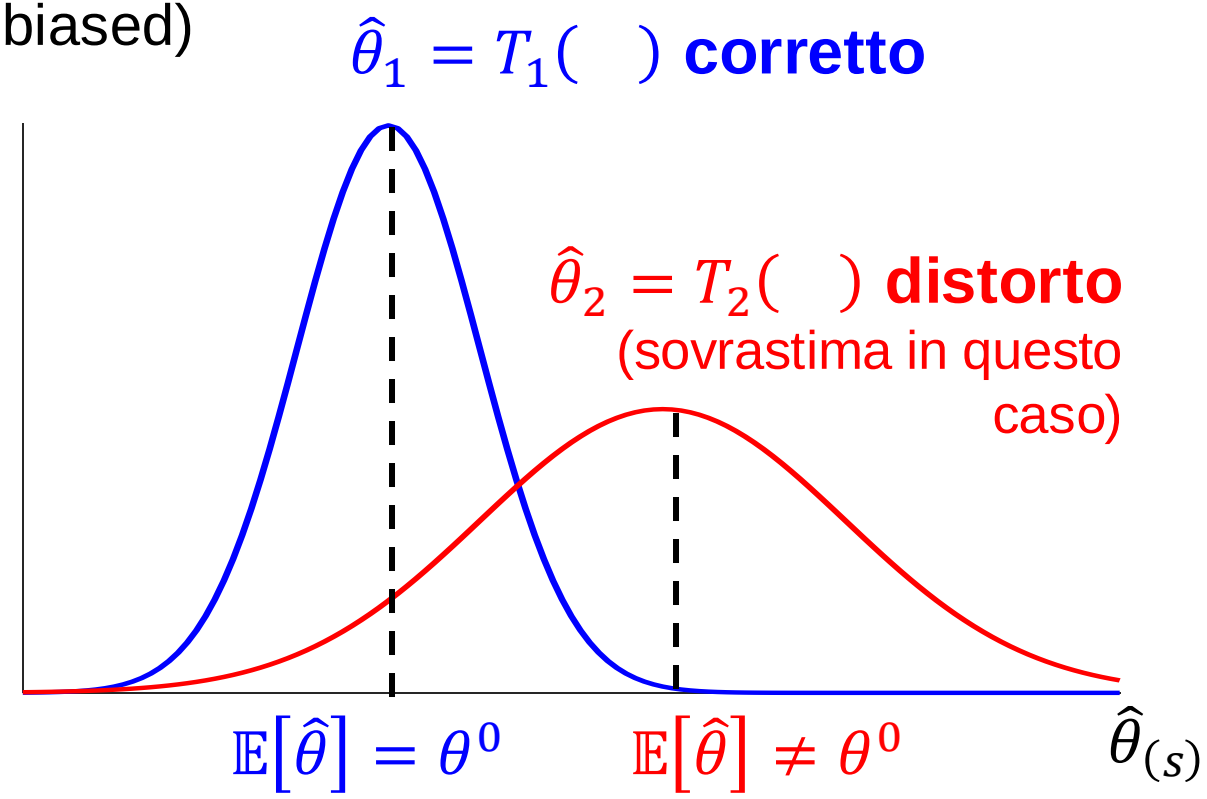
La **«bontà» di uno stimatore** non si giudica da una singola stima, ma dalle caratteristiche della sua distribuzione di probabilità

Correttezza (non polarizzazione, non deviato, unbiased)

Uno stimatore (scalare) $\hat{\theta}$ si dice **corretto** se $\mathbb{E}[\hat{\theta}] = \theta^0$, dove θ^0 è il valore vero del parametro

«In media» lo stimatore mi stima il valore vero del parametro

$$\text{bias} = \mathbb{E}[\hat{\theta}] - \theta^0$$



Proprietà di uno stimatore

Correttezza asintotica

Uno stimatore (scalare) $\hat{\theta}$ si dice **asintoticamente corretto** se $\lim_{N \to +\infty} \mathbb{E}\left[\hat{\theta}\right] = \theta^0$

