

# Inferenza Statistica *Classica*: Verosimiglianza e Stima Puntuale

Patrizio Frederic

Dipartimento di Economia Politica,  
Università di Modena e Reggio Emilia,  
[patrizio.frederic@unimore.it](mailto:patrizio.frederic@unimore.it)

Biostat 2008

1 / 19

## Indice

Probabilità e Verosimiglianza  
Fatti e congetture

Modelli Discreti  
Il modello Binomiale  
Il modello di Poisson

Modelli per variabili continue  
Il modello normale

2 / 19

- ▶ E' anzitutto un problema semantico.
- ▶ Cioè dobbiamo metterci d'accordo su cosa *esattamente* significa probabile cosa significa verosimile.
- ▶ Le fondamenta della scienza statistica sono motivo di continua discussione e dividono gli studiosi in scuole di pensiero.
- ▶ Per comodità dividiamo le scuole in due macro categorie: *Classici e Bayesiani*.
- ▶ Nell'impostazione Classica:
  - ▶ Probabili sono gli effetti.
  - ▶ Verosimili sono le cause.

3 / 19

## Eventi, Verità e Falsità

- ▶ "Uscirà "6" dal lancio di questo dado"
- ▶ "Questo dado è truccato."
- ▶ "Sapendo che  $A$  ha contratto la patologia  $X$ ,  $A$  ha la febbre."
- ▶ "Sapendo che  $A$  ha la febbre,  $A$  ha contratto  $X$ ."
- ▶ "L'indice Down Jones tra 3 ore quoterà 13.100"
- ▶ "La riduzione del costo del denaro incide del  $p\%$  sul Down Jones."
- ▶ "So come è fatta l'urna, con quale probabilità ottengo una data sequenza"
- ▶ "E' uscita una una data sequenza, quanto è verosimile una data conformazione dell'urna"

4 / 19

## $\mathcal{O}$ attrae l'insetto $\mathcal{I}$ ?

- ▶ Si mette l'insetto  $\mathcal{I}$  davanti ad un bivio alla estremità sinistra (sx) viene messo un odore  $\mathcal{O}$  con una data intensità  $x$ , a destra (dx) nessun odore.
- ▶ Vengono rilasciati  $n = 10$  insetti.
  - ▶ ogni insetto  $i = 1, \dots, n$  può andare a sx  $y_i = 1$  oppure a dx  $y_i = 0$ .
  - ▶  $S_n = \sum_{i=1}^n y_i$  =somma(tutti gli  $y_i$ )=numero totale di insetti andati sx
  - ▶  $S_n = 7$
  - ▶  $\hat{p} = S_n/n = 7/10 = 0.7$  ovvero il 70% del campione è andato a sx.
- ▶ Cosa posso concludere sulla propensione di  $\mathcal{I}$  verso  $\mathcal{O}$

5 / 19

## Come se... un urna con composizione incognita

- ▶ Sia  $\mathcal{U}$  un urna che contiene una certa proporzione  $0 \leq \theta \leq 1$  di palline marcate con 1 e il rimanente  $1 - \theta$  di palline marcate con 0.
- ▶ Estraiamo *con reintroduzione*  $n = 10$  palline
  - ▶ Ogni traiettoria scelta da un insetto è COME SE estraessi una pallina da  $\mathcal{U}$ .
  - ▶ La variabilità è individuale di ogni specifico insetto e da tutte le altre condizioni non verificabili sperimentalmente (il resto del mondo).
  - ▶ La propensione di ogni insetto non cambia nel tempo.
  - ▶ Gli insetti non comunicano tra di loro.

6 / 19

- ▶ se conoscessi  $\theta = \theta_0$  con quale probabilità otterrei  $S_n = s_n$  su  $n$  estrazioni?

$$P(S_n = s_n; \theta = \theta_0) = \binom{n}{s_n} \theta_0^{s_n} (1 - \theta_0)^{n-s_n}$$

- ▶ se conoscessi  $\theta = 0.65$  con quale probabilità otterrei  $S_n = 7$  su 10 estrazioni?

$$P(S_n = 7; \theta = 0.65) = \binom{10}{7} (0.65)^7 (1 - 0.65)^{10-7} = 0.2522$$

## Ammesso che le ipotesi calzino, se sapessi... un problema di inferenza

- ▶ non conosco  $\theta$  ma so che ho ottenuto  $S_n = s_n$  successi su  $n$  estrazioni con reintroduzione. Quanto è verosimile che un dato  $\theta_0$ ?
- ▶ Cos'è una misura di verosimiglianza?
- ▶ Definiamo  $L(\theta; S_n = s_n)$  la funzione di verosimiglianza per  $\theta$

$$L(\theta; S_n = s_n) = \text{Const} \cdot P(S_n = s_n; \theta)$$

- ▶ Ad esempio, supposto  $S_n = 7$ ,  $n = 10$ , ecco la verosimiglianza per alcuni valori di  $\theta$

$$\begin{aligned} L(\theta = 0.00; S_n = 7) &= 0 \\ L(\theta = 0.10; S_n = 7) &< 10^{-5} \\ L(\theta = 0.65; S_n = 7) &= 0.2522 \\ L(\theta = 0.70; S_n = 7) &= 0.2668. \end{aligned}$$

## Problema di probabilità

So com'è fatta l'urna (conosco  $\theta = \theta_0$ ), con quale probabilità estraggo  $S_n = s_n$ ?

$$P(S_n = s_n; \theta = \theta_0)$$

## Problema di Inferenza

Ho ottenuto  $S_n = s_n$ , quanto è verosimile  $\theta = \theta_0$ ?