Proprietà più debole rispetto alla correttezza

Proprietà di uno stimatore

Consistenza

Uno stimatore (scalare) $\hat{\theta}$ si dice **consistente** se, per $N \to +\infty$, $\hat{\theta}$ **converge** a θ^0 in **probabilità**

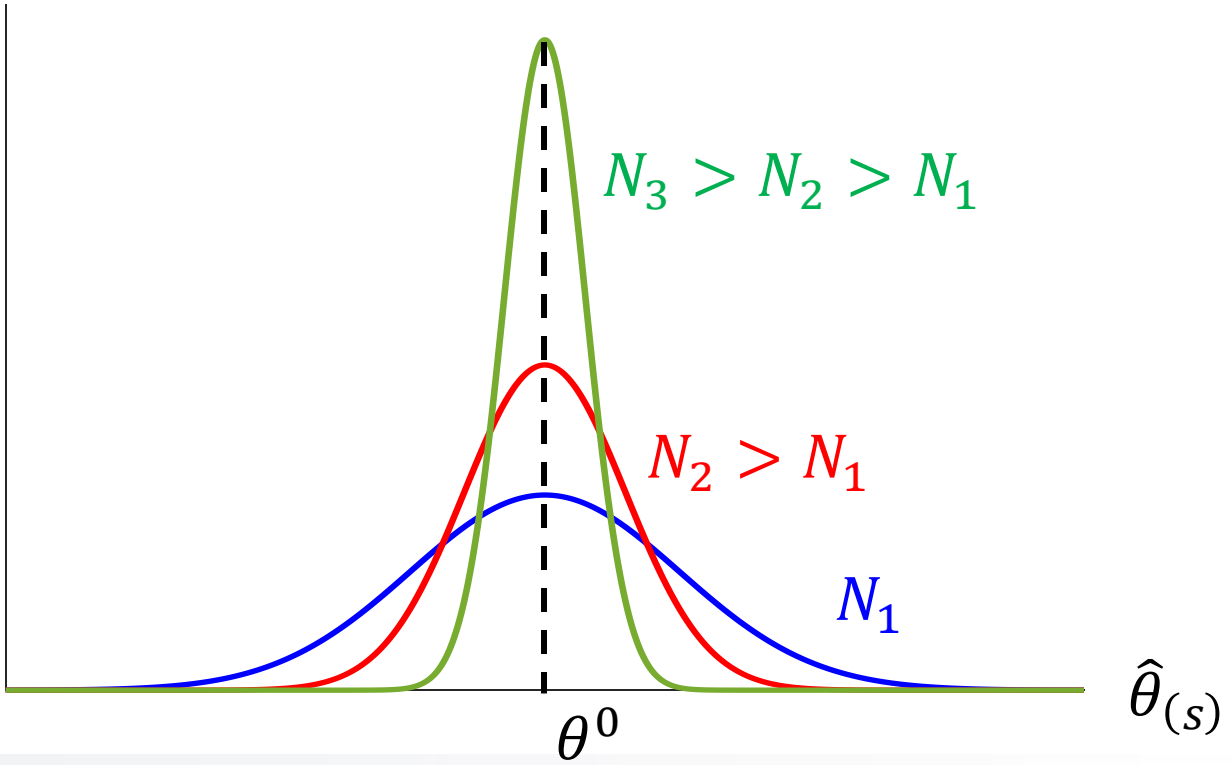
$$\lim_{N\to+\infty} P\left(\left|\hat{\theta} - \theta^0\right| \geq \varepsilon\right) = 0, \quad \forall \varepsilon > 0$$

Al crescere di N, la stima diventa sempre più precisa (la probabilità di commettere un errore $\geq \varepsilon$ tende a 0)

Convergenza in media quadratica

$$\lim_{N\to+\infty} P\left(\left|\hat{\theta} - \theta^0\right|^2\right) = 0$$

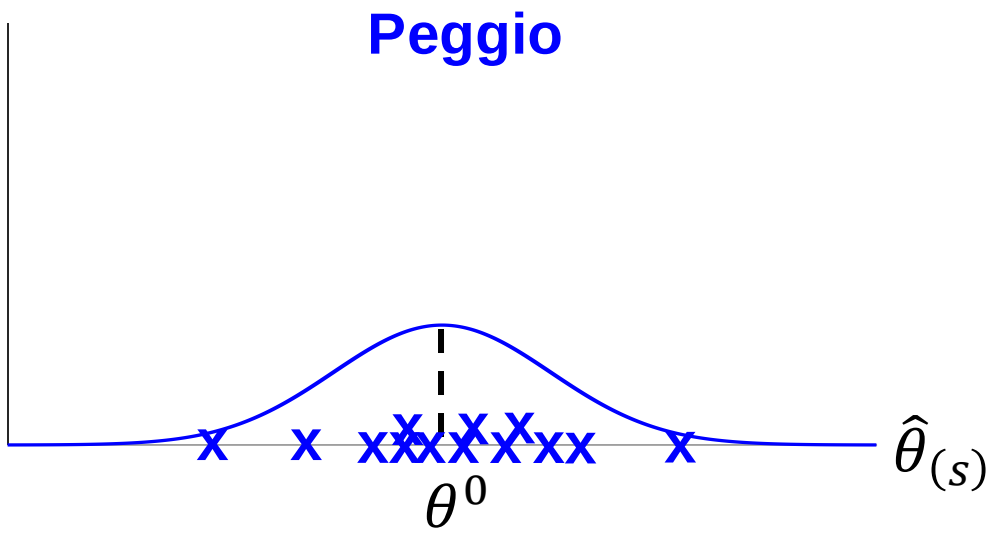
Implica la convergenza in probabilità

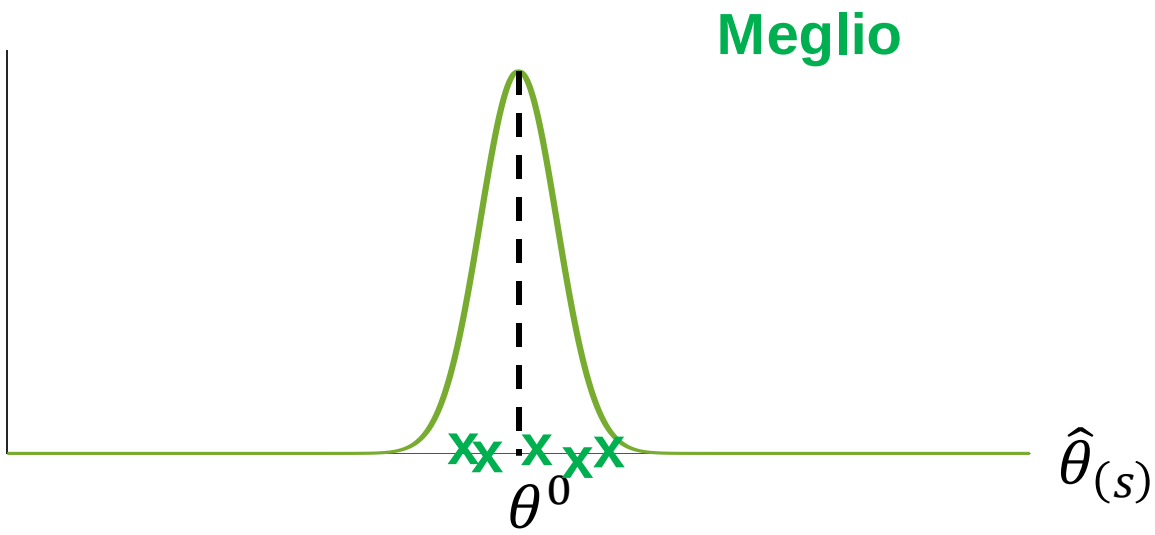


Proprietà di uno stimatore

Cerchiamo di valutare la bontà di uno stimatore senza per forza far riferimento a proprietà asintotiche come la consistenza, quindi per N **finito**

Se due stimatori sono entrambi **corretti**, qual è il migliore? ⇨ Quello a **minima varianza**





Quanto **piccola** può essere la varianza della stima?