$$L(\theta_0; S_n = s_n)$$

9 / 19

## Cos'è uno stimatore?

- ▶ Sia  $\mathcal{S}$  lo spazio dei campioni e sia  $\Omega$  lo spazio dei parametri;
- ▶ in questo caso:  $\mathcal{S} = \{0, 1, 2, \dots, 10\}$  e  $\Theta \equiv [0, 1]$
- ▶ Uno stimatore è una funzione  $h : \mathcal{S} \rightarrow \Theta$ .
- ▶ Una stima è la valutazione del singolo campione  $s \in \mathcal{S}$ :  
 $h(s) = \hat{\theta}$ .
- ▶ Nella statistica classica le proprietà degli stimatori vengono calcolate speculativamente considerando ogni possibile campione (quello osservato e quelli non osservati).
- ▶ Viene data molta importanza alle seguenti:
  - ▶ **CONSISTENZA**:  $h(s) \rightarrow \theta_{Vero}$  se  $n \rightarrow \infty$
  - ▶ **EFFICIENZA**:  $h_1$  è + efficiente di  $h_2$  se  $Var(h_1) < Var(h_2)$

10 / 19

E' quel valore di  $\theta$  che rende massima la funzione di verosimiglianza calcolata sui dati disponibili.

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} L(\theta; S_n = s_n)$$

ottenendo così:

- ▶  $\hat{\theta}$  consistente;
- ▶ per molti modelli è il più efficiente ( $\forall n$ );
- ▶ per molti modelli è il più efficiente asintoticamente ( $n$  sufficientemente grande);
- ▶ la curvatura della verosimiglianza è lo stimatore della varianza e dunque la sua precisione.

## AL LAVORO!

Cosa aspettiamo...

- ▶ Apriamo R!
- ▶ lo script si trova a questo indirizzo:  
<http://>

- ▶ The data give the ant species richness (number of ant species) found in 64 square meter sampling grids, in 22 bogs and 22 forests surrounding the bogs, in Connecticut, Massachusetts and Vermont (USA). The sites span a 3 dg of latitude in New England. Aaron M Ellison (2004). *Bayesian inference in ecology. Ecology Letters*, 7, 509-520
- ▶ ogni sito  $i = 1, \dots, n = 22 + 22$  può avere un numero casuale di specie diverse  $y_i \in \{0, 1, 2, \dots\}$ .
- ▶  $S_n = \sum_{i=1}^n y_i =$  somma(tutti gli  $y_i$ )=numero totale di specie negli  $n$  siti.
- ▶  $S_n = 309$
- ▶  $\bar{y} = S_n/n = 309/44 = 7.023$  numero medio di specie per sito.

13 / 19

## Come se... un'urna con composizione incognita

- ▶ Sia  $\mathcal{U}$  un'urna che contiene infinite palline ognuna marcata con un numero intero, tale che

$$P(Y_i = y; \theta) = \frac{\theta^y e^{-\theta}}{y!}, y = 0, 1, 2, \dots$$

dove  $\theta$  è il parametro incognito che definisce l'urna.

- ▶ Estraiamo *con reintroduzione*  $n = 10$  palline
  - ▶ Ogni sito ha un numero di specie COME SE estraessi una pallina da  $\mathcal{U}$ .
  - ▶ Il parametro  $\theta$  non cambia nel tempo e con l'osservazione.

14 / 19

- ▶ E' una funzione di  $\theta$

$$L(\theta; Y_i = y_i) \propto \prod_{i=1}^n \frac{\theta^{y_i} e^{-\theta}}{y_i!},$$
$$\propto \theta^{\sum_{i=1}^n y_i} e^{-n\theta}$$

dove  $\theta$  è il parametro incognito che definisce l'urna.

- ▶ Estraiamo *con reintroduzione*  $n = 10$  palline
  - ▶ Ogni sito ha un numero di specie COME SE estraessi una pallina da  $\mathcal{U}$ .
  - ▶ Il parametro  $\theta$  non cambia nel tempo e con l'osservazione.
- ▶ La log-verosimiglianza è:

$$\ell(\theta) = n\bar{y} \log \theta - n\theta$$

## AL LAVORO!

### Cosa aspettiamo...

- ▶ Apriamo R!
- ▶ lo script si trova a questo indirizzo:  
<http://>

Suppose you are a nutritionist interested in the relative merits of two diets, one featuring high protein, the other low protein. Do the two diets lead to differences in mean weight gain? Consider the data in Table 5.1, which shows the weight gains (in grams) for two lots of female rats, under the two diets.

*S-PLUS 6 for Windows Guide to Statistics, Volume 1, Insightful Corporation, Seattle, WA. Printed in the United States.*

High Protein	134	146	104	119	124	161	107	83	113	129	97	123
Low Protein	70	118	101	85	107	132	94					

## Come se... una sola urna con composizione incognita

- ▶ Sia  $\mathcal{U}$  un'urna che contiene infinite (ma con la potenza del continuo!!!) palline ognuna marcata con un numero intero, tale che

$$\begin{aligned} P(y_1 < Y_i \leq y_2; \theta) &= \int_{y_1}^{y_2} f(y; \theta) dy \\ &= F(y_2; \theta) - F(y_1; \theta), \end{aligned}$$

dove  $f(y; \theta)$  è la densità di una normale,  $F(y; \theta)$  la ripartizione e  $\theta = (\mu, \sigma^2)$  il parametro incognito che definisce l'urna.

- ▶ Estraiamo *con reintroduzione*  $n = 12 + 7$  palline
  - ▶ Ogni sito ha un numero di specie COME SE estraessi una pallina da  $\mathcal{U}$ .
  - ▶ Il parametro  $\theta$  non cambia nel tempo e con l'osservazione.

## Cosa aspettiamo...

- ▶ Apriamo R!
- ▶ lo script si trova a questo indirizzo:  
`http://`