Proprietà di uno stimatore

Limite di Cramer-Rao: Dato uno stimatore corretto $\hat{\theta}$, non possiamo rendere la sua varianza più piccola di una certa quantità

Caso scalare $\mathrm{Var}[\hat{\theta}] \geq 1/m$

Caso vettoriale $\mathrm{Var}[\hat{\boldsymbol{\theta}}] \succcurlyeq M^{-1}$

$\mathrm{Var}[\hat{\boldsymbol{\theta}}] - M^{-1}$ **semidefinita positiva**

La quantità m (o M) è detta **quantità (matrice) di informazione di Fisher**

Intuizione: avrò sempre un certo livello di incertezza sui dati che uso per fare la stima, che non posso rimuovere. Quindi, i dati non saranno mai «informativi al 100%» proprio perché affetti da rumore. Esistono dei limiti «strutturali» alla stima

Proprietà di uno stimatore

Efficienza e efficienza asintotica

Uno stimatore (scalare) $\hat{\theta}$ si dice **efficiente** se $\mathrm{Var}\left[\hat{\theta}\right] = 1/m$

Uno stimatore (scalare) $\hat{\theta}$ si dice **asintoticamente efficiente** se $\lim_{N\to+\infty} \mathrm{Var}\left[\hat{\theta}\right] = 1/m$

Minima varianza

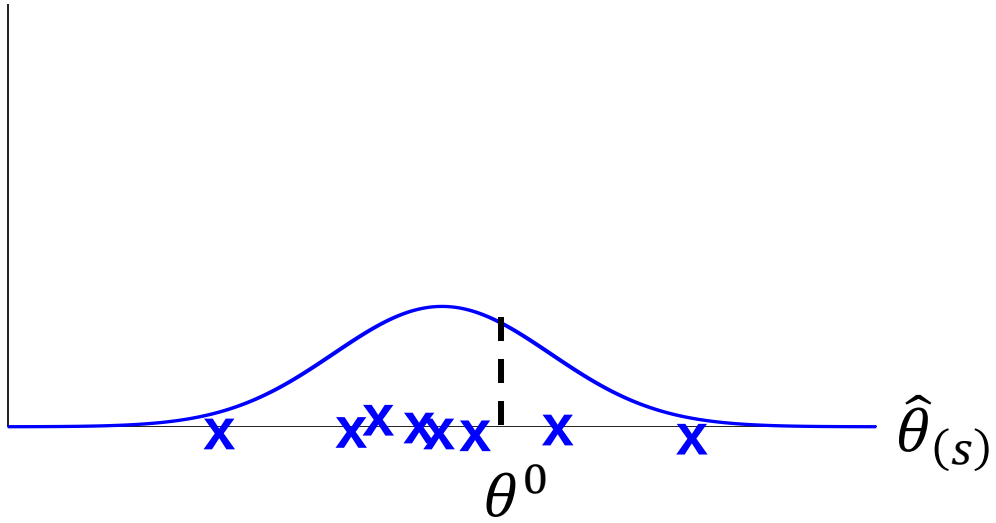
Uno stimatore (scalare) $\hat{\theta}^m$ corretto si dice **a minima varianza** se $\mathrm{Var}\left[\hat{\theta}^m\right] \leq \mathrm{Var}\left[\hat{\theta}\right]$, dove $\hat{\theta}$ è un qualsiasi stimatore corretto

- Se $\hat{\theta}$ è efficiente, allora è a minima varianza
- **Non vale il viceversa**. Ci sono casi in cui esistono stimatori a minima varianza che non sono efficienti. Questo accade quando non esistono stimatori che raggiungono il limite di Cramer-Rao

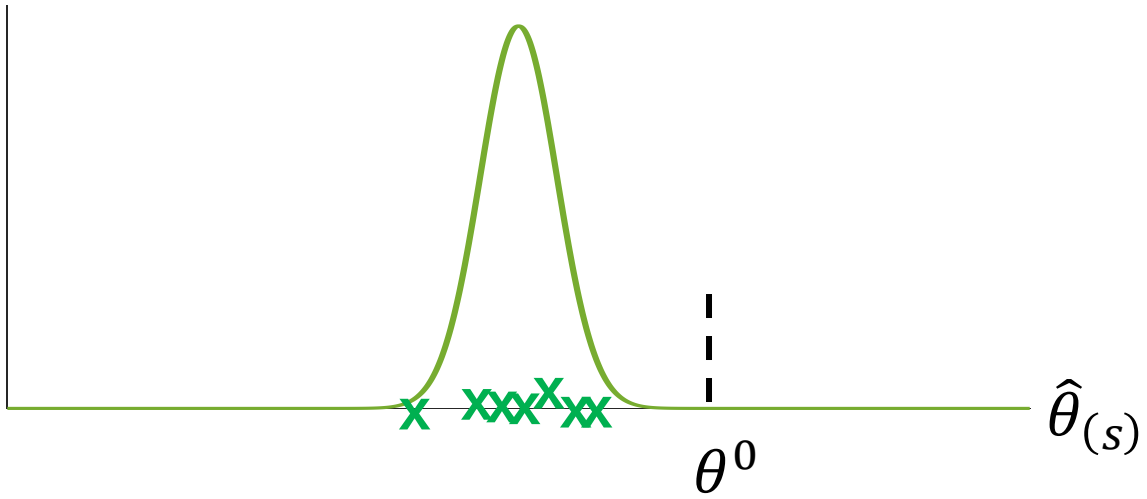
Proprietà di uno stimatore

Mean squared error

Per stimatori **non corretti** (distorti, polarizzati), la varianza, da sola, non è sufficiente come criterio di bontà



Varianza più piccola ma stimatore **«peggiore»** In che senso?



Proprietà di uno stimatore

Abbiamo bisogno di un indicatore «globale», che consideri sia il bias sia la varianza

Idea: uso come criterio **l'errore quadratico medio** (MSE – mean squared error)

$$\mathrm{MSE} = \mathbb{E}\left[\left(\hat{\theta} - \theta^0\right)^2\right]$$

- Caso θ^0 scalare

Proprietà

$$\mathrm{MSE} = \mathrm{bias}\left[\hat{\theta}\right]^2 + \mathrm{Var}\left[\hat{\theta}\right]$$

BIAS-VARIANCE dilemma

Questa proprietà tornerà utile quando vorremo stimare (identificare) modelli dai dati. In quel caso, il «soggetto» non sarà un parametro θ quanto piuttosto l'intero modello (che è di fatto uno stimatore di una funzione)

Esempio (stimatore della media)

Siano $\mathcal{D} = \{y(1), y(2), \dots, y(N)\}$ variabili casuali con media μ e varianza σ^2. Lo stimatore **media campionaria** $\hat{\mu}$ è **corretto** e **consistente**

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y(i)$$

In questo caso il parametro di interesse θ è la media μ della popolazione

Vogliamo dimostrare la correttezza, ovvero che $\mathbb{E}[\hat{\mu}] = \mu$

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} y(i)\right] = \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} y(i)\right] = \frac{1}{N}\mathbb{E}[y(1) + y(2) + \cdots + y(N)] = \frac{1}{N}\cdot N \cdot \mu = \mu$$

Esempio (stimatore della varianza)

Siano $\mathcal{D} = \{y(1), y(2), \dots, y(N)\}$ variabili casuali con media μ e varianza σ^2. Lo stimatore **varianza campionaria** S^2_{N-1} è **corretto**

$$S^2_{N-1} = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (y(i) - \hat{\mu})^2$$

Esercizio: dimostrare la correttezza di S^2_{N-1}. Suggerimenti:

- Usare la proprietà di linearità del valore atteso
- Usare la proprietà della varianza tale che $\mathrm{Var}[v] = \mathbb{E}[v^2] - \mathbb{E}[v]^2$

UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione