Appunti di Statistica Progredito

Daniele Zago

28 febbraio 2021

Indice

Lezione 1

Riferimenti Pace e Salvan (2001, § 1.1-1.4)
Azzalini (2001, § 1.1-1.3, 1.5, 2.1)

In generale chiamiamo ϑ_0 il *vero valore* del parametro, che identifica univocamente la *vera densità* generatrice dei dati

$$p^0(y) = p(y; \vartheta_0).$$

Nonostante si chiami vera densità, in tutte le applicazioni $\mathcal{F}$ viene considerato solo come un'approssimazione sufficientemente accurata per descrivere il fenomeno. Infatti, se si conoscesse esattamente il meccanismo generatore dei dati, non sarebbe necessaria alcuna inferenza statistica.

Un modello statistico si può definire sia tramite funzione di densità, sia attraverso altre specificazioni. Ad esempio, se Y è univariata:

1. Funzione di ripartizione $F(y) = P(Y \leq y)$
2. Funzione generatrice dei momenti $M(t) = \mathbb{E}[e^{tY}]$
3. Il tasso di guasto $r(y) = p(y)/\left(1 - F(y)\right)$, ovvero

$$r(y)dy = P\big(Y \in (y, y + dy) | Y > y\big),$$

da cui si ottiene la relazione

$$p(y) = r(y)R(y) = r(y) \exp\{-\int_0^y r(t)\, dt\}.$$

Quando necessario specificare a che variabile si riferiscono, le funzioni di Y verranno indicate con $p_Y(y), F_Y(y), \ldots$

Esempio (Modello binomiale)
I dati sul comportamento aggressivo sono costituiti da un campione di $n = 707$ adolescenti, di cui $y^{\text{oss}} = 159$ hanno avuto comportamento aggressivo.

Modello statistico: definiamo $\mathcal{F}$ come famiglia di distribuzioni binomiali, con ϑ probabilità di avere comportamento aggressivo, assumendo i.i.d.:

$$p(y; \vartheta) = \binom{707}{y} \vartheta^y (1 - \vartheta)^{707 - y},$$

che ovviamente può essere complicato, mettendo in relazione ϑ con le altre informazioni disponibili (esposizione alla TV, ...)

Esempio (Modello multinomiale)
Nell'esempio del genetic linkage, si ha un campione di $n = 197$ animali, classificati in categorie

$y = (125, 18, 20, 34)$. Indicando y_j, con $j = 1, \dots, 4$ il numero di categorie osservate, si ha

$$Y = (Y_1, Y_2, Y_3, Y_4)$$

che rappresenta il vettore delle frequenze in ogni categoria.

Modello statistico: assumiamo una distribuzione multinomiale $Y \sim \text{Mult}(197, \vartheta)$ con $\vartheta = (\vartheta_1, \vartheta_2, \dots, \vartheta_4)$, $\vartheta \in \Delta_4$ simplesso 4-dimensionale.

Da qui, si ottengono

$$Y = \{y : y_j \in \{0, \dots, n\}, \sum_{j=1}^{4} y_j = 197\}$$

$$\Theta = \{\vartheta : \vartheta \in (0,1), \sum_{j=1}^{4} \vartheta = 1\}$$

$$p(y; \vartheta) = \binom{197}{y_1\, y_2\, y_3\, y_4} \vartheta_1^{y_1} \vartheta_2^{y_2} \vartheta_3^{y_3} \vartheta_4^{y_4}.$$

Un modello alternativo tiene conto invece dell'*effetto repulsivo* tra le categorie A e B, imponendo dei vincoli sullo spazio parametrico:

$$Y \sim \text{Mult}(197, \pi(\vartheta)),$$

con ϑ parametro scalare, definito come

$$\pi(\vartheta) = \left(\frac{2+\vartheta}{4}, \frac{1-\vartheta}{4}, \frac{1-\vartheta}{4}, \frac{\vartheta}{4} \right), \quad \vartheta \in (0,1).$$

Lo spazio campionario rimane lo stesso, mentre la verosimiglianza diventa

$$p(y; \vartheta) = \binom{197}{y_1\, y_2\, y_3\, y_4} \left(\frac{2+\vartheta}{4} \right)^{y_1} \left(\frac{1-\vartheta}{4} \right)^{y_2} \left(\frac{1-\vartheta}{4} \right)^{y_3} \left(\frac{\vartheta}{4} \right)^{y_4}.$$

Osservazione

- $\sqrt{\vartheta}$ si dice *fattore di ricombinazione* con cui sono legati A e B, utile nell'interpretazione successiva.
- Nella teoria mendeliana, $\vartheta_0 = 1/4$ fornisce $\pi(\vartheta_0) = (9/16, 3/16, 5/16, 1/16)$.

Esempio (Modello esponenziale)

Nell'esempio della resistenza alla tensione, si hanno 10 osservazioni di uno stress, dunque $y^{\text{oss}} = (225, 171, \dots, 162)$.

Possiamo considerarle v.c. non negative e, anche se la misura è un numero naturale, possiamo

usare come approssimazione una v.c. con supporto $\mathbb{R}^+$, ad esempio $Y \sim \text{Exp}(\vartheta)$ con densità

$$p(y;\vartheta) = \vartheta e^{-\vartheta y}, \quad y > 0, \vartheta > 0.$$

Si ricorda che $\mathbb{E}[Y] = 1/\vartheta$, $\mathbb{V}[Y] = 1/\vartheta^2$, $r(Y) = \vartheta$. La proprietà che il tasso di guasto sia costante si chiama anche *assenza di memoria*, perché la probabilità condizionata non dipende dal momento in cui è stato osservato.

In questo caso,

$$\begin{aligned} p(y;\vartheta) &= \prod_{i=1}^{10} \vartheta e^{-\vartheta y_i} \\ &= \vartheta^{10} e^{-\vartheta \sum_{i=1}^{10} y_i}. \end{aligned}$$

Usando tutte le 60 osservazioni, abbiamo delle coppie (x_i, y_i), dove assumiamo che x_i siano delle costanti note (esperimento, valori prefissati).

Assumendo di nuovo l'indipendenza, possiamo modellare ϑ tramite x_i, ovvero $Y_i \sim \text{Exp}(\vartheta x_i)$, da cui

$$\mathbb{E}[Y_i] = (\vartheta x_i)^{-1}, \quad \mathbb{V}[Y_i] = (\vartheta x_i)^{-2},$$

per cui al decrescere della tensione salgono sia la media sia la varianza.

La densità congiunta è

$$\begin{aligned} p(y_1, \dots, y_{60}; \vartheta) &= \prod_{i=1}^{60} \vartheta x_i \exp\{-\vartheta x_i y_i\} \\ &= \vartheta^{60} \prod_{i=1}^{60} x_i \exp\{-\vartheta \sum_{i=1}^{60} x_i y_i\}, \end{aligned}$$

con spazio campionario $\mathcal{Y} = (\mathbb{R}^+)^{60}$ e spazio parametrico $\vartheta \in \Theta = \mathbb{R}^+$.

Osservazione

In questa modellazione, non sono state prese in considerazioni le censure a destra delle osservazioni.

1.1 Parametrizzazioni

Un modello statistico $\mathcal{F}$ si può parametrizzare in diversi modi, ad esempio

$$\begin{aligned} \mathcal{F} &= \{p(y;\vartheta) = \vartheta e^{-\vartheta y}, \vartheta > 0\} \\ &= \{p(y;\psi) = \frac{1}{\psi} e^{-y/\psi}\}. \end{aligned}$$

Def. (Riparametrizzazione)
In generale, dato un modello statistico $\mathcal{F}$ con spazio parametrico $\Theta \subseteq \mathbb{R}^p$, si dice riparametrizzazione di $\mathcal{F}$ una funzione

$$\psi : \Theta \to \Psi \quad \text{biunivoca,}$$

con $\Psi = \{y \in \mathbb{R}^p : \psi = \psi(\vartheta), \vartheta \in \Theta\}$.

Il modello parametrico tratta il parametro come un'etichetta, per cui è naturale chiedersi se le procedure inferenziali siano *invarianti* rispetto alla parametrizzazione degli elementi di $\mathcal{F}$.

Ad esempio, se $p^0(y) = 2e^{-2y}$, nelle due parametrizzazioni la densità corrisponde a $\vartheta = 2$ e a $\psi = 1/2$.

Alcune procedure inferenziali sono invarianti rispetto alla parametrizzazione del modello, per cui in linea di principio preferibili.

1.2 Informazioni ausiliarie

Informazioni ausiliarie devono essere incluse nel modello statistico, ad esempio lo *schema di osservazione* delle unità statistiche.

Un esempio importante è la **regola di arresto**:

Esempio (Regola di arresto)
Supponiamo che in 10 tiri liberi di un giocatore di basket, tiri indipendenti, ci sia la stessa probabilità ϑ di realizzare un canestro.

Usiamo un modello $Y \sim \text{Bin}(10, \vartheta)$, per cui formalmente si ha

$$\mathcal{F} = \{p(y; \vartheta) : y \in Y = \{0, \dots, 10\}, \vartheta \in \Theta = (0, 1)\},$$

$$p(y; \vartheta) = \binom{10}{y} \vartheta^y (1 - \vartheta)^{10-y}.$$

Si supponga che, invece, si facciano i tiri liberi fino a realizzare 6 canestri. In tal caso, bisogna usare una diversa modellazione, data da $Y \sim \text{Bineg}(6, \vartheta)$ con modello statistico

$$\mathcal{F} = \{p(y; \vartheta), y \in Y = \{6, 7, \dots\}, \vartheta \in \Theta = (0, 1)\}$$

e densità discreta

$$p(y; \vartheta) = \binom{y-1}{5} \vartheta^6 (1 - \vartheta)^{y-6}.$$

Si supponga che nel primo esperimento ci siano 6 realizzazioni e nel secondo esperimento si osservino 10 tiri.

In entrambi i casi ci sono 6 canestri su 10 tiri, ma i modelli statistici sono diversi, perché nel secondo caso è il numero di tiri ad essere variabile.

1.3 Considerazioni finali

Ci sono altri approcci che si possono utilizzare per estrarre informazione dai dati (*Machine Learning*, ...), non necessariamente legati alla probabilità.

Gli approcci algoritmici per la previsione o l'ottimizzazione di funzioni di utilità non per forza utilizzano modelli statistici. Analogamente, spesso non accompagnano le stime con l'errore probabilistico che si commette.

Lezione 2

Riferimenti Evans e Rosenthal (2006, § 7.1)
Efron e Hastie (2016)
Welsh (1996, §2.2)

2.1 Approcci all'inferenza

Non esiste un unico paradigma all'inferenza statistica, ma ci sono delle strutture generali basate sul calcolo delle probabilità. Esistono due paradigmi/filosofie/punti di vista:

1. **Inferenza bayesiana**
2. **Inferenza frequentista**

Entrambi gli approcci assumono un modello statistico $\mathcal{F} = \{p(y;\vartheta), \vartheta \in \Theta, y \in \mathcal{Y}\}$. Le differenze tra i paradigmi si riferiscono all'*interpretazione della probabilità* e agli obiettivi dell'inferenza statistica.

2.2 Inferenza bayesiana

In questo approccio, si intende la probabilità come incertezza sulle quantità osservabili y e future osservazioni y^*, ma anche su quelle non osservabili, come il parametro ϑ.

Oltre al modello statistico $\mathcal{F}$ per y, si assume che ϑ^0 sia una realizzazione di una variabile casuale, con distribuzione sullo spazio parametrico Θ e con *densità a priori* $\pi(\vartheta)$, che riassume l'*informazione preliminare* su ϑ.

Notazione La densità $p(y;\vartheta)$ si indica con $p(y|\vartheta)$, per sottolineare che la distribuzione per y è condizionata a ϑ.

Il *modello bayesiano* è dunque la coppia formata da $\mathcal{F}$ e $\pi(\vartheta)$.

L'aggiornamento dell'informazione a priori $\pi(\vartheta)$, una volta osservati i dati $(y_1^{\text{oss}}, y_2^{\text{oss}}, \ldots, y_n^{\text{oss}})$, si effettua attraverso il teorema di Bayes.

In statistica bayesiana, si usa per effettuare la trasformazione

$$\text{Distribuzione a priori} \quad \Longrightarrow \quad \text{Distribuzione a posteriori}$$

Si consideri un esperimento casuale $\mathcal{E}$ con spazio campionario $\mathcal{S}$ e sia $\mathcal{B}$ una σ-algebra su $\mathcal{S}$.

Teo. (Probabilità totali)
Sia A_i, $i \in I \subseteq \mathbb{N}$ una partizione di $\mathcal{S}$ in eventi non trascurabili, cioè $P(A_i) > 0$ per ogni i. Allora, per ogni $E \in \mathcal{B}$,

$$P(E) = \sum_{i \in I} P(E|A_i)P(A_i).$$

Le $P(A_i)$ sono *probabilità iniziali*, perché riflettono le conoscenze disponibili prima di effettuare l'esperimento casuale.

Supponiamo che ora sia noto che si è realizzato un evento E con associata probabilità $P(E)$. Alla luce di questa conoscenza, le probabilità delle ipotesi A_i vanno rivalutate in $P(A_i|E)$, dette anche *probabilità finali*.

Le probabilità finali sono legate alla probabilità iniziale tramite il *teorema di Bayes*:

Teo. (Teorema di Bayes)
Sia A_i, $i \in I \subseteq \mathbb{N}$ una partizione non trascurabile dello spazio $\mathcal{S}$, allora per ogni $E \in \mathcal{S}$ vale che

$$P(A_i|E) = \frac{P(A_i \cap E)}{P(E)}$$
$$= \frac{P(E|A_i)P(A_i)}{\sum_{j\in I} P(E|A_j)P(A_j)}.$$

Dal punto di vista dell'inferenza statistica, è importante per *invertire* l'ordine del condizionamento $P(A_i|E)$, in termini di $P(E|A_i)$ e della probabilità marginale $P(E)$.

Per v.c. continue, il teorema di Bayes assume la forma

$$p_{Z|Y=y}(z;y) = \frac{p_{Y|Z=z}(y;z)p_Z(z)}{\int_{\mathcal{Z}} p_{Y|Z=z}(y;z)p_Z(z)\,dz}.$$

Il modello bayesiano è costituito dalla coppia $\mathcal{F}$ e $\pi(\vartheta)$ e, da questo punto di vista, si può pensare che il meccanismo generatore dei dati sia

$$\vartheta \sim \pi(\vartheta)$$
$$Y|\vartheta \sim p(\cdot|\vartheta)$$

È importante notare che osserviamo solo y^{oss}, mentre l'esito del primo stadio è ignoto, ed è esattamente quello su cui si vuole fare inferenza.

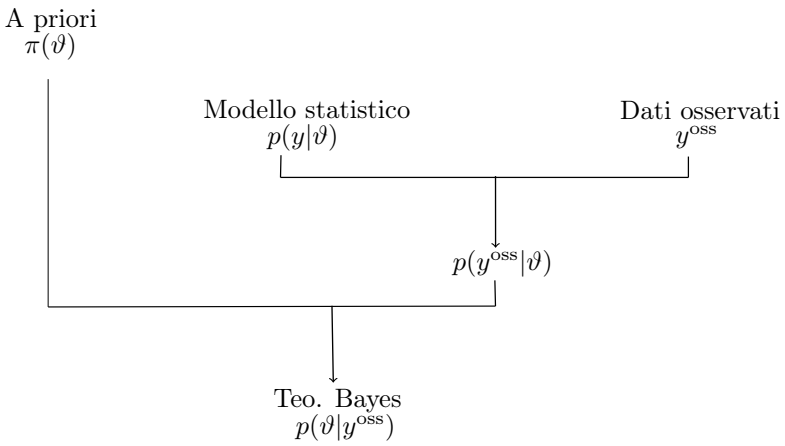


Figura 1: Aggiornamento della distribuzione a priori.

Usiamo il teorema di Bayes per ottenere la *distribuzione a posteriori* per ϑ dato y^{oss}

$$\pi(\vartheta|y^{\text{oss}}) = \frac{p(y^{\text{oss}}|\vartheta)\pi(\vartheta)}{\int_\Theta p(y^{\text{oss}}|\vartheta)\pi(\vartheta)\,d\vartheta}.$$

Ogni oggetto che si considera, in statistica bayesiana, è effettivamente trattato come una variabile casuale, e l'inferenza termina quando si ha in mano la distribuzione a posteriori.

Il problema è, quindi, calcolare $\pi(\vartheta|y^{\text{oss}})$ in modo da poterla utilizzare, riassumere e descrivere in modo adeguato.

Problema

Il denominatore è un integrale, per cui non è facilmente calcolabile per modelli complessi. Un ramo della statistica bayesiana si impegna nello sviluppo di metodi numerici per aggirare il calcolo dell'integrale (metodi MCMC, variazionali, ...).

Osservazioni

La distribuzione a posteriori dipende da y^{oss} solo attraverso $p(y^{\text{oss}}|\vartheta)$ come funzione di ϑ, una volta osservati i dati. Questa funzione, vista come funzione di ϑ, prende il nome di *verosimiglianza*:

$$L(\vartheta) = p(y|\vartheta).$$

La verosimiglianza può essere definita a meno di costanti positive moltiplicative, perché la distribuzione a posteriori viene normalizzata a 1 dal denominatore. Si può allora scrivere

$$\pi(\vartheta|y^{\text{oss}}) \propto L(\vartheta)\pi(\vartheta),$$

ovvero la distribuzione a posteriori è proporzionale al prodotto tra la distribuzione a priori e la verosimiglianza.

Infine, poiché il denominatore è costante in ϑ, il rapporto tra la distribuzione a posteriori in due valori distinti del parametro è

$$\underbrace{\frac{\pi(\vartheta_1|y)}{\pi(\vartheta_2|y)}}_{\text{posterior odds}} = \underbrace{\frac{\pi(\vartheta_1)}{\pi(\vartheta_2)}}_{\text{prior odds}} \cdot \underbrace{\frac{L(\vartheta_1)}{L(\vartheta_2)}}_{\text{likelihood ratio}},$$

cioè il rapporto tra gli odds a priori moltiplicato il rapporto delle verosimiglianze.

2.3 Interpretazione soggettiva e oggettiva

Nel paradigma bayesiano ci sono diversi punti di vista riguardo l'incertezza espressa da $\pi(\vartheta)$.

- Interpretazione **soggettiva** (Ramsey, de Finetti, Savage, Kadane, ...), nella quale si interpreta la probabilità come *grado di fiducia* di un soggetto sul determinarsi di un evento. In questa interpretazione non c'è frequentismo, ma stato di conoscenza sul parametro a priori del soggetto. Lo stesso insieme di dati, analizzato da diversi soggetti, può avere conclusioni differenti a seconda delle informazioni a priori.
- Interpretazione **oggettiva** (Jeffreys, Berger, ...), nella quale vengono usate distribuzioni a priori di default "non informative", allo scopo di non inserire opinione a priori nella distribuzione a posteriori. Per quanto filosoficamente sia condivisibile, ci sono dettagli nella costruzione della distribuzione a priori che rendono problematico questo approccio.

Lezione 3

3.1 Esempi di inferenza bayesiana

Esempio (Inferenza bayesiana su una proporzione)
Per i dati sul comportamento aggressivo, dato $n = 707$ di cui $y^{\text{oss}} = 159$, assumiamo un modello statistico con $\mathcal{F}$ famiglia di distribuzioni binomiali, ϑ probabilità di comportamento aggressivo. La verosimiglianza per i dati osservati è

$$L(\vartheta) = p(y^{\text{oss}}|\vartheta) = \binom{707}{159}\vartheta^{159}(1-\vartheta)^{707-159}.$$

Supponendo che uno psicologo ritenga che ci sia un'incidenza del 10% di comportamenti aggressivi, si sceglie di usare una distribuzioni a priori $\vartheta \sim \text{Beta}(\alpha_0, \beta_0)$, con $\alpha_0 = 1, \beta_0 = 9$:

$$p(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \quad \alpha,\beta > 0$$

con

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Inoltre, è noto che

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \quad \mathbb{V}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

In questo caso, la scelta dei due *iperparametri* α_0, β_0 implica che

$$\mathbb{E}[\vartheta] = 0.1, \quad \mathbb{V}[\vartheta] \approx 0.008$$

e

$$P(\vartheta < 0.3) = \int_0^{0.3} \pi(\vartheta)\, d\vartheta \approx 0.96.$$

La distribuzione a posteriori è dunque

$$\begin{aligned}\pi(\vartheta|y^{\text{oss}}) &= \frac{\pi(\vartheta)p(y^{\text{oss}}|\vartheta)}{\int_0^1 \pi(\vartheta)p(y^{\text{oss}}|\vartheta)\, d\vartheta} \\ &= \frac{\vartheta^{159}(1-\vartheta)^{556}}{\int_0^1 \vartheta^{159}(1-\vartheta)^{556}\, d\vartheta} \\ &\sim \text{Beta}(160, 557).\end{aligned}$$

Dunque, dopo aver osservato i dati, si ottiene un aggiornamento delle stime:

$$\mathbb{E}[\vartheta|y^{\text{oss}}] = \frac{160}{160+557} \approx 0.223$$

$$\mathbb{V}[\vartheta|y^{\text{oss}}] \approx 0.00024$$

$$P(0.193 < \vartheta < 0.254) \approx 0.95.$$

Nell'esempio, si è ottenuta una distribuzione a posteriori per ϑ, che è stata riassunta con degli indici di posizione e un intervallo di credibilità.

Una scelta più "oggettiva" è scegliere $\text{Beta}(1,1) = \text{Unif}(0,1)$, che porta a una distribuzione a posteriori semplicemente pari alla verosimiglianza normalizzata.

Con tale scelta, si ottengono diverse stime di valore atteso, varianza e intervallo di credibilità a posteriori, anche se non cambiano di molto.

Osservazioni

- In generale, nel caso beta-binomiale, si ha una distribuzione a posteriori $\text{Beta}(\alpha_0 + y^{\text{oss}}, \beta_0 + n - y^{\text{oss}})$, dunque

$$\mathbb{E}[\vartheta | y^{\text{oss}}] = \frac{\alpha_0 + y^{\text{oss}}}{\alpha_0 + \beta_0 + n}.$$

  Importante osservare che α_0, β_0 influenzano i dati come fossero osservazioni aggiuntive, per cui si ha una diversa influenza degli iperparametri a seconda della quantità di dati a disposizione.

- Attenzione quando si utilizza una distribuzione a priori con valore 0 in alcuni punti dello spazio parametrico, in quanto si ottiene una distribuzione a posteriori

$$p(\vartheta|y) = p(y|\vartheta)\pi(\vartheta) = \begin{cases} p(y|\vartheta)\pi(\vartheta) & \text{se } \pi(\vartheta) \neq 0 \\ 0 & \text{se } \pi(\vartheta) = 0 \end{cases}$$

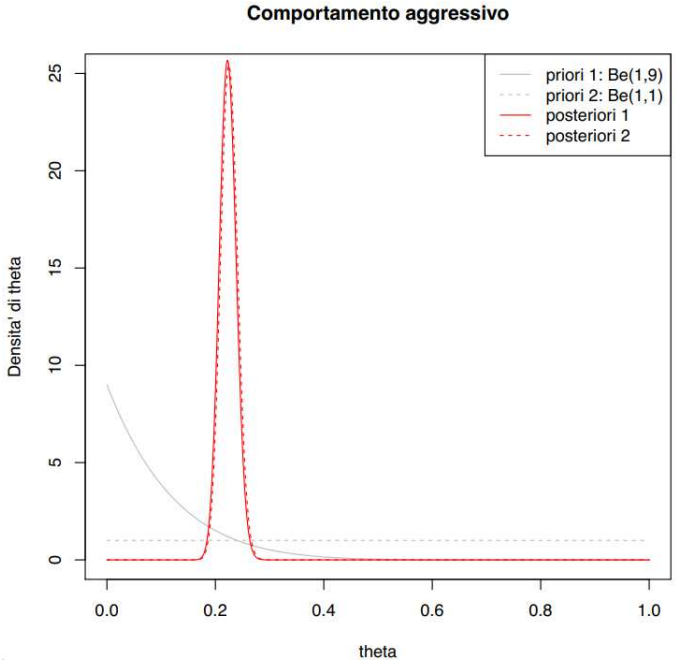


Figura 2: Distribuzioni a priori e a posteriori per la probabilità di comportamento aggressivo.

Esempio (Inferenza bayesiana sulla resistenza alla tensione)
Consideriamo $y = (225, 171, \dots, 162)$ con modello

$$\mathcal{F} = \{p(y|\vartheta) = \vartheta e^{-\vartheta y}, \vartheta \in \mathbb{R}^+, y \in \mathbb{R}^+\},$$

per cui scegliamo una distribuzione a priori $\vartheta \sim \text{Gamma}(\alpha_0, \lambda_0)$:

$$\pi(\vartheta|\alpha_0, \lambda_0) = \frac{\lambda_0^{\alpha_0}}{\Gamma(\alpha_0)} y^{\alpha_0 - 1} e^{-\lambda_0 y}.$$

La scelta degli iperparametri può essere dettata dai dati precedenti (contesto industriale), oppure dal parere di esperti (ingegneri, fisici, ...).

Scegliamo $\alpha_0 = 1$, $\beta_0 = 100$, da cui

$$\mathbb{E}[\vartheta] = \frac{\alpha_0}{\lambda_0} = 1/100, \quad \mathbb{V}[\vartheta] = \frac{\alpha_0}{\lambda_0^2} = 1/10000, \quad P(\vartheta < 0.001) \approx 0.095.$$

La verosimiglianza per i dati osservati è

$$L(\vartheta) = \prod_{i=1}^{n} \vartheta e^{-\vartheta y_i^{\text{oss}}} = \vartheta^n e^{-\vartheta \sum_{i=1}^n y_i^{\text{oss}}}.$$

Per cui, la distribuzione a posteriori è

$$\begin{aligned}
\pi(\vartheta|y^{\text{oss}}) &= \frac{\frac{\lambda_0^{\alpha_0}}{\Gamma(\alpha_0)} \vartheta^{\alpha_0 - 1} e^{-\lambda_0 y} \vartheta^n e^{-\vartheta \sum_{i=1}^n y_i^{\text{oss}}}}{\int_0^\infty \frac{(\lambda_0 + \sum_{i=1}^n y_i^{\text{oss}})^{\alpha_0 + n}}{\Gamma(\alpha_0 + n)} \vartheta^{\alpha_0 + n - 1} e^{-(\lambda_0 + \sum_{i=1}^n y_i^{\text{oss}})\vartheta} \, d\vartheta} \\
&= \frac{(\lambda_0 + \sum_{i=1}^n y_i^{\text{oss}})^{\alpha_0 + n}}{\Gamma(\alpha_0 + n)} \vartheta^{\alpha_0 + n - 1} e^{-(\lambda_0 + \sum_{i=1}^n y_i^{\text{oss}})y} \\
&\sim \text{Gamma}(\alpha_0 + n, \lambda_0 + \sum_{i=1}^n y_i^{\text{oss}}).
\end{aligned}$$

Con i dati osservati e la scelta di iperparametri, si ottiene $\vartheta|y^{\text{oss}} \sim \text{Gamma}(11, 1783)$. Analogamente, è come se avessi aggiunto un'osservazione di durata pari a 100.

In questo caso, si ottengono le stime

$$\mathbb{E}[\vartheta|y^{\text{oss}}] = 0.00617$$

$$\mathbb{V}[\vartheta|y^{\text{oss}}] \approx 3.5 \times 10^{-6}$$

$$P(\vartheta < 0.001|y^{\text{oss}}) \approx 3 \times 10^{-6}.$$

Osservazione

Se faccio tendere gli iperparametri $\alpha_0, \lambda_0 \longrightarrow 0$, si ottiene una distribuzione a priori

$$\pi(\vartheta) = \vartheta^{-1},$$

che ha integrale illimitato in $\mathbb{R}^+$.

Nonostante ciò, se uso questa distribuzione a priori, si ottiene comunque una distribuzione a posteriori propria

$$\vartheta | y^{\text{oss}} \sim \text{Gamma}(n, \sum_{i=1}^{n} y_i^{\text{oss}}),$$

il che non è una proprietà scontata.

Questo approccio si usa spesso, se si vogliono usare distribuzioni "oggettive", ma bisogna dimostrare caso per caso che ciò che si ottiene è una densità propria.

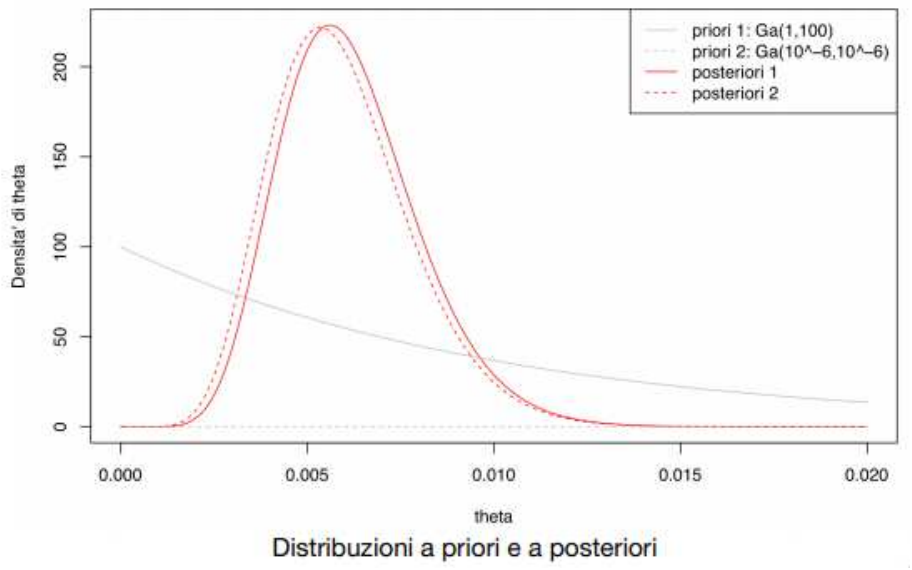


Figura 3: Distribuzioni a priori e a posteriori per la resistenza alla tensione.

Osservazione

In entrambi gli esempi, la distribuzione a posteriori appartiene alla stessa famiglia della distribuzione a priori. Questo accade per la particolare scelta di modello statistico e distribuzione a priori, ma in generale ciò **non vale**.

Lezione 4

Riferimenti Pace e Salvan (2001, §2.2-2.3)

4.1 Inferenza frequentista

Nell'inferenza bayesiana tutte le quantità sono v.c.: si parte da probabilità e si finisce in probabilità.

Nell'inferenza frequentista, la probabilità è un'idealizzazione di frequenze relative, legate ad *ipotetiche replicazioni* del processo generatore.

L'inferenza è una serie di procedure ad hoc, che hanno lo scopo di identificare un valore ϑ_0 all'interno di Θ_0, che è una *costante ignota*.

Assunzione

L'idea di partenza è che il processo inferenziale dovrebbe raramente condurre a gravi errori, qualunque sia il valore di ϑ all'interno di Θ.

Principio del campionamento ripetuto

L'inferenza ottenuta su y^{oss} deve essere valutata sul suo comportamento in ipotetiche replicazioni, sotto le stesse condizioni, dell'esperimento che ha generato y^{oss}.

Da questo punto di vista, tutte le proprietà di uno stimatore classico (distorsione, EQM), p-value, ecc. fanno tutte riferimento al principio del campionamento ripetuto.

Dal punto di vista formale, significa che una procedura inferenziale si basa su una funzione $t(\cdot)$, chiamata *statistica*, le cui proprietà vengono valutate studiando la distribuzione di $t(Y)$, con $Y \sim p(y; \vartheta)$.

Anche all'interno dell'inferenza frequentista, ci sono punti di vista distinti

- **Inferenza Fisheriana** (Fisher, Cox, ...)
  L'inferenza è basata sulla funzione di verosimiglianza (quando possibile), test di significatività e p-value. L'incertezza si valuta in base del P.C.R. applicato a replicazioni "rilevanti" del processo generatore, cioè condizionate rispetto a statistiche ancillari.
- **Inferenza frequentista decisionale** (Neyman, E. Pearson, Wald, Lehmann, ...)
  Al modello statistico $\mathcal{F}$ si aggiunge una *funzione di perdita*, che misura la "distanza" tra il risultato dell'inferenza e la verità.
  Vogliamo una procedura inferenziale che minimizzi la perdita attesa, detto *rischio*.

Esempio (Inferenza frequentista sul comportamento aggressivo)

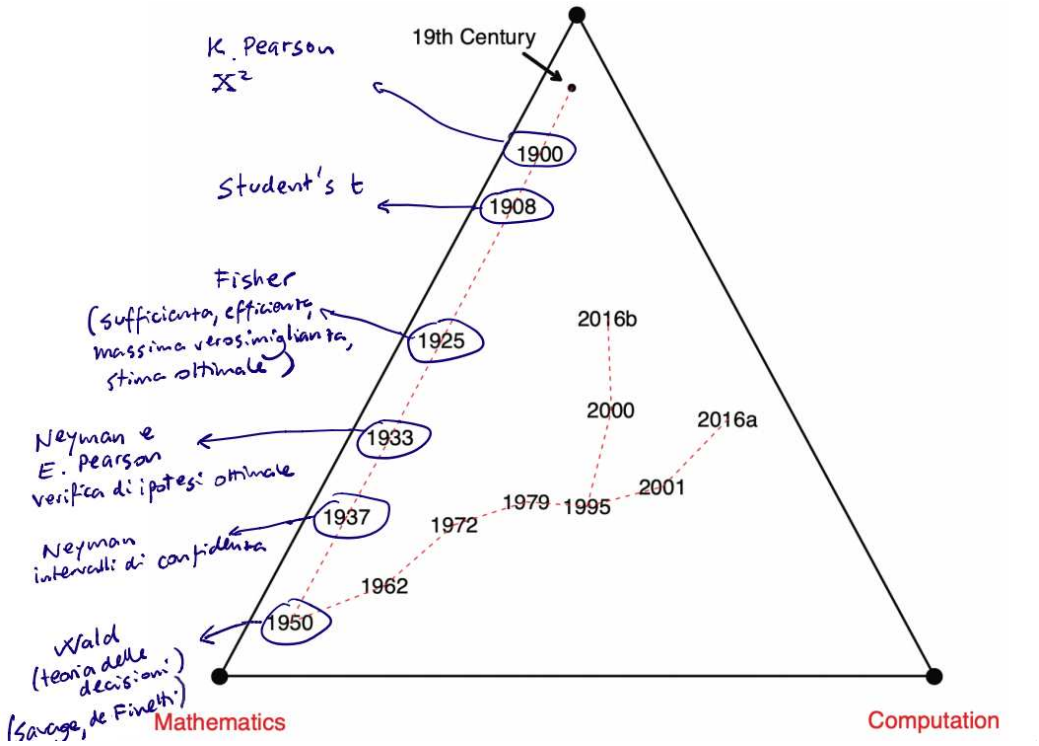


Figura 4: Sviluppo dell'inferenza statistica dalla fine del 1900 ad oggi.

Usiamo il modello statistico delle distribuzioni binomiali

$$\mathcal{F} = \{p(y;\vartheta), \vartheta \in (0,1)\}$$

$$p(y;\vartheta) = \binom{707}{y} \vartheta^y (1-\vartheta)^{707-y}, \quad y \in \mathcal{Y} = \{0,1,\dots,707\}$$

Si sottolinea che ϑ è fissato, da cui la notazione $p(\cdot;\vartheta)$ con il punto e virgola.

Utilizziamo la stima del parametro (qui ancora per analogia) con la proporzione osservata

$$t(y^{\text{oss}}) = \frac{y^{\text{oss}}}{n} = \frac{155}{707} = 0.225.$$

Notazione Una stima del parametro ϑ verrà indicata come $t(Y) = \hat{\vartheta} = \hat{\vartheta}(Y)$.

Per valutare la stima, studiamo le proprietà della v.c. $T = t(Y)$ che, in questo caso, ha distribuzione

$$p_T(t;\vartheta) = p_Y(nt;\vartheta) = \binom{n}{nt} \vartheta^{nt}(1-\vartheta)^{n-nt}, \quad t \in \{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}.$$

In particolare, si ha che

$$\mathbb{E}_\vartheta[T] = \frac{1}{n}\,\mathbb{E}_\vartheta[Y] = \frac{1}{n} n\vartheta = \vartheta,$$

per cui lo stimatore T è *non distorto*.

Notazione Con $\mathbb{E}_\vartheta[\cdot]$ si indica valore atteso rispetto alla distribuzione $p_Y(\cdot;\vartheta)$.

Inoltre, in media

$$\mathbb{V}_\vartheta[T] = \frac{1}{n^2}\,\mathbb{V}_\vartheta[Y] = \frac{1}{n^{\cancel{2}}} \cdot \cancel{n}\vartheta(1-\vartheta) = \frac{\vartheta(1-\vartheta)}{n}.$$

Come misura combinata di qualità dello stimatore, guardiamo

$$\begin{aligned}\mathrm{MSE}_\vartheta(T) = \mathbb{E}_\vartheta[(T-\vartheta)^2] &= \mathbb{V}_\vartheta[T] + (\mathbb{E}_\vartheta[T]-\vartheta)^2 \\ &= \mathbb{V}_\vartheta[T] + \mathrm{Bias}^2_\vartheta(T)\end{aligned}$$

Valutiamo ora l'incertezza della stima, stimando la varianza osservata

$$\hat{\mathbb{V}_\vartheta}[T] = \frac{\hat\vartheta(1-\hat\vartheta)}{n} \approx 0.00025.$$

L'errore standard è

$$\mathrm{se}(T) = \sqrt{\hat{\mathbb{V}_\vartheta}[T]} = 0.0157.$$

Non abbiamo detto nulla su come abbiamo costruito la stima, ma si vedrà che è la stima di massima verosimiglianza.

La stima proposta corrisponde anche a una stima ottima dal punto di vista decisionale sotto la funzione di perdita

$$d(t(Y),\vartheta) = \big(t(Y)-\vartheta\big)^2.$$

In particolare, il rischio da minimizzare corrisponde all'MSE che, sotto il vincolo di non distorsione $\mathbb{E}_\vartheta[(t(Y)-\vartheta)^2] = \vartheta$ fornisce lo stimatore $t(Y) = Y/n$.

Dunque, tra i non distorti, questo stimatore ha la varianza minore di tutti. Se avessimo scelto una diversa funzione di perdita, avremmo ottenuto un altro stimatore.

Commento

La stima frequentista $\hat\vartheta = 0.225$ è molto simile alla stima a posteriori $\mathbb{E}[\vartheta|y^{\text{oss}}] = 0.223$ e, analogamente, la varianza stimata è simile alla varianza a posteriori.

Nonostante l'interpretazione nei due casi sia molto diversa, è rassicurante che le due conclusioni numericamente siano sostanzialmente equivalenti.

Esempio (Resistenza alla tensione)

Andiamo ancora per analogia, con modello statistico dato dalle famiglie di distribuzioni esponenziali

$$p(y;\vartheta) = \vartheta e^{-\vartheta y},$$

che possiamo stimare empiricamente con

$$\hat\vartheta = t(y^{\text{oss}}) = \frac{1}{\overline{y}} \approx 0.00594,$$

ricordando che $\mathbb{E}_\vartheta[Y] = 1/\vartheta$.

Studiamo la distribuzione di $\hat{\vartheta} = t(Y) = 1/\overline{Y}$: dal calcolo delle probabilità sappiamo che $\overline{Y} \sim \text{Gamma}(n, n\vartheta)$. La distribuzione di $1/\overline{Y}$ è gamma inversa, con densità

$$p(t;\vartheta) = \dots$$

Si ottiene che

$$\mathbb{E}_\vartheta[\hat{\vartheta}] = \frac{n}{n-1}\vartheta \neq \vartheta$$

$$\mathbb{V}_\vartheta[\hat{\vartheta}] = \frac{(n\vartheta)^2}{(n-1)^2(n-2)} = \frac{n^2\vartheta^2}{(n-1)^2(n-2)} \sim O(\frac{1}{n})$$

L'MSE è dato da

$$\text{MSE}_\vartheta(\hat{\vartheta}) = \mathbb{V}_\vartheta[\hat{\vartheta}] + \left(\mathbb{E}_\vartheta[\hat{\vartheta} - \vartheta]\right)^2$$

$$\approx 5.9 \times 10^{-6}$$

Dal punto di vista decisionale, definendo una funzione di rischio quadratica con vincolo di non distorsione, la stima ottima è

$$\hat{\vartheta}_{\text{OP}}(Y) = \frac{n-1}{n} \cdot \frac{1}{\overline{Y}}.$$

In questo caso, diverse scelte della procedura di stima hanno portato a diverse valutazioni numeriche (stima) e di variabilità (incertezza), esattamente come nel caso bayesiano con la scelta di distribuzione a priori.

Di nuovo, si osserva che la stima $\hat{\vartheta} = 1/\overline{Y}$ coincide con $\mathbb{E}[\vartheta|y]$ nel caso si assuma una distribuzione a priori non informativa $\pi(\vartheta) = \vartheta^{-1}$.

4.2 Confronto tra inferenza frequentista e bayesiana

- Dal punto di vista bayesiano, l'inferenza è condizionata a y^{oss} e si rifiuta il principio del campionamento ripetuto.
- L'inferenza bayesiana si conclude quando ottengo la distribuzione a posteriori, ma sia il calcolo sia la scelta di priori non sono banali.
- L'inferenza bayesiana è *coerente*, visto che è sviluppata unicamente sulla base degli assiomi della probabilità.
- Verosimiglianze equivalenti portano alla stessa posteriori e lo schema di osservazione è irrilevante nell'inferenza bayesiana (no differenza nell'esempio del basket).
- Nell'inferenza frequentista, i parametri sono fissati e tutte le affermazioni hanno un'interpretazione sulla base del principio del campionamento ripetuto, anche ipoteticamente.
- La valutazione frequentista delle procedure bayesiane (ad esempio $\mathbb{E}[\vartheta|Y]$) sta diventando sempre più comune.
- La replicabilità dei risultati sta diventando sempre più importante negli ultimi anni.

Lezione 5

5.1 Specificazione del modello

La *specificazione del modello* è importante e tipicamente influenza di più le conclusioni, rispetto al paradigma inferenziale.

Di solito è legata al contesto applicativo (psicologico, genetico, ambientale, ...), non ci sono indicazioni esplicite nella teoria dell'inferenza.

Linee guida

- Bisogna tener conto della *natura dei dati*, ovvero se le variabili sono qualitative, discrete, funzioni, immagini, ...
- Le variabili di solito si possono suddividere in sottoinsiemi, ad esempio *risposta* ed *esplicative*.
- Il modello deve tenere conto delle informazioni sullo schema di osservazione:
  - c.c.s
  - randomizzazione (piano degli esperimenti)
  - censura delle osservazioni
  - dati mancanti
  - campionamento sequenziale
  - dipendenza temporale o spaziale
- Bisogna definire quali *aspetti* dei dati il modello deve essere in grado di descrivere: centralità, dipendenza da altre variabili, ...
- Sono importanti *aspetti complementari*, come dispersione, asimmetria, eteroschedasticità, ... Dunque, un modello deve essere in grado descrivere gli aspetti di interesse ma poter dare una descrizione realistica degli aspetti complementari.

A seconda del livello di informazione disponibile, sarà possibile estendere o meno la "dimensione" del modello statistico. In generale, ci sono tre livelli di specificazione:

1. **Parametrico**
$$\mathcal{F} = \{p(y;\vartheta), \vartheta \in \Theta \subseteq \mathbb{R}^p\}$$
2. **Semi-parametrico**
$$\mathcal{F} = \{p(y;\vartheta), \vartheta \in \Theta\},$$
dove $\vartheta = \big(\psi, h(\cdot)\big)$, $\psi \in \Psi \subseteq \mathbb{R}^p$ e la funzione $h(\cdot)$ non può essere indicizzata da un numero finito di parametri reali.
3. **Non parametrico**
Gli elementi di $\mathcal{F}$ non possono essere indicizzati da un numero finito di parametri, e anche l'inferenza non si riferisce a caratteristiche di dimensione finita.

Esempio (Modelli semi-parametrici)

- Considero tutte le distribuzioni simmetriche

$$p(y;\psi) = p_0(y-\psi), \quad \psi \in \Psi \subseteq \mathbb{R}$$

  dove $p_0(\cdot)$ è una funzione ignota e simmetrica.

- Considerato il modello di regressione lineare, uso il modello con le sole ipotesi del secondo ordine:

$$\mathbb{E}[Y_i] = \alpha + \beta x_i;$$

$$\mathbb{V}[Y_i] = \sigma^2.$$

  In generale, $\psi = (\alpha, \beta, \sigma^2)$ e la distribuzione di Y non è specificata.

- Considero Y_i durata di vita dell'i-esima unità e x_i variabile esplicativa, modello il tasso di guasto come

$$r_{Y_i}(y_i; x_i) = r_0(y_i) e^{\psi^\intercal x_i},$$

  con $r_0(\cdot)$ ignoto tasso di guasto di base e $\psi \in \Psi = \mathbb{R}^k$.

Esempio (Modello non parametrico)
Considero un modello

$$\mathcal{F} = \{ \text{ Tutte le distribuzioni con componenti i.i.d.}\},$$

e vado a ricostruire l'intera distribuzione di Y.

La specificazione di un modello è il risultato di un processo iterativo, e la scelta tra diversi modelli si può fare attraverso strumenti sia informali che formali: grafici, analisi dei residui, ...

Osservazioni sui modelli parametrici

È necessario un confronto tra il meccanismo che ha generato i dati e il processo probabilistico sottostante la distribuzione:

$$\text{Distribuzione binomiale} \iff \text{prove indip. a probabilità costanti}$$

$$\text{Distribuzione esponenziale} \iff \text{assenza di memoria}$$

In questo senso, sono utili *caratterizzazioni della densità* $p(\cdot)$, cioè risultati che valgono per una sola classe di distribuzioni.

Altre volte possono esserci argomenti di *natura asintotica*: somma di fattori tendono a una normale, distribuzioni per valori estremi, ...

5.2 Modelli statistici parametrici

Riferimenti Pace e Salvan (2001, §11.1-11.2)

Descriviamo alcuni modi opportuni per costruire dei modelli statistici parametrici.

Esempio (Dati censurati)
Nel caso della resistenza alla tensione, l'esperimento era terminato quando l'elastico non si era ancora rotto.

I dati sono quindi $Y_i = (Y_1, Y_2, \ldots, Y_n)$, con Y_i realizzazione di

$$Y_i = \min(T_i, C_0).$$

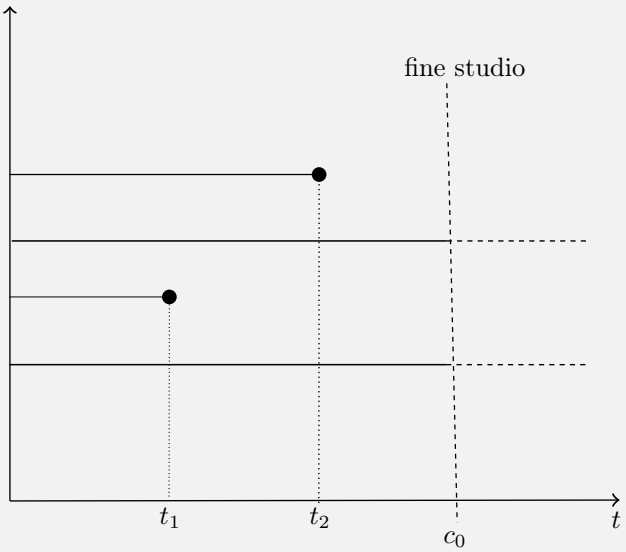


Figura 5: Esempio di censura delle osservazioni.

Lo spazio campionario è dunque $\mathcal{Y} = [0, c_0]^n$, mentre l'insieme delle distribuzioni nel modello statistico è $p(y; \vartheta)$, che dipende dall'assunzione sulla distribuzione di T_i.

In particolare,

$$\begin{aligned} F_{Y_i}(y_i; \vartheta) &= P_\vartheta\{\min(T_i, c_0) \leq y_i\} \\ &= \begin{cases} F_{T_i}(y_i, \vartheta) & \text{se } 0 \leq y_i < c_0 \\ 1 & \text{se } y_i \geq c_0 \end{cases} \end{aligned}$$

dunque si ottiene una variabile casuale mista tra v.c. continua e discreta. In particolare, la "densità" è

$$p_{Y_i}(y_i; \vartheta) = \begin{cases} p_{T_i}(y_i; \vartheta) & \text{se } 0 \leq y_i \leq c_0 \\ 1 - F_{T_i}(y_i; \vartheta) & \text{se } y_i = c_0 \\ 0 & \text{altrove} \end{cases}$$

Quindi, la verosimiglianza è data da

$$p_Y(y_1, \dots, y_n; \vartheta) = \prod_{\text{oss. non cens.}} p_{T_i}(y_i; \vartheta) \prod_{\text{oss. cens.}} \big(1 - F_{T_i}(y_i; \vartheta)\big).$$

Ipotizzando che $T_i \sim \text{Exp}(\vartheta)$, si ha

$$\begin{aligned} p_Y(y_1, \dots, y_n; \vartheta) &= \prod_{\text{oss. non cens.}} \vartheta e^{-y_i \vartheta} \prod_{\text{oss. cens.}} e^{-y_i \vartheta} \\ &= \vartheta^{n_u} e^{-\vartheta \sum_{i=1}^n y_i}. \end{aligned}$$

dove n_u è il numero di osservazioni non censurate. La verosimiglianza è la stessa, ma è come se si disponesse di meno osservazioni.

Estensione

In realtà potrebbe essere che il tempo di censura non sia lo stesso per ciascuno e sia invece c_i, variabile tra le unità. Allora, l'espressione per Y_i diventa

$$Y_i = \min(T_i, C_i).$$

Si ha allora $\mathcal{Y} = [0, c_1] \times [0, c_2] \times \ldots \times [0, c_n]$.

La verosimiglianza si ottiene introducendo la variabile indicatrice $d_i = \mathbb{1}_{[0,c_i]}(T_i)$:

$$p_Y(y_1, \dots, y_n; \vartheta) = \prod_{i=1}^n p_{T_i}(y_i; \vartheta)^{d_i} \big(1 - F_{T_i}(y_i; \vartheta)\big)^{1-d_i}.$$

La variabile indicatrice tiene conto se l'i-esima osservazione è censurata nel punto y_i o meno.

Nel caso esponenziale,

$$\begin{aligned} p_Y(y_1, \dots, y_n; \vartheta) &= \prod_{i=1}^n \left(\vartheta e^{-\vartheta y_i}\right)^{d_i} \left(e^{-\vartheta y_i}\right)^{1-d_i} \\ &= \vartheta^{\sum_{i=1}^n d_i} \exp\{-\vartheta \sum_{i=1}^n y_i\} \end{aligned}$$

Nota

Questo tipo di censura è di I tipo, ovvero in cui c_i sono fissi. Se invece l'esperimento proseguisse fino alla rottura di un numero $r < n$ prefissato di unità, sarebbe una censura di II tipo e si otterrebbe un modello statistico differente.

Lezione 6

Riferimenti Welsh (1996, §1.2-1.3)

6.1 Osservazioni non indipendenti

Assumiamo un vettore di dati $Y = (Y_1, Y_2, \ldots, Y_n)$, osservazioni di una sequenza temporale. Può aver senso pensare che siano una realizzazione di $(Y_1, Y_2, \ldots, Y_n)$, variabile casuale a *componenti dipendenti*:

$$\begin{aligned} p_Y(y_1, y_2, \ldots, y_n) &= p_{Y_n|Y_1,\ldots,Y_{n-1}}(y_n|y_1, \ldots, y_{n-1}) \cdot p_{Y_1,\ldots,Y_{n-1}}(y_1, \ldots, y_{n-1}) \\ &= p_{Y_n|Y_1,\ldots,Y_{n-1}}(y_n|y_1, \ldots, y_{n-1}) \cdot \ldots \\ &= p_{Y_n|Y_1,\ldots,Y_{n-1}}(y_n|y_1, \ldots, y_{n-1}) \cdot \ldots \cdot p_{Y_2|Y_1}(y_2|y_1) p_{Y_1}(y_1). \end{aligned}$$

Assumendo una struttura di *dipendenza Markoviana*, ovvero che

$$p_{Y_i|Y_{i-1},\ldots,Y_1}(y_i|y_{i-1}, \ldots, y_1) = p_{Y_i|Y_{i-1}}(y_i|y_{i-1}),$$

si ha la semplificazione

$$p_Y(y_1, \ldots, y_n) = p_{Y_1}(y_1) \prod_{j=2}^{n} p_{Y_i|Y_{i-1}}(y_i|y_{i-1}).$$

Esempio (Dati binari)
Assumiamo che $Y_i \in \{0, 1\}$ e che ci sia omogeneità nella probabilità di transizione

$$P(Y_i = 1|Y_{i-1} = y_{i-1}) = \begin{cases} \pi_{01} & \text{se } y_{i-1} = 0 \\ \pi_{11} & \text{se } y_{i-1} = 1 \end{cases}$$

da cui si ottengono le probabilità complementari

$$P(Y_i = 0|Y_{i-1} = y_{i-1}) = \begin{cases} \pi_{00} = 1 - \pi_{01} & \text{se } y_{i-1} = 0 \\ \pi_{10} = 1 - \pi_{11} & \text{se } y_{i-1} = 1 \end{cases}$$

La densità congiunta diventa allora

$$p_Y(y; \pi_{01}, \pi_{11}) = p_{Y_1}(y_1) \cdot \pi_{00}^{n_{00}} \cdot \pi_{01}^{n_{01}} \cdot \pi_{10}^{n_{10}} \cdot \pi_{11}^{n_{11}},$$

con n_{jk} numero di transizioni dallo stato j allo stato k. Lo spazio campionario è $\{0, 1\}^n$ e lo spazio parametrico è $\Theta = (0, 1)^2$.

Ci sono due strade per p_{Y_1}: si può fare un'ulteriore assunzione sulla distribuzione, oppure può condizionare l'inferenza alla prima osservazione e rimuoverla dalla densità congiunta.

Si potrebbe assumere questo modello nell'esempio dei tiri liberi del basket per valutare l'effetto "hot hand": è chiaro che è necessario avere l'intera sequenza di tiri liberi e non solamente il numero di tiri effettuati.

Con questo modello, a parità di tiri effettuati si possono avere densità congiunte differenti, ad esempio,

$$y_A = (0,0,0,0,1,1,1,1,1,1) \implies (n_{00}, n_{01}, n_{10}, n_{11}) = (3,1,0,5)$$

$$y_B = (0,1,0,1,0,1,1,0,1,1) \implies (n_{00}, n_{01}, n_{10}, n_{11}) = (0,4,3,2)$$

Il caso di indipendenza è un caso particolare con

$$\pi_{01} = \pi_{11} = \pi; \quad \pi_{00} = \pi_{10} = 1 - \pi.$$

Esempio (Modello normale autoregressivo)

Si consideri i dati del grafico, relativi alla misurazione di un ormone ogni 8h. È di interesse studiare la variazione giornaliera dell'ormone, osservando che c'è dipendenza tra le osservazioni.

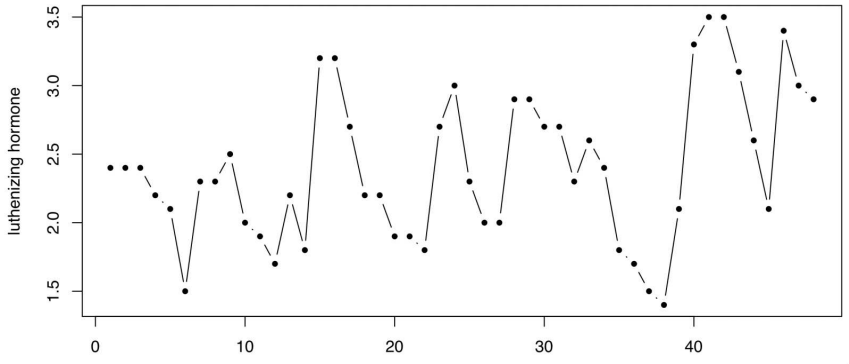


Figura 6: Osservazioni a distanza di 8h di un ormone.

Il modello statistico *autoregressivo* assume una relazione lineare tra i valori precedenti e successivi di Y:

$$Y_i | Y_{i-1} = y_{i-1} \sim \mathcal{N}(\mu + \rho y_{i-1}, \sigma^2), \quad i = 2, \ldots, n.$$

Il parametro di interesse è $\vartheta \sim (\mu, \rho, \sigma^2)$ e si ha

$$p_Y(y; \vartheta) = p_{Y_1}(y_1) \prod_{i=2}^{n} p_{Y_i|Y_{i-1}}(y_i | y_{i-1}; \vartheta)$$

e assumendo la prima osservazione nota, si ha

$$p_Y(y; \vartheta) = p_{Y_1}(y_1) \prod_{i=2}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - \mu - \rho y_{i-1})^2 \right\}.$$

Se si assume che marginalmente le Y_i abbiano la stessa distribuzione (processo stazionario),

si ha che

$$Y_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{1-\rho}\right), \quad |\rho^2| < 1,$$

condizione necessaria e sufficiente per la stazionarietà. La densità congiunta diventa allora

$$p_y(y;\vartheta) = \frac{1}{\sqrt{2\pi\sigma^2/(1-\rho^2)}} \exp\left\{-\frac{1}{2\sigma^2/(1-\rho^2)}(y_1-\mu)^2\right\} \times \prod_{i=2}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu - \rho y_{i-1})^2\right\}$$

Osservazione

Il modello è equivalente ad assumere che $Y = (Y_1, Y_2, \ldots, Y_n) \sim \mathcal{N}_n(\mu \mathbf{1}_n, \Omega)$, dove Ω è la matrice di varianza e covarianza con struttura autoregressiva, ovvero

$$\Omega = \frac{\sigma^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\ \rho & 1 & \rho & \ldots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{n-3} \\ \vdots & \vdots & \ddots & \vdots & \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1 \end{pmatrix}$$

e la densità congiunta è equivalente a

$$p_y(y;\vartheta) = \frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \exp\left\{-\frac{1}{2}(y-\mu\mathbf{1}_n)^\intercal \Omega^{-1}(y-\mu\mathbf{1}_n)\right\}.$$

6.2 Famiglie di posizione e scala

Def. (Famiglia di posizione)
Una *famiglia di posizione* è un insieme di distribuzioni con parametro $\mu \in \mathbb{R}$ per una singola osservazione scalare con densità

$$p_Y(y;\mu) = p_0(y-\mu),$$

dove $p_0(\cdot)$ è una densità fissata e μ è un *parametro di posizione*.

Osservazioni

- $Y = \mu + Y_0$, con $Y_0 \sim p_0(\cdot)$ e $p_0(\cdot)$ corrisponde a $\mu = 0$.
- Se Y_0 ha funzione generatrice dei momenti $M_0(t) = \mathbb{E}_\mu[e^{tY}]$, allora si ha che

  $$M_Y(t;\mu) = \mathbb{E}_\mu[e^{tY}] = \mathbb{E}_\mu[e^{t(\mu+Y_0)}] = e^{\mu t} M_0(t),$$

  in particolare $\mathbb{E}_\mu[Y] = \mu + \mathbb{E}[Y_0]$, $\mathbb{V}_\mu[Y] = \mathbb{V}[Y_0]$.

- Se $Y = (Y_1, Y_2, \ldots, Y_n)$ sono i.i.d. da una famiglia di posizione, allora

$$p_Y(y; \mu) = \prod_{i=1}^{n} p_0(y_i - \mu).$$

  Inoltre, la distribuzione di $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ apparterrà a sua volta ad una famiglia di posizione.

- Più in generale, il modello indotto da una qualunque funzione $t = t(y_1, \ldots, y_n)$ tale che

$$t(y_1 + a, \ldots, y_n + a) = a + t(y_1, \ldots, y_n)$$

  è ancora una famiglia di posizione.

Esempio (Famiglie di posizione)

1. $Y \sim \text{Unif}(\mu, \mu + 1),\ \mu \in \mathbb{R}$.

$$p_Y(y_i; \mu) = \mathbb{1}_{[\mu, \mu+1]}(y),$$

   sono traslazioni di $Y_0 \sim \text{Unif}(0, 1)$.

2. La famiglia di posizione generata da $Y_0 \sim \text{Exp}(1),\ \mu \in \mathbb{R}$

$$p_Y(y; \mu) = \mathbb{1}_{[\mu, +\infty)} e^{-(y-\mu)}.$$

3. $Y \sim \mathcal{N}(\mu, 1),\ \mu \in \mathbb{R}$

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2},$$

   generata da $Y \sim \mathcal{N}(0, 1)$.

4. $Y \sim \text{La}(\mu, 1),\ \mu \in \mathbb{R}$ distribuzione di Laplace

$$p_Y(y; \mu) = \frac{1}{2} \exp\{-|y - \mu|\},$$

   con distribuzione di base $p_0(y) = \frac{1}{2}\exp\{-|y|\}$.

5. $Y \sim \text{Cauchy}(\mu, 1),\ \mu \in \mathbb{R}$ distribuzione di Cauchy con

$$p(y; \mu) = \frac{1}{\pi\left\{1 + (y - \mu)^2\right\}},$$

   dove in questo caso μ non è il valore atteso della distribuzione, in quanto la Cauchy non ammette valore atteso.

6. $Y \sim \text{Lo}(\mu, 1),\ \mu \in \mathbb{R}$ distribuzione logistica con

$$p_Y(y; \mu) = \frac{e^{-(x-\mu)}}{\left\{1 + e^{-(y-\mu)}\right\}^2}$$

   con distribuzione di base $p_0(y) = e^{-y}/(1 + e^{-y})^2$.

Def. (Famiglia di scala)

Una *famiglia di scala* è un insieme di distribuzioni con parametro $\sigma \in \mathbb{R}^+$ con densità

$$p_Y(y;\sigma) = \frac{1}{\sigma} p_0\left(\frac{y}{\sigma}\right),$$

con $p_0(\cdot)$ densità fissata. Il parametro σ è detto *parametro di scala*, anche se a volte si preferisce parametrizzare con $\lambda = 1/\sigma$.

Osservazione

- Si ha che $Y = \sigma Y_0$, con $Y_0 \sim p_0(\cdot)$ e $p_0(\cdot)$ corrisponde a $\sigma = 1$.
- Se Y_0 ha funzione generatrice dei momenti $M_0(t)$, allora

  $$M_Y(t;\sigma) = \mathbb{E}_\sigma[e^{tY}] = \mathbb{E}[e^{t\sigma Y_0}] = M_0(\sigma t).$$

  In particolare, $\mathbb{E}_\sigma[Y] = \sigma\,\mathbb{E}[Y_0]$, $\mathbb{V}_\sigma[Y] = \sigma^2\,\mathbb{V}[Y_0]$.
- Se $Y = (Y_1, Y_2, \ldots, Y_n)$ sono i.i.d. da una famiglia di scala scalare, allora

  $$p_Y(y;\sigma) = \frac{1}{\sigma^n} \prod_{i=1}^{n} p_0(y_i/\sigma).$$

- La distribuzione di $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ apparterrà a sua volta a una famiglia di scala. Questo vale per qualunque funzione $t(Y_1, \ldots, Y_n)$ tale che

  $$t(by_1, \ldots, by_n) = bt(y_1, \ldots, y_n).$$

Esempio (Famiglie di scala)

- $Y \sim \text{Unif}(0,\sigma)$ con $Y_0 \sim \text{Unif}(0,1)$ e

  $$p_Y(y;\sigma) = \frac{1}{\sigma}\mathbb{1}_{[0,\sigma]}(y).$$

- $Y \sim \text{Exp}(\lambda)$ con

  $$p_Y(y;\lambda) = \lambda e^{-\lambda y}\mathbb{1}_{[0,+\infty)}(y).$$

- $Y \sim \mathcal{N}(0,\sigma^2)$, $\sigma \in \mathbb{R}^+$

  $$p_Y(y;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y)^2}.$$

- $Y \sim \text{La}(0,\sigma)$

  $$p_Y(y;\sigma) = \frac{1}{2\sigma}\exp\{-\frac{1}{\sigma}|y|\}.$$

- $Y \sim \text{Cauchy}(0,\sigma)$
- $Y \sim \text{Lo}(0,\sigma)$

- $Y \sim \text{Gamma}(\alpha_0, \lambda)$ con parametro di forma α_0 fissato, da cui

$$p_Y(y;\lambda) = \frac{\lambda^{\alpha_0}}{\Gamma(\alpha_0)} y^{\alpha_0 - 1} e^{-\lambda y}$$

  e distribuzione di base $Y_0 \sim \text{Gamma}(\alpha_0, 1)$.

Def. (Famiglia di posizione e scala)
Una *famiglia di posizione e scala* è una famiglia parametrica con parametro $\vartheta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$ per un'osservazione scalare y con densità

$$p_Y(y;\mu,\sigma) = \frac{1}{\sigma} p_0\left(\frac{y-\mu}{\sigma}\right).$$

Osservazioni

- $Y = \mu + \sigma Y_0$, con $Y_0 \sim p_0(\cdot)$, e $p_0(\cdot)$ corrisponde alla coppia di parametri $\mu = 0,\ \sigma = 1$.
- La funzione generatrice dei momenti è

$$M_Y(t;\mu,\sigma) = e^{\mu t} M_0(\sigma t),$$

  in particolare $\mathbb{E}_\vartheta[Y] = \mu + \sigma\,\mathbb{E}[Y_0]$, $\mathbb{V}_\vartheta[Y] = \sigma^2\,\mathbb{V}[Y_0]$.
- Un campione casuale semplice ha densità congiunta

$$p_Y(y;\mu,\sigma) = \frac{1}{\sigma^n} \prod_{i=1}^{n} p_0\left(\frac{y_i-\mu}{\sigma}\right).$$

- La distribuzione di $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ appartiene ad una famiglia di posizione e scala. Più in generale, il modello indotto da una funzione scalare $t = t(Y_1, Y_2, \ldots, Y_n)$ tale che

$$t(a + bY_1, \ldots, a + bY_n) = a + bt(Y_1, \ldots, Y_n)$$

  è ancora una famiglia di posizione e scala.

Lezione 7

Riferimenti Welsh (1996, §1.2-1.3)

7.1 Famiglie esponenziali monoparametriche

Def. (Famiglia esponenziale monoparametrica)
Una *famiglia esponenziale monoparametrica* è una famiglia di distribuzioni $\mathcal{F}$ per una osservazione Y univariata o multivariata con parametro $\vartheta \in \Theta \subseteq \mathbb{R}$ e densità

$$p_Y(y;\vartheta) = c(\vartheta)\exp\{\psi(\vartheta)t(y)\}h(y),$$

dove $h(\cdot) \geq 0$, $\psi(\vartheta) : \Theta \to \mathbb{R}$ e $t(\cdot) : \mathcal{Y} \to \mathbb{R}$.

Osservazioni

- Scelta una famiglia $\mathcal{F}$, le densità al suo interno sono o tutte continue o tutte discrete.
- La funzione $c(\vartheta)$ è la costante di normalizzazione della distribuzione.
- Il supporto di Y è la chiusura di $\{y \in \mathbb{R}^d : h(y) > 0\}$, quindi è uguale per qualunque valore di ϑ.
- Affinché $\mathcal{F}$ sia non banale e ϑ identificabile, Θ deve contenere almeno due elementi e $\psi(\cdot)$ deve essere iniettiva. Si dice che $\psi(\vartheta)$ è il *parametro canonico* e $t(y)$ è la *statistica canonica* di $\mathcal{F}$.
- Spesso Θ è un intervallo e $\psi(\cdot)$ è un diffeomorfismo di classe C^1.

Esempio (Famiglie esponenziali monoparametriche)

- $Y \sim \text{Bin}(m, \pi)$ con m fissato.
- $Y \sim \text{Pois}(\lambda)$.
- $Y \sim \text{Gamma}(\alpha, \lambda)$ con α fissato.
- $Y \sim \text{Gamma}(\alpha, \lambda)$ con λ fissato.
- $Y \sim \mathcal{N}(\mu, \sigma^2)$ con σ^2 fissato.
- $Y \sim \mathcal{N}(\mu, \sigma^2)$ con μ fissato.

Esempio (Distribuzione binomiale)
Sia $m \in \mathbb{N}^+$ fissato, consideriamo $Y \sim \text{Bin}(m, \pi)$, con densità

$$p_Y(y;\pi)\binom{m}{y}\pi^y(1-\pi)^{m-y}.$$

Scriviamo

$$p_Y(y;\pi) = (1-\pi)^m\binom{m}{y}\exp\left\{\log\left(\frac{\pi}{1-\pi}\right)y\right\},$$

da cui $\psi(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$, $t(y) = y$.

Esempio (Distribuzione normale con varianza nota)
Sia $\sigma^2 > 0$ fissata, la famiglia $\mathcal{N}(\mu, \sigma^2)$ ha densità

$$p_Y(y;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

che si può scrivere nella forma

$$p_Y(y;\mu) = e^{-\frac{\mu^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} \exp\left\{\mu\frac{y}{\sigma^2}\right\},$$

che è nella forma cercata con $\psi(\mu) = \mu$, $t(y) = y/\sigma^2$.

Chiusura rispetto al campionamento casuale semplice

Sia $y = (y_1, y_2, \ldots, y_n)$ un campione casuale semplice da una famiglia esponenziale monoparametrica. Allora, la densità congiunta del modello è

$$p_Y(y;\vartheta) = c(\vartheta)^n \left(\prod_{i=1}^n h(y_i)\right) \exp\left\{\psi(\vartheta)\sum_{i=1}^n t(y_i)\right\},$$

che è ancora una famiglia esponenziale monoparametrica.

7.2 Famiglie esponenziali multiparametriche

Def. (Famiglia esponenziale multiparametrica)
Una famiglia esponenziale multiparametrica è una famiglia parametrica di distribuzioni $\mathcal{F}$ per un'osservazione, univariata o multivariata, y e parametri $\vartheta \in \Theta \subseteq R^p$ con densità

$$p_Y(y;\vartheta) = c(\vartheta)\exp\left\{\psi(\vartheta)^\mathsf{T} t(y)\right\} h(y),$$

dove $h(\cdot) \geq 0$ e $\psi(\vartheta) = (\psi_1(\vartheta), \psi_2(\vartheta), \ldots, \psi_k(\vartheta))$ è una funzione con codominio $\psi(\Theta) = \mathbb{R}^k$.

Osservazioni

- Il supporto di Y sotto ϑ è la chiusura dell'insieme $\{y \in \mathbb{R}^d : h(y) > 0\}$.
- Anche qui si richiede che $\psi(\vartheta)$ sia iniettiva affinché il parametro $(\vartheta_1, \vartheta_2, \ldots, \vartheta_p)$ sia identificabile.
- La rappresentazione è *minimale* se coinvolge il minimo possibile di funzioni $\psi_j(\vartheta)$ e associate statistiche $t_j(y)$, ovvero se:
  1. Θ contiene almeno $k+1$ elementi, dove k è la dimensione della funzione $\psi(\cdot)$.
  2. Le $k+1$ funzioni reali $1, \psi_1(\vartheta), \ldots, \psi_k(\vartheta)$ sono linearmente indipendenti. Infatti, se avessimo
     $$\psi_k(\vartheta) = c_0 + c_1\psi_1(\vartheta) + \ldots + c_{k-1}\psi_{k-1}(\vartheta),$$
     allora si potrebbe riscrivere la densità usando solo le prime $k-1$ componenti di ψ e scegliendo $t'_j(y) = t_j(y) + c_j t_k(y)$.
  3. Analogamente, le $k+1$ funzioni $1, t_1(y), \ldots, t_k(y)$ devono essere linearmente indipendenti.

In una famiglia esponenziale a rappresentazione minimale, si dice *ordine della famiglia* il valore di k e la funzione $t(y) = (t_1(y), t_2(y), \ldots, t_k(y))$ si dice *statistica canonica* di $\mathcal{F}$.

Spesso, $k = p$ e $\psi(\vartheta)$ è biunivoca ed è una *riparametrizzazione* del modello, in tal caso $\psi = \psi(\vartheta)$ si chiama *parametro canonico*.

Nella parametrizzazione canonica ψ, la statistica canonica $t(Y)$ ha densità

$$p_T(t;\psi) = c(\vartheta(\psi))\tilde{h}(t)\exp\left\{\psi^\mathsf{T} t\right\},$$

dove $\tilde{h}(t)$ è una funzione opportuna. Ad esempio, nel caso discreto

$$\tilde{h}(t) = \sum_{\substack{y \in S_Y: \\ t(y)=t}}^{n} h(y).$$

Se Ψ è aperto e contiene tutti gli elementi di $\mathbb{R}^p$ per cui $\tilde{h}(t)$ è integrabile, si dice che la famiglia esponenziale è *regolare*.

Chiusura rispetto al campionamento casuale semplice

Analogamente al caso monoparametrico, in un campione casuale $y = (y_1, y_2, \ldots, y_n)$ da una famiglia esponenziale multiparametrica, la densità congiunta è

$$p_Y(y;\vartheta) = c(\vartheta)^n \left(\prod_{i=1}^n h(y_i)\right) \exp\left\{\psi(\vartheta)^\intercal \sum_{i=1}^n t(y_i)\right\},$$

e dunque è ancora una famiglia esponenziale multiparametrica, con statistica canonica

$$t(y) = \left(\sum_{i=1}^n t_1(y), \sum_{i=1}^n t_2(y), \ldots, \sum_{i=1}^n t_k(y)\right).$$

Esempio (Famiglie esponenziali multiparametriche)
Per y univariata:

- $Y \sim \mathcal{N}(\mu, \sigma^2)$ è famiglia esponenziale di ordine 2 con parametro $\vartheta = (\mu, \sigma^2)$.
- $Y \sim \text{Gamma}(\alpha, \lambda)$ è famiglia esponenziale di ordine 2 con parametro $\vartheta = (\alpha, \lambda)$.

Per y bivariata:

- Distribuzioni trinomiali con parametro $\vartheta = (\pi_x, \pi_y)$ è famiglia esponenziale di ordine 2 se $\pi_x + \pi_y < 1$.
- Distribuzioni normali bivariate con parametro $\vartheta = (\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ è una famiglia esponenziale di ordine 5.

Per y multivariata:

- Distribuzioni multinomiali $\text{Mn}_d(n, \pi)$ con n fissato e parametro $\vartheta = \pi(\pi_1, \ldots, \pi_d)$ con $\sum_{i=1}^d \pi_i = 1$ è famiglia esponenziale di ordine $d - 1$.
- Distribuzione normale multivariata $\mathcal{N}_d(\mu, \Sigma)$ con Σ simmetrica e definita positiva, è famiglia esponenziale di ordine $d + d + d(d-1)/2 = d(d+3)/2$

In tutti gli esempi, le famiglie esponenziali sono regolari.

Esempio (Normale univariata)
Considero la densità

$$\begin{aligned} p_Y(y;\vartheta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{\frac{\mu}{\sigma^2}y - \frac{1}{2\sigma^2}y^2\right\}, \end{aligned}$$

dunque l'ordine della famiglia è 2 con statistica canonica $t(y) = (y, y^2)$ e parametro canonico $\psi(\vartheta) = (\mu/\sigma^2, -1/2\sigma^2)$.

Esempio (Normale bivariata)
TODO

Esempio (Normale multivariata)
TODO

7.3 Momenti di famiglie esponenziali

Teo. (Momenti di una famiglia esponenziale)
Se Y ha una densità da famiglia esponenziale multiparametrica, allora per $h = 1, \dots, p$ vale che

$$\mathbb{E}_\vartheta \left[\frac{\partial \psi(\vartheta)}{\partial \vartheta_h}^\mathsf{T} t(Y) \right] = -\frac{\partial}{\partial \vartheta_h} \log c(\vartheta)$$

$$\mathbb{V}_\vartheta \left[\frac{\partial \psi(\vartheta)}{\partial \vartheta_h}^\mathsf{T} t(Y) \right] = -\frac{\partial^2}{\partial \vartheta_h^2} \log c(\vartheta) - \mathbb{E}_\vartheta \left[\frac{\partial^2 \psi(\vartheta)}{\partial \vartheta_h^2}^\mathsf{T} t(Y) \right]$$

Dim.
No.

□

Osservazione

Queste formule permettono il calcolo di valore atteso e varianza attraverso derivate, piuttosto che integrali o sommatorie, che in generale sono più complessi.

Esempio (Distribuzione binomiale)
Consideriamo

$$p_Y(y; \pi) = (1-\pi)^m \binom{m}{y} \exp\left\{ \log\left(\frac{\pi}{1-\pi} \right) y \right\},$$

dunque

$$\frac{\partial \psi(\pi)}{\partial \pi} = \frac{1}{\pi(1-\pi)},$$

per cui

$$\begin{aligned} \mathbb{E}_\pi \left[\frac{1}{\pi(1-\pi)} Y \right] &= -\frac{\partial}{\partial \pi} m \log(1-\pi) \\ &= \frac{m}{1-\pi}, \end{aligned}$$

da cui si ricava che $\mathbb{E}_\pi[Y] = m\pi$. Analogamente, per la varianza

$$\begin{aligned} \mathbb{V}_\pi \left[\frac{1}{\pi(1-\pi)} Y \right] &= \frac{m}{(1-\pi)^2} - \frac{1-2\pi}{\pi^2(1-\pi)^2} \\ &= \frac{m}{(1-\pi)^2} + \frac{1-2\pi}{\pi^2(1-\pi)^2} m\pi, \end{aligned}$$

da cui $\mathbb{V}_\pi[Y] = m\pi(1-\pi)$.

Esercizio: provare a usare la formula per calcolare i momenti di $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Osservazione

Se ϑ è il parametro canonico, ovvero $\vartheta = \psi$, si ha che

$$\frac{\partial \psi_j}{\partial \vartheta_h} = \delta_{jh} = \begin{cases} 1 & \text{se } j = h \\ 0 & \text{se } j \neq h \end{cases},$$

per cui si ottengono le formule semplificate

$$\mathbb{E}_\vartheta[t_j(Y)] = -\frac{\partial}{\partial \vartheta_j} \log c(\vartheta)$$

$$\mathbb{V}_\vartheta[t_j(Y)] = -\frac{\partial^2}{\partial \vartheta_j^2} \log c(\vartheta)$$

Esempio (Distribuzione Gamma)
Se $Y \sim \text{Gamma}(\alpha, \lambda)$, si ha

$$p_Y(y; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{1}{y} \exp\{\alpha \log y - \lambda y\},$$

con $c(\vartheta) = \lambda^\alpha / \Gamma(\alpha)$, $g(y) = 1/y$ e $\psi = \vartheta$. Dunque, dalla formula

$$\mathbb{E}_\vartheta[\log Y] = -\log \lambda + \frac{\partial}{\partial \alpha} \log \Gamma(\alpha)$$

$$\mathbb{E}_\vartheta[-Y] = \frac{\alpha}{\lambda}$$

$$\mathbb{V}_\vartheta[\log Y] = \frac{\partial^2}{\partial \alpha^2} \log \Gamma(\alpha)$$

$$\mathbb{V}_\vartheta[-Y] = \frac{\alpha}{\lambda^2}.$$

Lezione 8

Riferimenti Welsh (1996, §2.2)

8.1 Elicitazione della priori

Dal punto di vista bayesiano è necessario specificare anche la distribuzione a priori per il parametro di interesse. Questo è un passo delicato, in quanto non c'è un metodo empirico per valutare la bontà della distribuzione a priori, a differenza del modello statistico.

Criteri per la costruzione della priori

- *Studi precedenti*: se si dispone di dati simili a quelli osservati, si può inferire da essi una a priori per i dati attuali.
- *Distribuzioni convenienti*: scegliere distribuzioni a priori che semplificano i conti (priori coniugate).
- *Priori non informative*: Jeffreys, ...
- *Priori soggettive*: distribuzioni che dipendono dall'opinione del soggetto.

Esempio (Distribuzioni a priori da studi precedenti)
Una donna incinta sa dall'ecografia che avrà due gemelli maschi.

Si chiede se due gemelli maschi saranno monozigoti ($\vartheta = 1$) o dizigoti ($\vartheta = 0$). Sia y il risultato dell'ecografia, con risultati teorici possibili MM, FF, FM. Condizionatamente a ϑ, Y ha distribuzione

Y	MM	FF	FM
$\vartheta = 0$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
$\vartheta = 1$	$\frac{1}{2}$	$\frac{1}{2}$	0

Da un database nazionale, si sa che $\frac{1}{3}$ dei gemelli sono monozigoti e $\frac{2}{3}$ sono dizigoti, quindi la probabilità a priori è

$$\pi(1) = \frac{1}{3}, \quad \pi(0) = \frac{2}{3};$$

Con il teorema di Bayes, la distribuzione a posteriori è

$$\pi(1|y^{\text{oss}} = \text{MM}) = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{2}{3}} = \frac{1}{2}.$$

Osservazione

La distribuzione a priori è stata costruita a partire dalla frequenza osservata nella popolazione dei gemelli omozigoti. In questo caso, anche un frequentista sarebbe d'accordo con il procedimento.

Def. (Priori coniugata)
Una distribuzione π a priori per il parametro ϑ si dice *coniugata* per il modello statistico parametrico $\mathcal{F}$ se

$$p(y|\vartheta) \in \mathcal{F} \implies \pi(\vartheta|y) = \frac{p(y|\vartheta)\pi(\vartheta)}{p(y)} \in \mathcal{F}$$

Con una distribuzione a priori coniugata, l'aggiornamento della densità si mantiene all'interno del modello statistico $\mathcal{F}$. In particolare, la verosimiglianza ha il solo effetto di *aggiornare* gli iperparametri della distribuzione iniziale.

Distribuzioni coniugate e simulazioni

Storicamente il ruolo delle distribuzioni coniugate è stato quello di evitare il calcolo dell'integrale al denominatore

La possibilità di effettuare *Markov Chain Monte Carlo* (MCMC) ha permesso di effettuare inferenza bayesiana senza necessariamente specificare una priori coniugata. Con questi metodi, è unicamente necessario usare il nucleo della posteriori

$$\pi(\vartheta|y) \propto p(y|\vartheta)\pi(\vartheta).$$

Attualmente, le priori coniugate sono usate per rendere matematicamente conveniente la costruzione di modelli più complessi, ad esempio nel *Gibbs sampling*, dove si usano priori coniugate per le singole componenti di un modello multivariato.

8.2 Famiglie esponenziali e priori coniugate

Sia $y = (y_1, y_2, \ldots, y_n)$ un c.c.s. da una famiglia esponenziale, con densità marginale

$$p_{Y_i}(y_i; \vartheta) = \exp\left\{\psi(\vartheta)^\mathsf{T} t(y_i) - C(\vartheta)\right\} h(y_i),$$

dove $C(\vartheta) = -\log c(\vartheta)$. La verosimiglianza ha quindi forma

$$L(\vartheta; y) = \exp\left\{\psi(\vartheta)^\mathsf{T} \sum_{i=1}^{n} t(y) - nC(\vartheta)\right\} \prod_{i=1}^{n} h(y_i).$$

Se si considera una distribuzione a priori con iperparametri (ν, ξ) della forma

$$\pi(\vartheta|\nu, \xi) = \exp\left\{\psi(\vartheta)^\mathsf{T}\nu - \xi C(\vartheta) + D(\nu, \xi)\right\},$$

con $D(\cdot)$ costante di normalizzazione, allora si ottiene la distribuzione a posteriori

$$\pi(\vartheta|y) = \exp\left\{\psi(\vartheta)\left(\sum_{i=1}^{n} t(y_i) + \nu\right) - (n+\xi)\, C(\vartheta) + D(\nu, \xi, t^{(n)}(y))\right\}.$$

Purché la funzione sia integrabile, questa distribuzione appartiene alla stessa classe della priori, con parametri aggiornati

$$(\nu, \xi) \longmapsto \left(\nu + \sum_{i=1}^{n} t(y_i),\ n + \xi \right).$$

Esempio (Modello Gamma-Poisson)
Sia $y^{\text{oss}} = (y_1^{\text{oss}}, y_2^{\text{oss}}, \dots, y_n^{\text{oss}})$ un c.c.s. da una Poisson di media ϑ. Allora,

$$p(y^{\text{oss}}|\vartheta) = \prod_{i=1}^{n} e^{-\vartheta} \frac{\vartheta^{y_i^{\text{oss}}}}{y_i^{\text{oss}}!} = \prod_{i=1}^{n} \frac{1}{y_i^{\text{oss}}!} \exp\left\{ \log\vartheta \cdot \sum_{i=1}^{n} y_i^{\text{oss}} - n\vartheta \right\},$$

dunque si può usare la distribuzione a priori

$$\begin{aligned} \pi(\vartheta|\nu,\xi) &\propto \exp\{\log\vartheta \cdot \nu - \xi \cdot \vartheta\} \\ &\propto \vartheta^{\nu} e^{-\xi\vartheta}, \end{aligned}$$

dunque $\vartheta \sim \text{Gamma}(\nu, \xi)$ e la posteriori sarà

$$\vartheta|y^{\text{oss}} \sim \text{Gamma}\left(\nu + \sum_{i=1}^{n} y_i^{\text{oss}}, \xi + n \right).$$

Osservazione

Il valore atteso a posteriori è

$$\mathbb{E}\left[\vartheta|y^{\text{oss}}\right] = \frac{\nu + \sum_{i=1}^{n} y_i^{\text{oss}}}{\xi + n},$$

per cui scegliere ν e ξ vicini a 0 è come far sparire il loro effetto. In questo caso, ξ si può interpretare come "numero di osservazioni" aggiuntive.

Nota

È spesso vero che gli iperparametri si possono interpretati come *pseudo-osservazioni*, utile per scegliere valori coerenti con l'informazione a priori disponibile.

Esempio (Modello normale-normale)
Sia $y^{\text{oss}} = (y_1^{\text{oss}}, y_2^{\text{oss}}, \dots, y_n^{\text{oss}})$ un c.c.s. da $\mathcal{N}(\mu, \sigma^2)$ con σ^2 fissato.

Siccome σ^2 è fisso, lo si può portare fuori dall'esponente e ottenere la densità congiunta

$$\begin{aligned} p(y^{\text{oss}}|\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i^{\text{oss}} - \mu)^2\right\} \\ &\propto \exp\left\{\mu \cdot \sum_{i=1}^{n} \frac{y_i}{\sigma^2} - \frac{n}{2\sigma^2}\mu^2\right\} \\ &= \exp\left\{\mu \cdot \frac{n\overline{y}^{\text{oss}}}{\sigma^2} - \frac{n}{2\sigma^2}\mu^2\right\}. \end{aligned}$$

Allora, si può scegliere una distribuzione a priori con densità

$$\pi(\mu|\nu,\xi) \propto \exp\left\{\nu \cdot \mu - \xi \cdot \mu^2\right\},$$

che ha statistica canonica (μ, μ^2) ed è dunque la densità di una normale come visto in precedenza. Infatti, se si assume $\mu \sim \mathcal{N}(\nu, \tau^2)$, si ottiene la densità a priori cercata

$$\pi(\mu|\nu\xi) \propto \exp\left\{\frac{\nu}{\tau^2}\mu - \frac{1}{2\tau^2}\mu^2\right\}.$$

Si ottiene allora la densità a posteriori

$$\pi(\mu|y^{\text{oss}}) \propto \exp\left\{\mu \cdot \left(\frac{n\overline{y}^{\text{oss}}}{\sigma^2} + \frac{\nu}{\tau^2}\right) - \left(\frac{n}{2\sigma^2} + \frac{1}{2\tau^2}\right)\mu^2\right\},$$

dunque è ancora una distribuzione normale e $\mu|y^{\text{oss}} \sim \mathcal{N}(a,b)$, dove

$$\begin{aligned} \frac{1}{b} &= \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) \\ \frac{a}{b} &= \left(\frac{n\overline{y}^{\text{oss}}}{\sigma^2} + \frac{\nu}{\tau^2}\right) \end{aligned}$$

e, ricavando i parametri,

$$\mu|y^{\text{oss}} \sim \mathcal{N}\left(\frac{\frac{n\overline{y}^{\text{oss}}}{\sigma^2} + \frac{\nu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

Osservazione

Il valore atteso a posteriori è

$$\begin{aligned} \mathbb{E}\left[\mu|y^{\text{oss}}\right] &= \frac{\frac{n\overline{y}^{\text{oss}}}{\sigma^2} + \frac{\nu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \\ &= \frac{n\overline{y}^{\text{oss}} + (\sigma^2/\tau^2)\nu}{n + \sigma^2/\tau^2}, \end{aligned}$$

per cui la distribuzione a priori introduce un'informazione equivalente a σ^2/τ^2 pseudo-osservazioni con media ν e sposta la media campionaria $\overline{y}^{\text{oss}}$ verso il valore atteso della

distribuzione a priori.

1. Se $n \to \infty$ o $\tau^2 \to \infty$, ovvero se aumenta l'informazione dei dati rispetto alla priori, la posteriori tende a $\mathcal{N}(\overline{y}^{\text{oss}}, \sigma^2/n)$.
2. Se $\tau^2 \to 0$, la distribuzione a posteriori tende a $\mathcal{N}(\nu, \tau^2)$, che al limite è una variabile degenere in ν.

Tabella 1: Scelte comuni di distribuzioni a priori e a posteriori

Distribuzione	Parametro	Priori
$Bin(n, \pi)$	π	Beta(α, β)
Pois(λ)	λ	Gamma(α, β)
Exp(λ)	λ	Gamma(α, β)
$\mathcal{N}(\mu, \sigma^2)$	μ	$\mathcal{N}(\nu, \tau^2)$
$\mathcal{N}(\mu, \sigma^2)$	σ^2	$\mathcal{IG}(\alpha, \beta)$
Mult$_d(n, \pi)$	π	$\mathcal{D}(\alpha_1, \ldots, \alpha_d)$

Ci sono molti altri casi di distribuzioni coniugate, che si possono trovare su Wikipedia.

Lezione 9

9.1 Famiglie esponenziali e priori coniugate (cont.)

Esempio (Normale-Normale-Gammainversa)
Consideriamo un c.c.s. $y = (y_1, y_2, \ldots, y_n)$ da una v.c. $\mathcal{N}(\mu, \sigma^2)$ con verosimiglianza

$$L(\vartheta; y) \propto \left(\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}$$

$$\propto \left(\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left[nv_n^2 + n(\overline{y}-\mu)^2\right]\right\},$$

dove $v_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2$. Svolgendo il quadrato,

$$L(\vartheta; y) \propto \left(\sigma^2\right)^{-n/2} \exp\left\{n\overline{y}\frac{\mu}{\sigma^2} - \frac{1}{2\sigma^2}(nv_n^2 + n\overline{y}^2) - n\frac{\mu^2}{2\sigma^2}\right\}.$$

La distribuzione a priori coniugata è una distribuzione *normale-gamma inversa* $\vartheta \sim \mathcal{N} - \mathcal{IG}(\nu, \lambda, \alpha, \beta)$ che si può definire con una scomposizione

$$\pi(\vartheta) = \pi(\mu|\sigma^2)\pi(\sigma^2),$$

dove

$$\mu|\sigma^2 \sim \mathcal{N}(\nu, \frac{\sigma^2}{\lambda})$$

$$\sigma^2 \sim \mathcal{IG}(\alpha, \beta)$$

Quindi, la densità congiunta a priori è

$$\pi(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}}\left(\frac{\lambda}{\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda}{2\sigma^2}(\mu-\nu)^2\right\} \times \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left\{-\frac{\beta}{\sigma^2}\right\}$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\alpha+1+\frac{1}{2}} \exp\left\{-\frac{\lambda}{\sigma^2}(\mu-\nu)^2 - \frac{\beta}{\sigma^2}\right\}$$

$$= \left(\frac{1}{\sigma^2}\right)^{\alpha+\frac{3}{2}} \exp\left\{\lambda\nu\frac{\mu}{\sigma^2} - \frac{1}{2\sigma^2}(\lambda\nu^2 + 2\beta) - \lambda\frac{\mu^2}{2\sigma^2}\right\}.$$

La distribuzione a posteriori, allora, non è altro che

$$\vartheta|Y \sim \mathcal{IG}(\nu^*, \lambda^*, \alpha^*, \beta^*),$$

dove gli iperparametri aggiornati sono

$$\alpha^* = \alpha + \frac{n}{2}$$

$$\lambda^* = \lambda + n$$

$$\nu^* = \ldots = \frac{\lambda\nu + n\overline{y}}{\lambda + n}$$

$$\beta^* = \ldots = \beta + \frac{n}{2}v_n^2 + \frac{n\lambda}{2(\lambda+n)}(\overline{y} - \nu)^2$$

Osservazione

A volte si considera la parametrizzazione $\psi = 1/\sigma^2$, in qual caso la distribuzione a priori coniugata è una normale-gamma.

Partendo dalla distribuzione congiunta $\mathcal{N}$-$\mathcal{IG}(\nu, \lambda, \alpha, \beta)$, è interessante calcolare la distribuzione marginale per la media μ:

$$\begin{aligned}
\pi(\mu|y^{\text{oss}}) &= \int_0^{+\infty} \pi(\mu, \sigma^2|y^{\text{oss}})\, d\sigma^2 \\
&= \frac{\lambda^{1/2}}{\sqrt{2\pi}}\frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} \underbrace{\left(\frac{1}{\sigma^2}\right)^{\alpha+\frac{1}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{\lambda}{2}(\mu-\nu)^2 + \beta\right]\right\}}_{\text{kernel di una IG}} d\sigma^2 \\
&= \frac{\lambda^{1/2}}{\sqrt{2\pi}}\frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\frac{1}{2})}{\left(\frac{\lambda}{2}(\mu-\nu)^2 + \beta\right)^{\alpha+\frac{1}{2}}} \\
&= \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)}\frac{\lambda^{\frac{1}{2}}}{\sqrt{2\pi}}\beta^\alpha \left[\beta + \frac{\lambda}{2}(\mu-\nu)^2\right]^{-(\alpha+\frac{1}{2})}
\end{aligned}$$

che è una distribuzione t di Student con $k = 2\alpha$ gradi di libertà, parametro di posizione $m = \nu$ e parametro di scala $s = (\beta/\lambda\alpha)^{1/2}$, ovvero con densità

...

In particolare, se gli iperparametri della distribuzione a priori tendono a 0, si ha che

$$\alpha = \frac{n}{2}; \quad \lambda = n; \quad \nu = \overline{y}; \quad \beta = \frac{n}{2}v_n^2,$$

dunque $\mu|y^{\text{oss}}$ è una t di Student con n gradi di libertà, parametro di posizione $\overline{y}$ e parametro di scala $v_n/\sqrt{n}$.

9.2 Misture di distribuzioni coniugate

Le distribuzioni a priori coniugate, per quanto siano flessibili negli iperparametri, spesso sono troppo restrittive per poter rappresentare alcuni tipi di informazione a priori (e.g. binomiale).

Si può però dimostrare che una *mistura* di densità coniugate è ancora coniugata a priori, e questo estende considerevolmente la classe di densità con posteriori in forma chiusa.

Una mistura di m densità $f_1(y), f_2(y), \dots, f_m(y)$ con pesi $\pi_1, \pi_2, \dots, \pi_m$, $p_j > 0$ e $\sum_{j=1}^m p_j = 1$ ha densità

$$f_M(y) = \sum_{j=1}^m \pi_j f_j(y).$$

Esempio (Mistura di distribuzioni beta)
Consideriamo $Y \sim \text{Bin}(n, \vartheta)$ con distribuzione a priori una mistura di due distribuzioni beta, con probabilità di mistura p. La distribuzione a priori è dunque

$$\pi(\vartheta) = p\frac{1}{B(\alpha_1, \beta_1)}\vartheta^{\alpha_1 - 1}(1-\vartheta)^{\beta_1 - 1} + (1-p)\frac{1}{B(\alpha_2, \beta_2)}\vartheta^{\alpha_2 - 1}(1-\vartheta)^{\beta_2 - 1},$$

mentre la verosimiglianza è

$$L(\vartheta; y^{\text{oss}}) \propto \vartheta^{y^{\text{oss}}}(1-\vartheta)^{n - y^{\text{oss}}}.$$

La distribuzione a posteriori diventa

$$\begin{aligned}\pi(\vartheta | y^{\text{oss}}) \propto \quad & p\frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}\vartheta^{\alpha_1 + y^{\text{oss}} - 1}(1-\vartheta)^{\beta_1 + n - y^{\text{oss}} - 1} + \\ & (1-p)\frac{\Gamma(\alpha_2 + \beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)}\vartheta^{\alpha_2 + y^{\text{oss}} - 1}(1-\vartheta)^{\beta_2 + n - y^{\text{oss}} - 1}.\end{aligned}$$

Moltiplicando e dividendo per il fattore che normalizza il kernel a posteriori, si ottiene che la distribuzione a posteriori è una mistura di due distribuzioni beta, con parametri $(\alpha_j + y^{\text{oss}}, \beta_j + n - y^{\text{oss}})$ e pesi

$$\tilde{\pi}_j = \frac{\pi_j \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \frac{\Gamma(\alpha_j + y^{\text{oss}})\Gamma(\beta_j + n - y^{\text{oss}})}{\Gamma(\alpha_j + \beta_j + n)}}{\sum_{j=1}^m \pi_j \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \frac{\Gamma(\alpha_j + y^{\text{oss}})\Gamma(\beta_j + n - y^{\text{oss}})}{\Gamma(\alpha_j + \beta_j + n)}}.$$

Osservazione

La distribuzione a posteriori rimane una mistura di due distribuzioni beta, in cui non solo vengono aggiornati i parametri delle distribuzioni della mistura, ma anche i pesi che la compongono.

Applicazione

Applicando questo tipo di mistura ai dati dei tiri liberi nel basket, in particolare assumendo

i) $0.75 \cdot \text{Beta}(3, 9) + 0.25 \cdot \text{Beta}(9, 3)$

ii) $0.25 \cdot \text{Beta}(3, 9) + 0.75 \cdot \text{Beta}(9, 3)$

si ottengono le seguenti distribuzioni a posteriori:

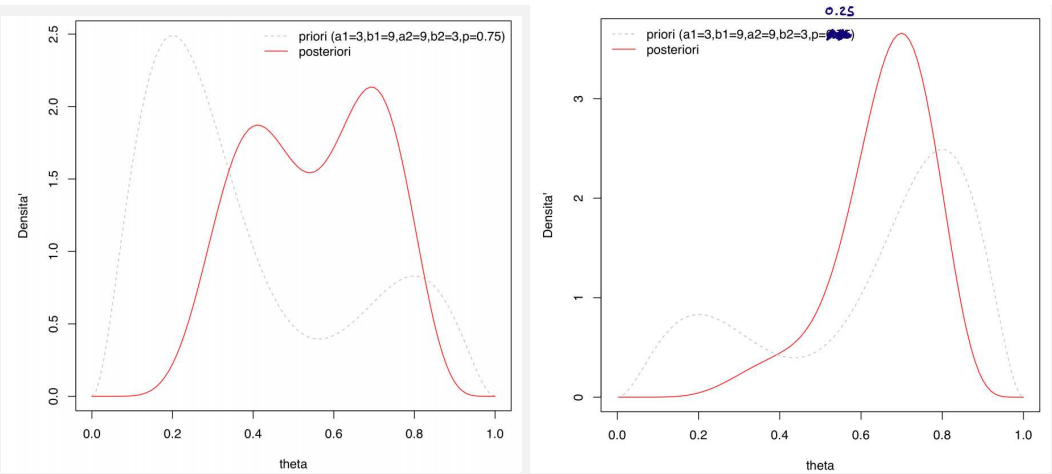


Figura 7: Distribuzioni a priori e a posteriori nei casi (i) e (ii), per i tiri liberi nel basket.

Una distribuzione a priori di questo tipo potrebbe derivare dal fatto che disponiamo di una squadra composta in parte da giocatori bravi e in parte da giocatori scarsi, e non sappiamo chi ha fatto i tiri liberi.

Lezione 10

Riferimenti Liseo (2010, §3.1-3.2)

10.1 Distribuzioni a priori non informative

A volte le distribuzioni a priori devono rappresentare ignoranza o *assenza di informazione* sul parametro, ad esempio quando abbiamo bisogno di una distribuzione "baseline" con cui confrontarsi.

Un'altra situazione è il caso in cui si voglia valutare la sensibilità della distribuzione a posteriori al cambiamento della distribuzione a priori. Questo può essere importante per capire se la distribuzione a priori sta influenzando eccessivamente i risultati dell'analisi.

Famiglie di posizione

Quando il parametro ha supporto limitato, una priori uniforme

$$\pi(\vartheta) \propto 1,$$

è una scelta ovvia.

Se il supporto è illimitato, una priori costante ha integrale illimitato ed è quindi *impropria*; ciò nonostante, a volte una priori impropria può portare a una distribuzione a posteriori propria, proprietà che va verificata caso per caso.

Nell'esempio della normale con varianza nota, se si assume $\tau^2 \longrightarrow \infty$ nella priori, si ottiene $\pi(\vartheta) \propto 1$ e la distribuzione a posteriori è comunque propria e pari a $\mathcal{N}(\overline{y}_n, \sigma^2/n)$.

Famiglie di scala

In una famiglia di scala con densità $p(y|\sigma) = \frac{1}{\sigma} p_0(\frac{y}{\sigma})$, una tipica scelta di priori non informativa è la densità impropria

$$\pi(\sigma) \propto \sigma^{-1}, \quad \sigma > 0.$$

Caso generale

Dato un parametro generico, tuttavia, non è immediato rappresentare è un'ignoranza a priori. In un certo senso, l'obiettivo è capire *rispetto a cosa* si vuole rappresentare l'ignoranza.

Esempio (Distribuzione binomiale)
Se $Y \sim \text{Bin}(n, \vartheta)$, $\vartheta \in (0, 1)$ potremmo usare la priori non informativa $\pi(\vartheta) \propto 1$.

Tuttavia, se siamo ignoranti su ϑ dovremmo esserlo anche sulla riparametrizzazione

$\psi = \log(\vartheta/(1-\vartheta)) \in \mathbb{R}$. In questa parametrizzazione, la densità a priori diventa

$$\pi(\psi) = \pi(\vartheta(\psi)) \left| \frac{d\vartheta(\psi)}{d\psi} \right|$$

$$= \frac{e^{\psi}}{(1+e^{\psi})^2},$$

ovvero $\psi \sim \text{Logis}(0,1)$, che non è completamente non informativa:

$$P(|\psi| < 3) = \frac{e^3}{1+e^3} - \frac{e^{-3}}{1+e^{-3}} \approx 0.905.$$

Analogamente, se consideriamo $\tilde{\pi}(\psi) \propto 1$, la trasformazione inversa è

$$\tilde{\pi}(\vartheta) = \tilde{\pi}(\psi(\vartheta)) \left| \frac{d\psi(\vartheta)}{d\vartheta} \right| \propto \frac{1}{\vartheta(1-\vartheta)},$$

che è molto informativa (vedi Figura 8).

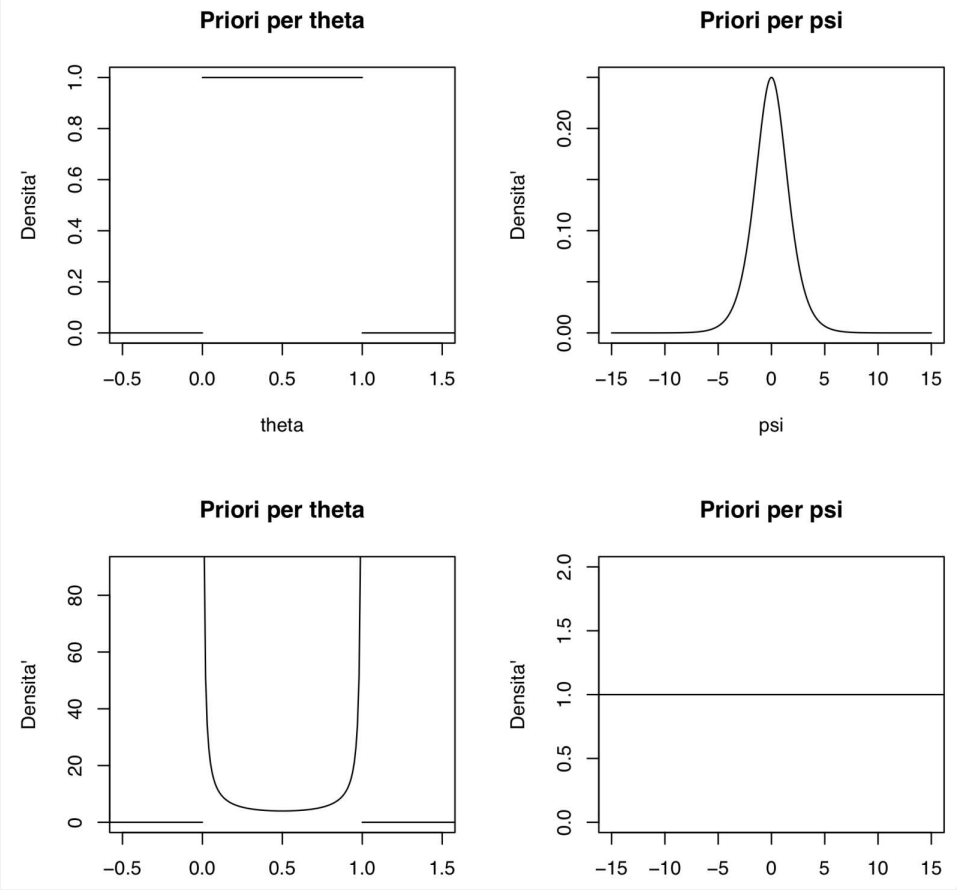


Figura 8: invariancePrior

Nota

Ci sono approcci più formali per costruire distruzioni a priori non informative, anche in modo automatico. Questi approcci si basano su nozioni di *invarianza*, *informazione minima* o *matching* tra proprietà di copertura bayesiana e confidenza frequentista.

10.2 Distribuzioni a priori soggettive

Queste distribuzioni a priori dovrebbero inglobare l'informazione a priori, intesa soprattutto come *opinioni di esperti*.

Il processo di traduzione di informazione a priori in una distribuzione di probabilità si chiama *elicitazione* ed è tutt'altro che semplice, specialmente se il parametro ha dimensione elevata.

Anche quando il parametro ha dimensione ridotta, ci possono essere diverse scelte di priori compatibili con l'informazione a priori: è dunque buona prassi accompagnare l'inferenza con un'*analisi di sensitività* rispetto alla scelta di $\pi(\vartheta)$.

Esempio (Differenze nelle posteriori)
Supponiamo che y sia un'unica realizzazione da $Y \sim \text{Pois}(\vartheta)$ e che sia disponibile a priori l'informazione che 2 e 4 sono mediana e terzi quartile per ϑ. Si potrebbero usare allora

1. $\vartheta \sim \text{Exp}(\alpha = 2^{-1} \log 2)$
2. $\log \vartheta \sim \mathcal{N}(\log 2, (\log 2/0.675)^2)$
3. $\log \vartheta \sim \text{Cauchy}(\log 2, \log 2)$

Gli iperparametri sono stati scelti in modo che le probabilità $P(\vartheta \in (0, 2)) = 0.5$ e $P(\vartheta \in (2, 4)) = 0.75$.

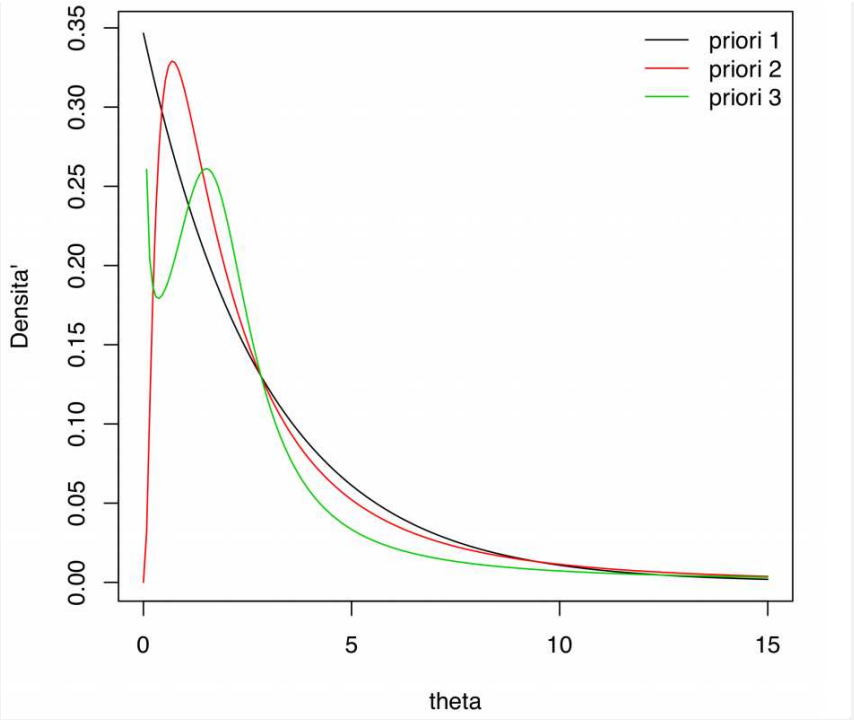


Figura 9: diffPriors

Le distribuzioni a posteriori nei tre casi sono

1. $$\pi(\vartheta|y^{\text{oss}}) = \dots$$
2. $$\pi(\vartheta|y^{\text{oss}}) = \dots$$
3. $$\pi(\vartheta|y^{\text{oss}}) = \dots$$

Vedendo le distribuzioni a posteriori per diverse osservazioni di y, è chiaro che diverse distribuzioni a priori, tutte compatibili con l'informazione a priori, possono risultare in distribuzioni diverse.

Commento

Ovviamente le distribuzioni a posteriori sono sostanzialmente diverse, in quanto $n = 1$. Quando n è elevato, l'impatto della distribuzione a priori sulla verosimiglianza è più lieve.

Lezione 11

11.1 Scambiabilità

Il concetto di *scambiabilità* è una generalizzazione dell'indipendenza e identica distribuzione per una successione di v.c.

È importante l'assunzione di scambiabilità sui parametri, meno restrittiva dell'ipotesi di indipendenza, per giustificare l'approccio bayesiano come base della costruzione di modelli gerarchici.

Def. (Scambiabilità finita)
Le variabili casuali $Y_1, Y_2, \dots, Y_n$ si dicono *finitamente scambiabili* se la distribuzione congiunta di qualunque permutazione coincide con quella di $Y_1, Y_2, \dots, Y_n$, ovvero

$$P(Y_{\sigma(1)}, Y_{\sigma(2)}, \dots, Y_{\sigma(n)}) = P(Y_1, Y_2, \dots, Y_n) \quad \forall \sigma : \{1, 2, \dots, n\} \longrightarrow \{1, 2, \dots, n\}.$$

Esempio (Scambiabilità $\not\Longrightarrow$ indipendenza)
Sia $Y = (Y_1, Y_2, \dots, Y_n) \sim \mathcal{N}_n(\mu \mathbf{1}_n, \sigma^2 \Omega)$, con

$$\Omega = (\omega)_{ij} = \begin{cases} 1 & \text{se } i = j \\ \rho & \text{se } i \neq j \end{cases}$$

con determinante

$$|\Omega| = (1-\rho)^{n-1}(1+\rho(n-1)) > 0 \implies -\frac{1}{n-1} < \rho < 1.$$

Le variabili $Y_1, \dots, Y_n$ sono infinitamente scambiabili (si può verificare trasformando con una matrice di permutazione), con $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. Tuttavia, non sono indipendenti, a meno che $\rho = 0$.

Def. (Scambiabilità)
La successione $(Y_n)_{n \in \mathbb{N}}$ si dice *scambiabile* se per ogni n-pla di v.c. risulta finitamente scambiabile.

Commenti

- V.c. scambiabili sono necessariamente identicamente distribuite (somiglianti).
- V.c. indipendenti sono anche scambiabili, ma non vale il viceversa.
- La scambiabilità è una *simmetria* della distribuzione rispetto agli argomenti della densità congiunta.

Teo. (Rappresentazione di de Finetti)
Sia $Y_1, Y_2, \ldots$ una successione scambiabile di v.c. binarie, che assumono valori $y_i \in \{0,1\}$. Allora, per ogni n esiste una distribuzione G tale che le Y_i sono condizionatamente indipendenti dato G:

$$p(y_1, y_2, \ldots, y_n) = \int_0^1 \underbrace{\prod_{j=1}^{n} \vartheta^{y_j}(1-\vartheta)^{1-y_j}}_{Y_i|\vartheta \overset{i.i.d.}{\sim} Ber(\vartheta)} \, dG(\vartheta),$$

dove

$$G(\vartheta) = \lim_{m\to\infty} P(m^{-1}(Y_1 + \ldots + Y_m) \leq \vartheta)$$

e

$$m^{-1}(Y_1 + \ldots + Y_m) \xrightarrow{P} \vartheta.$$

In particolare, si possono vedere i dati come realizzazione di un modello parametrico, con distribuzione a priori sul parametro ϑ.

Una generalizzazione di questo risultato si ha per variabili casuali scambiabili non necessariamente binarie.

Commenti

- Se si parte da variabili casuali binarie scambiabili, si possono vedere come v.c. Bernoulli indipendenti condizionatamente alla prob. di successo ϑ con distribuzione G, interpretabile come proporzione di successi osservati al crescere di n.
- Giustificazione bayesiana, in quanto se si assume scambiabilità delle osservazioni vuol dire che $P(Y_{n+1}|Y_n, \ldots, Y_1)$ è il rapporto di due integrali di quel tipo, che è la *densità predittiva bayesiana* usando una priori per ϑ.
- La scambiabilità su insiemi di parametri sta alla base dei *modelli gerarchici bayesiani*: se si assume che $\vartheta_1, \ldots, \vartheta_n$ sono a priori scambiabili, allora possono essere pensati come un campione casuale da una distribuzione ignota.

Esempio (Modello normale gerarchico)
Assumiamo che $Y_1, Y_2, \ldots, Y_n$ siano tali che

$$\mu \sim \mathcal{N}(\nu_0, \tau_0^2)$$

$$\vartheta_1, \ldots, \vartheta_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma_0^2)$$

$$Y_i|\vartheta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\vartheta_i, v_i^2)$$

Gli iperparametri ν_0, τ_0^2 controllano l'incertezza inserita al livello più alto della gerarchia.

Come esempio, si può pensare che Y_i siano i voti degli studenti dopo un esame, ϑ_i, v_i^2 la

bravura media e la variabilità dello studente. Si assume che la bravura media provenga da una popolazione con media μ, che ha a sua volta una distribuzione normale.

Nota

Versioni più generali mantengono anche v_i^2 e σ_0^2 ignote, assegnando delle distribuzioni a priori (spesso coniugate con gamma-inversa).

Una rappresentazione utile per questo tipo di modelli è quella grafica, dove i cerchi sono variabili casuali e quadrati sono costanti. Per questo esempio, si ha la seguente rappresentazione:

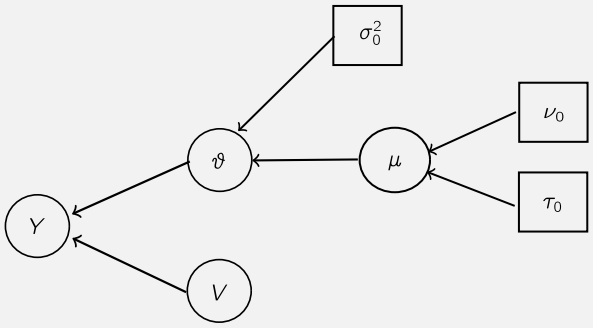


Figura 10: Rappresentazione grafica del modello gerarchico.

Con molta algebra, si ottiene una distribuzione per $\mu, \vartheta | y^{\text{oss}}$ e i valori attesi a posteriori sono

$$\mathbb{E}\left[\mu | y^{\text{oss}}\right] = \frac{\frac{\nu_0}{\tau_0^2} + \sum_{i=1}^{n} \frac{y_i}{\sigma_0^2 + v_i^2}}{\frac{1}{\tau_0^2} + \sum_{i=1}^{n} \frac{1}{\sigma_0^2 + v_i^2}}$$

$$\mathbb{E}\left[\vartheta_i | y^{\text{oss}}\right] = \mathbb{E}\left[\mu | y^{\text{oss}}\right] + \frac{\sigma_0^2}{\sigma_0^2 + v_i^2} \left(y_i^{\text{oss}} - \mathbb{E}\left[\mu | y^{\text{oss}}\right]\right)$$

Quindi la media a posteriori di μ è una media ponderata tra la priori e i dati osservati. La media a posteriori di ϑ_i è pesata tra quelle di y_i^{oss} e $\mathbb{E}\left[\mu | y^{\text{oss}}\right]$, mostrando un effetto di "*shrinkage*" verso la media della popolazione.

Questa è una caratteristica dei modelli gerarchici, in cui c'è un effetto di "borrowing of strength": per stimare le informazioni di un soggetto uso i dati degli altri soggetti.

11.2 Inferenza statistica

Riferimenti Pace e Salvan (2001, §2) (Inferenza frequentista)
Liseo (2010, §4) (Inferenza bayesiana)

Supponiamo di avere assegnato un modello statistico parametrico

$$\mathcal{F} = \{p_Y(y; \vartheta), \vartheta \in \Theta \subseteq \mathbb{R}^p\}$$

per i dati $y = y^{\text{oss}}$. Si supponga inoltre che $\mathcal{F}$ sia correttamente specificato e che, in base all'approccio di analisi,

- *Frequentista*: il parametro abbia vero valore ϑ_0.
- *Bayesiano*: il parametro ϑ sia generato da $\pi(\vartheta)$.

Problema di inferenza

Vogliamo utilizzare i dati y per rispondere a quesiti su ϑ e su osservazioni future y^*, associando alla risposta una *valutazione dell'incertezza*.

L'obiettivo è individuare procedure di inferenza appropriate *qualunque sia* $\vartheta_0 \in \Theta$, in quanto non sappiamo quale sia il valore specifico del parametro.

Notazione a meno di ambiguità, il vero valore ϑ_0 si indicherà con ϑ.

I quesiti standard su ϑ possono essere

- Stima puntuale: sulla base di y, quale elemento di $\mathcal{F}$ è più "appropriato" per i dati?
- Stima per regioni: (con $p = 1$ stima intervallare) come sopra, ma si desidera individuare un insieme di elementi di $\mathcal{F}$, ossia un sottoinsieme Θ.
- Verifica di ipotesi: duale della stima intervallare, ci si chiede se i dati y sono compatibili con l'ipotesi $\vartheta \in \Theta_0 \subset \Theta$.
- Previsione: previsione di valori, o insieme di valori, di un'osservazione futura y^* generata dallo stesso modello che ha generato y, o da una sua componente.

In generale l'inferenza si basa su funzioni $T : \mathcal{Y} \longrightarrow \mathbb{R}^k$ che sintetizzano i dati, dette *statistiche*.

> Sia l'individuazione di procedure inferenziali sia la quantificazione dell'incertezza associata avvengono in modo diverso a seconda dell'approccio, bayesiano o frequentista.

Le soluzioni, tuttavia, spesso sono molto simili tra gli approcci o comunque collegate.

Come organizzazione del corso, in primo luogo richiameremo le nozioni relative all'inferenza frequentista e successivamente le corrispondenti procedure di inferenza bayesiana.

11.3 Inferenza frequentista: stima puntuale

Riferimenti Pace e Salvan, (2001, §2.2)

Def. (Stima)
Una *stima* è definita da una funzione $\tilde{\vartheta} : \mathcal{Y} \longrightarrow \Theta$, ed è indicata con $\tilde{\vartheta} = \tilde{\vartheta}(y)$.

Prima di osservare i dati, chiaramente $\tilde{\vartheta}$ è una funzione di Y, per cui è una variabile casuale e si dice *stimatore*.
Lo stimatore deve godere di alcune proprietà per essere considerato di buona qualità:

- Lo stimatore $\tilde{\vartheta}(Y)$ è detto *non distorto (in media)* se

$$\mathbb{E}_\vartheta\left[\tilde{\vartheta}(Y)\right] = \vartheta,$$

per ogni $\vartheta \in \Theta$. Ripetendo l'esperimento infinite volte, in media si deve ottenere il vero valore del parametro. Si ricorda infatti che

$$\mathbb{E}_{\vartheta_0}\left[\tilde{\vartheta}(Y)\right] = \int_{\mathcal{Y}} \tilde{\vartheta}(y)\, dP(y;\vartheta_0)$$

- La *distorsione (bias)* di $\tilde{\vartheta}(Y)$ è

$$b(\vartheta) = \mathbb{E}_\vartheta\left[\tilde{\vartheta}(Y)\right] - \vartheta.$$

- L'*errore quadratico medio* è

$$\mathbb{E}_\vartheta\left[\left\{\tilde{\vartheta}(Y) - \vartheta\right\}^2\right] = \mathbb{V}_\vartheta\left[\tilde{\vartheta}(Y)\right] + b(\vartheta)^2.$$

- Uno stimatore basato su $Y = (Y_1, Y_2, \ldots, Y_n)$, $\tilde{\vartheta}_n = \tilde{\vartheta}(Y_1, Y_2, \ldots, Y_n)$ è detto *(debolmente) consistente* per ϑ se $\tilde{\vartheta}_n \xrightarrow{P} \vartheta$, cioè se

$$\lim_{n\to\infty} P(|\tilde{\vartheta}_n - \vartheta| > \varepsilon) = 0.$$

In pratica, più informazione si ha da parte del campione, meglio si comporta lo stimatore.

Nota Se $EQM(\tilde{\vartheta}_n) \xrightarrow{n\to\infty} 0$, allora lo stimatore è anche debolmente consistente.

- $\tilde{\vartheta}_n$ si dice *(fortemente) consistente* per ϑ se $\tilde{\vartheta}_n \xrightarrow{\text{a.s.}} \vartheta$.

Lezione 12

Riferimenti Pace e Salvan, (2001, §2.4)

12.1 Inferenza frequentista: verifica di ipotesi

Def. (Ipotesi nulla)
Un'*ipotesi nulla* $H_0 : \vartheta \in \Theta_0 \subset \Theta$ corrisponde all'assunzione che $p_Y(y;\vartheta)$ appartenga al sottomodello $\mathcal{F}_0$ di $\mathcal{F}$ con spazio parametrico Θ_0.

Tipicamente H_0 è un'assunzione semplificatrice, ad esempio nel caso del genetic linkage può interessare $H_0 : \vartheta = \frac{1}{4}$, che corrisponde alla teoria mendeliana.

Se $\mathcal{F}_0$ ha un solo elemento, H_0 si dice *semplice*, altrimenti è detta *composita*.

Def. (Test)
Un *test* (o *statistica test*) è una funzione $t : \mathcal{Y} \longrightarrow \mathbb{R}$ che identifica in $\mathcal{Y}$ due regioni

- La *regione di accettazione* A_{Θ_0}, in cui i dati non indicano evidenza contro H_0.
- La *regione di rifiuto* $R_{\Theta_0} = \mathcal{Y} \setminus A_{\Theta_0}$ in cui i dati indicano evidenza contro H_0.

Se $y \in R_{\Theta_0}$, si dice che il test è *significativo contro* H_0.

Si dice che il test ha *livello* α se la più alta probabilità di rifiutare H_0 quando essa è vera è pari ad α

$$\sup_{\vartheta \in \Theta_0} P_\vartheta(Y \in R_{\Theta_0}) = \alpha,$$

mentre si dice che il test ha *livello costante* α se $P_\vartheta(Y \in R_{\Theta_0}) = \alpha$ per ogni $\vartheta \in \Theta_0$.

Tabella 2: Errori nel test di ipotesi: α è la prob. di errore di I tipo, β è la prob. di errore di II tipo

	Θ_0	$\Theta \setminus \Theta_0$
A_{Θ_0}	✓	β
R_{Θ_0}	α	✓

Pensiamo ad α come la più alta probabilità di errore di I tipo, a meno che il test sia a livello costante.

Le distribuzioni di $t(Y)$ sotto $\vartheta \in \Theta_0$ sono dette *distribuzioni nulle*, mentre le stesse distribuzioni sotto $\vartheta \in \Theta \setminus \Theta_0$ sono dette *distribuzioni non nulle*.

Se le distribuzioni nulle di $t(Y)$ sono tutte somiglianti, o si ha un'unica distribuzione nulla, si dice che il test è *simile*. Questo è banale quando l'ipotesi nulla è semplice, mentre nel caso composito avere un test simile produce un test di livello costante.

Def. (Ipotesi alternativa)
Un'*ipotesi alternativa* H_1 è l'affermazione che $\vartheta \in \Theta_1$, $\Theta \subset \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$, ed a sua volta può essere semplice o composita.

Def. (Funzione di potenza)
La funzione di potenza di un test è definita come

$$\pi(\vartheta) = P_\vartheta(Y \in R_{\Theta_0}),$$

per ogni valore di $\vartheta \in \Theta$. Se per i valori specificati dall'ipotesi nulla $\sup_{\vartheta \in \Theta_0} \pi(\vartheta) = \alpha$, il test ha livello α.

Vogliamo che la potenza sia più alta possibile quando siamo in $\vartheta \notin \Theta_0$, perché significa che si fa giusto con buona probabilità.

In particolare, un test per $H_0 : \vartheta \in \Theta_0$ basato su $Y_1, Y_2, \ldots, Y_n$ con distribuzione in $\mathcal{F}$ si dice *consistente* se

$$\lim_{n\to\infty} \pi(\vartheta) = 1 \quad \forall \vartheta \in \Theta \setminus \Theta_0.$$

Una statistica test $t(Y)$ con valori in $\mathbb{R}$ è detta *unilaterale* se

$$R_{\Theta_0} = \{y \in \mathcal{Y} : t(y) > c_\alpha\}$$

oppure

$$R_{\Theta_0} = \{y \in \mathcal{Y} : t(y) < c_\alpha\}$$

Def. (*p*-value)
Per una statistica test con valori grandi significativi contro H_0, il *livello di significatività osservato* (*p-value*) per i dati $y = y^{\text{oss}}$ è

$$\alpha^{\text{oss}} = \sup_{\vartheta \in \Theta_0} P_\vartheta\{t(Y) \geq t(y^{\text{oss}})\}$$

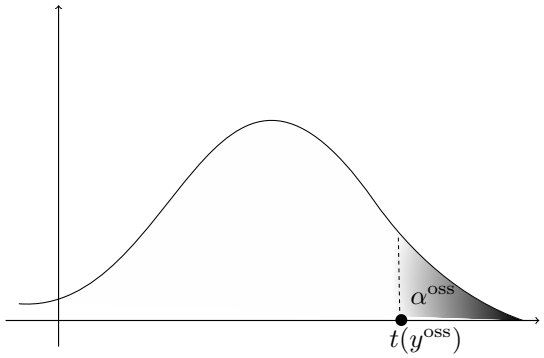


Figura 11: Tanto più piccolo è il p-value osservato, tanto maggiore è l'evidenza contro l'ipotesi nulla.

Una statistica test è detta *bilaterale* se

$$R_{\Theta_0} = \{y \in \mathcal{Y} : t(y) < c_{\alpha'}\} \cup \{y \in \mathcal{Y} : t(y) > c_{\alpha''}\}, \quad c_{\alpha'} < c_{\alpha''}.$$

In tal caso, il p-value osservato è definito come

$$\alpha^{\text{oss}} = 2 \sup_{\vartheta \in \Theta_0} \min \left\{ P_\vartheta\big(t(Y) < t(y^{\text{oss}})\big), P_\vartheta\big(t(Y) > t(y^{\text{oss}})\big) \right\}.$$

Osserviamo che α^{oss} è a sua volta una statistica, ovvero $\alpha^{\text{oss}} = \alpha^{\text{oss}}(Y)$.

Si può inoltre dimostrare che, se t è un test simile con distribuzione nulla continua, la variabile casuale $\alpha^{\text{oss}}(Y)$ ha distribuzione Unif(0,1) sotto H_0 (Pace e Salvan (2001, p. 17)).

Dim.
Nel caso di regione critica destra, vale che

$$\alpha^{\text{oss}} = P(T \geq t | H_0) = 1 - F_T(T(Y)).$$

Sia $X = F_T(T(Y))$, allora

$$P(X \leq x) = P\big(F_T(T(Y)) \leq x\big) = P\big(T(Y) \leq F_T^{-1}(x)\big) = F_T(F_T^{-1}(x)) = x,$$

per cui $X \sim \text{Unif}(0,1)$ e dunque $\alpha^{\text{oss}} = 1 - X \implies \alpha^{\text{oss}} \sim \text{Unif}(0,1)$.

□

Grazie a questa proposizione, un test di livello fissato si può costruire dopo aver calcolato α^{oss}, semplicemente confrontandolo con un α fissato.

12.2 Inferenza frequentista: regioni di confidenza

Riferimenti Pace e Salvan, (2001, p. 2.5)

Def. (Regione di confidenza)
Una *regione di confidenza* per ϑ con livello $1 - \alpha$ è una regione $\hat{\Theta}(y) \subset \Theta$ tale che, per ogni $\vartheta \in \Theta$.

$$P_\vartheta\left(\hat{\Theta}(Y) \ni \vartheta\right) = 1 - \alpha.$$

Nota bene Il valore ϑ è una costante fissata, mentre è la regione di confidenza ad essere aleatoria: da qui, l'utilizzo del simbolo $\ni$.

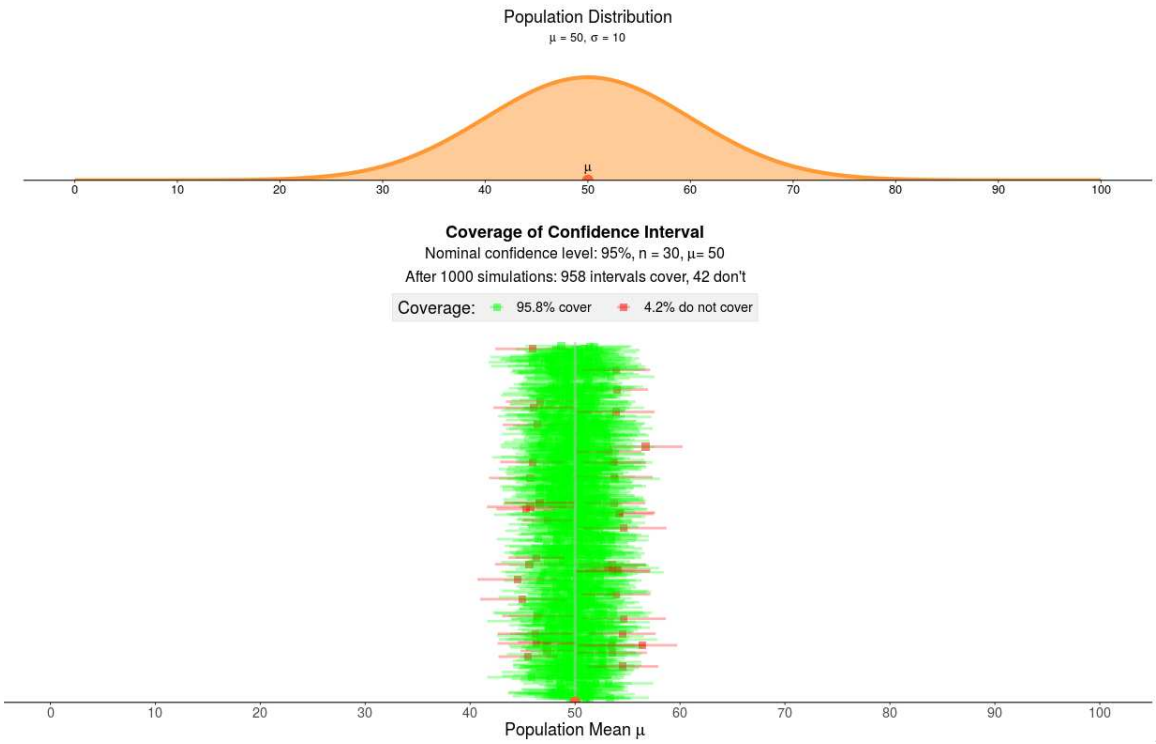


Figura 12: Intervalli di confidenza con copertura nominale 95%, 1000 simulazioni.

La *probabilità di copertura non nulla*, che vorremmo minimizzare, è la probabilità che la procedura includa valori sbagliati di ϑ

$$P_\vartheta \left\{ \hat{\Theta}(y) \ni \vartheta' \right\}, \quad \vartheta', \vartheta \in \Theta,\ \vartheta' \neq \vartheta.$$

Costruzione di regioni di confidenza

Def. (Quantità pivotale)
Si definisce *quantità pivotale* per $\mathcal{F}$ è una funzione $q(y, \vartheta)$ tale che, sotto ϑ, abbia distribuzione indipendente da ϑ

$$q(Y, \vartheta) \sim p_Q(q)$$

Esempio (Normale con varianza nota)
Se $Y \sim \mathcal{N}(\vartheta, 1)$, allora

$$Y - \vartheta \sim \mathcal{N}(0, 1)$$

per cui la funzione $q(y, \vartheta) = y - \vartheta$ è una quantità pivotale per il modello.

Una quantità pivotale scalare permette di

1. Definire una statistica test per l'ipotesi semplice $H_0 : \vartheta = \vartheta_0$, in quanto $t(Y) = q(Y, \vartheta_0)$ ha distribuzione nulla $p_Q(q)$ sotto H_0.

2. Costruire una regione di confidenza per ϑ con livello $1-\alpha$: se per ogni $\vartheta \in \Theta$

$$P_\vartheta \{q(Y, \vartheta) \in E_{1-\alpha}\} = 1 - \alpha,$$

allora

$$\hat{\Theta}(y) = \{\vartheta \in \Theta : q(y, \vartheta) \in E_{1-\alpha}\}$$

è una regione di confidenza per ϑ con livello $1-\alpha$ (Pace e Salvan (2001, §2.5.2)).

Esempio (Normale con varianza nota)
Nel caso precedente di $Y \sim \mathcal{N}(\vartheta, 1)$, poiché $Y - \vartheta \sim \mathcal{N}(0, 1)$ si ha che

$$P_\vartheta \left(Y - \vartheta \in (z_{\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}) \right) = 1 - \alpha,$$

da cui si può ottenere la regione di confidenza

$$\hat{\Theta}(Y) = \left\{ \vartheta \in \Theta : Y - z_{1-\frac{\alpha}{2}} \leq \vartheta \leq Y - z_{\frac{\alpha}{2}} \right\}.$$

Regioni di confidenza basate su una classe di test

C'è una sorta di *dualità* tra verifica di ipotesi e regioni di confidenza: una regione con livello $1-\alpha$ può essere ottenuta a partire da una regione di accettazione A_{ϑ_0} di una classe di test con livello α per $H_0 : \vartheta = \vartheta_0$.

Per ogni $y \in \mathcal{Y}$, sia

$$\hat{\Theta}(y) = \{\vartheta \in \Theta : y \in A_\vartheta\},$$

allora si ha che

$$P_\vartheta \left\{ \hat{\Theta}(Y) \ni \vartheta \right\} = P_\vartheta \{Y \in A_\vartheta\} = 1 - \alpha.$$

La probabilità di copertura non nulla è quindi legata alla potenza del test:

$$P_\vartheta \left\{ \hat{\Theta}(Y) \ni \vartheta' \right\} = P_\vartheta \{Y \in A_{\vartheta'}\} = 1 - \pi_{\vartheta'}(\vartheta),$$

dove $\pi_{\vartheta'}(\vartheta)$ è la potenza del test per $H_0 : \vartheta = \vartheta'$. Se la potenza è elevata, la probabilità di copertura non nulla diventa molto piccola.

12.3 Inferenza frequentista: previsione

Riferimenti Pace e Salvan, (2001, §2.3)

Nei problemi di previsione, ci si limiterà ad osservare gli *intervalli di previsione* per un valore futuro y^*, realizzazione di Y^*, con stessa distribuzione delle variabili osservate $Y = (Y_1, Y_2, \ldots, Y_n)$ osservato. Assumiamo che Y^* sia indipendente da $Y_1, Y_2, \ldots, Y_n$, anche se ovviamente è possibile includere dipendenza.

Un intervallo $(\hat{y}_l^*, \hat{y}_u^*) = \left(\hat{y}_l^*(y_1, \ldots, y_n),\ \hat{y}_u^*(y_1, \ldots, y_n)\right)$ è detto *intervallo di previsione con livello predittivo* $1 - \alpha$ se

$$P_\vartheta \left\{\hat{y}_l^*(Y_1, \ldots, Y_n) \leq Y^* \leq \hat{y}_u^*(Y_1, \ldots, Y_n)\right\} = 1 - \alpha \quad \forall \vartheta \in \Theta.$$

Nota La probabilità calcolata è riferita alla distribuzione congiunta di $(Y_1, Y_2, \ldots, Y_n, Y^*)$.

Quando calcoliamo questo intervallo relativamente a $y_1, y_2, \ldots, y_n$, l'intervallo è una misura della nostra fiducia che Y^* cada in (y_l^*, y_u^*).

Intervalli previsivi con livello esatto $1 - \alpha$

In alcuni rari casi, possiamo trovare $g(Y, Y^*)$ con distribuzione nota e indipendente da ϑ. In tal caso, se g_α è il quantile di livello α di g,

$$(\hat{y}_l^*, \hat{y}_u^*) = \left\{y^* : g_{\frac{\alpha}{2}} \leq g(y^{\text{oss}}, y^*) \leq g_{1-\frac{\alpha}{2}}\right\}.$$

Intervalli di previsione con livello approssimato $1 - \alpha$

Possiamo ottenerlo dai quantili di $F_{Y^*}(y^*, \hat{\vartheta}_n)$, dove $\hat{\vartheta}_n$ è uno stimatore consistente di ϑ, calcolato a partire da $y_1, y_2, \ldots, y_n$.

L'intervallo di previsione è approssimato, in quanto la stima non è mai equivalente al vero valore di ϑ.

Nota

Fino a qui non si è detto nulla su come costruire stimatori, test o quantità pivotali. Una via generale, che sotto opportune condizioni fornisce stimatori consistenti, quantità pivotali per test consistenti e regioni di confidenza, sono i *metodi di verosimiglianza*.

Per qualche particolare modello e problema sarà possibile individuare *procedure ottime*, ad esempio stimatori non distorti a minima varianza oppure test di livello α con potenza massima.

Lezione 13

13.1 Inferenza bayesiana: stima puntuale

Riferimenti Liseo, (2010, §4.1)

Siano assegnati un modello statistico $\mathcal{F} = \{p(y|\vartheta), \vartheta \in \Theta \subseteq \mathbb{R}^p\}$ e una distribuzione a priori $\pi(\vartheta)$. L'obiettivo principale è ottenere la *distribuzione a posteriori*

$$\pi(\vartheta|y^{\text{oss}}) = \frac{\pi(\vartheta)p(y^{\text{oss}}|\vartheta)}{\int_\Theta \pi(\vartheta)p(y^{\text{oss}}|\vartheta)\,d\vartheta}.$$

Tutta l'informazione sul parametro è contenuta nella distribuzione a posteriori, ma è comunque possibile formulare l'analogo di stima puntuale, intervallare e test di ipotesi in ambito bayesiano.

Nella stima puntuale, si tratta di riassumere la distribuzione a posteriori $\pi(\vartheta|y^{\text{oss}})$ attraverso dei singoli valori, come:

- *Media a posteriori*: $\mathbb{E}[\vartheta|y^{\text{oss}}] = \int_\Theta \vartheta\pi(\vartheta|y^{\text{oss}})\,d\vartheta$.
- *Moda a posteriori*: $\text{argmax}_\theta\, \pi(\vartheta|y^{\text{oss}})$.

  Poiché la moda a posteriori non dipende dal denominatore, si può massimizzare la distribuzione a posteriori non normalizzata.
- *Mediana a posteriori*: $\text{Me}(\vartheta|y^{\text{oss}}) \;:\; \int_{-\infty}^{\text{Me}(\vartheta|y^{\text{oss}})} \pi(\vartheta|y^{\text{oss}})\,d\vartheta = 0.5$.

  Nel caso multidimensionale si usa il vettore delle mediane marginali come definizione di mediana della distribuzione congiunta.

13.2 Inferenza bayesiana: stima per regioni

Riferimenti Liseo, (2010, §4.2)

Def. (Regione di credibilità)
Un sottoinsieme $C \subset \Theta$ è detto *regione di credibilità* con livello $1-\alpha$ se

$$P(\vartheta \in C|y^{\text{oss}}) = 1-\alpha.$$

Se la regione è un intervallo $(\vartheta_L, \vartheta_U)$, si dice *bilanciato* (*equi-tailed*) se

$$P(\vartheta < \vartheta_L|y^{\text{oss}}) = P(\vartheta > \vartheta_U|y^{\text{oss}}) = \frac{\alpha}{2},$$

per cui ϑ_L e ϑ_U sono i quantili di livello $\frac{\alpha}{2}$ e $\alpha_{1-\frac{\alpha}{2}}$ di $\pi(\vartheta|y^{\text{oss}})$.

Alternativamente, si può scegliere la regione di *massima credibilità a posteriori* (*HPD*), tramite la risoluzione del problema di minimo

$$\underset{a,b}{\operatorname{argmin}} f(a,b) = b - a$$

$$\text{s.t.} \int_a^b \pi(\vartheta|y^{\text{oss}}) = 1 - \alpha.$$

Nota

È importante osservare che le regioni di credibilità forniscono effettivamente la probabilità che il parametro appartenga alla regione. In questo caso, la variabile casuale è il parametro ϑ e non gli estremi dell'intervallo.

13.3 Inferenza bayesiana: verifica delle ipotesi

Riferimenti Liseo, (2010, §4.3)

Si consideri il problema di verifica di ipotesi con due insiemi $\Theta_0 \subset \Theta,\ \Theta_1 \subset \Theta,\ \Theta_0 \cap \Theta_1 = \emptyset$:

$$\begin{cases} H_0 : \vartheta \in \Theta_0 \\ H_1 : \vartheta \in \Theta_1 \end{cases}$$

La procedura bayesiana rifiuta H_0 se $P(H_0|y^{\text{oss}}) > P(H_0|y^{\text{oss}})$, ovvero se

$$P(\vartheta \in \Theta_0|y^{\text{oss}}) > P(\vartheta \in \Theta_1|y^{\text{oss}}).$$

Osservazioni

1. Nell'inferenza bayesiana, $P(H_0|y^{\text{oss}})$ è effettivamente la probabilità che sia vera l'ipotesi nulla, condizionata ai dati. Questo non vale nel caso dell'inferenza frequentista.
2. Per la verifica di ipotesi bayesiana, bisogna che H_1 sia completamente specificata, mentre ciò non è necessario per calcolare α^{oss}.

Il test bayesiano rifiuta l'ipotesi nulla se

$$\frac{P(H_1|y^{\text{oss}})}{P(H_0|y^{\text{oss}})} > 1.$$

Nel caso più semplice possibile, cioè $\Theta_0 = \{\vartheta_0\}$ e $\Theta_1 = \{\vartheta_1\}$, allora

$$\frac{P(H_1|y^{\text{oss}})}{P(H_0|y^{\text{oss}})} = \frac{\pi(\vartheta_1)}{\pi(\vartheta_0)} \underbrace{\frac{p(y^{\text{oss}}|\vartheta_1)}{p(y^{\text{oss}}|\vartheta_0)}}_{\text{Bayes factor}}.$$

Con ipotesi composite, si ha che

$$\begin{aligned} P(H_j|y^{\text{oss}}) &= P(\vartheta \in \Theta_j)\frac{\int_{\Theta_j} p(y^{\text{oss}}|\vartheta)\pi(\vartheta)\,d\vartheta}{P(\vartheta \in \Theta_j)} \\ &= P(\vartheta \in \Theta_j)p(y^{\text{oss}}|\vartheta \in \Theta_j) \\ &= P(\vartheta \in \Theta_j)p(y^{\text{oss}}|\vartheta \in H_j) \end{aligned}$$

Dunque, il test di ipotesi ora diventa

$$\frac{P(H_1|y^{\text{oss}})}{P(H_0|y^{\text{oss}})} = \frac{P(H_1)}{P(H_0)}\underbrace{\frac{p(y^{\text{oss}}|H_1)}{p(y^{\text{oss}}|H_0)}}_{\text{Bayes factor}}.$$

Scrivendo

$$\frac{\pi(\vartheta)}{P(\vartheta \in \Theta_j)} = \pi(\vartheta|\vartheta \in \Theta_j),$$

si ottiene

$$p(y^{\text{oss}}|H_j) = \int_{\Theta_j} p(y^{\text{oss}}|\vartheta)\pi(\vartheta|\vartheta \in \Theta_j)\,d\vartheta.$$

Poiché nel rapporto eventuali costanti di proporzionalità si semplificano, il fattore di Bayes è anche pari a

$$\frac{p(y^{\text{oss}}|H_1)}{p(y^{\text{oss}}H_0)} = \frac{\int_{\Theta_1} L(\vartheta; y^{\text{oss}})\pi(\vartheta|\vartheta \in \Theta_1)\,d\vartheta}{\int_{\Theta_0} L(\vartheta; y^{\text{oss}})\pi(\vartheta|\vartheta \in \Theta_0)\,d\vartheta}.$$

Il problema nel caso bayesiano sorge quando si ha un'ipotesi alternativa composita

$$\begin{cases} H_0 : \vartheta = \vartheta_0 \\ H_1 : \vartheta \neq \vartheta_1 \end{cases}$$

Ovviamente il denominatore diventa complicato da gestire, per cui tipicamente si sceglie una distribuzione a priori particolare, che assegni una probabilità positiva a un singolo punto. In particolare, si sceglie

$$\pi(\vartheta) = \pi_0 \mathbb{1}_{\vartheta_0}(\vartheta) + (1-\pi_0)\pi_1(\vartheta)\mathbb{1}_{\Theta\setminus\{\vartheta_0\}}(\vartheta),$$

con $\pi_0 = P(H_0)$. Sostituendo, si ottiene

$$\frac{P(H_1|y^{\text{oss}})}{P(H_0|y^{\text{oss}})} = \frac{1-\pi_0}{\pi_0}\frac{\int_{\Theta\setminus\vartheta_0} L(\vartheta; y^{\text{oss}})\pi_1(\vartheta)\,d\vartheta}{L(\vartheta_0; y^{\text{oss}})}.$$

Questa soluzione presenta però aspetti problematici (Liseo, 2010, §4.3.2).

13.4 Inferenza bayesiana: previsione

Riferimenti Liseo, (2010, §4.4)

Sia y^* un'osservazione futura e sia $p(y^*, y|\vartheta)$ la densità congiunta di Y e Y^* dato ϑ.

La *densità predittiva a posteriori* di Y^* è

$$\begin{aligned} p(y^*|y^{\text{oss}}) &= \frac{\int_\Theta p(y^*, y^{\text{oss}}|\vartheta)\pi(\vartheta)\,d\vartheta}{\int_\Theta p(y^{\text{oss}}|\vartheta)\pi(\vartheta)\,d\vartheta}. \\ &= \frac{\int_\Theta p(y^*|y^{\text{oss}}, \vartheta)p(y^{\text{oss}}|\vartheta)\pi(\vartheta)\,d\vartheta}{\int_\Theta p(y^{\text{oss}}|\vartheta)\pi(\vartheta)\,d\vartheta}. \\ &= \int_\Theta p(y^*|y^{\text{oss}}, \vartheta)\pi(\vartheta|y^{\text{oss}})\,d\vartheta. \end{aligned}$$

Assumendo inoltre che, condizionatamente a ϑ, valga l'indipendenza tra Y^* e Y, si può semplificare la densità predittiva come

$$p(y^*|y^{\text{oss}}) = \int_\Theta p(y^*|\vartheta)\pi(\vartheta|y^{\text{oss}})\,d\vartheta.$$

Analogamente, per ottenere una regione previsiva A, assumendo che Y^* sia continua e che si possa scambiare l'ordine di integrazione tra ϑ e Y^*, si ottiene

$$\begin{aligned} P(Y^* \in A|y^{\text{oss}}) &= \int_A \int_\Theta p(y^*|\vartheta)\pi(\vartheta|y^{\text{oss}})\,d\vartheta\,dy^* \\ &= \int_\Theta \left(\int_A p(y^*|\vartheta)\pi(\vartheta|y^{\text{oss}})\,dy^* \right) d\vartheta \\ &= \mathbb{E}\left[P_\vartheta(Y^* \in A)|y^{\text{oss}} \right]. \end{aligned}$$

Lezione 14

14.1 Esempi di inferenza statistica

Esempio (Inferenza frequentista sulla media di $\mathcal{N}(\mu, \sigma_0^2)$)
È un esempio poco realistico, ma importante per valutare le differenze tra procedure frequentiste e bayesiane.

Sia $y = (y_1, y_2, \dots, y_n)$ un c.c.s. da una v.c. $\mathcal{N}(\mu, \sigma_0^2)$, con $\sigma_0^2 = 40^2$ noto.

Inferenza frequentista

Una stima di μ è $\bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i = 20.93$, il cui stimatore corrispondente è $\overline{Y}$ media campionaria, tale che

$$\mathbb{E}_\mu\left[\overline{Y}_n\right] = \mu$$

$$\mathbb{V}_\mu\left[\overline{Y}_n\right] = \frac{\sigma_0^2}{n} = \frac{40^2}{15}.$$

Lo stimatore è non distorto, quindi $\mathrm{EQM}(\overline{Y}_n) = \mathbb{V}_\mu\left[\overline{Y}_n\right]$. Inoltre, è anche consistente per μ sotto l'assunzione di correttezza del modello. Più in generale, è consistente per $\mathbb{E}[Y_i]$ per qualunque distribuzione con media finita.

Verifica di ipotesi per μ

Infine, si dispone di una quantità pivotale esatta sotto μ

$$\frac{\overline{Y} - \mu}{\sqrt{\sigma_0^2/n}} \sim \mathcal{N}(0, 1).$$

Si può usare la quantità pivotale per verificare $H_0 : \mu \leq \mu_0$ contro $H_1 : \mu > \mu_0$, utilizzando la statistica test

$$t = t(Y) = \frac{\overline{Y}_n - \mu_0}{\sqrt{\sigma_0^2/n}} \sim \mathcal{N}\left(\frac{\mu - \mu_0}{\sqrt{\sigma_0^2/n}}, 1\right), \quad \mu \leq \mu_0.$$

Sotto H_1, la statistica T ha distribuzioni *stocasticamente più grandi*, per cui si considerano significativi contro H_0 valori grandi di t.

Il test con livello α fissato rifiuta H_0 per valori $t > z_{1-\alpha}$, con z_α quantile α di $\mathcal{N}(0, 1)$. Infatti,

$$P_\mu\left(t(Y) > z_{1-\alpha} | H_0\right) = P\left(t(Y) - \frac{\mu - \mu_0}{\sqrt{\sigma_0^2/n}} > z_{1-\alpha} - \frac{\mu - \mu_0}{\sqrt{\sigma_0^2/n}} | H_0\right)$$

$$= 1 - \Phi\left(z_{1-\alpha} - \frac{\mu - \mu_0}{\sqrt{\sigma_0^2/n}}\right),$$

che è una funzione monotona crescente in μ ed è pari ad α per $\mu = \mu_0$. In particolare,

$$\begin{aligned}\alpha^{\text{oss}} &= \sup_{\mu \le \mu_0} P_\mu \left(t(Y) \ge t^{\text{oss}}\right) \\ &= \sup_{\mu \le \mu_0} \left\{1 - \Psi\left(t^{\text{oss}} - \frac{\mu - \mu_0}{\sqrt{\sigma_0^2/n}}\right)\right\} \\ &= 1 - \Phi\left(t^{\text{oss}}\right).\end{aligned}$$

Interpretazione Questa è la massima probabilità sotto H_0 di osservare una statistica $t(Y) \ge t^{\text{oss}}$.

Con i dati a disposizione, supponendo di voler verificare $H_0 : \mu \le \mu_0 = 10$ contro $H_1 : \mu > \mu_0$, si ottiene $t^{\text{oss}} = 1.058$. Il valore di soglia $z_{0.95} = 1.64$ indica che non si rifiuta l'ipotesi nulla al livello $\alpha = 0.05$.

La funzione di potenza del test a livello α è la probabilità di rifiutare il test, come funzione del vero valore del parametro

$$\begin{aligned}\pi(\mu) &= P_\mu\big(t(Y) > z_{1-\alpha}\big) \\ &= 1 - \Phi\big(z_{1-\alpha} - \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}}\big)\end{aligned}$$

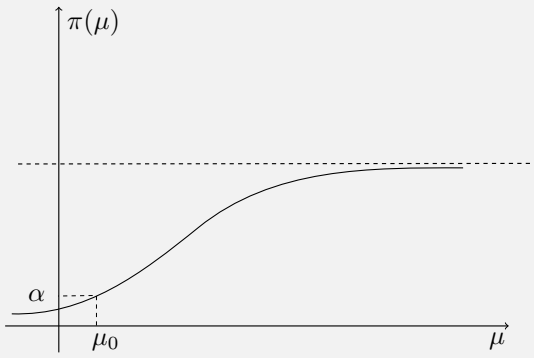


Figura 13: Funzione di potenza al variare del parametro μ.

Osserviamo che la funzione di potenza, al variare di n, ha limite

$$1 - \lim_{n\to\infty} \Phi\left(z_{1-\alpha} - \sqrt{n} \cdot \frac{\mu - \mu_0}{\sigma_0}\right) = 1, \quad \forall \mu > \mu_0$$

per cui si dice che il test è *consistente*.

Determinazione della numerosità campionaria

La funzione di potenza si può usare per determinare la potenza assegnata in $\mu = \mu_1,\ \mu_1 > \mu_0$. In questo modo, si può cercare di limitare anche l'errore di secondo tipo: si desidera

$\pi(\mu_1) \geq \gamma_0$ (ad esempio $\gamma = 0.9$), ovvero

$$1 - \Phi\left(z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}}\right) \geq \gamma_0$$

$$\Phi\left(z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}}\right) \leq 1 - \gamma_0$$

$$z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sqrt{\sigma^2/n}} \leq z_{1-\gamma_0}$$

$$\sqrt{n} \geq \frac{\sqrt{\sigma_0^2}}{\mu_1 - \mu_0}\left(z_{1-\sigma} - z_{1-\gamma_0}\right)$$

$$n \geq \frac{\sigma_0^2}{(\mu_1 - \mu_0)^2}\left(z_{1-\sigma} - z_{1-\gamma_0}\right)^2$$

Con un c.c.s da $\mathcal{N}(\mu, 40^2)$, $\mu_0 = 10$, $\mu_1 = 20$, $\alpha = 0.05$, $\gamma_0 = 0.9$, necessitiamo di almeno

$$n \geq \frac{40^2}{10^2}\left(1.6449 + 1.2816\right)^2 \approx 137.03,$$

dunque $n \geq 138$.

Per casa: svolgere tutti i conti precedenti per il problema di verifica di ipotesi bilaterale $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$.

Intervallo di confidenza per μ

La quantità pivotale $\frac{\overline{Y}_n - \mu}{\sqrt{\sigma_0^2/n}}$ è tale che

$$P_\mu\left(\frac{|\overline{Y}_n - \mu|}{\sqrt{\sigma_0^2/n}} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha,$$

e dunque l'intervallo

$$\bar{y}_n \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_0^2}{n}}$$

ha livello esatto $1 - \alpha$. Infatti,

$$P_\mu\left(\overline{Y}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_0^2}{n}} \leq \mu \leq \overline{Y}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_0^2}{n}}\right) = 1 - \alpha.$$

Lo stesso intervallo si ottiene invertendo la regione di accettazione del test con livello α per $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$, che risulta (esercizio)

$$A_{\mu_0} = \left\{y \in \mathcal{Y} : \left|\frac{\bar{y}_n - \mu_0}{\sqrt{\sigma_0^2/n}}\right| \leq z_{1-\frac{\alpha}{2}}\right\},$$

per cui la regione corrispondente è

$$\{\mu \in \mathbb{R} : y \in A_\mu\} = \left\{ \left| \frac{\bar{y}_n - \mu}{\sqrt{\sigma_0^2/n}} \right| \leq z_{1-\frac{\alpha}{2}} \right\}$$

Intervallo di previsione per Y^*

Sia Y^* un'osservazione futura dallo stesso modello, ovvero $Y^* \sim \mathcal{N}(\mu, \sigma_0^2)$ indipendente da $Y_1, Y_2, \ldots, Y_n$.

Poiché $Y^* - \overline{Y}_n \sim \mathcal{N}\left(0, \sigma_0^2\left(1 + \frac{1}{n}\right)\right)$, questa è una quantità pivotale per la previsione. Un intervallo di previsione per Y^* con livello $1 - \alpha$ è

$$P_\mu \left\{ \left| Y^* - \overline{Y}_n \right| \leq z_{1-\frac{\alpha}{2}} \sqrt{\sigma_0^2 \left(1 + \frac{1}{n}\right)} \right\},$$

dunque è pari a

$$\bar{y}_n \pm z_{1-\frac{\alpha}{2}} \sqrt{\sigma_0^2 \left(1 + \frac{1}{n}\right)}.$$

Esempio (Inferenza bayesiana con distribuzione a priori coniugata $\mathcal{N}(\nu_0, \tau_0^2)$)
Scelta degli iperparametri per l'esempio delle piante di mais: una possibilità è usare

Tabella 3: caption

$\nu_0 = 0$	Non si ha informazione sulla media delle differenze.
$\tau_0^2 \approx 33^2$	Varianza tale che con probabilità quasi 1 la media delle differenze assolute sia ≤ 100: $$\frac{100}{\tau_0} = 3$$

La distribuzione a priori è allora

$$\pi(\mu) \propto \exp\left\{ -\frac{1}{2\tau_0^2}(\mu - \nu_0)^2 \right\}.$$

Dai conti che abbiamo fatto in precedenza, la distribuzione a posteriori è

$$\mu | y^{\text{oss}} \sim \mathcal{N}\left(\frac{\frac{\nu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma_0^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}}, \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}} \right),$$

per cui la media a posteriori è

$$\mathbb{E}\left[\mu | y^{\text{oss}}\right] = \frac{n\bar{y} + \frac{\sigma_0^2}{\tau_0^2}\nu_0}{n + \frac{\sigma_0^2}{\tau_0^2}}.$$

Siccome la media a priori è 0, ci sarà uno shrinkage verso valori piccoli di differenza.

Stima puntuale

La stima puntuale è immediata, in quanto media, mediana e moda a posteriori coincidono. In particolare, con i dati osservati,

$$\mathbb{E}\left[\mu|y^{\text{oss}}\right] = 19.07.$$

Se invece $\tau_0^2 \longrightarrow \infty$, dalla formula del valore atteso a posteriori si ottiene la media campionaria.

La valutazione della precisione basata sulla varianza a posteriori è

$$\mathbb{V}\left[\mu|y^{\text{oss}}\right] = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1} \approx 97.15.$$

Se si assumesse invece $\tau_0^2 = 100^2$, si otterrebbe

$$\mathbb{V}\left[\mu|y^{\text{oss}}\right] = 105.2,$$

prossima alla varianza frequentista $\sigma_0^2/n = 106.67$.

Regioni di credibilità

In questo caso si hanno intervalli bilanciati e HPD che coincidono, in quanto la distribuzione a posteriori è una normale (e quindi simmetrica):

$$\mu_L = \mathbb{E}\left[\mu|y^{\text{oss}}\right] - z_{1-\frac{\alpha}{2}}\sqrt{\mathbb{V}\left[\mu|y^{\text{oss}}\right]};$$

$$\mu_U = \mathbb{E}\left[\mu|y^{\text{oss}}\right] + z_{1-\frac{\alpha}{2}}\sqrt{\mathbb{V}\left[\mu|y^{\text{oss}}\right]}.$$

Nel nostro caso, l'intervallo di credibilità è $(\mu_L, \mu_U) = (-0.25, 38.38)$. Aumentando l'incertezza a priori, $\tau_0^2 = 100^2$, si ottiene l'intervallo $(0.58, 40.85)$.

Verifica di ipotesi

Si consideri $H_0 : \mu \leq \mu_0$ contro $H_1 : \mu > \mu_0$, il test bayesiano rifiuta H_0 se

$$\frac{P(\mu > \mu_0|y^{\text{oss}})}{P(\mu \leq \mu_0|y^{\text{oss}})} = \frac{P(\mu > \mu_0|y^{\text{oss}})}{1 - P(\mu > \mu_0|y^{\text{oss}})} > 1.$$

Per questi dati,

$$P(\mu \leq \mu_0|y^{\text{oss}}) = \Phi\left(\frac{\mu_0 - \mathbb{E}\left[\mu|y^{\text{oss}}\right]}{\sqrt{\mathbb{V}\left[\mu|y^{\text{oss}}\right]}}\right) = \Phi(-0.92) \approx 0.18.$$

Con $\tau_0^2 = 100^2$, si otterrebbe $P(H_0|y^{\text{oss}}) \approx 0.15$, prossimo a α^{oss} frequentista.

L'ipotesi alternativa ha $\frac{1-0.18}{0.18} \approx 4.6$ volte più probabilità di essere vera.

Il fattore di Bayes è il rapporto delle verosimiglianze sotto le diverse ipotesi

$$\frac{P(H_1|y^{\text{oss}})}{P(H_0|y^{\text{oss}})} = \frac{P(H_1)}{P(H_0)} \cdot \frac{p(y^{\text{oss}}|H_1)}{p(y^{\text{oss}}|H_0)}.$$

Poiché a priori si aveva un prior odds di 0.61 a favore dell'ipotesi nulla, il fattore di bayes è pari a $\frac{4.6}{0.61} = 7.4$ e porta a rifiutare con forza H_0.

Previsione di un'osservazione futura

Vogliamo calcolare la densità predittiva a posteriori

$$p(y^*|y^{\text{oss}}) = \int_{\mathbb{R}} p_{Y^*}(y^*|\mu)\pi(\mu|y^{\text{oss}})\, d\mu.$$

A meno di costanti moltiplicative, l'integrale è della forma

$$\int_{\mathbb{R}} e^{-\frac{1}{2\sigma_0^2}(\mu-y^*)^2 - \frac{1}{2\,\mathbb{V}[\mu|y^{\text{oss}}]}(\mu-\mathbb{E}[\mu|y^{\text{oss}}])^2}\, d\mu.$$

Usando la formula $a(x-b)^2 + c(x-d)^2 = c + f(x-g)^2$ già ottenuta in precedenza, si ottiene (esercizio)

$$TODO$$

Si ottiene quindi

$$p(y^*|y^{\text{oss}}) \propto \exp\left\{-\frac{1}{2}\frac{(y^* - \mathbb{E}\left[\mu|y^{\text{oss}}\right])^2}{\sigma_0^2 + \mathbb{V}\left[\mu|y^{\text{oss}}\right]}\right\},$$

cioè

$$Y^*|y^{\text{oss}} \sim \mathcal{N}\left(\mathbb{E}\left[\mu|y^{\text{oss}}\right], \sigma_0^2 + \mathbb{V}\left[\mu|y^{\text{oss}}\right]\right).$$

Un intervallo di previsione con probabilità $1-\alpha$ è ovviamente, per simmetria,

$$\mathbb{E}\left[\mu|y^{\text{oss}}\right] \pm z_{1-\frac{\alpha}{2}}\sqrt{\sigma_0^2 + \mathbb{V}\left[\mu|y^{\text{oss}}\right]}.$$

Per τ_0^2, gli intervalli di previsione frequentista e bayesiano tendono a coincidere.

Lezione 15

15.1 Esempi di inferenza statistica (cont.)

Esempio (Inferenza sul tasso di guasto di Exp(ϑ))
Consideriamo $y_1, y_2, \ldots, y_n$ realizzazioni di $Y \sim \text{Exp}(\vartheta)$. Per l'inferenza bayesiana si considera una priori Gamma(α_0, β_0), che fornisce la distribuzione a posteriori Gamma$(\alpha_0 + n, \beta + n\bar{y}_n)$.

Come iperparametri avevamo scelto $\alpha_0 = 1,\ \beta = 100$, alternativamente possiamo usare una distribuzione a posteriori non informativa.

Tabella 4: caption

Stima frequentista	Stima bayesiana
$\hat{\vartheta} = \frac{1}{\bar{y}_n} = 0.00594$	$\mathbb{E}_\vartheta\left[\vartheta \mid y^{\text{oss}}\right] = \frac{\alpha_0+n}{\beta_0+n\bar{y}_n} = 0.00617$
$\text{MSE}(\hat{\vartheta}) = \frac{n^2+n-2}{(n-1)^2(n-2)}\vartheta^2 = 0.00243^2$	$\mathbb{V}\left[\vartheta \mid y^{\text{oss}}\right] = \frac{\alpha_0+n}{\beta_0+n\bar{y}_n} = 0.00186^2$

Se invece si usasse la distribuzione impropria non informativa, si otterrebbe lo stesso valore dello stimatore a posteriori, ma con varianza a posteriori diversa (MSE tiene conto anche della distorsione).

Osservazione

1. La distribuzione gamma non è simmetrica, in particolare l'indice di asimmetria è $2/\sqrt{\alpha}$. Si potrebbe allora usare un altro indice di sintesi a posteriori (esercizio):

$$\text{moda}(\pi(\vartheta \mid y^{\text{oss}})) = \frac{\alpha_0 + n - 1}{\beta_0 + n\bar{y}_n} \approx 0.0056$$

$$\text{med}(\pi(\vartheta) \mid y^{\text{oss}}) \approx 0.0060$$

2. Parametrizzando con la media $\psi = \vartheta^{-1}$, si dovrebbe usare la distribuzione a priori IG(α_0, β_0), con distribuzione a posteriori IG$(\alpha_0 + n, \beta_0 + n\bar{y}_n)$. Ricordiamo che, nel caso di una v.c. X gamma inversa, con $X = V^{-1}$,

$$\mathbb{E}\left[X^r\right] = \mathbb{E}\left[V^{-r}\right] = \frac{\Gamma(\alpha - r)}{\Gamma(\alpha)}\beta^r, \quad r < \alpha.$$

   In tal caso, la stima frequentista è $\hat{\psi} = \bar{y}_n$, $\text{se}(\hat{\psi}) = \frac{\bar{y}_n}{\sqrt{n}}$. Per la distribuzione a posteriori, invece,

$$\mathbb{E}\left[\psi \mid y^{\text{oss}}\right] = \frac{\beta_0 + n\bar{y}_n}{\alpha_0 + n - 1}$$

$$\mathbb{V}\left[\psi \mid y^{\text{oss}}\right] = \frac{(\beta_0 + n\bar{y}_n)^2}{(\alpha_0 + n - 1)^2(\alpha_0 + n - 2)}$$

Stima intervallare e verifica di ipotesi frequentiste

Anche in questo caso si dispone di una quantità pivotale:

$$\vartheta \overline{Y}_n \sim \text{Gamma}(n, n)$$

- Test: considero $H_0 : \vartheta \leq \vartheta_0$ contro $H_1 : \vartheta > \vartheta_0$, posso utilizzare la statistica test $t = t(Y) = \vartheta_0 \overline{Y}_n = \frac{\vartheta_0}{\vartheta} \vartheta \overline{Y}_n$ con distribuzioni nulle $\text{Gamma}(n, n\frac{\vartheta}{\vartheta_0})$, $\vartheta \leq \vartheta_0$.

  Sotto H_1, $t(Y)$ ha distribuzioni stocasticamente più piccole rispetto all'ipotesi nulla

  $$P_\vartheta\big(t(Y) \leq t\big) = P\big(\frac{\vartheta_0}{\vartheta} X \leq t\big) = P\big(X \leq \frac{\vartheta}{\vartheta_0} t\big),$$

  monotona crescente in ϑ. Perciò, $F_T(t; \vartheta) \geq F_T(t; \vartheta')$ per $\vartheta > \vartheta'$ e sono quindi significativi contro H_0 valori piccoli di t.

  Il test con livello di significatività α rifiuta se $t(y) < x_\alpha$, con x_α quantile di livello α di $\text{Gamma}(n, n)$. Ricordiamo anche che $2n \cdot \text{Gamma}(n, n) = \text{Gamma}(\frac{2n}{2}, \frac{1}{2}) = \chi_{2n}$, per cui possiamo prendere

  $$x_\alpha = \chi_{2n;\,\alpha}/2n.$$

  Il livello di significatività osservato è

  $$\begin{aligned} \alpha^{\text{oss}} &= \sup_{\vartheta \leq \vartheta_0} F_\vartheta(t(Y) \leq t^{\text{oss}}) \\ &= F_X(t^{\text{oss}}), \end{aligned}$$

  con $X \sim \text{Gamma}(n, n)$.

  Si può trovare una numerosità campionaria che garantisca potenza in $\vartheta_1 = 0.002$ con livello α maggiore o uguale a un dato valore $\gamma \geq 0.8$. Allora

  $$\begin{aligned} \pi(\vartheta_1) &= P_{\vartheta_1}\left\{\vartheta_0 \overline{Y}_n \leq x_\alpha\right\} \\ &= P_{\vartheta_1}\left\{\vartheta_1 \overline{Y}_n \leq \frac{\vartheta_1}{\vartheta_0} x_\alpha\right\} \\ &= F_X\big(\frac{\vartheta_1}{\vartheta_0} x_\alpha\big) \\ &\geq 0.8 \iff n \geq 96. \end{aligned}$$

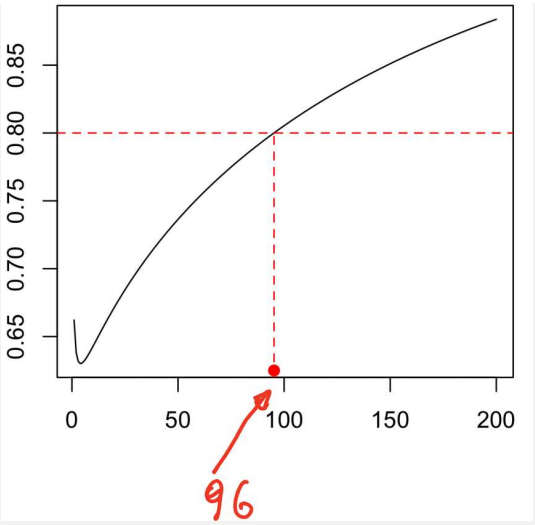


Figura 14: Funzione di potenza al variare di n.

- Intervallo di confidenza: sia x_α il quantile di Gamma(n, n), allora si può prendere un intervallo di confidenza esatto $1-\alpha$ usando

$$P_\vartheta \left\{ x_{\frac{\alpha}{2}} \leq \vartheta \overline{Y}_n \leq x_{1-\frac{\alpha}{2}} \right\} = 1-\alpha,$$

da cui si ottiene

$$IC(\vartheta) = \left(\frac{x_{\frac{\alpha}{2}}}{\bar{y}_n}, \frac{x_{1-\frac{\alpha}{2}}}{\bar{y}_n} \right).$$

Esempio (Inferenza basata su q.tà pivotale approssimata)
Per n sufficientemente grande, nell'esempio precedente, sappiamo che

$$\overline{Y}_n \dot{\sim} \mathcal{N}\left(\frac{1}{\vartheta}, \frac{1}{n\vartheta^2} \right),$$

per cui

$$\sqrt{n\vartheta^2} \left(\overline{Y} - \frac{1}{\vartheta} \right) \dot{\sim} \mathcal{N}(0,1)$$

$$\sqrt{n} \left(\vartheta \overline{Y} - 1 \right) \dot{\sim} \mathcal{N}(0,1)$$

ed è approssimativamente una quantità pivotale. In particolare, per verificare il test di ipotesi $H_0 : \vartheta \leq \vartheta_0$ contro $H_1 : \vartheta > \vartheta_0$, si rifiuta per valori piccoli della statistica test. In particolare, si rifiuta H_0 se

$$\sqrt{n}(\vartheta_0 \bar{y}_n - 1) < z_\alpha$$

e il livello di significatività osservato è

$$\alpha^{\text{oss}} \doteq \Phi\left(\sqrt{n}(\vartheta_0 \bar{y}_n - 1) \right).$$

L'intervallo di confidenza con livello approssimato $1-\alpha$ è dunque

$$\left\{\vartheta > 0 : \left|\sqrt{n}(\vartheta \bar{y}_n - 1)\right| \leq z_{1-\frac{\alpha}{2}}\right\},$$

che risulta

$$\frac{1}{\bar{y}_n}\left(1 \pm \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right).$$

Esempio (Previsione di un'osservazione futura)

Sempre nel caso precedente, utilizziamo entrambi gli approcci per prevedere una osservazione futura.

Approccio frequentista

Considerando y^* realizzazione di $Y^* \sim \text{Exp}(\vartheta)$, indipendentemente da $Y_1, Y_2, \ldots, Y_n$. Si può trovare una funzione di Y e Y^* con distribuzione esatta:

$$\frac{Y^*}{\overline{Y}_n} \sim F_{2,2n}.$$

Infatti,

$$\frac{Y^*}{\overline{Y}_n} \sim \frac{\frac{1}{\vartheta}\text{Exp}(1)}{\frac{1}{\vartheta}\text{Gamma}(n,1)/n} \sim \frac{\text{Gamma}(1,1)}{\text{Gamma}(n,1)/n},$$

con numeratore e denominatore indipendenti. Raccogliendo da entrambi i membri un 2, si ottiene

$$\frac{Y^*}{\overline{Y}_n} \sim \frac{2 \cdot \text{Gamma}(1, \frac{1}{2})}{2 \cdot \text{Gamma}(n,1)/n} \sim \frac{\chi_2/2}{\chi_{2n}/2n} \sim F_{2,2n}.$$

Si ha allora

$$P\left(F_{2,2n;\ \frac{\alpha}{2}} \leq \frac{Y^*}{\overline{Y}_n} \leq F_{2,2n;\ 1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

dunque un intervallo di previsione per i dati futuri è

$$\left(\bar{y}_n \cdot F_{2,2n;\ \frac{\alpha}{2}},\ \bar{y}_n \cdot F_{2,2n;\ 1-\frac{\alpha}{2}}\right)$$

Approccio bayesiano

Dalla distribuzione a posteriori

$$\vartheta | y^{\text{oss}} \sim \text{Gamma}(\alpha_0 + n, \beta_0 + n\bar{y}_n),$$

per ottenere un intervallo bilanciato basta calcolare i quantili di livello $\frac{\alpha}{2}$ e $1-\frac{\alpha}{2}$. Scegliendo una distribuzione non informativa a priori, si ottiene un intervallo di credibilità praticamente coincidente con l'intervallo di confidenza.

Poiché la gamma non è simmetrica, si può calcolare un intervallo HPD $(\vartheta_L, \vartheta_U)$ per via numerica tramite la libreria `TeachingDemos`.

Il test si basa sul rapporto

$$\frac{P(H_1|y^{\text{oss}})}{P(H_0|y^{\text{oss}})}$$

e valori maggiori di 1 indicano di rifiutare H_0. Con i dati,

$$P(H_0|y^{\text{oss}}) = \int_0^{0.001} \pi(\vartheta|y^{\text{oss}})\, d\vartheta \approx 2.86 \times 10^{-6},$$

nettamente a favore di H_1. A priori, gli odds erano 9.5 a favore di H_1, mentre il fattore di Bayes è dell'ordine di 3.67×10^5.

Per prevedere un'osservazione futura, si ha che

$$\begin{aligned}
p(y^*|y^{\text{oss}}) &= \int_0^\infty \vartheta e^{-\vartheta y^*} \pi(\vartheta|y^{\text{oss}})\, d\vartheta \\
&= \frac{(\beta_0 + n\bar{y}_n)^{\alpha_0+n}}{\Gamma(\alpha_0+n)} \int_0^\infty \underbrace{\vartheta^{\alpha_0+n} e^{-\vartheta(y^*+\beta_0+n\bar{y}_n)}\, d\vartheta}_{\text{Kernel Gamma}(\alpha_0+n+1, y^*+\beta_0+n\bar{y}_n)} \\
&= \frac{(\beta_0 + n\bar{y}_n)^{\alpha_0+n}}{\Gamma(\alpha_0+n)} \cdot \frac{\Gamma(\alpha_0+n+1)}{(y^*+\beta_0+n\bar{y}_n)^{\alpha_0+n+1}} \\
&= \frac{(\alpha_0+n)\cdot(\beta_0+n\bar{y}_n)^{\alpha_0+n}}{(y^*+\beta_0+n\bar{y}_n)^{\alpha_0+n+1}},
\end{aligned}$$

che è una distribuzione di Pareto traslata.

Lezione 16

16.1 Esempi di inferenza statistica (cont. bis)

Esempio (Inferenza su ϑ nel modello $\text{Bin}(n, \vartheta)$)
Esempio già considerato prima, valutiamo la verifica di ipotesi e la stima intervallare.

Inferenza frequentista

Test e intervalli di confidenza sono basati sulle quantità pivotali approssimate

$$q_1(\hat{\vartheta}, \vartheta) = \frac{\hat{\vartheta} - \vartheta}{\sqrt{\frac{\hat{\vartheta}(1-\hat{\vartheta})}{n}}}, \quad q_2(\hat{\vartheta}, \vartheta) = \frac{\hat{\vartheta} - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}},$$

con $\hat{\vartheta} = \overline{Y}$, aventi distribuzione nulla approssimata $\mathcal{N}(0,1)$. Se $n\vartheta < 5$ o $n(1-\vartheta) < 5$, è preferibile basarsi sulla distribuzione esatta $n\hat{\vartheta} \sim \text{Bin}(n, \vartheta)$.

Un test per verificare $H_0 : \vartheta \leq \vartheta_0$ contro $H_1 : \vartheta > \vartheta_0$ rifiuta H_0 se

$$q_1^{\text{oss}} > z_{1-\alpha}$$

con livello di significatività osservato $\alpha^{\text{oss}} \doteq 1 - \Phi(q_1^{\text{oss}})$. In modo analogo si può utilizzare $q_2(\hat{\vartheta}, \vartheta_0)$, per la quale ci si attende che l'approssimazione normale sia più accurata. Infatti, per ogni n

$$\mathbb{E}_{\vartheta_0}\left[q_2(\hat{\vartheta}, \vartheta_0)\right] = 0,$$

$$\mathbb{V}_{\vartheta_0}\left[q_2(\hat{\vartheta}, \vartheta_0)\right] = 1.$$

Nel caso non si possa usare l'approssimazione normale, si può utilizzare invece il test esatto, che rifiuta se

$$n\hat{\vartheta} > c_{1-\alpha},$$

con $c_{1-\alpha}$ quantile di livello α della distribuzione $\text{Bin}(n, \vartheta_0)$. Poiché

$$c_{1-\alpha} = \inf\left\{y \in S_Y : F_Y(y; \vartheta_0) \geq 1 - \alpha\right\},$$

nel caso discreto non si ottiene sempre il valore esatto di quantile. In particolare, il test ha livello effettivo minore o uguale al livello nominale α visto che

$$P_{\vartheta_0}\left\{Y > c_{1-\alpha}\right\} = 1 - F_Y(c_{1-\alpha}) \leq \alpha.$$

Il problema non si pone se si calcola α^{oss}, perché si possono solo assumere i valori interi di supporto. In particolare, siccome $\text{Bin}(n, \vartheta)$ è stocasticamente ordinata rispetto a ϑ, il livello di significatività osservato è

$$\alpha^{\text{oss}} = \sup_{\vartheta \leq \vartheta_0} P_\vartheta\left\{Y \geq y^{\text{oss}}\right\} = P_{\vartheta_0}\left\{Y \geq y^{\text{oss}}\right\}.$$

È spesso consigliato l'uso del *mid-p-value*

$$\begin{aligned}\alpha_{mid}^{\text{oss}} &= \sup_{\vartheta\in\Theta_0}\left\{\frac{1}{2}P_\vartheta(T > t^{\text{oss}}) + \frac{1}{2}P_\vartheta(T \geq t^{\text{oss}})\right\}\\ &= \sup_{\vartheta\in\Theta_0}\left\{P_\vartheta(T > t^{\text{oss}}) + \frac{1}{2}P_\vartheta(T = t^{\text{oss}})\right\}.\end{aligned}$$

Gli intervalli di confidenza con livello approssimato $1-\alpha$ basati su $q_1(\hat\vartheta,\vartheta)$ sono

$$\left\{\vartheta\in(0,1) : \frac{|\hat\vartheta-\vartheta|}{\sqrt{\hat\vartheta(1-\hat\vartheta)/n}} < z_{1-\alpha}\right\},$$

cioè gli usuali intervalli $\hat\vartheta \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat\vartheta(1-\hat\vartheta)}{n}}$. Basandoci invece su $q_2(\hat\vartheta,\vartheta)$, si ottiene un'equazione di secondo grado da risolvere in ϑ. L'intervallo che si ottiene non è degenere anche se $\hat\vartheta = 0$ oppure $\hat\vartheta = 1$.

Usando la distribuzione esatta e la relazione tra test e intervalli di confidenza, si possono ottenere intervalli con livello $\geq 1-\alpha$, noti come intervalli di Clopper-Pearson:

$$\hat\Theta(y^{\text{oss}}) = \left\{\vartheta\in(0,1) : \underbrace{P_\vartheta(Y \leq y^{\text{oss}}) \geq \frac{\alpha}{2}}_{\text{Unilat. sx}}, \underbrace{P_\vartheta(Y \geq y^{\text{oss}}) \geq \frac{\alpha}{2}}_{\text{Unilat. dx}}\right\}.$$

Se $y = 0$, si ha che $P_\vartheta(Y \geq y) = 1$, mentre $P_\vartheta(Y \leq y) = P(Y = 0) = (1-\vartheta)^n$. Allora,

$$(1-\vartheta)^n = \frac{\alpha}{2} \iff \vartheta = 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{n}}$$

e l'intervallo di Clopper-Pearson è $\left(0, 1-(\frac{\alpha}{2})^{1/n}\right)$. Usando invece il *mid-p-value*,

$$\frac{1}{2}P_\vartheta(Y \leq 0) = \frac{\alpha}{2} \iff (1-\vartheta)^n = \alpha$$

e si ottiene $\left(0, 1-(\alpha)^{1/n}\right)$

Inferenza bayesiana

Consideriamo la distribuzione a priori $\text{Beta}(\alpha_0,\beta_0)$, da cui la distribuzione a posteriori $\text{Beta}(\alpha_0+y, \beta_0+n-y)$. Come intervalli di credibilità, si possono utilizzare quello bilanciato $(q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}})$ oppure quello di massima densità a posteriori risolvendo

$$\begin{cases}\pi(\vartheta_l|y^{\text{oss}}) = \pi(\vartheta_u|y^{\text{oss}})\\ \int_{\vartheta_L}^{\vartheta_U}\pi(\vartheta|y^{\text{oss}})\,d\vartheta = 1-\alpha.\end{cases}$$

Il test bayesiano vuole verificare $H_0 : \vartheta \leq 0.2$ contro $H_1 : \vartheta > 0.2$, per cui si possono calcolare

$$P(H_0|y^{\text{oss}}) = 0.065.$$

Il test è allora fortemente in favore di H_1.

Lezione 17

Riferimenti Casella e Berger, (2001, §6.1-6.2)
Pace e Salvan (2001, §5.1-5.3)
Azzalini (2001, §2.3-2.4)

17.1 Statistiche sufficienti

Dato il modello $\mathcal{F}$, con eventualmente distribuzione a priori $\pi(\vartheta)$, vediamo come si possono sintetizzare i dati senza perdere l'informazione in essi contenuta. I dati sono affetti da variabilità accidentale, per cui necessitiamo di *filtri* o *statistiche*, funzioni misurabili che riducono la componente aleatoria.

$$\text{Due estremi: } \begin{cases} a. \text{ Statistiche che estraggono solo aspetti aleatori da } y \\ b. \text{ Statistiche che estraggono tutta l'informazione su } \vartheta_0 \text{ da } y \end{cases}$$

Statistiche e partizioni indotte nello spazio campionario

Tipicamente una *statistica* è una funzione misurabile non biunivoca, $t = t(y)$, da cui si possono evidenziare gli elementi che mandano nello stesso valore di t

$$\mathcal{Y}_t = \{y \in \mathcal{Y} : t(y) = t\}.$$

Ad esempio, se $\mathcal{Y} = \mathbb{R}^2$ e $y = (y_1, y_2)$ e $t(y) = y_1^2 + y_2^2$, l'insieme $\mathcal{Y}_t$ è l'insieme dei punti di $\mathbb{R}^2$ sulla circonferenza con centro in $(0, 0)$ e raggio $\sqrt{t}$

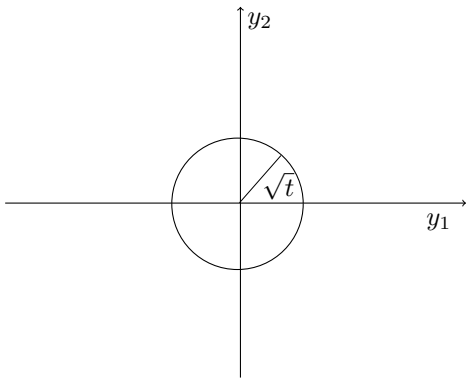


Figura 15: Partizione indotta dalla statistica $t(y) = y_1^2 + y_2^2$.

Al variare di $t \in \mathcal{T}$, con $\mathcal{T}$ spazio campionario di t, definiamo una *partizione di* $\mathcal{Y}$:

- $\bigcup_{t \in \mathcal{T}} \mathcal{Y}_t = \mathcal{Y}$
- $\mathcal{Y}_t \cap \mathcal{Y}_{t'} = \emptyset \quad \text{se } t \neq t'$

Proprietà

- Una funzione *biunivoca* di t induce in $\mathcal{Y}$ la stessa partizione di t, per cui non cambia la partizione dello spazio campionario.
- Una funzione *non biunivoca* di t induce in $\mathcal{Y}$ una partizione *meno fine* di t, cioè in cui ci sono "meno fette". Sia $v = v(t)$ una funzione non biunivoca di t, allora

$$\mathcal{Y}_v = \bigcup_{t:v(t)=v} \mathcal{Y}_t,$$

per cui $\mathcal{Y}_t \subseteq \mathcal{Y}_v$ per ogni t e valore corrispondente di v.

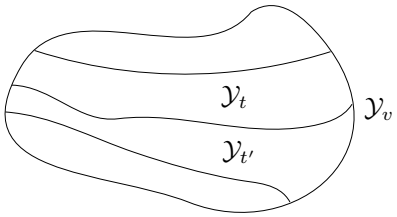


Figura 16: La partizione indotta da $v = v(t)$ è meno fine della partizione indotta da t, se v non è biunivoca.

Sia $t(y)$ una statistica, allora a $T = t(Y)$ è associato il modello indotto

$$\mathcal{F}_T = \{p_T(t;\vartheta),\ \vartheta \in \Theta\},$$

tipicamente indicizzato anch'esso da ϑ.

Esempio (Modello indotto da un'esponenziale)
Se $y = (y_1, y_2, \ldots, y_n)$ è un c.c.s. da $\text{Exp}(\vartheta)$ e $t(y) = \sum_{i=1}^n y_i$, il modello indotto associato a T è la famiglia $\text{Gamma}(n, \vartheta)$.

Per Y tale che $t(y) = t$, sotto deboli condizioni vale la fattorizzazione

$$p_Y(y;\vartheta) = p_T(t;\vartheta) p_{Y|T=t}(y;t,\vartheta),$$

che permette di pensare a y come generato da un esperimento in 2 stadi:

1. Viene generata t da $p_T(t;\vartheta) \implies$ si sceglie la partizione a cui y appartiene.
2. Viene generato y da $p_{Y|T=t}(y;t,\vartheta) \implies$ si definisce y entro $\mathcal{Y}_t$

Ci sono allora due casi estremi:

a) Il modello indotto contiene un solo elemento, cioè $p_T(t;\vartheta) = p_T(t)$. $t(y)$ allora non contiene informazione su ϑ_0 e T si dice *costante in distribuzione*.

b) $p_{Y|T=t}(y;t,\vartheta) = p_{Y|T=t}(y;t)$ indipendentemente da ϑ. Allora, si dice che T è *sufficiente*, perché basta solamente conoscere T per avere tutta l'informazione.

Def. (Sufficienza)
$T = t(Y)$ è detta *sufficiente* per l'inferenza su ϑ, nel modello $\mathcal{F}$, se la distribuzione di Y condizionata a $T = t$ non dipende da ϑ.

Osservazioni

- $t(y)$ opera una riduzione dei dati e del modello, da $\mathcal{F}$ a $\mathcal{F}_T$.
- La definizione non richiede che $\mathcal{F}$ sia un modello parametrico.

Esempio (C.c.s. da una distribuzione continua)
Sia $y = (y_1, (y_2, \dots, (y_n)$ realizzazione di Y_i con densità comune $p_0(y_i)$ e densità ignota. Il modello statistico è dunque non parametrico:

$$\mathcal{F} = \left\{ p_Y(y_i) = \prod_{i=1}^{n} p_0(y_i),\ p_0(\cdot) \text{ densità su } \mathbb{R} \right\}.$$

Sia $t = t(y) = \left(y_{(1)}, \dots, y_{(n)}\right)$ la *statistica ordinata* tale che

$$y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}.$$

Vale che $p_T(t; p_0) = n! \prod_{i=1}^{n} p_0(t_i)$ (Azzalini, 2001, A.7), con $t_1 \leq t_2 \leq \dots \leq t_n$. La densità condizionata sicuramente non è costante, in quanto entra $p_0(\cdot)$ nell'espressione della densità.

La densità condizionata di Y a $T = t$ è allora

$$p_{Y|T=t}(y; t, p_0) = \frac{p_Y(y; p_0)}{p_T(t; p_0)} = \frac{1}{n!},$$

quindi $p_{Y|T=t}$ è costante in distribuzione. Nota t, tutte le permutazioni sono equiprobabili e la continuità assicura che con probabilità 1 si abbiano componenti distinte.

Esempio (C.c.s. da una Bin$(1, \vartheta)$)
Se $y = (y_1, y_2, \dots, y_n)$ è un c.c.s. da $\text{Bin}(1, \vartheta)$, allora $t = \sum_{i=1}^{n} y_i$ è realizzazione di $T \sim \text{Bin}(n, \vartheta)$ e, per y tale che $\sum_{i=1}^{n} y_i = t$

$$\begin{aligned} p_{Y|T=t}(y; t, \vartheta) &= \frac{\prod_{i=1}^{n} \vartheta^{y_i}(1-\vartheta)^{(1-y_i)}}{\binom{n}{t}\vartheta^t(1-\vartheta)^t} \\ &= \frac{\vartheta^t(1-\vartheta)^t}{\binom{n}{t}\vartheta^t(1-\vartheta)^{n-t}} \\ &= \frac{1}{\binom{n}{t}} \end{aligned}$$

Dunque, noto il numero totale t di successi, non sono necessarie altre informazioni per l'inferenza sul modello $\mathcal{F}$.

Nota La riduzione da y a t non comporta perdita di informazioni solo se le Y_i sono indipendenti, altrimenti la collocazione dei t successi nelle n prove può risultare informativa.

Esempio (Modello di regressione Poisson)
Consideriamo $Y_i \sim \text{Pois}(\vartheta x_i)$ indipendenti, x_i costanti note positive e $\vartheta > 0$ ignoto. Allora, $T = \sum_{i=1}^n Y_i \sim \text{Pois}(\vartheta \sum_{i=1}^n x_i)$ e

$$\begin{aligned} p_{Y|T=t}(y; t, \vartheta) &= \frac{\prod_{i=1}^n e^{\vartheta x_i}(\vartheta x_i)^{y_i}/y_i!}{e^{\vartheta \sum_{i=1}^n x_i}(\vartheta \sum_{i=1}^n x_i)^t/t!} \\ &= \frac{\cancel{e^{\vartheta \sum_{i=1}^n x_i}}\ \cancel{\vartheta^{\sum_{i=1}^n y_i}} \prod_{i=1}^n x_i^{y_i}/y_i!}{\cancel{e^{\vartheta \sum_{i=1}^n x_i}}\ \cancel{\vartheta^t} \left(\sum_{i=1}^n x_i\right)^t/t!} \\ &= \frac{t!}{y_1! \dots y_n!} \prod_{i=1}^n \left(\frac{x_i}{\sum_{i=1}^n x_i}\right)^{y_i}, \end{aligned}$$

dunque $Y|T = t \sim \text{Mn}_n\left(t, \left(\frac{x_1}{\sum_{i=1}^n x_i}, \dots, \frac{x_n}{\sum_{i=1}^n x_i}\right)\right)$ e la statistica T è sufficiente.

17.2 Statistiche sufficienti e inferenza bayesiana

La sufficienza è un concetto pervasivo nell'inferenza statistica, presente anche nell'inferenza bayesiana e negli approcci decisionali.

Se $t(y)$ è sufficiente per $\mathcal{F}$, nel modello bayesiano specificato con $\pi(\vartheta)$, si ha

$$p(y|\vartheta) = p_{T|\vartheta}(t|\vartheta) p_{Y|T=t}(y; t),$$

per cui la distribuzione di $Y|T, \vartheta$ coincide con quella di $Y|T$. Nella distribuzione a posteriori,

$$\begin{aligned} \pi(\vartheta|y) &= \frac{\pi(\vartheta) p_{T|\vartheta}(t|\vartheta)\ \cancel{p_{Y|T=t}(y; t)}}{\int_\Omega \pi(\vartheta) p_{T|\vartheta}(t|\vartheta)\ \cancel{p_{Y|T=t}(y; t)}\, d\vartheta} \\ &= \frac{\pi(\vartheta) p_{t|\vartheta}(t|\vartheta)}{\int_\Theta \pi(\vartheta) p_{t|\vartheta}(t|\vartheta)\, d\vartheta} \\ &= \pi(\vartheta|t) \end{aligned}$$

17.3 Criterio di fattorizzazione di Neyman-Fisher

Per individuare una statistica sufficiente tramite la definizione è necessario

i. Avere un'idea di quale sia la statistica sufficiente $t(y)$.

ii. Verificare la sufficienza calcolando $p_T(t; \vartheta)$ e $p_{Y|T=t}(y; t, \vartheta)$.

Teo. (Criterio di fattorizzazione di Neyman-Fisher)
Assegnato un modello statistico $\mathcal{F} = \{p(y;\vartheta), \vartheta \in \Theta\}$, una statistica t è sufficiente se e solo se

$$p(y;\vartheta) = h(y)k\left(t(y);\vartheta\right).$$

Dim.

$\boxed{\Longrightarrow}$: Se t è sufficiente, vale il teorema prendendo $h(y) = p_{Y|T=t}(y;t)$ e $k\left(t(y);\vartheta\right) = p_T(t;\vartheta)$.

$\boxed{\Longleftarrow}$: Supponendo che Y sia discreta, se vale la scomposizione, allora

$$\begin{aligned} p_T(t;\vartheta) &= \sum_{y:t(y)=t} p(y;\vartheta) \\ &= \sum_{y:t(y)=t} h(y)k\left(t(y);\vartheta\right) \\ &= k(t;\vartheta) \sum_{y:t(y)=t} h(y), \end{aligned}$$

e quindi per y tale che $t(y) = t$, la distribuzione condizionata è

$$p_{Y|T=t}(y;t,\vartheta) = \frac{h(y)\ \cancel{k\left(t(y);\vartheta\right)}}{\cancel{k(t;\vartheta)}\ \sum_{y:t(y)=t} h(y)}.$$

Si può estendere al caso continuo, supponendo di poter identificare una statistica complementare V tale che $Y \mapsto (T,V)$ sia biunivoca, con determinante dello Jacobiano della trasformazione inversa $y = y(t,v)$ pari a $|J|$. Allora, (t,v) è equivalente ai dati y e la densità è

$$\begin{aligned} p_{T,V}(t,v;\vartheta) &= p_Y(y(t,v);\vartheta)|J| \\ &\overset{Hp.}{=} h(y(t,v))k(t;\vartheta)|J| \end{aligned}$$

da cui la distribuzione marginale

$$\begin{aligned} p_T(t;\vartheta) &= \int_{V_t} h(y(t,v))k(t;\vartheta)|J|\,dv \\ &= k(t;\vartheta)\int_{V_t} h(y(t,v))|J|\,dv. \end{aligned}$$

Dunque, facendo il rapporto delle densità, si ottiene che la distribuzione a posteriori è indipendente da ϑ:

$$p_{Y|T=t}(y;t,\vartheta) = \frac{h(y)\ \cancel{k\left(t(y);\vartheta\right)}}{\cancel{k(t;\vartheta)}\ \int_{V_t} h(y(t,v))|J|\,dv}.$$

□

Osservazioni

1. Se $t = t(y)$ è sufficiente, lo è anche ogni funzione biunivoca di t.

2. I dati y sono sempre statistica sufficiente per i dati stessi.

In un modello parametrico $\mathcal{F}$ con *verosimiglianza* $L(\vartheta; y) = c(y)p(y; \vartheta)$, il criterio comporta che anche la verosimiglianza si può esprimere come funzione di ϑ solo tramite $t = t(y)$.

Esempio (C.c.s. da Unif$(0, \vartheta)$)
Sia $y_1, y_2, \dots, y_n$ da Unif$(0, \vartheta)$, allora si ha che la verosimiglianza è

$$\begin{aligned} L(\vartheta; y) &= \frac{1}{\vartheta^n} \prod_{i=1}^{n} \mathbb{1}_{[0,\vartheta]}(y_i) \\ &= \frac{1}{\vartheta^n} \underbrace{\mathbb{1}_{[0,+\infty)}(y_{(1)})}_{h(y)} \mathbb{1}_{[0,\vartheta)}(y_{(n)}) \end{aligned}$$

Esempio (Modello biparametrico)
Consideriamo un campione $y_1, y_2, \dots, y_n$ da $Y_i \sim$ Gamma(α, λ), allora

$$\begin{aligned} p_Y(y; \alpha, \lambda) &= \prod_{i=1}^{n} \lambda^\alpha y_i^{\alpha-1} e^{-\lambda y_i} / \Gamma(\alpha) \\ &= \lambda^{n\alpha} \left(\prod_{i=1}^{n} y_i \right)^{\alpha-1} e^{-\lambda \sum_{i=1}^{n} y_i} / \Gamma(\alpha)^n, \end{aligned}$$

che è già scritta come funzione di una statistica t. Per il criterio di fattorizzazione di Neyman-Fisher, $t = (\prod_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i)$ è una statistica sufficiente o, in modo equivalente, $t' = (\sum_{i=1}^{n} \log y_i, \sum_{i=1}^{n} y_i)$, così come qualunque altra funzione biunivoca di t.

Lezione 18

18.1 Statistiche sufficienti minimali

Per un modello $\mathcal{F}$, esiste in genere una pluralità di statistiche sufficienti:

- y e qualunque trasformazione biunivoca di y
- A volte qualche statistica $t(y)$ con dimensione inferiore a quella di y, e ogni loro funzione biunivoca.

Esempio (Campionamento)
Dalla densità

$$p(y;\vartheta) = \vartheta^{\sum_{i=1}^n y_i}(1-\vartheta)^{n-\sum_{i=1}^n y_i},$$

si osserva che sia $t = \sum_{i=1}^n y_i$ che $t' = \left(\sum_{i=1}^m y_i, \sum_{i=m+1}^n y_i\right)$ sono statistiche sufficienti. Naturalmente, si preferisce t, che è funzione non biunivoca di t' e la cui partizione indotta è anche meno fine.

Def. (Sufficienza minimale)
Una statistica sufficiente $s = s(y)$ è detta *minimale* se è funzione di ogni altra statistica sufficiente $t = t(y)$.

In altre parole, la partizione definita da s è pari o meno fine di quella di qualunque altra statistica sufficiente.

Nota Se s è sufficiente minimale, qualunque funzione biunivoca $u = u(s)$ è sufficiente minimale, in quanto s e u inducono la medesima partizione di $\mathcal{Y}$.

Sufficienza minimale tramite Neyman-Fisher

Si può identificare una statistica sufficiente minimale sfruttando il criterio di fattorizzazione di Neyman-Fisher: sia $t = t(y)$ una qualsiasi statistica sufficiente, per cui $p(y;\vartheta) = h(y)k\big(t(y);\vartheta\big)$. Se y e y' danno lo stesso valore di t, $t(y) = t(y')$ e si ha che

$$p(y;\vartheta) = h(y)k\big(t(y);\vartheta\big)p(y';\vartheta) = h(y')k\big(t(y');\vartheta\big)$$

dunque

$$\frac{p(y;\vartheta)}{p(y';\vartheta)} = \frac{h(y)}{h(y')} = c(y,y').$$

Per qualunque statistica sufficiente, allora

$$t(y) = t(y') \implies \frac{p(y;\vartheta)}{p(y';\vartheta)} = c(y,y').$$

Prop. (Caratterizzazione della sufficienza minimale)
Se $s = s(y)$ è statistica sufficiente minimale, vale anche il viceversa, ovvero

$$\frac{p(y;\vartheta)}{p(y';\vartheta)} = c(y,y') \implies s(y) = s(y').$$

In termini di partizioni indotte,

$$\underbrace{\{y' \in \mathcal{Y} : t(y) = t(y')\}}_{\mathcal{Y}_t} \subseteq \underbrace{\{y' \in \mathcal{Y} : s(y') = s(y)\}}_{\mathcal{Y}_s}.$$

Hanno lo stesso valore di s tutti e soli i punti dello spazio campionario che hanno rapporto di densità indipendente da ϑ. Siccome si può scrivere che

$$\left.\begin{array}{rl} p(y;\vartheta) = & c(y;y')p(y';\vartheta) \\ & \Updownarrow \\ L(\vartheta;y) = & d(y,y')L(\vartheta,y') \end{array}\right\} \implies s(y) = s(y')$$

si dice anche **criterio della partizione di verosimiglianza**.

Per l'inferenza bayesiana, se la verosimiglianza si può scomporre in quel modo, la distribuzione a posteriori è equivalente nei punti y e y'.

Esempio (Sufficienza minimale da $\mathcal{N}(\mu, \sigma^2)$)
Sia $y = (y_1, y_2, \ldots, y_n)$ da $\mathcal{N}(\mu, \sigma^2)$. Si ha che

$$\begin{aligned} p(y;\vartheta) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2\right\} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i^2 + \mu^2 - 2\mu y_i)\right\} \end{aligned}$$

per cui il criterio della partizione di verosimiglianza equipara tutti e soli i campioni per cui

$$\frac{p(y;\vartheta)}{p(y';\vartheta)} = \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - \sum_{i=1}^n (y_i')^2\right) + \frac{\mu}{\sigma^2}\left(\sum_{i=1}^n y_i - \sum_{i=1}^n y_i'\right)\right\}.$$

La dipendenza dal parametro $\vartheta = (\mu, \sigma^2)$ non vale solamente se e solo se

$$\left(\sum_{i=1}^n y_i, \sum_{i=1}^n y_i^2\right) = \left(\sum_{i=1}^n y_i', \sum_{i=1}^n (y_i')^2\right),$$

per cui la statistica $t = \left(\sum_{i=1}^n y_i, \sum_{i=1}^n y_i^2\right)$ è sufficiente minimale.

Esempio (C.c.s. da una Cauchy)

Consideriamo un c.c.s. da una distribuzione Cauchy$(\vartheta, 1)$ con densità

$$p(y_i;\vartheta) = \frac{1}{\pi\left(1+(y_i-\vartheta)^2\right)}.$$

Per applicare il criterio di partizione è necessario che

$$\frac{p(y;\vartheta)}{p(y';\vartheta)} = \frac{\prod_{i=1}^n \left[1+(y_i'-\vartheta)^2\right]}{\prod_{i=1}^n \left[1+(y_i-\vartheta)^2\right]}$$

non dipenda da ϑ. Si tratta del rapporto tra due polinomi in ϑ di grado $2n$: con $n=2$, si ha che

$$\prod_{i=1}^2 \left[1+(y_i-\vartheta)^2\right] = a + b\vartheta + c\vartheta^2 + d\vartheta^3 + \vartheta^4.$$

Si ha l'indipendenza da ϑ solo se tutti i coefficienti coincidono, ovvero se posso scambiare gli indici delle y_i. L'unica cosa in comune tra le densità sono le statistiche ordinate, ovvero si ha coincidenza se e solo se

$$\left(y_{(1)}, y_{(2)}, \ldots, y_{(n)}\right) = \left(y'_{(1)}, y'_{(2)}, \ldots, y'_{(n)}\right).$$

Più in generale, lo è anche per il modello parametrico di c.c.s. dalla classe di tutte le distribuzioni continue con supporto $\mathbb{R}$.

Esempio (Statistica sufficiente minimale nel modello normale autoregressivo)
Si consideri il modello normale autoregressivo, con $Y_1 \sim \mathcal{N}(0, (1-\rho^2)^{-1})$, $|\rho| < 1$ e $Y_i | Y_{i-1} = y_{i-1} \sim \mathcal{N}(\rho y_{i-1}, 1)$ per $i = 2, \ldots, n$.

Si ha la densità congiunta

$$\begin{aligned}
p(y;\rho) &= \frac{\sqrt{1-\rho^2}}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-\rho^2)y_1^2} \prod_{i=2}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i-\rho y_{i-1})^2} \\
&= (2\pi)^{-\frac{n}{2}} \sqrt{1-\rho^2} \exp\left\{-\frac{1}{2}\left[(1-\rho^2)y_1^2 + \sum_{i=2}^n \left(y_i^2 + \rho^2 y_{i-1}^2 - 2\rho y_i y_{i-1}\right)\right]\right\} \\
&= (2\pi)^{-\frac{n}{2}} \sqrt{1-\rho^2} \exp\left\{\rho \sum_{i=2}^n y_i y_{i-1} - \frac{1}{2}\rho^2 \sum_{i=2}^{n-1} y_i^2 - \frac{1}{2}\sum_{i=1}^n y_i^2\right\}.
\end{aligned}$$

Sono allora equiparati tutti i campioni il cui rapporto è indipendente da ρ, ovvero

$$\frac{p(y;\rho)}{p(y';\rho)} = \exp\left\{\rho\left(\sum_{i=2}^n y_i y_{i-1} - \sum_{i=2}^n y_i' y_{i-1}'\right) - \frac{1}{2}\rho^2\left(\sum_{i=2}^{n-1} y_i^2 - \sum_{i=2}^{n-1}(y_i')^2\right) - \frac{1}{2}\left(w(y) - w(y')\right)\right\},$$

che è indipendente da ρ quando i coefficienti di ρ e ρ^2 sono 0. Quindi, la statistica sufficiente

minimale è

$$s = \left(\sum_{i=2}^{n} y_i y_{i-1}, \sum_{i=2}^{n-i} y_i^2 \right).$$

In particolare, la statistica sufficiente minimale è di dimensione maggiore della dimensione del parametro.

Nota In termini di partizione indotta, la statistica sufficiente minimale

- Esiste sempre, in quanto il campione è sempre statistica sufficiente.
- È unica, a meno di trasformazioni biettive.

18.2 Famiglie esponenziali e sufficienza

Quando il modello statistico è una famiglia esponenziale, è particolarmente semplice individuare una statistica sufficiente e stabilire se è minimale.

Sia $y = (y_1, y_2, \dots, y_n)$ un c.c.s. da una famiglia esponenziale monoparametrica, allora

$$p(y; \vartheta) = c(\vartheta)^n \prod_{i=1}^{n} h(y_i) \exp\left\{ \psi(\vartheta) t^{(n)}(y) \right\},$$

per cui $t^{(n)}(y)$ è statistica sufficiente per il criterio di fattorizzazione. È anche sufficiente minimale, in quanto

$$\frac{p(y; \vartheta)}{p(y'; \vartheta)} = z(y, y') \exp\left\{ \psi(\vartheta) \big(t^{(n)}(y) - t^{(n)}(y') \big) \right\},$$

per cui è indipendente da ϑ sse $t^{(n)}(y) = t^{(n)}(y')$. Questo vale anche se i dati non sono indipendenti, quando la densità si può scrivere in forma esponenziale (e.g. modello autoregressivo).

Esercizi Verificare le statistiche sufficienti minimali per $Y_i \sim \mathcal{N}(\vartheta x_i, 1)$ e $Y_i \sim \text{Pois}\big(e^{\vartheta x_i}\big)$.

Famiglia esponenziale multiparametrica

Analogamente, scriviamo la densità congiunta

$$p(y; \vartheta) = c(\vartheta)^n \prod_{i=1}^{n} h(y_i) \exp\left\{ \psi(\vartheta)^\intercal t^{(n)}(y) \right\}.$$

Vediamo subito che vale il teorema di fattorizzazione, e la statistica sufficiente è

$$t^{(n)}(y) = \left(\sum_{i=1}^{n} t_1(y_i), \dots, \sum_{i=1}^{n} t_k(y_i) \right).$$

È statistica sufficiente minimale se gli insiemi di funzioni $\{1, \psi_1(\vartheta), \dots, \psi_k(\vartheta)\}$ e $\{1, t_1(y_i), \dots, t_k(y_i)\}$ sono linearmente indipendenti, ovvero se la famiglia è in forma minimale.

Esempio (GLM con parametro di dispersione noto)

Supponiamo un modello lineare generalizzato, ovvero

$$p_{Y_i}(y_i;\vartheta_i) = \exp\left\{\frac{\vartheta_i y_i - b(\vartheta_i)}{a_i(\varphi_0)} - c(y_i,\varphi_0)\right\},$$

dove $\vartheta_i = \sum_{i=1}^{p}\beta_j x_{ij}$ e sono note $a_i(\varphi_0), c(y_i,\varphi_0)$. Assumiamo inoltre che la matrice $X_{n\times p}$ abbia rango p. Vediamo facilmente che

$$p_Y(y_i;\beta,\boldsymbol{x}_i) = \exp\left\{\beta^{\mathsf{T}}\sum_{i=1}^{n}\frac{\boldsymbol{x}_i y_i}{a_i(\varphi_0)} - \sum_{i=1}^{n}\frac{b(\beta^{\mathsf{T}}\boldsymbol{x}_i)}{a_i(\varphi_0)} - \sum_{i=1}^{n}c(y_i,\varphi_0)\right\}.$$

La statistica sufficiente minimale per l'inferenza su β è allora

$$t(y) = \left(\sum_{i=1}^{n}\frac{1}{a_i(\varphi_0)}x_{i1}y_i, \ldots, \sum_{i=1}^{n}\frac{1}{a_i(\varphi_0)}x_{ip}y_i\right).$$

Lezione 19

19.1 Famiglie esponenziali curve

È importante osservare che, nei casi standard di famiglia esponenziale, la dimensione della statistica sufficiente minimale è pari a quella del parametro. In altri casi, l'ordine della famiglia è superiore alla dimensione del parametro, ad esempio nel caso normale autoregressivo si ha la statistica sufficiente minimale

$$s = \left(\sum_{i=2}^{n} y_i y_{i-1}, \sum_{i=2}^{n-1} y_i^2 \right).$$

Questa famiglia non è regolare e si dice *famiglia esponenziale curva*, in quanto il parametro ψ non è libero di variare ma è vincolato a una curva

$$\psi_1(\vartheta) = \rho$$

$$\psi_2(\vartheta) = -\frac{1}{2}\rho^2$$

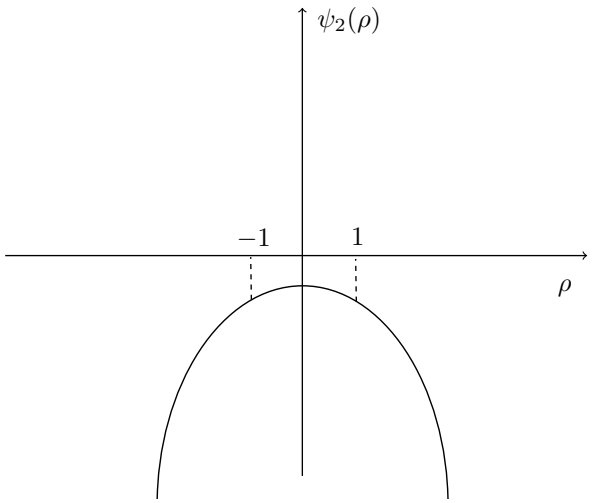


Figura 17: curveExp

Esempio (Genetic linkage)
Dati l'esempio delle quattro categorie (AB, Ab, aB, ab), si osserva $y = (y_1, y_2, y_3, y_4)$ e si ipotizza un modello $Y \sim \mathrm{Mn}_4(n, \pi(\vartheta))$, con

$$\pi(\vartheta) = \left(\frac{2+\vartheta}{4}, \frac{1-\vartheta}{4}, \frac{1-\vartheta}{4}, \frac{\vartheta}{4} \right).$$

Il modello è allora

$$\begin{aligned} p(y;\vartheta) &= \frac{n!}{y_1!\cdots y_4!}\frac{1}{4^n}(2+\vartheta)^{y_1}(1-\vartheta)^{y_2+y_3}\vartheta^{n-y_1-y_2-y_3} \\ &= h(y)\vartheta^n \exp\left\{ y_1 \log\left(\frac{2+\vartheta}{\vartheta}\right) + (y_2+y_3)\log\left(\frac{1-\vartheta}{\vartheta}\right)\right\}. \end{aligned}$$

Si tratta di una famiglia esponenziale di ordine $k=2$ con $p=\dim(\vartheta)=1$.

19.2 Statistiche complete

Resta da chiarire il ruolo delle statistiche costanti in distribuzione sotto il modello $\mathcal{F}$, ovvero statistiche $c=c(y)$ tali che

$$p_Y(y;\vartheta) = p_C(c)\cdot p_{Y|C=c}(y;c,\vartheta).$$

Pensando ad un esperimento in due stadi, la costanza in distribuzione della statistica c corrisponde al fatto che la scelta della partizione

$$\mathcal{Y}_c = \{y\in\mathcal{Y} : c(y)=c\}$$

non è informativa su ϑ.

Esempio (c.c.s. da una famiglia di posizione)
Consideriamo $y=(y_1,y_2,\dots,y_n)$ un c.c.s. da una variabile casuale con densità

$$p(y;\mu) = p_0(y-\mu).$$

Osserviamo che l'insieme delle $n-1$ differenze

$$c = (y_2-y_1,\dots,y_n-y_1)$$

è costante in distribuzione, poiché $Y_i=\mu+Y_i^0$ e

$$\begin{aligned} C &= \left(\cancel{\mu}+Y_2^0-(\cancel{\mu}+Y_1^0),\dots,\cancel{\mu}+Y_n^0-(\cancel{\mu}+Y_1^0)\right) \\ &= \left(Y_2^0-Y_1^0,\dots,Y_n^0-Y_1^0\right). \end{aligned}$$

Il modello condizionato $Y_1,\dots,Y_n|C=c$ è essenzialmente unidimensionale e, nota $p_0(\cdot)$, potrebbe essere ottenuto tramite gli usuali strumenti del calcolo delle probabilità.

Si noti che, se il c.c.s. è da una $\mathcal{N}(\mu,1)$, sembrano aprirsi due possibilità

1. Riduzione al modello unidimensionale indotto dalla statistica sufficiente minimale $s=\sum_{i=1}^n Y_i$.
2. Riduzione al modello unidimensionale $Y_1,Y_2,\dots,Y_n|C=c$.

Nel primo caso, fissate le $n-1$ differenze $c_1,c_2,\dots,c_{n-1}$, si ha la corrispondenza biunivoca

$(y_1, y_2, \dots, y_n) \mapsto (y_1, c_1, \dots, c_{n-1})$, dove

$$\begin{cases} y_2 = c_1 + y_1 \\ \vdots \\ y_n = c_{n-1} + y_1 \end{cases}$$

e quindi la statistica sufficiente è $s = y_1 + \sum_{i=1}^{n-1}(c_i + y_1) = ny_1 + \sum_{i=1}^{n-1} c_i$. Dunque, data $C = c$, s e y_1 sono in corrispondenza biunivoca. La riduzione per condizionamento può quindi considerare il modello per $S|C = c$ equivalentemente.

Nota Dal momento in cui s contiene tutta e sola l'informazione su ϑ in y, mentre c non contiene informazione, è naturale aspettarsi che vi sia indipendenza tra S e C. Questo è vero solamente se il modello indotto dalla statistica sufficiente minimale S soddisfa la proprietà di *completezza*.

Def. (Completezza)
La famiglia di distribuzioni $\mathcal{F}_s = \{p_s(s;\vartheta), \vartheta \in \Theta\}$ è detta *completa* se, per qualunque funzione $u(s)$ misurabile con

$$\mathbb{E}_\vartheta\left[u(S)\right] = 0 \quad \forall \vartheta \in \Theta,$$

implica che $P_\vartheta\big(u(S) = 0\big) = 1$ per ogni $\vartheta \in \Theta$. Si dice anche in tal caso che S è *completa*.

Osservazione

Questo significa che si ammette valore atteso 0 (o analogamente pari a k) indipendentemente da ϑ solo se la funzione di S è degenere.

A maggior ragione, dunque, non si ammettono funzioni non banali di S che siano né costanti in distribuzione, né con media costante in ϑ.

Prop.
Se s è una statistica sufficiente per $\mathcal{F}$ e completa, allora s è sufficiente minimale.

Non è molto agevole mostrare la completezza di una famiglia $\mathcal{F}$, se non utilizzando un teorema generale per le famiglie esponenziali.

Teo. (Completezza e famiglie esponenziali)
Sia Y una famiglia esponenziale della forma

$$p_Y(y;\vartheta) = h(y)c(\vartheta)\exp\left\{\psi^\mathsf{T} t(y)\right\},$$

tale che $\Psi = \psi(\Theta)$ ha interno non vuoto, ovvero contiene un rettangolo p-dimensionale. Allora, $t(y)$ è una statistica sufficiente completa.

Dim.
Per la dimostrazione, vedere Pace e Salvan (1996, Teo. 5.7).

□

In particolare, Y può essere della forma $Y_1, Y_2, \dots, Y_n$ sono i.i.d, per cui la chiusura rispetto al campionamento casuale semplice garantisce che la statistica continui ad essere sufficiente e completa.

Esempio
Negli esempi, non è necessario che siano i.i.d, ad esempio se $Y_i \sim \text{Pois}(\vartheta x_i)$, si può scrivere

$$p_Y(y;\vartheta) = c(\vartheta)h(y)\exp\Big\{\log\vartheta \sum_{i=1}^{n} y_i\Big\}.$$

Poiché $\psi = \log\vartheta$ e $\Psi = \mathbb{R}$, certamente contiene un rettangolo. Dunque, la statistica sufficiente minimale

$$t(y) = \sum_{i=1}^{n} y_i$$

è una statistica completa.

Per casa Mostrare la completezza nel caso $Y_i \sim \text{Pois}(e^{\vartheta x_i})$.

Esempio (Normale con media e varianza ignota)
Considero il parametro canonico $\psi(\vartheta) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$: siccome è uno spazio prodotto cartesiano, cioè $\Psi = \mathbb{R} \times (-\infty, 0)$, allora contiene un rettangolo in $\mathbb{R}^2$.

Non soddisfano la condizione di completezza quelle famiglie esponenziali di ordine k con $k > p$, ad esempio il normale autoregressivo (esercizio: cercare una funzione non banale che abbia media costante in ϑ).

Vediamo un altro teorema, che garantisce l'indipendenza di una statistica completa da qualunque statistica costante in distribuzione (Basu, 1955).

Teo. (Teorema di Basu)
Se sotto il modello $\mathcal{F}$ abbiamo $s = s(y)$ sufficiente e completa e $c = c(y)$ è costante in distribuzione, allora $S = s(Y)$ e $C = c(Y)$ sono stocasticamente indipendenti.

Dim.
Per la sufficienza di s, $C|S = s$ ha distribuzione indipendente da ϑ. Inoltre, per ogni insieme E nello spazio campionario di C, la quantità

$$P(C \in E) - P(C \in E|S = s)$$

è funzione solo di s e per ogni $\vartheta \in \Theta$, per cui si può applicare la regola del valore atteso iterato

$$\mathbb{E}_\vartheta \big[\underbrace{P(C \in E) - P(C \in E|S)}_{u(s) \text{ banale per def.}}\big] = 0.$$

Allora, per la definizione di completezza di s, con probabilità 1 per ogni ϑ vale che

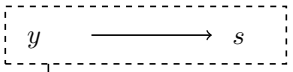
$$P(C \in E | S = s) = P(C \in E),$$

e dunque C e S sono stocasticamente indipendenti

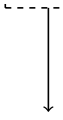
□

Conseguenze

- s sufficiente e completa $\implies$
  - non esistono funzioni di s costanti in distribuzione del modello,
  - s è indipendente da qualunque funzione di y costante in distribuzione.



Nessuna funzione di s è costante in distrib.



Costante in distrib.

- Il teorema permette di mostrare facilmente alcuni risultati di indipendenza, ad esempio tra media e devianza campionaria sotto c.c.s. da $\mathcal{N}(\mu, \sigma^2)$.

Esempio (Indipendenza tra $\overline{Y}$ e $\sum_{i=1}^n (Y_i - \overline{Y})^2$ da $\mathcal{N}(\mu, \sigma^2)$)
La v.c. $\sum_{i=1}^n Y_i$, e dunque $\overline{Y}$, è statistica sufficiente e completa per l'inferenza su μ nel modello con σ^2 nota (vedremo che non fa differenza).

D'altra parte, sempre nel modello con σ^2 nota, la v.c. $\sum_{i=1}^n (Y_i - \overline{Y})^2$ è costante in distribuzione, in quanto

$$Y_i = \mu + Z_i$$

$$\overline{Y} = \mu + \overline{Z}$$

e dunque le differenze $\left(Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y}\right)$ hanno la stessa distribuzione delle variabili $\left(Z_1 - \overline{Z}, \ldots, Z_n - \overline{Z}\right)$ con $Z_1, \ldots, Z_n$ i.i.d $\mathcal{N}(0, \sigma^2)$. Dunque, per il teorema di Basu le v.c. $\overline{Y}$ e $\sum_{i=1}^n (Y_i - \overline{Y})^2$ sono indipendenti.

Vediamo in ultimo un caso particolare di famiglia non esponenziale con statistica sufficiente bidimensionale non completa.

Esempio (c.c.s. da Unif$(\vartheta, \vartheta + 1)$)

Se $y = (y_1, y_2, \ldots, y_n)$ da $\text{Unif}(\vartheta, \vartheta + 1)$, si ha

$$\begin{aligned} p(y; \vartheta) &= \prod_{i=1}^{n} \mathbb{1}_{[\vartheta, \vartheta+1]}(y_i) \\ &= \mathbb{1}_{[\vartheta, +\infty)}\big(y_{(1)}\big)\, \mathbb{1}_{(-\infty, \vartheta+1]}\big(y_{(n)}\big). \end{aligned}$$

In particolare, per il criterio di fattorizzazione, la statistica $\big(y_{(1)}, y_{(n)}\big)$ è statistica sufficiente per l'inferenza su ϑ. Inoltre, per il criterio di fattorizzazione della verosimiglianza, è anche statistica sufficiente minimale.

Tuttavia, la famiglia $\text{Unif}(\vartheta, \vartheta + 1)$ è anche famiglia di posizione, in quanto $Y_i = \vartheta + U_i$, $U_i \sim \text{Unif}(0, 1)$ e allora $Y_{(n)} - Y_{(1)}$ è costante in distribuzione:

$$Y_{(n)} - Y_{(1)} = \cancel{\vartheta} + U_{(n)} - (\cancel{\vartheta} + U_{(1)}) = U_{(n)} - U_{(1)}.$$

Indicando con τ il valore atteso $\mathbb{E}_\vartheta \left[Y_{(n)} - Y_{(1)}\right]$, allora la statistica

$$\mathbb{E}_\vartheta \left[Y_{(n)} - Y_{(1)} - \tau\right] = 0,$$

ma non è una v.c. degenere, dunque la statistica sufficiente minimale non è completa.

In questi casi, cioè quando la statistica sufficiente minimale non è completa, si apre la possibilità di una ulteriore riduzione inferenziale tramite condizionamento (Pace e Salvan, 1997) del tipo

$$p_S(s; \vartheta) = p_V(v) p_{S|V}(s; v, \vartheta).$$

Lezione 20

20.1 Inferenza di verosimiglianza

Introdotta da R.A. Fisher in una serie di lavori dal 1921 al 1935, l'inferenza di verosimiglianza è una metodologia basata sulla *funzione di verosimiglianza*, elemento chiave sia per l'inferenza frequentista sia per quella bayesiana.

In inferenza frequentista, la verosimiglianza fornisce

1. Procedure per stima puntuale, per regioni e test di ipotesi.
2. Soluzioni basate su risultati asintotici per problemi di distribuzione delle procedure inferenziali.

Def. (Verosimiglianza)
Si dice *verosimiglianza* (*likelihood*) che l'osservazione di Y attribuisce a ϑ, indicata con $L(\vartheta)$, la funzione $L : \Theta \longrightarrow [0, +\infty)$ definita da

$$L(\vartheta) = c(y)p_Y(y; \vartheta),$$

ovvero la densità congiunta di Y al variare del parametro ϑ.

Osservazione

- Non si può interpretare come una probabilità, in quanto lo sarebbe solo in funzione di y.

Esempio (Modello binomiale)
Dati 6 successi su 10 tiri, assumiamo $Y \sim \text{Bin}(n, \vartheta)$ con $\vartheta \in (0, 1)$ e distribuzione

$$p(y; \vartheta) = \binom{n}{y} \pi^y (1 - \pi)^{n-y},$$

dunque la verosimiglianza è

$$L(\vartheta) = \binom{10}{6} \vartheta^6 (1 - \vartheta)^4.$$

Visto che usualmente si realizzano valori campionari con probabilità elevata, è ragionevole pensare che ϑ_0 sia collocato nella regione di valori di ϑ con $L(\vartheta)$ elevato.

$L(\vartheta)$ si interpreta come il grado di accordo tra ϑ e y^{oss}, da cui la terminologia "è più *verosimile* un valore del parametro rispetto a un altro" se $L(\vartheta) > L(\vartheta')$, o equivalentemente, se

$$\frac{L(\vartheta)}{L(\vartheta')} \gg 1.$$

Poiché solo confronti relativi hanno senso, tutti i fattori che non dipendono da ϑ si possono trascurare in $L(\vartheta)$.

Se si è interessati ad identificare un valore di ϑ più ragionevole di un altro, si può scegliere

$$\hat{\vartheta} = \underset{\theta \in \Theta_0}{\operatorname{argmax}}\, L(\vartheta),$$

che si interpreta come il valore più supportato dai dati y^{oss}. Si può alternativamente ragionare in termini di *verosimiglianza relativa*, $L(\vartheta)/L(\hat{\vartheta}) \in [0,1]$, anche se comunque non si può interpretare come probabilità.

Per quanto riguarda la stima intervallare, si scelgono valori del parametro con verosimiglianza superiore ad una certa soglia, che può essere definita in modo rigoroso e tale da soddisfare certe proprietà asintotiche di copertura intervallare.

Analogamente, per quanto riguarda la verifica di ipotesi, è necessario definire *quando* la differenza di verosimiglianza è sufficientemente grande da portare al rifiuto dell'ipotesi nulla.

Nota Al crescere della numerosità campionaria, più in generale dell'informazione, la verosimiglianza tende a concentrarsi di più e, dunque, discriminare maggiormente tra diversi valori del parametro.

Nel caso continuo, la funzione di verosimiglianza non esprime più la massa di probabilità, bensì la densità di probabilità di y^{oss} al variare di $\vartheta \in \Theta$.

Esempio (Modello Weibull)
Assumendo $Y = (Y_1, Y_2, \ldots, Y_n)$ realizzazioni i.i.d da $Y \sim \text{Weibull}(\beta, \gamma)$, con distribuzione marginale

$$p_{Y_i}(y_i; \vartheta) = \gamma \beta^{-\gamma} y_i^{\gamma-1} e^{\left(-\frac{y_i}{\beta}\right)^{\gamma}}, \quad \gamma > 0,\ \beta > 0.$$

In particolare, il massimo $\hat{\vartheta} = (\hat{\beta}, \hat{\gamma})$ si può trovare tramite massimizzazione numerica della funzione $L(\vartheta)$.

Spesso è conveniente considerare la funzione di *log-verosimiglianza*

$$\ell(\vartheta) = \ell(\vartheta; y) = \begin{cases} \log L(\vartheta) & \text{se } L(\vartheta) > 0 \\ -\infty & \text{se } L(\vartheta) = 0 \end{cases},$$

ovviamente definita a meno di costanti additive. Per osservazioni indipendenti,

$$\ell(\vartheta) = \sum_{i=1}^{n} \log p_{Y_i}(y_i; \vartheta),$$

per cui la log-verosimiglianza ha il vantaggio di ridurre la complessità della funzione e di favorire l'applicazione dei teoremi limite.

Def. (Stima di massima verosimiglianza)
Un valore $\hat{\vartheta}$ che massimizza $L(\vartheta)$, o equivalentemente $\ell(\vartheta)$, in Θ si dice *stima di massima verosimiglianza*. Se visto come funzione di Y, $\hat{\vartheta}(Y)$ è detto *stimatore di massima verosimiglianza.*

Siccome molte proprietà di $\hat{\vartheta}(Y)$ sono asintotiche, in genere ci si limita a richiedere che $\hat{\vartheta}(Y)$ esista con probabilità che tende a 1 per $n \to \infty$.

Esercizio: confrontare lo stimatore $\hat{\vartheta} = Y_{(n)}$, per il modello uniforme $\text{Unif}(0, \vartheta)$, con lo stimatore ottenuto basandosi sul metodo dei momenti $\tilde{\vartheta}$.

Lezione 21

21.1 Modello statistico regolare

Nelle applicazioni, spesso si considerano modelli parametrici che soddisfano determinate caratteristiche, in modo da rendere possibili le procedure inferenziali standard.

Def. (Modello parametrico regolare)
Sia $\mathcal{F}$ un modello statistico parametrico con parametro $\vartheta \in \Theta$. $\mathcal{F}$ si dice *regolare* se

- Θ è un sottoinsieme aperto di $\mathbb{R}^p$.
- $\ell(\vartheta)$ è continua e differenziabile almeno tre volte in Θ.

Osservazione

Una condizione necessaria affinché il modello sia regolare è che tutte le distribuzioni abbiano *supporto indipendente da* ϑ. Le famiglie esponenziali hanno verosimiglianza regolare, a meno che le funzioni ψ_j non siano derivabili.

In un modello regolare, indichiamo le derivate di $\ell(\vartheta)$ come

$$\ell_r(\vartheta) = \frac{\partial}{\partial \vartheta_r}\ell(\vartheta) \qquad \text{vettore } p \times 1$$

$$\ell_{rs}(\vartheta) = \frac{\partial^2}{\partial \vartheta_r \vartheta_s}\ell(\vartheta) \qquad \text{matrice } p \times p$$

$$\ell_{rst}(\vartheta) = \frac{\partial^3}{\partial \vartheta_r \vartheta_s \vartheta_t}\ell(\vartheta) \qquad \text{tensore } p \times p \times p$$

Il vettore delle derivate prime è detto *funzione punteggio* (*score function*) e si indica con

$$\ell_*(\vartheta) = (\ell_1, \ell_2, \ldots, \ell_n)^{\mathsf{T}}.$$

Nel seguito, se non specificato, si assumerà che $\hat{\vartheta}$ sia unico e sia l'unica soluzione dell'*equazione di verosimiglianza*

$$\ell_*(\vartheta) = 0.$$

La matrice $p \times p$ delle derivate seconde cambiata di segno è chiamata *informazione osservata*

$$j(\vartheta) = -l_{**}(\vartheta) = \begin{pmatrix} -\ell_{11} & \ldots & -\ell_{1p} \\ \vdots & \ldots & \vdots \\ -\ell_{p1} & \ldots & -\ell_{p} \end{pmatrix}$$

o in modo compatto $j = [-\ell_{rs}]$.

Esempio (Famiglia esponenziale)
Supponiamo che $Y_1, Y_2, \dots, Y_n$ provenga da

$$p_{Y_i}(y_i;\vartheta) = c(\vartheta)h(y_i)\exp\left\{\psi(\vartheta)^{\mathsf{T}}t(y_i)\right\}, \quad \vartheta \in \Theta \subseteq \mathbb{R}^p,$$

la cui densità congiunta è

$$p(y;\vartheta) = c(\vartheta)^n \prod_{i=1}^{n} h(y_i)\exp\left\{\psi(\vartheta)^{\mathsf{T}}\sum_{i=1}^{n} t(y_i)\right\},$$

dunque la log-verosimiglianza assume la forma

$$\begin{aligned}\ell(\vartheta) &= \psi^{\mathsf{T}}t^{(n)}(y) + n\log c(\vartheta)\\ &= \sum_{j=1}^{k}\psi_j(\vartheta)t_j^{(n)}(y) + n\log c(\vartheta).\end{aligned}$$

La funzione score è data da

$$\ell_r(\vartheta) = \sum_{j=1}^{k}\frac{\partial\psi_j(\vartheta)}{\partial\vartheta_r}t_j^{(n)}(y) + n\frac{\partial}{\partial\vartheta_r}\log c(\vartheta).$$

Ricordando che $\sum_{j=1}^{k}\frac{\partial}{\partial\vartheta_r}\psi_j(\vartheta)\,\mathbb{E}_\vartheta\left[t_j^{(n)}(Y)\right] = -n\frac{\partial}{\partial\vartheta_r}\log c(\vartheta)$, si può sostituire al secondo addendo per ottenere

$$\ell_r(\vartheta) = \sum_{j=1}^{k}\frac{\partial\psi_j(\vartheta)}{\partial\vartheta_r}\left(t_j^{(n)}(y) - \mathbb{E}_\vartheta\left[t_j^{(n)}(Y)\right]\right).$$

Se la famiglia è regolare ($p = k$) e scritta in parametrizzazione canonica ψ, definendo $K(\psi) = -\log c(\vartheta(\psi))$,

$$\ell(\psi) = \sum_{j=1}^{p}\psi_j t_j^{(n)}(y) - nK(\psi), \quad l_*(\psi) = \left[t_r^{(n)} - n\frac{\partial K(\psi)}{\partial\psi_r}\right]_{r=1,\dots,p}.$$

In particolare, sostituendo $\mathbb{E}_\psi\left[t_r^{(n)}(Y)\right] = n\frac{\partial K(\psi)}{\partial\psi_r}$, l'equazione di verosimiglianza diventa

$$l_*(\psi) = \vec{0} \iff \mathbb{E}_\psi\left[t_r^{(n)}(Y)\right] = t_r^{(n)}(Y).$$

La matrice di informazione osservata vale

$$j(\psi) = \left[n\frac{\partial^2 K(\psi)}{\partial\psi_r\partial\psi_s}\right]_{r,s=1,\dots,p}.$$

Nota essendo $t_r^{(n)}(y)$ statistica sufficiente minimale, lo è anche la stima di massima verosimiglianza $\hat{\psi}$ e c'è corrispondenza biunivoca

$$t^{(n)}(y) \longleftrightarrow \hat{\psi}$$

21.2 Informazione osservata

Maggiore è $j(\hat{\vartheta})$, tanto più la log-verosimiglianza è concentrata attorno a $\hat{\vartheta}$, per cui tanto meglio si distinguono regioni con elevata verosimiglianza da regioni con verosimiglianza minore.

Considerando lo sviluppo di Taylor di $\ell(\vartheta)$ attorno a $\hat{\vartheta}$ per $p = 1$, si ha

$$\begin{aligned}\ell(\vartheta) &= l(\hat{\vartheta}) + (\vartheta - \hat{\vartheta}) \overbrace{\ell_*(\hat{\vartheta})}^{=0} + \frac{1}{2}(\vartheta - \hat{\vartheta}) l_{**}(\hat{\vartheta}) + \dots \\ &= l(\hat{\vartheta}) - \frac{1}{2}(\vartheta - \hat{\vartheta}) j(\hat{\vartheta}) + \dots\end{aligned}$$

e dunque $l(\vartheta) - l(\hat{\vartheta}) = -\frac{1}{2}(\vartheta - \hat{\vartheta}) j(\hat{\vartheta})$, dove $j(\hat{\vartheta}) > 0$ in quanto massimo. Quindi, tanto $j(\hat{\vartheta})$ è grande, tanto più rapidamente valori di ϑ lontani da $\hat{\vartheta}$ perdono sostegno empirico: $j(\hat{\vartheta})$ è una misura dell'*informazione* che i dati forniscono sul parametro ϑ. Il discorso vale ugualmente se $p > 1$, considerando $\det j(\hat{\vartheta})$ e l'espansione

$$\ell(\vartheta) - \ell(\hat{\vartheta}) = -\frac{1}{2}(\vartheta - \hat{\vartheta})^{\mathsf{T}} j(\hat{\vartheta})(\vartheta - \hat{\vartheta}).$$

In sintesi, in un modello regolare, $\ell(\vartheta) - \ell(\hat{\vartheta})$ ha un comportamento approssimativamente parabolico in un intorno di $\hat{\vartheta}$.

Esempio (Modello binomiale)
(?) TODO

Esercizio: calcolare le quantità di verosimiglianza per il modello esponenziale.

Esercizio: mostrare che per un c.c.s. da $\mathcal{N}(0, \sigma_0^2)$ con σ_0^2 nota, l'approssimazione quadratica di $\ell(\vartheta) - \ell(\hat{\vartheta})$ è esatta.

Esempio (Modello Weibull)
Dato il modello Weibull, la verosimiglianza è

$$\begin{aligned}L(\vartheta) &= \gamma^n \beta^{-n\gamma} \exp\left\{(\gamma - 1)\sum_{i=1}^n \log y_i - \sum_{i=1}^n \left(\frac{y_i}{\beta}\right)^\gamma\right\} \\ &= c(y)\gamma^n \beta^{-n\gamma} \exp\left\{\gamma \sum_{i=1}^n \log y_i - \beta^{-\gamma} \sum_{i=1}^n y_i^\gamma\right\}.\end{aligned}$$

La log-verosimiglianza è dunque

$$\ell(\vartheta) = n \log \gamma - n\gamma \log \beta + \gamma \sum_{i=1}^n \log y_i - \beta^{-\gamma} \sum_{i=1}^n y_i^\gamma.$$

La funzione punteggio è un vettore di dimensione 2

$$\ell_*(\vartheta) = \big(\ell_\gamma(\vartheta), \ell_\beta(\vartheta)\big)^\mathsf{T},$$

dove

$$\ell_\gamma(\vartheta) = \frac{\partial \ell(\vartheta)}{\partial \gamma} = \frac{n}{\gamma} - n\log\beta + \sum_{i=1}^n \log y_i - \sum_{i=1}^n \frac{y_i}{\beta}^\gamma \log\left(\frac{y_i}{\beta}\right).$$

$$\ell_\beta(\vartheta) = \frac{\partial \ell(\vartheta)}{\partial \beta} = -\frac{n\gamma}{\beta} + \frac{\gamma}{\beta^{\gamma+1}} \sum_{i=1}^n y_i^\gamma.$$

L'equazione di verosimiglianza non ha soluzione analitica e va quindi risolta numericamente. Si può in realtà risolvere in β tenendo γ fissato, ovvero

$$\hat\beta_\gamma = \left(\sum_{i=1}^n y_i^\gamma / n\right)^{1/\gamma},$$

che sostituendo in ℓ_γ, dà la seconda condizione

$$\frac{n}{\gamma} + \sum_{i=1}^n \log y_i - n\frac{\sum_{i=1}^n y_i^\gamma \log y_i}{\sum_{i=1}^n y_i^\gamma} = 0.$$

Quest'ultima va risolta numericamente, ad esempio con la funzione `uniroot` in R.

Esercizio: Mostrare che la condizione per γ ha un'unica soluzione, a meno che le y_i non siano tutte coincidenti.

Per l'informazione osservata, si ha la matrice delle derivate seconde cambiate di segno

$$j_{\gamma\gamma}(\vartheta) = -\frac{\partial^2 \ell(\vartheta)}{\partial \gamma^2} = \frac{n}{\gamma^2} + \sum_{i=1}^n \left(\frac{y_i}{\beta}^\gamma \log^2\left(\frac{y_i}{\beta}\right)\right);$$

$$j_{\gamma\beta}(\vartheta) = -\frac{\partial \ell(\vartheta)}{\partial \gamma \beta} = \frac{n}{\beta} - \sum_{i=1}^n \frac{y_i}{\beta^{\gamma+1}} \left[\gamma \log\left(\frac{y_i}{\beta}\right) - 1\right];$$

$$j_{\beta\beta}(\vartheta) = -\frac{\partial \ell(\vartheta)}{\partial \beta} = -\frac{n\gamma}{\beta^2} + \frac{\gamma(\gamma+1)}{\beta^2} \sum_{i=1}^n \left(\frac{y_i}{\beta}\right)^\gamma.$$

Esercizio: Ipotizzando i dati per le piante di mais, trovare le quantità di verosimiglianza ipotizzando $\mathcal{N}(\mu, \sigma^2)$ con $\vartheta = (\mu, \sigma^2)$.

Esercizio: Considerare il modello esponenziale con censura e variabile esplicativa x, calcolare le quantità di verosimiglianza.

Esempio (Censura di II tipo)

Nella censura di II tipo si inverte lo schema di osservazione, in particolare si fissa il nume-

ro di guasti r da osservare e si termina l'osservazione delle n unità una volta osservati r guasti.

Siano $Y_1, Y_2, \ldots, Y_n$ le v.c. i.i.d che rappresentano i tempi di durata. L'osservazione è dei tempi ordinati $Y_{(1)}, \ldots, Y_{(n)}$ con densità

$$p_{Y_{(1)},\ldots,Y_{(n)}}(y_1, y_2, \ldots, y_n; \vartheta) = \underbrace{\frac{n!}{(n-r)!}}_{n-r \text{ permut.}} \big(1 - F_Y(y_r)\big)^{n-r} \prod_{i=1}^{r} p_Y(y_i; \vartheta),$$

dove $y_1 < y_2 < \ldots < y_r$ e le densità p, F sono relative alla generica Y_i. Nella verosimiglianza, il termine combinatorio è trascurabile e

$$L(\vartheta) = \big(1 - F_Y(y_{(r)}; \vartheta)\big)^{n-r} \prod_{i=1}^{r} p_Y(y_{(i)}; \vartheta).$$

Se nella censura di I tipo si avesse $t_0 = t_{(r)}$ e se $n - r$ osservazioni risultassero maggiori di t_0, allora la verosimiglianza sarebbe la stessa.

In particolare, *la verosimiglianza non dipende dallo schema di osservazione*, proprietà che non è condivisa dal modello statistico.

Nota Nella censura di I tipo, la statistica sufficiente nel caso esponenziale è $\big(\sum_{i=1}^{n} d_i, \sum_{i=1}^{n} y_i\big)$, mentre nel caso della censura di II tipo è solamente y_{tot}.

Lezione 22

22.1 Verosimiglianza profilo

Quando il parametro è vettoriale, spesso alcune sue componenti sono di maggiore interesse, mentre altre hanno l'effetto di rendere più complicata l'inferenza.

Si supponga che $\vartheta = (\psi, \lambda)$, con $\Theta = \Psi \times \Lambda$, dove

- ψ è un parametro *di interesse* di dimensione $k < p$
- λ è un parametro *di disturbo* di dimensione $p - k$.

Nota Per "disturbo" si intende il fatto che, se λ fosse noto, l'inferenza per ψ sarebbero in generale più semplici ed efficaci.

La stima di massima verosimiglianza di ψ è per definizione $\hat{\psi}$, dove $\hat{\vartheta} = (\hat{\psi}, \hat{\lambda})$.

La *verosimiglianza profilo* rimuove il problema della presenza di λ, utilizzando il valore più favorevole di λ fissato ψ:

$$L_p(\psi) = L(\psi, \hat{\lambda}_\psi),$$

dove $\hat{\lambda}_\psi$ è la stima di massima verosimiglianza di λ per ψ fissato. Analogamente, la log-verosimiglianza profilo è

$$\ell_p(\psi) = \log L_p(\psi).$$

È immediato verificare che $L_p(\psi)$ è massimizzata in $\hat{\psi}$.

Nota Nel calcolare le derivate, si utilizzerò la regola della derivata composta (*chain rule*), che nel caso unidimensionale diventa

$$\frac{\partial f(x_1(t), x_2(t))}{\partial t} = \frac{\partial x_1(t)}{\partial t} \left.\frac{\partial f(x_1, x_2)}{\partial x_1}\right|_{(x_1,x_2)=(x_1(t),x_2(t))} + \frac{\partial x_2(t)}{\partial t} \left.\frac{\partial f(x_1, x_2)}{\partial x_2}\right|_{(x_1,x_2)=(x_1(t),x_2(t))}$$

Indicando con ℓ_ψ, ℓ_λ i blocchi di ℓ_* e $\ell_{\psi\psi}$, $\ell_{\lambda\lambda}$, $\ell_{\psi\lambda}$ i blocchi di l_{**}, si può scrivere la *funzione di punteggio profilo*

$$\frac{\partial \ell_p(\psi)}{\partial \psi} = \ell_\psi(\psi, \hat{\lambda}_\psi) + \underbrace{\frac{\partial \hat{\lambda}_\psi}{\partial \psi} \ell_\lambda\big(\psi, \hat{\lambda}_\psi\big)}_{=0 \text{ eq. veros.}}$$

$$= l_\psi(\psi, \hat{\lambda}_\psi).$$

Nel caso in cui ψ sia k-dimensionale, si ha che

$$\ell_p(\psi) = \ell(\psi_1, \ldots, \psi_p, \hat{\lambda}_{\psi,1}, \hat{\lambda}_{\psi,p-k})$$

l'analogo della formula della derivata composta

$$\frac{\partial}{\partial \psi_r} \ell_p(\psi) = \ell_{\psi_r}(\psi, \hat{\lambda}_\psi) + \sum_{a=1}^{p-k} \frac{\partial \hat{\lambda}_{a,\psi}}{\partial \psi_r} \ell_{\lambda_a}(\psi, \hat{\lambda}_\psi).$$

In forma matriciale,

$$\frac{\partial}{\partial\psi}\ell_p(\psi) = \underbrace{l_\psi(\psi,\hat\lambda_\psi)}_{k\times 1} + \underbrace{\left(\frac{\partial\hat\lambda_\psi}{\partial\psi}\right)^{\mathsf{T}}}_{k\times(p-k)}\underbrace{\ell_\lambda(\psi,\hat\lambda_\psi)}_{(p-k)\times 1}$$

$$= l_\psi(\psi,\hat\lambda_\psi) \qquad \text{(regolarità)}$$

Con un'ulteriore derivata si ottiene l'*informazione osservata profilo* $j_p(\psi) = -\frac{\partial^2}{\partial\psi\partial\psi^{\mathsf{T}}}\ell_p(\psi)$ con generico elemento

$$\frac{\partial^2}{\partial\psi_r\partial\psi_s}\ell_p(\psi) = \ell_{\psi_r\psi_s}(\psi,\hat\lambda_\psi) + \sum_{a=1}^{p-k}\ell_{\psi_r\lambda_a}\frac{\partial\hat\lambda_{\psi,a}}{\partial\psi},$$

quindi in forma matriciale si ha

$$\frac{\partial^2}{\partial\psi\partial\psi^{\mathsf{T}}}\ell_p(\psi) = \ell_{\psi\psi}(\psi,\hat\lambda_\psi) + \ell_{\psi\lambda}(\psi,\hat\lambda_\psi)\frac{\partial\hat\lambda_\psi}{\partial\psi}.$$

In questo caso, si può utilizzare il fatto che $\ell_\lambda(\psi,\hat\lambda_\psi) = 0$ derivando entrambi i lati, per otteneere

$$\frac{\partial\ell_\lambda(\psi,\hat\lambda_\psi)}{\partial\psi} = \ell_{\lambda,\psi}(\psi,\hat\lambda_\psi) + \ell_{\lambda\lambda}(\psi,\hat\lambda_\psi)\frac{\partial\hat\lambda_\psi}{\partial\psi} = 0,$$

dunque sostituendo si ha

$$j_p(\psi) = -\frac{\partial^2}{\partial\psi\partial\psi^{\mathsf{T}}}\ell_p(\psi) = -\ell_{\psi\psi} + \ell_{\psi\lambda}\ell_{\lambda\lambda}^{-1}\ell_{\lambda\psi}.$$

Nella stima di massima verosimiglianza $j_p(\hat\psi)$, si ottiene la matrice di informazione osservata profilo

$$j_p(\hat\psi) = j_{\psi\psi} - j_{\psi\lambda}j_{\lambda\lambda}^{-1}j_{\lambda\psi},$$

dove $j_{\psi\psi}, j_{\psi\lambda}, j_{\lambda\lambda}$ sono i blocchi dell'informazione osservata $j(\vartheta)$ calcolati in $\hat\vartheta$.

Nei modelli regolari, la verosimiglianza profilo si può usare per costruire procedure inferenziali come se fosse, almeno approssimativamente, una verosimiglianza dedotta dal modello con il solo ψ.

$$\ell_p(\psi) \text{ non è in genere pari a } \log p\big(t(y);\psi\big).$$

Esempio (Modello Weibull)
Riprendendo l'esempio Weibull con parametro di interesse γ, si ottiene la stima di massima verosimiglianza

$$l_\beta(\vartheta) = 0 \implies \hat\beta_\gamma = \left(\frac{\sum_{i=1}^n y_i^\gamma}{n}\right)^{1/\gamma}.$$

Quindi,

$$\begin{aligned}\ell_p(\gamma) = \ell(\gamma, \hat{\beta}_\gamma) &= n\log\gamma - n\gamma\log\hat{\beta}_\gamma + \gamma\sum_{i=1}^n \log y_i - \hat{\beta}_\gamma^{-\gamma}\sum_{i=1}^n y_i^\gamma \\ &= n\log\gamma - n\cancel{\gamma}\frac{1}{\cancel{\gamma}}\log\sum_{i=1}^n y_i^\gamma/n + \gamma\sum_{i=1}^n \log y_i - \frac{n}{\cancel{\sum_{i=1}^n y_i^\gamma}}\cancel{\sum_{i=1}^n y_i^\gamma} \\ &= n\log\gamma - n\log\sum_{i=1}^n y_i^\gamma + \gamma\sum_{i=1}^n \log y_i - n + n\log n.\end{aligned}$$

La funzione punteggio profilo è

$$\frac{\partial\ell_p(\gamma)}{\partial\gamma} = \frac{n}{\gamma} - \frac{n}{\sum_{i=1}^n y_i^\gamma}\sum_{i=1}^n y_i^\gamma \log y_i + \sum_{i=1}^n \log y_i.$$

L'informazione osservata profilo si può ottenere dai blocchi di $j(\vartheta)$, oppure utilizzando la derivata della funzione punteggio profilo. Siccome

$$j(\hat{\vartheta}) = \begin{pmatrix} 0.52 & -0.025 \\ -0.025 & 0.011 \end{pmatrix},$$

si ha $j_p(\hat{\gamma}) = 0.52 - 0.011^{-1}(0.0245)^2 = 0.4694$. Infine, vale l'approssimazione

$$\ell_p(\gamma) - \ell_p(\hat{\gamma}) \approx -\frac{1}{2} j_p(\hat{\gamma})(\gamma - \hat{\gamma})^2.$$

22.2 Verosimiglianza e sufficienza

La verosimiglianza è direttamente legata al concetto di sufficienza, come mostrato nell'esempio seguente.

Esempio (Sufficienza del rapporto di verosimiglianza)
Si consideri il modello statistico più semplice possibile $\mathcal{F} = \{p(y;0), p(y;1)\}$, con uguale supporto per le due densità. Sotto il modello $\mathcal{F}$, il rapporto di verosimiglianza

$$t(y) = \frac{p(y;1)}{p(y;0)}$$

è una statistica sufficiente. Infatti, vale il criterio di fattorizzazione di Neyman-Fisher

$$\begin{aligned} p(y;\vartheta) &= \left(\frac{p(y;1)}{p(y;0)}\right)^\vartheta p(y;0), \quad \vartheta = 0, 1. \\ &= \begin{cases} p(y;1) & \text{se } \vartheta = 1 \\ p(y;0) & \text{se } \vartheta = 0 \end{cases}\end{aligned}$$

Inoltre, è anche minimale ma (più difficile) non è completa.

Più in generale, se $t = t(y)$ è sufficiente, si ha

$$L(\vartheta; y) = c(y) p_T\big(t(y); \vartheta\big),$$

per cui la verosimiglianza riferita al modello indotto da t e quella ottenuta dai dati originali sono equivalenti. Inoltre, in un modello parametrico una statistica t è sufficiente se e solo se la funzione di verosimiglianza dipende dai dati esclusivamente attraverso t.

Una statistica sufficiente minimale s esiste sempre ed è fornita dalla funzione di verosimiglianza: il procedimento costruttivo è la partizione di verisimiglianza, ovvero quando il rapporto di verosimiglianza è indipendente dal parametro:

$$L(\vartheta; y)/L(\vartheta; y') = c(y, y').$$

Se s è sufficiente minimale, la verosimiglianza dipende da y solo tramite s, dunque anche $\hat{\vartheta}$, se esiste, dipende da y solo attraverso s. Tuttavia, può non essere funzione biunivoca di s e quindi non essere a sua volta sufficiente minimale (e.g. modello autoregressivo, genetic linkage).

In questi casi, se si può scrivere s come $(\hat{\vartheta}, a)$ dove $a = a(y)$ è una statistica costante in distribuzione (*ancillare*), si può utilizzare una riduzione per condizionamento al modello $\hat{\vartheta}|a$ (Azzalini, 2001, §3.3.7).

22.3 Principi di verosimiglianza

La stretta connessione tra sufficienza e verosimiglianza porta al seguente principio:

Prop. (Principio debole di verosimiglianza)
Dati un modello statistico parametrico $\mathcal{F}$, due osservazioni y e y' con verosimiglianza equivalente, ovvero tali che $L(\vartheta; y) = c(y, y')L(\vartheta, y')$ per ogni $\vartheta \in \Theta$, devono portare alle stesse conclusioni inferenziali su ϑ.

L'inferenza frequentista basata sulla verosimiglianza soddisfa il principio debole di verosimiglianza, detto anche principio di sufficienza.

Anche l'inferenza bayesiana soddisfa automaticamente il principio, visto che la distribuzione a posteriori dipende dai dati unicamente attraverso la verosimiglianza. In realtà, soddisfa un principio più forte:

Prop. (Principio forte di verosimiglianza)
Dati due modelli parametrici $\mathcal{F}_1$ e $\mathcal{F}_2$ possibilmente diversi, con verosimiglianze $L_{\mathcal{F}_1}$ e $L_{\mathcal{F}_2}$. Due osservazioni, y da $\mathcal{F}_1$ e z da $\mathcal{F}_2$ tali che

$$L_{\mathcal{F}_1}(\vartheta; y) = c(y, z) L_{\mathcal{F}_2}(\vartheta; z) \quad \text{per ogni } \vartheta,$$

devono portare alle medesime conclusioni inferenziali su ϑ.

Questo principio non viene accettato nell'inferenza frequentista, in quanto essa dipende anche dallo schema di campionamento (e.g. regola di arresto). In ambito bayesiano, vale a meno che l'apriori non dipenda dal modello (e.g. a priori non informative à la Jeffreys).

22.4 Invarianza ed equivarianza di $L(\vartheta)$

Riferimenti Pace e Salvan, (1997, §2.11)

Invarianza rispetto a trasformazioni biunivoche dei dati

Sia $t = g(y)$ una trasformazione biunivoca di y $t : \mathcal{Y} \to \mathcal{T}$ e sia $\mathcal{F}_T = \{p_T(t;\vartheta), \vartheta \in \Theta \subseteq \mathbb{R}^p\}$ il modello indotto da $T = g(Y)$. L'informazione su ϑ contenuta in y è equivalente a quella contenuta in t e la verosimiglianza è coerente con questo fatto intuitivo:

$$p_T(t;\vartheta) = \begin{cases} p_Y(y(t);\vartheta) \cdot \left|\frac{\partial y(t)}{\partial t}\right| & y \text{ continua} \\ p_Y(y(t);\vartheta) & y \text{ discreta} \end{cases}$$

In entrambi i casi, visto che lo jacobiano non dipende da ϑ, vale che

$$L_{\mathcal{F}_T}(\vartheta; t(y)) = L_{\mathcal{F}}(\vartheta; y)c(y).$$

Equivarianza rispetto a riparametrizzazioni

Sia $\omega = \omega(\vartheta)$ una riparametrizzazione di $\mathcal{F}$, $\omega : \Theta \to \Omega$ diffeomorfismo di classe C^∞, allora questo corrisponde a un cambio di coordinate per il modello statistico. In particolare,

$$L^\Omega(\omega) = L^\Theta\big(\vartheta(\omega)\big) \implies \hat{\omega} = \omega(\hat{\vartheta}).$$

In generale, l'inferenza di verosimiglianza su ω è la traduzione nella nuova parametrizzazione delle conclusioni inferenziali su ϑ. In particolare, questo corrisponde alla richiesta che il seguente diagramma per le conclusioni inferenziali C^Θ e C^Ω sia commutativo:

$$\begin{array}{ccc} \Theta & \xrightarrow{\omega(\vartheta)} & \Omega \\ {\scriptstyle \pi(\vartheta)}\downarrow & & \downarrow{\scriptstyle \pi(\omega)} \\ C^\Theta & \xrightarrow{\omega(\vartheta)} & C^\Omega \end{array}$$

Lezione 23

23.1 Invarianza e parametri di disturbo

Pensando a una riparametrizzazione globale, non si mantiene in genere la distinzione tra componente di interesse e di disturbo. Per questo motivo, l'invarianza è richiesta alle sole riparametrizzazioni *che non alterano l'interesse*, ovvero della forma

$$\omega = \omega(\psi, \lambda) = (\tau, \zeta),$$

tali che $\tau = \tau(\psi)$ è una funzione biunivoca e $\zeta = \zeta(\psi, \lambda)$.

La funzione di verosimiglianza profilo è invariante per costruzione a riparametrizzazioni che non alterano l'interesse, cioè

$$L_p(\tau) = L_p\big(\psi(\tau)\big).$$

Esempio (Modello weibull)
La log-verosimiglianza profilo per $\psi = \gamma$ era

$$\ell_p(\gamma) = n \log \gamma - n \log \sum_{i=1}^{n} y_i^{\gamma} + \gamma \sum_{i=1}^{n} \log y_i.$$

La riparametrizzazione $\omega = (\tau, \zeta) = (\log \gamma, \beta^{-\gamma}) \in \mathbb{R} \times \mathbb{R}^+$. Dunque, corrisponde a una riparametrizzazione che non altera l'interesse, in quanto $\tau = \log \gamma \implies \gamma = e^{\tau}$. La log-verosimiglianza profilo per τ è

$$\ell_p(\tau) = \ell(\tau, \hat{\zeta}_\tau) = \ell\big(\gamma(\tau)\big) = n\tau - n \log \sum_{i=1}^{n} y_i^{e^{\tau}} + e^{\tau} \sum_{i=1}^{n} \log y_i.$$

Invece, la riparametrizzazione $\omega' = \left(\log \frac{\gamma}{\beta}, \beta^{-\gamma}\right)$ non è una riparametrizzazione che non altera l'interesse, perché entrambi sono funzioni di entrambi i parametri.

Aspetti computazionali

Nella maggior parte dei casi, la SMV deve essere ottenuta per via numerica, nel senso che $l_*(\vartheta) = 0$ è un sistema di p equazioni non lineari in p incognite.

L'algoritmo di Newton-Raphson costituisce un semplice metodo iterativo basato sul gradiente e sull'informazione osservata: dato lo sviluppo di Taylor attorno a ϑ_0, si ha

$$\ell_*(\vartheta) \approx \ell_*(\vartheta_0) + \ell_{**}(\vartheta_0)(\vartheta - \vartheta_0),$$

da cui si può ottenere la soluzione dell'equazione $\ell_*(\vartheta) = 0$ data da

$$\begin{aligned} \vartheta &= \vartheta_0 - \ell_{**}(\vartheta_0)^{-1} \ell_*(\vartheta_0) \\ &= \vartheta_0 + j(\vartheta_0)^{-1} \ell_*(\vartheta_0). \end{aligned}$$

Iterando la procedura, si ottiene la generica soluzione

$$\hat{\vartheta}_{(k)} = \hat{\vartheta}_{(k-1)} + j(\vartheta_{(k-1)})^{-1}\ell_{**}(\vartheta_{(k-1)}).$$

Come criterio di convergenza si usa in genere $|\ell_*(\vartheta_{(k+1)}) - \ell_*(\vartheta_{(k)})| < \varepsilon$, per una soglia ε scelta. Come valore iniziale, in genere si possono scegliere valori ottenuti da un metodo alternativo (e.g. metodo dei momenti), oppure si utilizzano tanti valori iniziali e si controlla la stabilità della stima.

Esempio (NR per un modello Gamma)
Dato $Y_1, Y_2, \ldots, Y_n$ da Gamma$(\vartheta, 1)$, la log- verosimiglianza è

$$\ell(\vartheta) = \vartheta \sum_{i=1}^{n} \log y_i - n \log \Gamma(\vartheta),$$

per cui $l_*(\vartheta) = \sum_{i=1}^{n} \log y_i - n\Psi(\vartheta)$, con $\Psi(\vartheta)$ e $\Psi'(\vartheta)$ funzioni digamma e trigamma.

L'equazione di verosimiglianza $l_*(\vartheta) = 0$ non ha soluzione esplicita e la generica iterazione di NR è

$$\hat{\vartheta}_{(k+1)} = \hat{\vartheta}_{(k)} + \frac{\sum_{i=1}^{n} \log y_i - n\Psi(\hat{\vartheta}_{(k)})}{n\Psi'(\hat{\vartheta}_{(k)})}.$$

In particolare, scegliendo il punto iniziale con il metodo dei momenti si converge più velocemente alla stima di massima verosimiglianza.

Osservazione

Tutte le quantità di verosimiglianza sono di interesse anche per l'inferenza bayesiana, in particolare la SMV $\hat{\vartheta}$ può essere vista come moda a posteriori quando si utilizza una distribuzione a priori costante per ϑ.

23.2 Proprietà campionarie esatte

Riferimenti Pace e Salvan (2001, §5.4-5.5, §7.2-7.3)
Azzalini (2001, §3.2, §4.2.1)
Liseo (2010, §3.2.1)

Si consideri un modello statistico parametrico $\mathcal{F}$ con densità $p(y; \vartheta)$, $\vartheta \in \Theta \subseteq \mathbb{R}^p$ e funzione di verosimiglianza $L(\vartheta)$. Abbiamo visto che il rapporto di verosimiglianza è sufficiente minimale se $\Theta = \{\vartheta_1, \vartheta_2\}$.

È importante osservare che tutte le quantità di verosimiglianza sono funzioni dei dati osservati

$$\hat{\vartheta}(y^{\text{oss}})$$

$$\ell(\vartheta; y^{\text{oss}})$$

$$j(\vartheta; y^{\text{oss}})$$

$$\ell_p(\psi; y^{\text{oss}})$$

$$\vdots$$

La funzione di verosimiglianza osservata rappresenta una sintesi dei dati a una statistica sufficiente minimale, e mostra un comportamento coerente per trasformazioni biettive dei dati o di ϑ.

La funzione di verosimiglianza osservata è anche la base per calcolare la distribuzione a posteriori, in particolare

$$\log \pi(\vartheta|y^{\text{oss}}) = \ell(\vartheta; y^{\text{oss}}) + \log \pi(\vartheta).$$

Le quantità di verosimiglianza osservate sono utili per studiare $\pi(\vartheta|y^{\text{oss}})$, in particolare quando la numerosità campionaria è elevata e $\log \ell(\vartheta)$ domina $\log \pi(\vartheta)$ in $\log \pi(\vartheta|y^{\text{oss}})$.

Per l'inferenza frequentista, le procedure inferenziali si basano sul principio del campionamento ripetuto. Questo vale per

(a) *Inferenza fisheriana*, dove l'interesse principale è studiare la distribuzione di quantità basate su $L(\vartheta)$, come $\hat{\vartheta}(Y)$, $L(\vartheta; Y)/L\big(\hat{\vartheta}(Y); Y\big)$ etc.

(b) *Inferenza frequentista decisionale*, dove l'obiettivo è individuare procedure con proprietà di ottimalità, come regioni di confidenza e test.

Tanto in (a) quanto in (b), si possono valutare

- Proprietà esatte: proprietà valide sotto condizioni di regolarità per qualunque dimensione campionaria n.
- Proprietà asintotiche: proprietà valide per $n \to \infty$, utili per approssimare distribuzioni che sarebbero altrimenti intrattabili.

Esempio (Modello binomiale)
Nell'esempio del basket e dei tiri liberi, si supponga che $\vartheta^0 = 0.5$ sia il vero valore del parametro. Allora,

$$P_{\vartheta^0}(Y = y) = \binom{10}{y} (\vartheta^0)^y (1 - \vartheta^0)^{10-y} = \binom{10}{y} 0.5^{10}.$$

In particolare, al variare di y secondo la distribuzione di Y, varia anche $L(\vartheta) = L(\vartheta; Y)$, e dunque l'andamento casuale di $L(\vartheta; Y)$ dipende dalla distribuzione di Y. In questo caso, ci

sono 11 possibili valori di Y e dunque 11 possibili funzioni $L(\vartheta; Y)$. In questo esempio,

$$P_{\vartheta^0}\big(\hat{\vartheta}(y) = \vartheta^0\big) > P_{\vartheta^0}\big(\hat{\vartheta}(Y) = \vartheta\big), \quad \vartheta \neq \vartheta^0.$$

Quando si scrive una funzione di verosimiglianza, il suo comportamento casuale varia al variare del vero valore del parametro.

Lezione 24

24.1 Test ottimo in un modello con due elementi

In primo luogo, mostriamo che in un modello statistico con 2 soli elementi, il rapporto di verosimiglianza è anche lo strumento migliore per scegliere tra i due elementi sulla base di y. La statistica rapporto di verosimiglianza, oltre a essere una statistica sufficiente, ha anche in sé delle proprietà di *efficienza*.

Sia $\mathcal{F} = \{p(y;0), p(y;1)\}$, con entrambe le distribuzioni aventi lo stesso supporto $\mathcal{Y}$. Il problema di inferenza entro $\mathcal{F}$ si può vedere come problema di verifica di ipotesi

$$\begin{cases} H_0 : \vartheta = 0 \\ H_1 : \vartheta = 1 \end{cases}$$

Sia R la regione di $\mathcal{Y}$ per cui si sceglie ϑ_1, mentre $A = \mathcal{Y} \setminus R$ per cui si sceglie ϑ_0. Gli errori possibili sono

Tabella 5: caption

Scelta	ϑ_0	ϑ_1
A	✓	β
R	α	✓

Guardiamo allora due criteri per valutare la scelta di R sulla base del principio del campionamento ripetuto

i) $\min\limits_{\alpha,\beta} \alpha + \beta = \min P_{\vartheta_0}(Y \in R) + P_{\vartheta_1}(Y \in A)$;

ii) $\min P_{\vartheta_1}(Y \in A)$ vincolato a $P_{\vartheta_0}(Y \in R) \leq \alpha$.

Adottando il criterio i) e assumendo Y continua, la quantità da minimizzare è

$$\begin{aligned} \min_R \{P_{\vartheta_0}(Y \in R) + P_{\vartheta_1}(Y \in A)\} &= \min_R \{P_{\vartheta_0}(Y \in R) + 1 - P_{\vartheta_1}(Y \in R)\} \\ &= \min_R \left\{1 + \int_R p_0(y)\, dy - \int_R p_1(y)\, dy\right\} \\ &= \min_R \left\{1 + \int_R \big(p_0(y) - p_1(y)\big)\, dy\right\} \end{aligned}$$

La minimizzazione si ottiene includendo in R tutti i punti tali che $p_0(y) < p_1(y)$, scegliendo cioè

$$R^{\text{smv}} = \{y \in \mathcal{Y} : p_0(y) < p_1(y)\}\,.$$

Infatti, se anche un solo punto in R fosse tale che $p_0(y) > p_1(y)$, allora la quantità nell'integrale

sarebbe positiva e il valore della funzione obiettivo aumenterebbe. Il test sceglie dunque

$$\begin{cases} \vartheta_1 & \text{se } \dfrac{p_1(y)}{p_0(y)} = \dfrac{L(\vartheta_1; y)}{L(\vartheta_0; y)} > 1 \implies \hat{\vartheta} = \vartheta_1 \\ \vartheta_0 & \text{se } \dfrac{p_1(y)}{p_0(y)} = \dfrac{L(\vartheta_1; y)}{L(\vartheta_0; y)} < 1 \implies \hat{\vartheta} = \vartheta_0 \end{cases}$$

Pertanto, la scelta basata sulla SMV è quella più efficiente secondo il criterio i). L'adozione del criterio *ii*) porta al risultato fondamentale sui test ottimi (Neyman e Pearson, 1933):

Lemma (di Neyman e Pearson)

Per un test che verifica $H_0 : \vartheta = \vartheta_0$ contro $H_1 : \vartheta = \vartheta_1$, un test con livello α e regioni di rifiuto della forma

$$R^* = \left\{ y \in \mathcal{Y} : \frac{p_{\vartheta_1}(y)}{p_{\vartheta_0}(y)} > k \right\}$$

ha potenza $P_{\vartheta_1}(Y \in R)$ massima tra tutti i test con livello non superiore ad α.

Dim.

Si considerino la regione R^* definita nel teorema sopra e un qualunque altro test con regione di rifiuto R e livello $\leq \alpha$:

$$P_{\vartheta_0}(Y \in R) \leq P_{\vartheta_0}(Y \in R^*) = \alpha$$

Bisogna dimostrare, da questa ipotesi, che $P_{\vartheta_1}(Y \in R^*) \geq P_{\vartheta_1}(Y \in R)$. Siano

$$\overline{R}^* = R^* \setminus R \subseteq R^* \qquad \text{(arancione)}$$

$$\overline{A}^* = R \setminus R^* \subseteq A^* \qquad \text{(blu)}$$

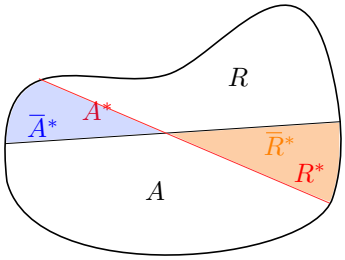


Figura 18: partitionNeyman

Si ha che

$$R^* = (R^* \cap R) \cup \overline{R}^*$$

$$R = (R^* \cap R) \cup \overline{A}^*$$

Poiché per ipotesi $P_{\vartheta_0}(Y \in R) \leq P_{\vartheta_0}(Y \in R^*)$, si ha necessariamente

$$P_{\vartheta_0}(Y \in \overline{A}^*) \leq P_{\vartheta_0}(Y \in \overline{R}^*).$$

Moltiplicando per $k > 0$ entrambi i membri,

$$kP_{\vartheta_0}(Y \in \overline{A}^*) \leq kP_{\vartheta_0}(Y \in \overline{R}^*) \qquad (*)$$

Dal momento che $\overline{A}^* \subseteq A^*$ ed essendo $p_1(y) < kp_0(y)$ per ogni $y \in A^*$, vale che

$$P_{\vartheta_1}(Y \in \overline{A}^*) \leq kP_{\vartheta_0}(Y \in \overline{A}^*). \qquad (1)$$

Analogamente, essendo $\overline{R}^* \subseteq R^*$ e $p_1(y) > kp_0(y)$ per $y \in R^*$,

$$P_{\vartheta_1}(Y \in \overline{R}^*) \geq kP_{\vartheta_0}(Y \in \overline{R}^*). \qquad (2)$$

Mettendo insieme le disuguaglianze $(*)$, (1) e (2),

$$P_{\vartheta_1}(Y \in \overline{A}^*) \leq P_{\vartheta_1}(Y \in \overline{R}^*),$$

ma queste regioni sono esattamente quelle che servono per passare dalla regione R^* a R, in particolare sommando $P\big(Y \in (R^* \cap R)\big)$ a entrambi i membri, si ottiene

$$P_{\vartheta_1}(Y \in R) \leq P_{\vartheta_1}(Y \in R^*),$$

che dimostra la tesi.

□

Commento

Anche secondo il criterio *ii*), la regione di rifiuto del test ottimo è determinata dal rapporto di verosimiglianza

$$R^* = \left\{ y \in \mathcal{Y} : \frac{p_1(y)}{p_0(y)} = \frac{L(1;y)}{L(0;y)} > k \right\},$$

per una certa soglia k, fissata in modo che il test abbia livello α. Il test così ottenuto è il più potente (*MP, Most Powerful*) con livello α per H_0 semplice contro H_1 semplice.

Esempio (Esercizio 7.1 Pace e Salvan, 2001)
Sia y osservazione da Y, si consideri il test che verifica il sistema di ipotesi $H_0 : Y \sim \text{Pois}(1)$ contro $H_1 : \underbrace{Y-1}_{\mathcal{Y}=\mathbb{N}} \sim \text{Geom}(\frac{1}{2})$.

Si ottenga l'espressione della statistica rapporto di verosimiglianza, si mostri che $R^* = \{4, 5, \ldots\}$ e si calcolino livello di significatività e potenza del test.

Si ha $p_0(y) = e^{-1}/y!$ e $p_1(y) = 1/2^{y+1}$, per cui il rapporto di verosimiglianza è

$$t^*(y) = \frac{e}{2}\frac{y!}{2^y} = c \cdot t(y).$$

Dalla tabella dei valori di $t(y)$, si osserva che, includendo $y = 3$ nell'insieme di rifiuto, si contraddirebbe la costruzione del test rapporto di verosimiglianza: $t^*(3) < t^*(1)$, quindi non

si conterrebbero tutti i punti con rapporto superiore ad una certa soglia.

Tabella 6: caption

y	0	1	2	3	4	5	6
$t(y)$	1	0.5	0.5	0.75	1.5	3.75	11.25

Sotto H_0 e H_1, la distribuzione di $t(Y)$ e rispettivamente $p_T(t;0)$ e $p_T(t;1)$:

Tabella 7: caption

t	0.5	0.75	1	1.5	3.75	11.25	...
$p_T(t;0)$	0.5518	0.0613	0.3679	0.0153	0.0031	0.0005	...
$p_T(t;1)$	0.375	0.0625	0.5	0.03125	0.0156	0.0078	...

Il livello di significatività è $P_0(Y \in R^*) = P\big(\text{Pois}(1) \geq 4\big) = P_0\big(t(Y) \geq 1.5\big) \approx 0.019$. Per la potenza, si ha lo stesso calcolo sotto H_1, $P\big(\text{Geom}(1/2) - 1 \geq 3\big) \approx 0.0625$.

Nota 1 Spesso, si possono individuare funzioni monotone $t(y) = g(t^*)$, per cui è equivalente scrivere

$$t^*(y) \geq k \implies t(y) \geq g^{-1}(k) = k'.$$

Tabella 8: caption

y	0	1	2	3	4	5	6
$t(y)$	1	0.5	0.5	0.75	1.5	3.75	11.25

Nota 2 Nel caso delle distribuzioni discrete, la funzione di ripartizione è a gradini e in generale non si possono ottenere tutti i livelli di significatività $\alpha \in (0, 1)$.

Esercizio: Verificare che per ogni α raggiungibile, il test ottimo è *non distorto*, ovvero

$$P_0(Y \in R^*) \leq P_1(Y \in R^*).$$

Teo. (Non distorsione del test ottimo)
Il test ottimo con livello α è sempre non distorto, cioè

$$P_1(Y \in R^*) \geq P_0(Y \in R^*).$$

Dim.
Consideriamo i tre casi: $k = 1$, $k > 1$, $0 < k < 1$ e Y discreta, senza perdita di generalità.
$\boxed{k = 1}$: $p_1(y) > p_0(y)$, dunque

$$P_0(Y \in R^*) = \sum_{y \in R^*} p_0(y) < \sum_{y \in R^*} p_1(y) = P_1(Y \in R^*).$$

$\boxed{k > 1}$: $p_1(y) > kp_0(y) > p_0(y)$, dunque

$$P_0(Y \in R^*) = \sum_{y \in R^*} p_0(y) < \sum_{y \in R^*} kp_0(y) < \sum_{y \in R^*} p_1(y) = P_1(Y \in R^*).$$

$\boxed{k \in (0,1)}$: in A^* vale che $p_1(y) \leq kp_0(y) < p_0(y)$.

$$P_0(Y \in A^*) = \sum_{y \in A^*} p_0(y) > \sum_{y \in A^*} kp_0(y) \geq \sum_{y \in R^*} p_1(y) = P_1(Y \in A^*).$$

□

Esempio (Test ottimo in $\mathcal{F}$ con leggi continue)

Sia y osservazione da Y e supponiamo di voler verificare $H_0 : Y \sim \text{Exp}(1)$ contro $H_1 : Y \sim \text{Unif}(0,1)$. Mostrare che il test ottimo ha regione di rifiuto $R^* = (c, 1]$, $0 \leq c < 1$. Calcolare inoltre il livello di significatività e mostrare che per $c = 0$, il test ha livello $1 - e^{-1}$. Mostrare anche che per un livello $\alpha > 1 - e^{-1}$ si possono definire più regioni di rifiuto con potenza 1, quindi più test ottimi.

Si noti che i due modelli non hanno lo stesso supporto:

$$S_0 = [0, +\infty)$$

$$S_1 = [0, 1]$$

Se $y > 1$, non può valere H_1 e quindi osservazioni in $(1, +\infty)$ portano a scegliere H_0 con contributo all'errore di secondo tipo nullo:

$$P_1(Y > 1) = 0.$$

Il rapporto di verosimiglianza è

$$t^*(y) = \frac{p_1(y)}{p_0(y)} = e^y \mathbb{1}_{[0,1]}(y)$$

che è funzione crescente di y e quindi valori di y in $(c, 1] = R^*$ portano a valori grandi di $t^*(y)$.

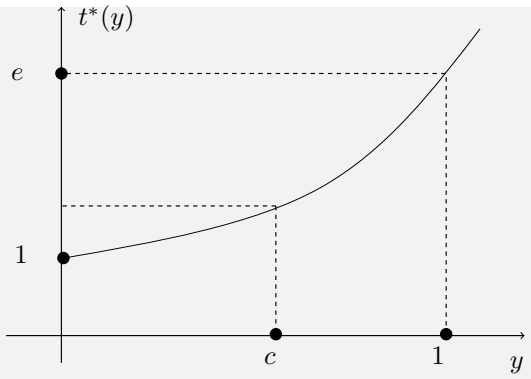


Figura 19: exTRV

Il livello del test è

$$P_0\big(Y \in (c,1]\big) = F_0(1) - F_0(c) = e^{-c} - e^{-1}.$$

La potenza è

$$P_1\big(Y \in (c,1]\big) = 1 - c.$$

Se $c = 0$, si ha livello $\alpha = P_0\big(Y \in (0,1]\big) = 1 - e^{-1}$. Per l'ultima parte, si può scegliere

$$R = [0,1] \cup [a,b], \quad 1 < a < b,$$

in particolare risulta

$$P_0(Y \in R) > 1 - e^{-1}, \quad P_1(Y \in R) = P_1\big(Y \in [0,1]\big) + P_1\big(Y \in [a,b]\big)$$

Lezione 25

Esempio (Test ottimo per famiglie esponenziali)
Sia $y = (y_1, y_2, \ldots, y_n)$ un c.c.s. con densità congiunta

$$p(y;\vartheta) = c(\vartheta)^n \prod_{i=1}^{n} h(y_i) \exp\left\{\psi(\vartheta) t^{(n)}(y)\right\}, \quad \vartheta \in \Theta.$$

Si desidera verificare $H_0 : \vartheta = \vartheta_0$ contro $H_1 : \vartheta = \vartheta_1$. Risulta allora il rapporto di verosimiglianza

$$t^*(y) = \frac{c(\vartheta_1)^n}{c(\vartheta_0)^n} \exp\left\{\big(\psi(\vartheta_1) - \psi(\vartheta_0)\big) t^{(n)}(y)\right\}.$$

Supponendo che $\vartheta_1 > \vartheta_0$ e che $\psi(\cdot)$ monotona crescente, allora $\psi(\vartheta_1) - \psi(\vartheta_0) > 0$ e $t^*(y)$ è funzione monotona crescente di $t^{(n)}(y)$. Il test ottimo con livello α è allora identificato dalla regione di rifiuto

$$R^* = \left\{y \in \mathcal{Y} : t^{(n)} > k\right\},$$

dove k è tale che $P_{\vartheta_0}\left\{t^{(n)}(Y) > k\right\} = \alpha$. Se $t^{(n)}(Y)$ è discreta, non si possono raggiungere esattamente tutti i livelli α.

Nota 1 Se $\psi(\cdot)$ è monotona decrescente, la zona di rifiuto sarà per valori piccoli di $t^{(n)}(Y)$.

Nota 2 Con $\psi(\cdot)$ monotona, la definizione di R^* non dipende da ϑ_1, ma solo dal fatto che ϑ_1 sia maggiore o minore di ϑ_0.

Dall'esempio, con $\psi(\cdot)$ monotona, il test con regione di rifiuto

$$R^* = \left\{y \in \mathcal{Y} : t^{(n)} > (<) k\right\}$$

è più potente per $H_0 : \vartheta = \vartheta_0$ contro l'ipotesi composita $H_1 : \vartheta > \vartheta_0$. Il test è allora *uniformemente più potente* (*UMP*) tra i test di livello α.

In realtà, il test è UMP anche per $H_0 : \vartheta \leq \vartheta_0$ contro $H_1 : \vartheta > \vartheta_0$. Infatti, per $\vartheta < \vartheta_0$ e $\psi(\cdot)$ crescente, si ha

$$P_\vartheta\big(t^{(n)}(Y) > k\big) \leq P_{\vartheta_0}\big(t^{(n)}(Y) > k\big)$$

per la non distorsione del test. Analogamente, vale per le disuguaglianze in senso opposto se $H_0 : \vartheta \geq \vartheta_0$ contro $H_1 : \vartheta < \vartheta_0$.

Determinare k nella pratica richiede di conoscere la distribuzione di $t^{(n)}(Y)$ quando $\vartheta = \vartheta_0$.

Esercizio: Per ciascuna delle famiglie esponenziali per cui è riportata la statistica sufficiente minimale a pagina 5.24, indicare come calcolare la soglia critica k.

Esempio (Test per Unif$(0, \vartheta)$)
Dato un c.c.s. da Unif$(0, \vartheta)$, sia $H_0 : \vartheta = 1$ e $H_1 : \vartheta = 2$. Si mostri che il test ottimo rifiuta H_0 per valori grandi di $y_{(n)}$ e si dica se il test è UMP per $H_0 : \vartheta \leq 1$ contro $H_1 : \vartheta > 1$. Determinare n affinché la potenza in $\vartheta = 2$ del test con livello 0.05 sia almeno pari a 0.95.

La densità del campione è

$$p(y; \vartheta) = \frac{1}{\vartheta^n} \mathbb{1}_{[0,+\infty)}(y_{(1)}) \mathbb{1}_{[0,\vartheta]}(y_{(n)}).$$

Con il modello correttamente specificato, entrambe le funzioni indicatrici sono pari a 1 e il rapporto di verosimiglianza è

$$t^*(y) = \begin{cases} \frac{1}{2^n} & \text{se } y_{(n)} \leq 1 \\ +\infty & \text{se } y_{(n)} > 1 \end{cases}$$

funzione monotona crescente di $y_{(n)}$ qualunque valore di ϑ sotto H_1 purché maggiore di $\vartheta_0 = 1$. Dunque il test è UMP (stesso discorso di prima).

Il test ottimo rifiuta H_0 per $y_{(n)} > c$, da determinare per avere livello α: per $z \in (0, 1)$, si ha

$$P_{\vartheta_0}(Y_{(n)} \leq z) = z^n \implies 1 - c^n = \alpha \iff c = (1 - \alpha)^{1/n}.$$

Per n generico, la potenza è

$$P_{\vartheta_1}(Y_{(n)} > (1 - \alpha)^{1/n}) = 1 - \Big[\frac{(1 - \alpha)^{1/n}}{\vartheta_1}\Big]^n.$$

Con $1 - \alpha = 0.95$ e $\vartheta_1 = 2$, si trova

$$1 - \frac{0.95}{2^n} \geq 0.95 \iff 2^n \geq \frac{0.95}{0.05} \iff n \geq \frac{\log 19}{\log 2} \approx 4.25.$$

Il test ha livello α anche per $H_0 : \vartheta \leq 1$, in quanto

$$\begin{aligned} P_\vartheta(Y_{(n)} \geq c) &= \begin{cases} 0 & \text{se } c \geq \vartheta \\ 1 - \left(\frac{c}{\vartheta}\right)^n & \text{se } c < \vartheta \end{cases} \\ &\leq 1 - c^n \\ &= \alpha \quad \text{per } \vartheta \leq 1. \end{aligned}$$

25.1 Test ottimi (UMP) per ipotesi composite unilaterali

Si può generalizzare il risultato al di fuori delle famiglie esponenziali, per $H_0 : \vartheta \leq \vartheta_0$ contro $H_1 : \vartheta > \vartheta_0$, in modelli monoparametrici in cui

$$\vartheta_1 < \vartheta_2 \implies \frac{p(y;\vartheta_2)}{p(y;\vartheta_1)} = t^*(y;\vartheta_1,\vartheta_2) \quad \text{funzione monotona di } t(y).$$

In tal caso, la famiglia $\mathcal{F} = \{p(y;\vartheta), \vartheta \in \Theta \subseteq \mathbb{R}\}$ è detta *con rapporto di verosimiglianza monotono.*

Se ad esempio $t^*(y;\vartheta_1,\vartheta_2) = h\big(t(y);\vartheta_1,\vartheta_2\big)$ monotona crescente, il test MP ha regione di rifiuto

$$\begin{aligned} R^* &= \{y \in \mathcal{Y} : t^*(y;\vartheta_1,\vartheta_2) > k\} \\ &= \{y \in \mathcal{Y} : t(y) > k'\} \end{aligned}$$

con k' tale da avere livello α. Il test con R^* è UMP anche per

$$H_0 : \vartheta = \vartheta_1, \quad H_1 : \vartheta > \vartheta_1$$

$$H_0 : \vartheta \leq \vartheta_1, \quad H_1 : \vartheta > \vartheta_1$$

Oltre alle famiglie esponenziali con $\psi(\cdot)$ monotona, questo risultato vale per Unif$(0,\vartheta)$, l'ipergeometrica, t Student, χ^2, F non centrale con gdl fissati, ...

25.2 Test ottimi per $H_0 : \vartheta = \vartheta_0$ contro $H_1 : \vartheta \neq \vartheta_0$

Anche in una famiglia esponenziale monoparametrica, non esiste test UMP con livello α per questo tipo di test. Infatti, con $\psi(\cdot)$ monotona crescente,

$$H_0 : \vartheta = \vartheta_0, \quad H_1 : \vartheta > \vartheta_0 \implies R^*_{dx} = \Big\{y \in \mathcal{Y} : t^{(n)}(y) > k_{dx}\Big\}$$

$$H_0 : \vartheta = \vartheta_0, \quad H_1 : \vartheta < \vartheta_0 \implies R^*_{sx} = \Big\{y \in \mathcal{Y} : t^{(n)}(y) < k_{sx}\Big\}$$

Quindi, essendo $R^*_{dx} \neq R^*_{sx}$, non esiste il test UMP con livello α.

La soluzione "matematica" è aggiungere la condizione di non distorsione:

$$\min \pi(\vartheta) = P_\vartheta(Y \in R^*), \quad \pi(\vartheta_0) = \alpha.$$

Si dimostra allora che nelle famiglie esponenziali monoparametriche esiste il test UMPU (*Uniformly Most Powerful Unbiased*) e ha regione di rifiuto

$$R^* = \Big\{y \in \mathcal{Y} : t^{(n)} > t_{1-\alpha'}\Big\} \cup \Big\{y \in \mathcal{Y} : t^{(n)} > t_{\alpha''}\Big\}$$

tale che $\alpha' + \alpha'' = \alpha$, $\pi(\vartheta_0) = 0$. La costruzione è molto complicato, a meno che $t(Y)$ non abbia distribuzione simmetrica sotto H_0, in qual caso si sceglie il test bilanciato.

In una famiglia esponenziale monoparametrica, questo test è equivalente a quello basato sulla verosimiglianza che vedremo più avanti.

25.3 Disuguaglianza di Wald

Teo. (Disuguaglianza di Wald)
Sotto tenui condizioni di regolarità, se ϑ è identificabile per $\mathcal{F}$ e ϑ^0 è il vero valore del parametro, allora

$$\mathbb{E}_{\vartheta^0}\left[\ell(\vartheta;Y)\right] \leq \mathbb{E}_{\vartheta^0}\left[\ell(\vartheta^0;Y)\right], \quad \vartheta \neq \vartheta^0.$$

Dim.
ϑ identificabile $\implies P_{\vartheta^0}\left(\frac{p(Y;\vartheta)}{p(Y;\vartheta^0)} = 1\right) < 1$, dunque la v.c. $\frac{p(Y;\vartheta)}{p(Y;\vartheta^0)}$ è non degenere per ogni coppia (ϑ^0, ϑ), $\vartheta \neq \vartheta^0$.

Inoltre, poiché $\log(\cdot)$ è una funzione strettamente concava, per la disuguaglianza di Jensen

$$\mathbb{E}_{\vartheta^0}\left[\log\frac{p(Y;\vartheta)}{p(Y;\vartheta^0)}\right] < \log\mathbb{E}_{\vartheta^0}\left[\frac{p(Y;\vartheta)}{p(Y;\vartheta^0)}\right] = \log 1 = 0.$$

□

Questo implica anche che il valore atteso di $\ell(\vartheta;Y)$ è massimo nel vero valore del parametro.

Esempio (Distribuzione binomiale)
Si consideri $\text{Bin}(10,\vartheta)$ per Y, allora la funzione di log-verosimiglianza attesa è

$$\begin{aligned}\mathbb{E}_{\vartheta^0}\left[\ell(\vartheta;Y)\right] &= \mathbb{E}_{\vartheta^0}\left[Y\log\vartheta + (10-Y)\log(1-\vartheta)\right]\\ &= 10\left(\vartheta^0\log\vartheta + (1-\vartheta^0)\log(1-\vartheta)\right).\end{aligned}$$

Se $\vartheta^0 = 0.5$, $\mathbb{E}_{0.5}\left[\ell(\vartheta;Y)\right] = 5\big(\log\vartheta + \log(1-\vartheta)\big)$.

Nota La quantità

$$D(\vartheta,\vartheta^0) = -\,\mathbb{E}_{\vartheta^0}\left[\log\frac{p(Y;\vartheta)}{p(Y;\vartheta^0)}\right] = \mathbb{E}_{\vartheta^0}\left[\log\frac{p(Y;\vartheta^0)}{p(Y;\vartheta)}\right]$$

è positiva per $\vartheta \neq \vartheta^0$ e si chiama *divergenza di Kullback-Leibler* di $p(y;\vartheta)$ da $p(y;\vartheta^0)$. Si chiama divergenza in quanto non è una funzione simmetrica.

Il vero valore del parametro minimizza $D(\vartheta,\vartheta^0)$ rispetto a ϑ, in quanto

$$D(\vartheta,\vartheta^0) = \underbrace{\mathbb{E}_{\vartheta^0}\left[\log p(Y;\vartheta^0)\right]}_{\text{const.}} - \mathbb{E}_{\vartheta^0}\left[\log p(Y;\vartheta)\right],$$

il che equivale a dire che massimizza $\mathbb{E}_{\vartheta^0}\left[\log p(Y;\vartheta)\right]$, il cui equivalente empirico è

$$\frac{1}{n}\sum_{i=1}^{n}\log p(y_i;\vartheta) = \frac{1}{n}\ell(\vartheta;y).$$

Dunque, il metodo della massima verosimiglianza massimizza a livello campionario la divergenza di Kullback-Leibler.

Lezione 26

26.1 Proprietà esatte della verosimiglianza

Le proprietà discusse in seguito richiedono che il modello $\mathcal{F}$ sia con verosimiglianza regolare e due ulteriori assunzioni:

- $\mathcal{Y}$ indipendente da ϑ.
- Si possono scambiare le operazioni di derivazione rispetto a ϑ e integrazione rispetto a y.

Nel seguito, si indica con ϑ il vero valore del parametro.

Teo. (Valore atteso di $\ell_*(\vartheta; Y)$)
Nel vero valore del parametro, il valore atteso dello score è nullo, cioè

$$\mathbb{E}_\vartheta \left[l_*(\vartheta; Y) \right] = 0.$$

Dim.
Nel caso continuo, per ogni $\vartheta \in \Theta$

$$\int_{\mathcal{Y}} p(y; \vartheta)\, dy = 1.$$

Derivando rispetto a ϑ_r, $r = 1, \ldots, p$ e scambiando integrale con derivata,

$$\frac{\partial}{\partial \vartheta_r} \int_{\mathcal{Y}} p(y; \vartheta)\, dy = \int_{\mathcal{Y}} \frac{\partial}{\partial \vartheta_r} p(y; \vartheta)\, dy = 0.$$

Usando il fatto che

$$\frac{\partial}{\partial \vartheta_r} p(y; \vartheta) = \left(\frac{\partial}{\partial \vartheta_r} \log p(y; \vartheta) \right) p(y; \vartheta),$$

si ha che

$$\int_{\mathcal{Y}} \left(\frac{\partial}{\partial \vartheta_r} \log p(y; \vartheta) \right) p(y; \vartheta)\, dy = \mathbb{E}_\vartheta \left[l_*(\vartheta; Y) \right] = 0.$$

□

Nota L'equazione di verosimiglianza $l_*(\vartheta; y) = 0$ si dice *equazione di stima non distorta*, in quanto vale $\mathbb{E}_\vartheta \left[l_*(\vartheta; Y) \right] = 0$. Più in generale, si dice equazione di stima non distorta qualunque equazione

$$q(\vartheta; y) = 0, \quad \text{con } \mathbb{E}_\vartheta \left[q(\vartheta; Y) \right] = 0.$$

Teo. (Identità dell'informazione)

$$\mathbb{E}_\vartheta \left[\ell_*(\vartheta; Y) \ell_*(\vartheta; Y)^\mathsf{T} \right] = i(\vartheta) = \mathbb{E}_\vartheta \left[j(\vartheta; Y) \right],$$

dove $i(\vartheta)$ si dice ***informazione attesa di Fisher****.*

Osservazione

Con questa proprietà si collega lo score alla matrice di informazione osservata; inoltre, $i(\vartheta)$ è semidefinita positiva, dunque nel vero valore del parametro c'è in media un massimo globale.

Dim.
Riprendendo quanto dimostrato in precedenza e derivando nuovamente, si ottiene

$$\int_{\mathcal{Y}} \frac{\partial}{\partial \vartheta_s} \big[\ell_r(\vartheta; y) p(y; \vartheta)\big]\, dy = 0.$$

Espandendo la derivata del prodotto,

$$\frac{\partial}{\partial \vartheta_s} = \ell_{rs}(\vartheta) p(y; \vartheta) + \ell_r(\vartheta) \underbrace{\ell_s(\vartheta) p(y; \vartheta)}_{\frac{\partial}{\partial \vartheta_s} p(y;\vartheta)},$$

per cui

$$\int_{\mathcal{Y}} \ell_{rs}(\vartheta) p(y; \vartheta)\, dy + \int_{\mathcal{Y}} \ell_r(\vartheta; y) \ell_s(\vartheta; y) p(y; \vartheta)\, dy = 0$$

$$\iff i_{rs}(\vartheta) = \mathrm{Cov}_\vartheta \big(\ell_r(\vartheta; Y) \ell_s(\vartheta; Y)\big).$$

□

26.2 Quantità di verosimiglianza per famiglie esponenziali

Usando queste due proprietà, si possono ricavare proprietà interessanti per le famiglie esponenziali.

Esercizio: Mostrare il teorema sul valore atteso della famiglia esponenziale dato a pag. ...

Si consideri la distribuzione della statistica sufficiente minimale $t = t^{(n)}(y)$ con densità

$$p_T(t; \psi) = c\big(\vartheta(\psi)\big)^n \tilde{h}(t) \exp\left\{\psi^\mathsf{T} t\right\},$$

con log-verosimiglianza

$$\ell(\psi; t) = \psi^\mathsf{T} t - nK(\psi),$$

con $K(\psi) = -\log c\big(\vartheta(\psi)\big)$. La funzione score è

$$l_*(\psi) = \frac{\partial}{\partial \psi} \ell(\psi; t) = \left[t_r - n \frac{\partial K(\psi)}{\partial \psi_r}\right]_r,$$

da cui l'informazione osservata

$$j(\psi) = \left[n \frac{\partial^2 K(\psi)}{\partial \psi_r \partial \psi_s}\right]_{r,s} = i(\psi).$$

Dall'identità sulla funzione di punteggio,

$$\mathbb{E}_\psi \left[T_r\right] = n \frac{\partial K(\psi)}{\partial \psi_r}, \quad r = 1, \ldots, p.$$

Dall'identità dell'informazione,

$$\mathrm{Cov}_\psi(T_r, T_s) = n\frac{\partial^2 K(\psi)}{\partial\psi_r\partial\psi_s}.$$

Dal momento che una matrice di covarianza è sempre semidefinita positiva e si assume l'indipendenza lineare delle componenti di T, è definita strettamente positiva. Inoltre,

$$\frac{\partial^2\ell(\psi)}{\partial\psi\partial\psi^\intercal} = -\big[\,\mathrm{Cov}(T_r, T_s)\big]_{r,s} < 0.$$

In particolare, in una famiglia esponenziale si può evitare di valutare la derivata seconda nel calcolo della SMV, in quanto se ammette soluzione questa è unica ed è punto di massimo.

Teo. (Esistenza della SMV)
L'equazione di stima $\ell_*(\psi; t) = 0$ ammette soluzione se e solo se t è un punto interno dell'involucro convesso chiuso del supporto di T, cioè

$$\operatorname{conv}\mathcal{T} = \bigcap_{C \text{ chiuso},\ \mathcal{T}\subseteq C} C.$$

Ad esempio, $\mathcal{T} = \{0, 1, \dots, n\} \implies \operatorname{conv}\mathcal{T} = [0, n]$. Se invece $\mathcal{T} = \{(0,0), (1,0), (1,-1), (2,0)\}$, l'involucro convesso è il quadrato con vertici $(0,0), (1,1), (2,0), (1,-1)$.

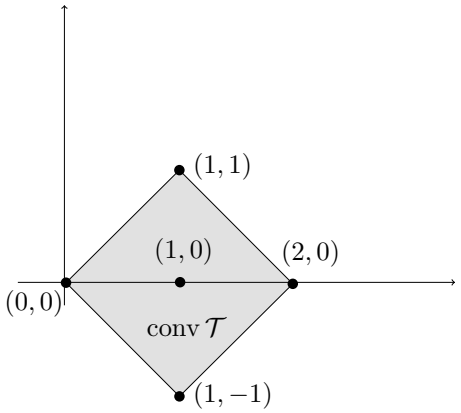


Figura 20: convTcal

In particolare, è possibile che non esista finita la SMV con probabilità positiva, se ci si trova sulla frontiera di $\operatorname{conv}\mathcal{T}$, ad esempio in modelli esponenziali regolari discreti.

Riparametrizzazione in media

L'equazione di verosimiglianza è $l_*(\psi) = 0$, per cui dal campione osservato si impone

$$t_r - \mathbb{E}_\psi\left[T_r\right] = 0 \quad \forall r = 1, \dots, p.$$

Dunque, $\hat{\psi}$ rende il valore osservato di T uguale al suo valore atteso, cioè

$$\mathbb{E}_{\hat{\psi}}\left[T_r\right] = n\left.\frac{\partial K(\psi)}{\partial \psi_r}\right|_{\psi=\hat{\psi}} = t_r.$$

Ora, poiché per l'identità dell'informazione

$$n\frac{\partial^2 K(\psi)}{\partial \psi \partial \psi^{\mathsf{T}}} = \mathbb{V}_{\psi}\left[T\right] > 0,$$

la funzione $t \to \hat{\psi}$ è biunivoca all'interno di $\operatorname{conv}\mathcal{T}$. Di conseguenza, $\hat{\psi}$ è statistica sufficiente minimale e si può scrivere la log-verosimiglianza attraverso $\hat{\psi}$

$$\ell(\psi;\hat{\psi}) = \psi^{\mathsf{T}}K(\hat{\psi}) - nK(\psi),$$

dove si pone $K_r(\psi) = n\frac{\partial K(\psi)}{\partial \psi_r}$.

Inoltre, $\mu(\psi) = \mathbb{E}_{\psi}\left[T/n\right]$ definisce una riparametrizzazione della famiglia, detta riparametrizzazione con la media. Per $\mu = \mu(\psi)$ si ha $\hat{\mu} = t/n$ e lo stimatore corrispondente è non distorto.

Esercizi: ...

26.3 Riparametrizzazioni e quantità di verosimiglianza

Sia $\omega = \omega(\vartheta)$ una riparametrizzazione di $\mathcal{F}$, valgono le relazioni

$$L^{\Omega}(\omega) = L^{\Theta}(\vartheta(\omega))$$

$$\ell^{\Omega}(\omega) = \ell^{\Theta}(\vartheta(\omega))$$

$$\hat{\omega} = \omega(\hat{\vartheta})$$

Derivando rispetto a ω, si ottiene la relazione tra le funzioni di punteggio nelle due parametrizzazioni:

$$\boxed{p=1}: \ell^{\Omega}_*(\omega) = \vartheta'(\omega) l^{\Theta}_*(\vartheta(\omega)).$$

$$\boxed{p>1}: \ell^{\Omega}_*(\omega) = D(\omega)^{\mathsf{T}} l^{\Theta}_*(\vartheta(\omega)),$$

dove

$$D(\omega)_{p\times p} = \left[\frac{\partial \vartheta_r(\omega)}{\partial \omega_s}\right].$$

Derivando una seconda volta e cambiando di segno,

$$\boxed{p=1}: j^{\Omega}(\omega) = -\vartheta''(\omega) l^{\Theta}_*(\vartheta(\omega)) + \left(\vartheta'(\omega)\right)^2 j^{\Omega}\big(\vartheta(\omega)\big).$$

Nella stima di massima verosimiglianza, il primo addendo si annulla e

$$j^{\Omega}(\hat{\omega}) = \left(\vartheta'(\hat{\omega})\right)^2 j^{\Omega}(\hat{\vartheta}).$$

Anche per l'informazione attesa vale questa proprietà, in quanto $\mathbb{E}_\vartheta\left[l_*(\vartheta;Y)\right]=0$ e

$$i^\Omega(\omega)=\left(\vartheta(\omega)\right)^2 i^\Omega\big(\vartheta(\omega)\big).$$

$$\boxed{p>1}: j^\Omega(\hat{\omega})=D(\hat{\omega})^\mathsf{T} j^\Theta(\vartheta(\hat{\omega}))D(\hat{\omega}),$$

e l'informazione attesa è

$$i^\Omega(\omega)=D(\omega)^\mathsf{T} i^\Theta(\vartheta(\omega))D(\omega).$$

La dipendenza della matrice delle derivate seconde, rispetto al cambio di coordinate, dalla sola derivata prima si dice *comportamento tensoriale*.

26.4 Informazione attesa e statistiche sufficienti

Se $T=t(Y)$ è statistica sufficiente per l'inferenza su ϑ entro $\mathcal{F}$,

$$p_Y(y;\vartheta)=p_T(t;\vartheta)p_{Y|T=t}(y;t)$$

per y tali che $t(y)=t$. La funzione di log-verosimiglianza è equivalente dai dati y ai dati t

$$\ell^Y(\vartheta;y)=\ell^T(\vartheta;t(y))$$

$$j^Y(\vartheta;y)=j^T(\vartheta;t(y))$$

Dunque, per l'informazione attesa

$$\begin{aligned} i^Y(\vartheta) &= \int_{\mathcal{Y}} j^T(\vartheta;t(y))p(y;\vartheta)\,dy \\ &= \int_{\mathcal{T}} j^T(\vartheta;t)p_T(t;\vartheta)\,dt \\ &= i^T(\vartheta). \end{aligned}$$

Osservazione

La definizione di informazione attesa è coerente con l'assenza di perdita di informazione nel passaggio da Y a T.

Lezione 27

27.1 Informazione attesa e stimatori efficienti

I test basati sulla verosimiglianza sono, per modelli con due elementi o monoparametrici con rapporto di verosimiglianza monotoni, uniformemente più potenti per $H_0 : \vartheta \underset{(\geq)}{\leq} \vartheta_0$ vs. $H_1 : \vartheta \underset{(<)}{>} \vartheta_0$.

Anche per il problema di *stima puntuale*, lo stimatore di massima verosimiglianza ha buone proprietà campionarie. Le proprietà desiderabili sono:

Centratura	**Modesta dispersione**
$\mathbb{E}_\vartheta\left[\hat{\vartheta}\right] = \vartheta$	$\mathbb{V}_\vartheta\left[\hat{\vartheta}\right]$ minima

La teoria classica degli stimatori ottimi è

- strettamente legata al concetto di sufficienza;
- in conflitto con equivarianza rispetto a riparametrizzazione:

$$\mathbb{E}_\vartheta\left[\hat{\vartheta}(Y)\right] = \vartheta \implies \mathbb{E}_\vartheta\left[\omega(\hat{\vartheta})\right] \neq \omega(\vartheta).$$

  In generale, saranno solamente *asintoticamente ottimali*.

Con $p > 1$ la condizione di minima varianza viene posta sugli stimatori intesi come *combinazioni* lineari.

Esempio (c.c.s da Pois(ϑ))
Siano $Y_i \overset{i.i.d}{\sim} \text{Pois}(\vartheta)$, allora

$$\overline{Y}_n = \hat{\vartheta} \quad \text{e} \quad S_n^2 = \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2/(n-1)$$

sono entrambi stimatori non distorti di ϑ come pure qualunque loro combinazione lineare

$$a\overline{Y}_n + (1-a)S_n^2.$$

Lemma (Covarianza e stimatori non distorti)
In un modello monoparametrico con condizioni regolari, se $\tilde{\vartheta}_n(Y)$ è uno stimatore non distorto di ϑ, allora

$$\text{Cov}_\vartheta\left(\ell_*(\vartheta; Y), \hat{\vartheta}_n(Y)\right) = 1$$

Dim.
Per la non distorsione,

$$\begin{aligned}
&\int_{\mathcal{Y}} \tilde{\vartheta}(y)p(y;\vartheta)\,dy = \vartheta \\
&\implies \int_{\mathcal{Y}} \frac{\partial}{\partial\vartheta}\tilde{\vartheta}(y)p(y;\vartheta)\,dy = 1 && \text{(regolarità)} \\
&\implies \int_{\mathcal{Y}} \tilde{\vartheta}(y)\ell_*(\vartheta;y)p(y;\vartheta)\,dy = 1 && \text{(trucco derivata)} \\
&\implies \mathrm{Cov}\left(\hat{\vartheta}(Y), \ell_*(\vartheta;Y)\right) = 1 && \left(\mathbb{E}_\vartheta\left[l_*(\vartheta;Y)\right] = 0\right)
\end{aligned}$$

□

Nota 1 Se $\mathbb{E}_\vartheta\left[\tilde{\vartheta}_n(Y)\right] = a(\vartheta)$, si ha che

$$\mathrm{Cov}_\vartheta\left(l_*(\vartheta;Y), \tilde{\vartheta}_n(Y)\right) = a'(\vartheta).$$

Nota 2 Con $p > 1$, la relazione vale componente per componente e

$$\mathrm{Cov}_\vartheta\left(\ell_r(\vartheta;Y), \tilde{\vartheta}_r(Y)\right) = 1.$$

Teo. (Diseguaglianza di Cramér-Rao)
Sia $\tilde{\vartheta}_n(Y)$ uno stimatore non distorto di ϑ in un modello monoparametrico $\mathcal{F}$ che soddisfa le condizioni di regolarità, per cui $i(\vartheta) > 0$. Allora

$$\mathbb{V}_\vartheta\left[\tilde{\vartheta}_n\right] \geq i(\vartheta)^{-1}.$$

Dim.
Se $\mathbb{V}_\vartheta\left[\tilde{\vartheta}_n\right] = +\infty$, la disuguaglianza è certamente soddisfatta. Se invece $\mathbb{V}_\vartheta\left[\tilde{\vartheta}_n\right] < +\infty$, utilizzando la diseguaglianza di Cauchy-Schwarz

$$\underbrace{\mathrm{Cov}^2_\vartheta\left(\tilde{\vartheta}_n, \ell_*(\vartheta;Y)\right)}_{=1} \leq \mathbb{V}_\vartheta\left[\tilde{\vartheta}_n\right]\mathbb{V}_\vartheta\left[\ell_*(\vartheta;Y)\right],$$

dunque

$$\mathbb{V}_\vartheta\left[\tilde{\vartheta}_n\right] \geq \frac{1}{\mathbb{V}_\vartheta\left[\ell_*(\vartheta;Y)\right.} = i(\vartheta)^{-1}.$$

□

Osservazione

Con $p \geq 1$, vale comunque il teorema, nel senso che la differenza tra le matrici di varianza è definita non negativa:

$$\mathbb{V}_\vartheta\left[\tilde{\vartheta}_n\right] - i(\vartheta)^{-1} \succeq 0.$$

In particolare, se $\psi = c^\mathsf{T}\vartheta$, lo stimatore $\tilde{\psi} = c^\mathsf{T}\tilde{\vartheta}$ ha varianza

$$c^\mathsf{T}\,\mathbb{V}_\vartheta\left[\tilde{\vartheta}_n\right]c \geq c^\mathsf{T} i(\vartheta)^{-1}c,$$

in quanto

$$c^\mathsf{T}\Big(\mathbb{V}_\vartheta\left[\tilde{\vartheta}_n\right] - i(\vartheta)^{-1}\Big)c \geq 0$$

per definizione di matrice definita non negativa.

Sotto c.c.s con numerosità n da $Y_1 \sim p_{Y_1}(y;\vartheta)$, l'informazione attesa è $i^Y(\vartheta) = n i_1(\vartheta)$, dove

$$i_1(\vartheta) = \mathbb{E}_\vartheta\Big[-\frac{\partial^2}{\partial\vartheta\partial\vartheta^\mathsf{T}}\log p_{Y_1}(y;\vartheta)\Big],$$

per cui

$$\mathbb{V}_\vartheta\left[\tilde{\vartheta}_n\right] \geq \big(n i_1(\vartheta)\big)^{-1}.$$

Conseguenze

Se in un modello regolare si individua uno stimatore non distorto di ϑ con varianza pari a $i(\vartheta)$, lo stimatore trovato è *efficiente* e non migliorabile per conseguenza del teorema.

Quando si raggiunge la soglia di efficienza? L'uguaglianza stretta vale se

$$\mathrm{Cov}^2_\vartheta\left(\tilde{\vartheta}_n(Y), \ell_*(\vartheta;Y)\right) = 1, \quad \vartheta \in \Theta,$$

cioè quando

$$\ell_*(\vartheta;Y) = a(\vartheta) + b(\vartheta)\tilde{\vartheta}_n(Y),$$

per $a(\cdot), b(\cdot)$ opportune. Inoltre, siccome $\mathbb{E}_\vartheta\left[\ell_*(\vartheta;Y)\right] = 0$, deve essere che

$$a(\vartheta) + b(\vartheta)\vartheta = 0 \implies a(\vartheta) = -\vartheta b(\vartheta),$$

per cui la funzione di punteggio deve essere della forma

$$\ell_*(\vartheta) = b(\vartheta)\big[\tilde{\vartheta}_n(y) - \vartheta\big].$$

Il modello deve allora essere

- una famiglia esponenziale;
- nella parametrizzazione con la media.

Esempio (Normale con varianza nota)
Se $Y_i \sim \mathcal{N}(\mu, \sigma_0^2)$, allora

$$\ell_*(\mu) = \frac{1}{\sigma_0^2}\sum_{i=1}^n (y_i - \mu)^2$$

e la varianza di $\hat{\mu} = \overline{Y}_n$ è

$$\mathbb{V}_\mu\left[\hat{\mu}\right] = \frac{\sigma_0^2}{n} = \big(n \cdot i_1(\mu)\big)^{-1}$$

che è dunque uno stimatore efficiente.

Esempio (Normale con varianza ignota)
La funzione di punteggio è

$$\ell_*(\mu, \sigma^2) = \begin{pmatrix} \ell_\mu(\mu, \sigma^2) \\ \ell_{\sigma^2}(\mu, \sigma^2) \end{pmatrix} = \begin{pmatrix} \frac{n}{\sigma^2}(\overline{Y}_n - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \mu)^2 \end{pmatrix}$$

e la matrice di informazione osservata è

$$j(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{n}{\sigma^4}(\overline{Y}_n - \mu) \\ \frac{n}{\sigma^4}(\overline{Y}_n - \mu) & -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^n (Y_i - \mu)^2 \end{pmatrix}$$

$$\implies i(\vartheta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Dunque, $i(\mu, \sigma^2)^{-1} = \operatorname{diag}\left(\frac{\sigma^2}{n}, \frac{2\sigma^4}{n}\right)$. Lo stimatore di massima verosimiglianza è distorto per (μ, σ^2) in quanto $\mathbb{E}_{\mu,\sigma^2}\left[\hat{\sigma}^2\right] = \frac{n-1}{n}\sigma^2$. Lo stimatore non distorto è la coppia $(\overline{Y}_n, S_n^2)$ con

$$S_n^2 = \frac{n}{n-1}\hat{\sigma}^2,$$

che ha matrice di covarianza

$$\mathbb{V}_{\mu,\sigma^2}\left[\overline{Y}_n, s_n^2\right] = \operatorname{diag}\left(\frac{\sigma^2}{n}, \frac{2\sigma^4}{n-1}\right)$$

che non raggiunge la limitazione inferiore di Cramér-Rao, dunque lo stimatore è efficiente solo per μ e non per σ^2.

Esempio (Esponenziale)
Se $Y_1, Y_2, \ldots, Y_n \overset{i.i.d}{\sim} \operatorname{Exp}(\vartheta)$, allora

$$\ell(\vartheta) = n \log \vartheta - n\vartheta\bar{y}_n;$$

$$\ell_*(\vartheta) = \frac{n}{\vartheta} - n\bar{y}_n;$$

$$j(\vartheta) = i(\vartheta) = \frac{n}{\vartheta^2}.$$

Lo SMV è $\hat{\vartheta} = \frac{1}{n}$, che è distorto: $\mathbb{E}_\vartheta\left[\hat{\vartheta}\right] = \frac{n}{n-1}\vartheta$ ma si può correggere con $\tilde{\vartheta}_n = \frac{n-1}{n}\hat{\vartheta}$, che ha varianza

$$\mathbb{V}_\vartheta\left[\tilde{\vartheta}\right] = \frac{\vartheta^2}{n-2} > i(\vartheta)^{-1} = \frac{\vartheta^2}{n}.$$

Nella parametrizzazione con la media, se $\mu = \frac{1}{\vartheta}$, lo SMV è efficiente:

$$\hat{\mu} = \bar{y}, \mathbb{V}_\mu\left[\hat{\mu}\right] = \frac{\mu^2}{n}.$$

In particolare, se $\vartheta(\mu) = \frac{1}{\mu}$,

$$\begin{aligned} i^M(\mu) &= \left(\vartheta'(\mu)\right)^2 i\left(\vartheta(\mu)\right) \\ &= \left(\frac{1}{\mu^4}\right) n\mu^2 \\ &= \frac{n}{\mu^2} \\ &= \mathbb{V}_\mu\left[\hat{\mu}\right]. \end{aligned}$$

La proprietà di ottimalità vale sempre in una famiglia esponenziale parametrizzata con la media, come mostrato nell'esempio seguente.

Esempio (Famiglia esponenziale monoparametrica)
In una famiglia esponenziale regolare di ordine 1, si ha

$$\ell(\psi; t) = \psi t - nK(\psi),$$

con $i(\psi) = nK''(\psi)$. Nella parametrizzazione con la media, $\mu(\psi) = K'(\psi)$, per cui

$$i^M(\mu) = \left[\psi'(\mu)\right]^2 i\big(\psi(\mu)\big).$$

Poiché

$$\begin{aligned} i\big(\psi(\mu)\big) &= nK''(\psi)|_{\psi=\psi(\mu)} \\ \mu(\psi) &= K'(\psi) \\ \psi'(\mu) &= \frac{1}{\mu'(\psi)}\bigg|_{\psi=\psi(\mu)} = \frac{1}{K''(\psi)}\bigg|_{\psi=\psi(\mu)} \end{aligned}$$

Dunque,

$$i^M(\mu) = \frac{n}{K''(\psi)}\bigg|_{\psi=\psi(\mu)}$$

Lo stimatore di massima verosimiglianza di μ è $K'(\hat{\psi}) = \frac{T}{n}$, con media μ e varianza $K''(\psi)/n$, che è pari a $i^M(\mu)^{-1}$ in $\psi = \psi(\mu)$.

Sorgono alcune domande:

- In modelli regolari in cui gli stimatori non raggiungono il limite inferire, si può sapere se esiste/quale stimatore è ottimo, ovvero con varianza minima tra i non distorti? (UMVU, *Uniformly Minimum Variance Unbiased Estimator*).

- È possibile reperire uno stimatore UMVU in modelli non regolari? Ad esempio sotto c.c.s da Unif$(0, \vartheta)$, possiamo avere lo stimatore non distorto

$$\tilde{\vartheta}_n = \frac{n+1}{n} Y_{(n)}.$$

  Questo stimatore è UMVU?

Si noti che negli esempi precedenti, gli stimatori non distorti sono funzione della SMV e quindi della statistica sufficiente minimale, proprietà molto importante.

Teo. (Rao-Blackwell)
Sia $\mathcal{F}$ un modello statistico monoparametrico per cui $s(Y)$ è statistica sufficiente minimale. Se $\tilde{\vartheta}_n(Y)$ è uno stimatore non distorto di ϑ, allora lo stimatore

$$\tilde{\vartheta}^*(S) = \mathbb{E}\left[\tilde{\vartheta}_n(Y)|S\right]$$

è tale che

1) $\mathbb{E}_\vartheta\left[\tilde{\vartheta}^*(S)\right] = \vartheta$

2) $\mathbb{V}_\vartheta\left[\tilde{\vartheta}^*(S)\right] \leq \mathbb{V}_\vartheta\left[\tilde{\vartheta}_n(Y)\right]$

Dim.
Usando il valore atteso iterato:

$$\mathbb{E}_\vartheta\left[\tilde{\vartheta}_n(S)\right] = \mathbb{E}_\vartheta\left[\mathbb{E}\left[\tilde{\vartheta}_n(Y)\right]|S\right] = \mathbb{E}_\vartheta\left[\tilde{\vartheta}_n(Y)\right] = \vartheta.$$

Usando la formula della varianza totale:

$$\begin{aligned}\mathbb{V}_\vartheta\left[\tilde{\vartheta}_n(Y)\right] &= \mathbb{V}_\vartheta\left[\mathbb{E}\left[\tilde{\vartheta}_n(Y)|S\right]\right] + \mathbb{E}_\vartheta\left[\mathbb{V}\left[\tilde{\vartheta}_n(Y)|S\right]\right] \\ &= \mathbb{V}_\vartheta\left[\tilde{\vartheta}^*_n(S)\right] + \mathbb{E}_\vartheta\left[\mathbb{V}\left[\tilde{\vartheta}_n(Y)|S\right]\right] \\ &\geq \mathbb{V}_\vartheta\left[\tilde{\vartheta}^*_n(S)\right].\end{aligned}$$

□

Osservazione

Nonostante il teorema dia anche una costruzione per uno stimatore migliore, resta un dubbio circa più stimatori non distorti basati su s. Tale possibilità è esclusa se s è *completa*.

Teo. (Rao-Blackwell + Lehmann-Scheffé)
Sia $\mathcal{F}$ un modello statistico monoparametrico con S statistica sufficiente completa, allora lo stimatore $\tilde{\vartheta}^*(S)$ è l'unico stimatore non distorto di ϑ con varianza minima.

Lo stimatore è anche caratterizzato da essere l'unica funzione non distorta $T = \varphi(S)$ della statistica completa.

Dim.
Se $\tilde{\vartheta}_1^*$ e $\tilde{\vartheta}_2^*$ sono due stimatori non distorti di ϑ, la funzione

$$g(s) = \tilde{\vartheta}_1^*(s) - \tilde{\vartheta}_2^*(s)$$

ha media nulla e dunque $\tilde{\vartheta}_1 = \tilde{\vartheta}_2$ quasi certamente per la completezza di S.

□

Negli esempi precedenti gli stimatori non distorti sono UMVU, in quanto la statistica sufficiente minimale è completa. Bisogna invece dimostrarlo per l'uniforme.

Esempio (Stimatore UMVU da Unif$(0, \vartheta)$)
Lo stimatore $\hat{\vartheta} = Y_{(n)}$ è statistica sufficiente minimale. Per mostrare che è completa, bisogna mostrare che

$$\int_0^{\vartheta} g(s) p_{Y_{(n)}}(s;\vartheta)\, ds = 0 \quad \forall \vartheta > 0 \implies g(s) = 0 \quad \text{a.s.}$$

La densità è

$$\begin{aligned} p_{Y_{(n)}}(s;\vartheta) &= \frac{\partial}{\partial s}\left(\frac{s}{\vartheta}\right)^n \mathbb{1}_{[0,\vartheta]}(s) \\ &= (n s^{n-1}/\vartheta^n)\mathbb{1}_{[0,\vartheta]}(s) \end{aligned}$$

e

$$\int_0^{\vartheta} g(s) n \frac{s^{n-1}}{\vartheta^n}\, ds = \frac{n}{\vartheta^n}\int_0^{\vartheta} g(s) s^{n-1}\, ds \stackrel{?}{=} 0$$

implica che

$$\forall \vartheta:\ g(\vartheta)\vartheta^{n-1} = 0 \iff g(\vartheta) \equiv 0 \implies Y_{(n)} \text{ completa.}$$

Lezione 28

28.1 Distribuzione a priori di Jeffreys

Riferimenti Liseo (2010, §3.2)

Sono stati considerati diversi criteri per la specificazione della distribuzione a priori per un modello $\mathcal{F}$:

1. Studi precedenti.
2. Distribuzioni coniugate.
3. A priori soggettive
4. A priori non informative.

In generale, una distribuzione a priori non informativa non è facilmente specificabile, per i problemi di non invarianza rispetto a riparametrizzazioni. L'informazione osservata è molto utile in tal senso, perché permette di definire una classe di distribuzioni a priori non informativa in qualunque riparametrizzazione.

Def. (Distribuzione a priori di Jeffreys univariata)
Per ϑ scalare, in un modello regolare, si definisce la distribuzione a priori di Jeffreys come

$$\pi(\vartheta) \propto |i(\vartheta)|^{1/2}.$$

Osservazione

Se $\omega = \omega(\vartheta)$ è una riparametrizzazione del modello $\mathcal{F}$, per la regola di trasformazione si ha che

$$i^{\Omega}(\omega) = \left(\vartheta'(\omega)\right)^2 i\left(\vartheta(\omega)\right).$$

Di conseguenza la distribuzione a priori diventa

$$\begin{aligned}\pi^{\Omega}(\omega) &= \pi\left(\vartheta(\omega)\right)\left|\vartheta'(\omega)\right| \\ &= i\left(\vartheta(\omega)\right)^{1/2}\left|\vartheta'(\omega)\right| \\ &= |i^{\Omega}(\omega)|^{1/2}.\end{aligned}$$

Cioè che è *invariante* è la struttura della distribuzione a priori: rimane la radice quadrata dell'informazione attesa in qualunque parametrizzazione.

Def. (Distribuzione a priori di Jeffreys)
Con $p > 1$, la distribuzione a priori di Jeffreys è

$$\pi(\vartheta) \propto |\det i(\vartheta)|^{1/2}.$$

Osservazione

La distribuzione non rappresenta in realtà uno stato di "ignoranza a priori", ma è una scelta convenzionale che produce a volte un accordo con le procedure frequentiste.

Esempio (Modello normale con varianza nota)
Se $Y_i \sim \mathcal{N}(\mu, \sigma_0^2)$, si ha $i(\mu) = \frac{n}{\sigma_0^2}$, per cui la distribuzione di Jeffreys è

$$\pi(\mu) \propto C \quad \text{cost. su } \mathbb{R}$$

per cui

$$\mu | y \sim \mathcal{N}\left(\bar{y}, \frac{\sigma_0^2}{n}\right).$$

L'intervallo di credibilità con livello $1 - \alpha$ è esattamente uguale a quello frequentista.

Esempio (Modello binomiale)
Se $Y_i \sim \text{Bin}(n, \vartheta)$, allora

$$i(\vartheta) = \frac{n}{\vartheta} + \frac{n}{1-\vartheta} = \frac{n}{\vartheta(1-\vartheta)}.$$

La distribuzione a priori di Jeffreys è

$$\pi(\vartheta) \propto \vartheta^{-\frac{1}{2}}(1-\vartheta)^{-\frac{1}{2}} \quad \sim \text{Beta}(1/2, 1/2).$$

La distribuzione a posteriori è allora $\vartheta | y \sim \text{Beta}(y + 1/2, n - y + 1/2)$, che corrisponde ad aggiungere ai dati una singola prova con $1/2$ successo e $1/2$ insuccesso.

Esempio (Famiglia di posizione)
Sia Y con densità $p_0(y - \mu)$, $\mu \in \mathbb{R}$ famiglia di posizione con $p_0(\cdot)$ densità nota. Allora,

$$\ell(\mu) = \log p_0(y - \mu) = q_0(y - \mu).$$

Allora,

$$\ell_*(\mu) = -q_0'(y - \mu)$$

$$\ell_{**}(\mu) = q_0''(y - \mu)$$

$$j(\mu) = -q_0''(y - \mu).$$

Siccome $Y - \mu = Z$ ha distribuzione $p_0(z)$, si ha

$$\begin{aligned} i(\mu) &= \mathbb{E}_\mu \left[j(\mu; Y) \right] \\ &= - \mathbb{E}_\mu \left[q_0''(Y - \mu) \right] \\ &= - \mathbb{E}_\mu \left[q_0''(Z) \right] \end{aligned}$$

costante rispetto a μ. Allora, la distribuzione a priori di Jeffreys è costante e pari a

$$\pi(\mu) \propto C \quad \mu \in \mathbb{R}.$$

Esempio (Famiglia di scala)
Se Y proviene da una famiglia di scala, con $p(y;\sigma) = \frac{1}{\sigma} p_0\left(\frac{y}{\sigma}\right)$, allora

$$\begin{aligned}
\ell(\sigma) &= -\log\sigma + q_0\left(\frac{y}{\sigma}\right) \\
\ell_*(\sigma) &= -\frac{1}{\sigma} + q_0'\left(\frac{y}{\sigma}\right)\cdot\left(-\frac{y}{\sigma^2}\right) \\
\ell_{**}(\sigma) &= \frac{1}{\sigma^2} + q_0''\left(\frac{y}{\sigma}\right)\left(\frac{y^2}{\sigma^4}\right) + q_0'\left(\frac{y}{\sigma}\right)\left(\frac{2y}{\sigma^3}\right) \\
&= \frac{1}{\sigma^2}\left[1 + q_0''\left(\frac{y}{\sigma}\right)\left(\frac{y}{\sigma}\right)^2 + 2q_0'\left(\frac{y}{\sigma}\right)\left(\frac{y}{\sigma}\right)\right],
\end{aligned}$$

ma poiché la variabile Y/σ ha distribuzione indipendente da σ, si ha che

$$i(\vartheta) \propto \sigma^{-2},$$

da cui la distribuzione coniugata di Jeffreys

$$\pi(\sigma) \propto \frac{1}{\sigma}.$$

Esercizio: Mostrare che la priori di Jeffreys per $\omega = \sigma^2$ è $\pi^\Omega(\omega) = \frac{1}{\omega}$.

Difficoltà

Nel caso multiparametrico, ci sono delle difficoltà nel calcolare l'informazione attesa. Si può verificare (esercizio), che

$$i(\mu,\sigma) = \frac{1}{\sigma^2}\cdot A,$$

con $A_{2\times 2}$ matrice costante, da cui si ottiene la priori

$$\pi(\mu,\sigma) \propto \frac{1}{\sigma^2},$$

che è diversa dalla a priori $\frac{1}{\sigma}$ che si ottiene assumendo μ, σ indipendenti con priori marginali di Jeffreys. In generale, si preferisce questo secondo approccio.

Si vede che l'inferenza su un parametro ψ di interesse dipende dalla presenza di ulteriori parametri di disturbo, anche assumendo indipendenza dei parametri.

Esistono altri criteri per definire priori non informative, basati su:

- Requisiti di invarianza per famiglie di gruppo;
- Massimizzazione di una divergenza tra distribuzione a priori e a posteriori: scegliere una distribuzione a priori che massimizzi il contenuto informativo dei dati y.

Reference priors

A priori di Berger-Bernardo o *reference priors*, sono definite a partire da $\mathrm{KL}(\pi(\vartheta), \pi(\vartheta|y))$:

$$\mathbb{E}_{p_Y}\left[\int_\Omega \pi(\vartheta|Y)\log\frac{\pi(\vartheta|Y)}{\pi(\vartheta)}\,d\vartheta\right].$$

Con condizioni di regolarità, si ottiene di nuovo la distribuzione a priori di Jeffreys, mentre per parametri di disturbo viene definita una procedura "stepwise".

- Requisiti di accordo con inferenza frequentista (*matching priors*).

28.2 Proprietà asintotiche e approssimazioni per distribuzioni

Riferimenti Pace e Salvan (2001, §6)
Casella e Berger (2001, §10)

Sono disponibili dei risultati distributivi asintotici, che valgono per un'ampia varietà di casi, e che permettono delle procedure inferenziali quasi automatiche.

Dato un modello statistico parametrico $\mathcal{F}$ e assunta la sua *corretta specificazione* di $\mathcal{F}$, cioè $p_0(\cdot) \in \mathcal{F}$, allora

- $L(\vartheta)$ sintetizza y in modo più conciso sia per l'inferenza frequentista sia per quella bayesiana (statistica suff. minimale)
- Fornisce lo strumento frequentista più efficiente per discriminare tra coppie di elementi di $\mathcal{F}$ (Neyman-Pearson).

Più nel dettaglio, la funzione di verosimiglianza suggerisce in modo naturale:

1. La *stima puntuale* di ϑ usando la smv
$$\hat{\vartheta} = \operatorname*{argmax}_{\theta\in\Theta} L(\vartheta).$$
2. La *regione di stima* usando l'insieme
$$\hat{\Theta}(y) = \left\{\vartheta \in \Theta : \frac{L(\vartheta)}{L(\hat{\vartheta})} \geq k\right\}, \quad k \in (0,1).$$
3. La *statistica test* per $H_0 : \vartheta \in \Theta_0,\ \Theta_0 \subset \Theta$ il rapporto
$$\frac{L(\hat{\vartheta})}{L(\hat{\vartheta}_0)},$$
o in modo equivalente $\ell(\hat{\vartheta}) - \ell(\hat{\vartheta}_0)$, dove $L(\hat{\vartheta}_0)$ indica il massimo di $L(\vartheta)$ sotto il vincolo $\vartheta \in \Theta$. Valori grandi della statistica danno evidenza *contro* H_0, in quanto significa che il massimo si abbassa notevolmente sotto il vincolo.

Parametri di disturbo

Questi concetti si possono estendere ad un sottoinsieme ψ di componenti di ϑ, con $\vartheta = (\psi, \lambda)$, a partire dalla verosimiglianza profilo

$$L_p(\psi) = L(\psi, \hat{\lambda}_\psi).$$

I tre problemi diventano automaticamente stima puntuale $\hat{\psi} = \text{argmax}_\psi L_p(\psi)$, regione di stima $\hat{\Psi}(y) = \left\{\psi \in \Psi : \frac{L_p(\psi)}{L_p(\hat{\psi})} \geq k\right\}$, $k \in (0,1)$, e statistica test $L_p(\hat{\psi})/L_p(\hat{\psi}_0) \equiv \ell_p(\hat{\psi}) - \ell_p(\hat{\psi}_0)$.

Distribuzioni approssimate

Supponendo $y_1, y_2, \ldots, y_n$, con n numerosità campionaria, ci si occuperà

(a) di giustificare formalmente l'uso della SMV come metodo di stima, studiando la *consistenza* di $\hat{\vartheta}_n$ per $n \to \infty$. Parallelamente, si studia la consistenza del test basato su

$$T_n = \frac{L(\hat{\vartheta}_n)}{L(\hat{\vartheta}_{0,n})};$$

- di ottenere approssimazioni per la distribuzione sotto ϑ di

$$\ell_*(\vartheta; Y), \quad \hat{\vartheta}_n(Y), \quad \frac{L(\hat{\vartheta}_n(Y); Y)}{L(\hat{\vartheta}_{0,n}(Y); Y)},$$

in modo da

1. poter definire regioni di confidenza con copertura asintotica pari a $1 - \alpha$.
2. valutare, almeno approssimativamente, il *p-value* osservato

$$\alpha^{\text{oss}} = \sup_{\vartheta \in \Theta_0} P_\vartheta \left(\frac{L(\hat{\vartheta}(Y); Y)}{L(\hat{\vartheta}_0(Y); Y)} \geq \frac{L(\hat{\vartheta}^{\text{oss}}; y^{\text{oss}})}{L(\hat{\vartheta}_0^{\text{oss}}; y^{\text{oss}})} \right)$$

Note

– Le proprietà di consistenza servono come garanzie di coerenza dei metodi utilizzati.

– I risultati di approssimazione si basano sui teoremi limite di Calcolo delle Probabilità, in particolare la convergenza in distribuzione.

– Si considera il risultato limite e lo si usa come approssimazione per l'n effettivamente disponibile.

– Non si ha una garanzia dell'accuratezza dell'approssimazione, che va valutata ad esempio tramite simulazione.

– n può, in generale, rappresentare un indice della quantità di informazione in y.

– L'approssimazione asintotica può essere sostituita dai metodi Monte Carlo e bootstrap.

Lezione 29

29.1 Consistenza di $\hat{\vartheta}_n$

La consistenza dello stimatore $\hat{\vartheta}_n$ si può dimostrare anche senza assumere la derivabilità della log-verosimiglianza.

Sia $\vartheta^0 \in \Theta$ il vero valore del parametro e $Y = (Y_1, Y_2, \dots, Y_n)$ campione i.i.d, allora la log-verosimiglianza è

$$\ell(\vartheta; Y) = \sum_{i=1}^{n} \ell(\vartheta; Y_i),$$

dunque $\mathbb{E}_{\vartheta^0}\left[\ell(\vartheta; Y)\right] = n\,\mathbb{E}_{\vartheta^0}\left[\ell(\vartheta; Y_i)\right]$. Per la legge dei grandi numeri, si ha che

$$\frac{1}{n}\sum_{i=1}^{n} \ell(\vartheta; Y_i) \xrightarrow{P} \mathbb{E}_{\vartheta^0}\left[\ell(\vartheta; Y_i)\right]$$

e inoltre

$$\frac{1}{n}\ell(\vartheta; Y) - \frac{1}{n}\ell(\vartheta^0; Y) \xrightarrow{P} \mathbb{E}_{\vartheta^0}\left[\ell(\vartheta; Y_i)\right] - \underbrace{\mathbb{E}_{\vartheta^0}\left[\ell(\vartheta^0; Y_i)\right]}_{\text{max per Wald}}.$$

Dunque, il limite è una quantità negativa e

$$\ell(\vartheta; Y) - \ell(\vartheta^0; Y) \xrightarrow{n\to\infty} -\infty,$$

per qualunque altro valore $\vartheta \neq \vartheta^0$. Dunque, questo ci fa supporre che $\hat{\vartheta}_n \xrightarrow{P} \vartheta^0$, poiché massimizza la verosimiglianza. La dimostrazione formale richiede alcune ipotesi tecniche aggiuntive.

Caso derivabile

Si assume che

1. $\ell(\vartheta; Y)$ sia derivabile e che $\hat{\vartheta}_n$ sia l'unica soluzione dell'equazione $\ell_*(\vartheta) = 0$.
2. $\ell_*(\vartheta; Y)$ e il suo valore atteso in ϑ^0 siano continue.

Sotto c.c.s, l'equazione di verosimiglianza è

$$\frac{1}{n}\sum_{i=1}^{n} \ell_*(\vartheta; Y_i) = 0.$$

Per la legge dei grandi numeri,

$$\frac{1}{n}\sum_{i=1}^{n} \ell_*(\vartheta; Y_i) \xrightarrow{P} \mathbb{E}_{\vartheta^0}\left[\ell_*(\vartheta; Y_i)\right],$$

dove il limite si annulla in $\vartheta = \vartheta^0$ per Wald. Inoltre, se il modello è regolare, la funzione $\mathbb{E}_{\vartheta^0}\left[\ell_*(\vartheta; Y_i)\right]$ è monotona decrescente in un intorno di ϑ^0, visto che

$$\frac{\partial}{\partial\vartheta}\,\mathbb{E}_{\vartheta^0}\left[\ell_*(\vartheta; Y_i)\right]\Big|_{\vartheta=\vartheta^0} = -i_1(\vartheta^0) < 0.$$

Siccome si assume (2) che $H(\vartheta) = \frac{\partial}{\partial\vartheta}\,\mathbb{E}_{\vartheta^0}\left[\ell_*(\vartheta; Y_i\right]$ è continua in ϑ, il segno rimane negativo in un intorno di ϑ^0.

Teo. (Consistenza di $\hat{\vartheta}_n$)
Sia $p = 1$ e si assuma un c.c.s da un modello statistico regolare con $H(\vartheta)$ continua in un intorno di ϑ^0. Se $^1/_n\ell_*(\vartheta; Y) = 0$ ha un'unica soluzione $\hat{\vartheta}_n$, allora sotto ϑ^0 $\hat{\vartheta}_n$ converge in probabilità a ϑ^0.

Dim.
Sia $\left(\vartheta^0 - \varepsilon, \vartheta^0 + \varepsilon\right)$ un intorno in cui $H(\vartheta)$ è continua. Poiché $\hat{\vartheta}_n$ è tale che $l_*(\vartheta) = 0$, allora

$$\begin{cases} \ell_*(\vartheta^0 - \varepsilon; Y) > 0 \\ \ell_*(\vartheta^0 + \varepsilon; Y) < 0 \end{cases} \implies \vartheta^0 - \varepsilon < \hat{\vartheta}_n < \vartheta^0 + \varepsilon,$$

dunque, usando il fatto che $(A \implies B) \implies P(A) < P(B)$,

$$P_{\vartheta^0}\Big(\frac{1}{n}\ell_*(\vartheta^0 - \varepsilon; Y) > 0,\ \frac{1}{n}\ell_*(\vartheta^0 + \varepsilon; Y) < 0\Big) \le P_{\vartheta^0}(\vartheta^0 - \varepsilon < \hat{\vartheta}_n < \vartheta^0 + \varepsilon).$$

Mostriamo che il membro di sinistra converge a 1, e quindi per forza anche quello di destra, usando la disuguaglianza di Bonferroni $\big(P(A \cap B) \ge P(A) + P(B) - 1\big)$:

$$\frac{1}{n}\ell_*(\vartheta^0 - \varepsilon; Y) \xrightarrow{P} \mathbb{E}_{\vartheta^0}\left[\ell_*(\vartheta^0 - \varepsilon; Y_i)\right] > 0 \implies P(A) \to 1;$$

$$\frac{1}{n}\ell_*(\vartheta^0 + \varepsilon; Y) \xrightarrow{P} \mathbb{E}_{\vartheta^0}\left[\ell_*(\vartheta^0 + \varepsilon; Y_i)\right] < 0 \implies P(B) \to 1,$$

quindi

$$P_{\vartheta^0}(\vartheta^0 - \varepsilon < \hat{\vartheta}_n < \vartheta^0 + \varepsilon) \xrightarrow{n\to\infty} 1 \implies \hat{\vartheta}_n \xrightarrow{P} \vartheta^0.$$

□

Nota Il teorema precedente è generalizzabile a stimatori $\tilde{\vartheta}_n(Y)$ soluzioni di equazioni di stima non distorte $q(\vartheta; Y) = 0$, purché $q(\vartheta; Y)$ soddisfi delle condizioni analoghe (o simili) a quelle richieste sopra per ℓ_*.

Nota 2 Spesso per verificare la consistenza di uno stimatore $\tilde{\vartheta}_n(Y)$ si può utilizzare la disuguaglianza di Chebyshev, per cui

$$P_\vartheta(|\tilde{\vartheta}_n(Y) - \vartheta| \ge \varepsilon) \le \frac{\mathbb{E}_\vartheta\left[|\tilde{\vartheta}_n(Y) - \vartheta|^2\right]}{\varepsilon^2}.$$

Se lo stimatore converge in media quadratica, allora è consistente; inoltre, siccome il numeratore è pari a $\mathrm{EQM}(\hat{\vartheta}_n(Y))$, è sufficiente (ma non necessario) che

$$\begin{cases} \mathbb{E}_\vartheta\left[\tilde{\vartheta}_n(Y)\right] \xrightarrow{n\to\infty} \vartheta \\ \mathbb{V}_\vartheta\left[\tilde{\vartheta}_n(Y)\right] \xrightarrow{n\to\infty} 0 \end{cases}$$

29.2 Distribuzioni asintotiche

Vediamo le distribuzioni asintotiche di $\ell_*(\vartheta)$, $\hat{\vartheta}$ e del log-rapporto di verosimiglianza, introdotto in precedenza.

Le condizioni di validità dei risultati sono le *condizioni di regolarità* per il modello: si assume $\mathcal{F} = \{p_Y(y;\vartheta),\ \vartheta \in \Theta \subseteq \mathbb{R}^p,\ y \in \mathcal{Y}\}$, tale che

- Il supporto $\mathcal{Y}$ non dipende da ϑ.
- Il modello è identificabile e correttamente specificato, con $p^0(y) = p_Y(y;\vartheta^0)$, ϑ^0 punto interno di Θ.
- $\ell(\vartheta)$ si può espandere in serie di Taylor in un intorno di ϑ^0 fino al secondo ordine, il cui resto è uniformemente limitato. In particolare, significa che

$$\left|\frac{\partial^3}{\partial\vartheta^3}\ell(\vartheta)\right| \leq C.$$

- Il valore atteso nullo di $\ell(\vartheta)$ e delle sue derivate fino alla terza è finito. In particolare, valgono le identità di Bartlett e dell'informazione, con $i(\vartheta^0)$ definita positiva.
- La funzione di Y che limita il valore assoluto della derivata terza ha valore atteso nullo.
- $\hat{\vartheta}_n$ è ben definito a.s. quando $n \to \infty$ ed è consistente.

Il senso di utilizzare $n \longrightarrow \infty$ è per considerare una quantità crescente di informazione, poiché $i(\vartheta) = O(n)$. In particolare, nel caso di c.c.s, $i(\vartheta) = ni_1(\vartheta)$, mentre se non fossero indipendenti si potrebbe intendere $i_1(\vartheta)$ come *informazione media per osservazione*:

$$i_1(\vartheta) = \lim_{n\to\infty} \frac{i(\vartheta)}{n}.$$

29.2.1 Distribuzione asintotica di $\ell_*(\vartheta)$

Nel caso $p = 1$, sotto c.c.s $\ell_*(\vartheta;Y) = \sum_{i=1}^n \ell_*(\vartheta;Y_i)$ e

$$\mathbb{E}_\vartheta\left[\ell_*(\vartheta;Y_i)\right] = 0$$

$$\mathbb{V}_\vartheta\left[\ell_*(\vartheta;Y_i)\right] = i_1(\vartheta)$$

Assumendo $i_1(\vartheta) > 0$ finita, quando ϑ è il vero valore del parametro, per il teorema del limite centrale

$$\frac{\ell_*(\vartheta;Y)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, i_1(\vartheta)).$$

per cui se n è sufficientemente grande, la distribuzione approssimata è

$$\ell_*(\vartheta;Y) \dot{\sim} \mathcal{N}(0, i(\vartheta)).$$

Nel caso multiparametrico, analogamente si ottiene

$$\ell_*(\vartheta;Y) \dot{\sim} \mathcal{N}_p(0, i(\vartheta)).$$

Nota In entrambi i casi vale se la funzione punteggio è valutata nel *vero* valore del parametro.

29.2.2 Distribuzione asintotica di $\hat{\vartheta}_n$

Sempre nel caso $p = 1$, si ha che $\hat{\vartheta}_n$ è tale che $\ell_*(\hat{\vartheta}_n) = 0$ e $\hat{\vartheta}_n - \vartheta = o_p(1)$ cioè converge in probabilità a 0.

Si consideri lo sviluppo della funzione score

$$0 = \ell_*(\hat{\vartheta}_n) = \ell_*(\vartheta) + (\hat{\vartheta}_n - \vartheta)\ell_{**}(\vartheta) + \frac{1}{2}(\hat{\vartheta}_n - \vartheta)^2\ell_{***}(\tilde{\vartheta}_n),$$

dove $|\tilde{\vartheta}_n - \vartheta| < |\hat{\vartheta}_n - \vartheta|$. Allora,

$$\ell_*(\vartheta) = (\hat{\vartheta}_n - \vartheta)j(\vartheta) - \frac{1}{2}(\hat{\vartheta}_n - \vartheta)^2\ell_{***}(\tilde{\vartheta}_n),$$

per cui se $i(\vartheta) > 0$ si può scrivere

$$\ell_*(\vartheta)i(\vartheta)^{-1} = (\hat{\vartheta}_n - \vartheta)j(\vartheta)i(\vartheta)^{-1} - \frac{1}{2}(\hat{\vartheta}_n - \vartheta)^2\ell_{***}(\tilde{\vartheta}_n)i(\vartheta)^{-1}.$$

Raccogliendo $\sqrt{n}$ a entrambi i membri,

$$\frac{\sqrt{n}\ell_*(\vartheta)}{i(\vartheta)} = \sqrt{n}(\hat{\vartheta}_n - \vartheta)\left[\frac{j(\vartheta)}{i(\vartheta)} - \frac{1}{2}(\hat{\vartheta}_n - \vartheta)\frac{\ell_{***}(\tilde{\vartheta}_n)}{i(\vartheta)}\right]$$

Per la legge dei grandi numeri,

$$\frac{j(\vartheta)}{i(\vartheta)} \xrightarrow{P} 1$$

e inoltre

$$\frac{\ell_{***}(\tilde{\vartheta}_n)}{ni_1(\vartheta)} \xrightarrow{P} \frac{\mathbb{E}_\vartheta\left[\ell_{***}(\vartheta; Y)\right]}{i_1(\vartheta)} < \infty \qquad \left(\tilde{\vartheta}_n \xrightarrow{P} \vartheta \text{ poiché } \hat{\vartheta} \xrightarrow{P} \vartheta\right),$$

per cui nella parentesi quadra il secondo termine è trascurabile:

$$\frac{1}{2}(\hat{\vartheta}_n - \vartheta)\frac{\ell_{***}(\tilde{\vartheta}_n)}{i(\vartheta)} = o_p(1).$$

Allora,

$$\left[\frac{j(\vartheta)}{i(\vartheta)} - \frac{1}{2}(\hat{\vartheta}_n - \vartheta)\frac{\ell_{***}(\tilde{\vartheta}_n)}{i(\vartheta)}\right] \xrightarrow{P} 1,$$

per cui per il teorema di Slutsky, quando ϑ è il vero valore del parametro,

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta) \dot{\sim} \sqrt{n}\ell_*(\vartheta)/i(\vartheta) \dot{\sim} \mathcal{N}(0, i_1(\vartheta)^{-1}),$$

da cui

$$\hat{\vartheta}_n \dot{\sim} \mathcal{N}_p(\vartheta, i(\vartheta)^{-1}).$$

Commenti

- La varianza asintotica $i(\vartheta)^{-1}$ si può sostituire con $i(\hat{\vartheta})^{-1}$ o $j(\hat{\vartheta})^{-1}$.

– $\hat{\vartheta}_n$ è asintoticamente non distorto, in generale la distorsione è dell'ordine di $O(n^{-1})$.

– $\hat{\vartheta}_n$ è asintoticamente efficiente, in quanto la sua varianza raggiunge il limite inferiore di Cramér-Rao.

– Visto che $\sqrt{n}(\hat{\vartheta}_n - \vartheta) \xrightarrow{P} \mathcal{N}(0,1)$, si ha che

$$\hat{\vartheta}_n - \vartheta = O_p(n^{-1/2}),$$

che è la "velocità di convergenza" rispetto ad n.

Def. (Ordine asintotico)
Si dice che la successione di v.c. Y_n è *asintoticamente di ordine n^α in probabilità*, e si indica con $O_p(n^\alpha)$, se

$$\forall \varepsilon > 0 \; \exists A_\varepsilon > 0, n_\varepsilon : \forall n \geq n_\varepsilon, \quad P\left(\left|\frac{Y_n}{n^\alpha}\right| < A_\varepsilon\right) > 1 - \varepsilon.$$

Inoltre, una successione è di ordine $O_p(1)$ se è limitata in probabilità. Infine, se $Y_n \xrightarrow{d} Y$, allora Y_n è $O_p(1)$.

29.2.3 Distribuzione asintotica di $\ell(\vartheta) - \ell(\hat{\vartheta})$

Si è visto che la verosimiglianza relativa $L(\vartheta)/L(\hat{\vartheta})$ si può utilizzare per definire regioni di stima e statistiche test; è di interesse quindi ottenere una distribuzione asintotica per questa quantità. Si considererà la quantità

$$W(\vartheta) = -2 \log \frac{L(\vartheta)}{L(\hat{\vartheta})} = 2\big\{\ell(\hat{\vartheta}) - \ell(\vartheta)\big\},$$

per cui regioni di stima $\left\{\vartheta : \frac{L(\vartheta)}{L(\hat{\vartheta})} \geq c\right\}$ saranno equivalenti a $\big\{\vartheta : W(\vartheta) \leq e^{-k/2}\big\}$.

In modo analogo, statistiche test che rifiutano per valori piccoli di $L(\vartheta)/L(\hat{\vartheta})$ sono equivalenti a test che rifiutano le stesse ipotesi per valori grandi di $W(\vartheta)$.

Si consideri di nuovo l'approssimazione di Taylor,

$$\ell(\vartheta_0) = \ell(\hat{\vartheta}_n) + (\vartheta_0 - \hat{\vartheta}_n)\cancelto{0}{\ell_*(\hat{\vartheta})_n} + \frac{1}{2}(\vartheta_0 - \hat{\vartheta}_n)^2 \ell_{**}(\hat{\vartheta}_n) + \frac{1}{6}(\vartheta_0 - \hat{\vartheta}_n)^3 \ell_{***}(\tilde{\vartheta}_n),$$

con $|\tilde{\vartheta}_n - \vartheta_0| < |\hat{\vartheta}_n - \vartheta_0|$. Ricavando l'espressione per $W(\vartheta_0)$, si ha

$$\begin{aligned} W(\vartheta_0) &= 2\big\{\ell(\hat{\vartheta}_n) - \ell(\vartheta_0)\big\} \\ &= (\vartheta_0 - \hat{\vartheta}_n)^2 j(\hat{\vartheta}_n) - \frac{1}{3}(\vartheta_0 - \hat{\vartheta}_n)^3 \ell_{***}(\tilde{\vartheta}_n) \\ &= (\vartheta_0 - \hat{\vartheta}_n)^2 i(\vartheta_0) \underbrace{\frac{j(\hat{\vartheta}_n)}{i(\vartheta_0)}}_{\xrightarrow{P} 1} - \frac{1}{3} \underbrace{n(\vartheta_0 - \hat{\vartheta}_n)^3}_{\xrightarrow{P} 0} \cdot \underbrace{\frac{\ell_{***}(\tilde{\vartheta}_n)}{n}}_{\xrightarrow{P} \; \mathbb{E}_{\vartheta_0}[\ldots] < \infty}. \end{aligned}$$

La convergenza a 0 di $n(\hat{\vartheta}_n - \vartheta_0)^3$ è data dal fatto che

$$n(\hat{\vartheta}_n - \vartheta_0)^3 = O_p(n^{-\frac{3}{2}} \cdot n) = o_p(n^{-\frac{1}{2}}) = o_p(1).$$

Dunque,

$$W(\vartheta_0) = (\hat{\vartheta}_n - \vartheta_0)^2 i(\vartheta_0) + o_p(1).$$

Poiché $(\hat{\vartheta} - \vartheta_0)\sqrt{i(\vartheta_0)} \xrightarrow{d} \mathcal{N}(0,1)$, si ha che

$$W(\vartheta_0) \xrightarrow{d} \chi^2_1 \quad \text{sotto } \vartheta_0.$$

Nel caso multiparametrico, passaggi analoghi mostrano che

$$W(\vartheta_0) = (\hat{\vartheta}_n - \vartheta_0)^\mathsf{T} i(\vartheta_0)(\hat{\vartheta}_n - \vartheta_0) + o_p(1) \xrightarrow{d} \chi^2_p \quad \text{sotto } \vartheta_0.$$

Nota Se $X \sim \mathcal{N}_p(0, \Sigma)$, allora $X^\mathsf{T}\Sigma^{-1}X \sim \chi^2_p$.

La statistica $W(\vartheta)$ può essere usata per verifiche di ipotesi su ϑ e per regioni di stima. Dato il sistema di ipotesi $H_0 : \vartheta = \vartheta_0$ contro $H_1 : \vartheta \neq \vartheta_0$, valori grandi di $W(\vartheta_0)$ rappresentano un allontanamento da H_0.

La regione di rifiuto di livello approssimato α avrà forma

$$R = \left\{ y \in \mathcal{Y} : W(\vartheta_0) > \chi^2_{p;1-\alpha} \right\}.$$

Invece, il livello di significatività approssimato è

$$\alpha^{\text{oss}} = P_{\vartheta_0}\big(W(\vartheta_0) \geq w^{\text{oss}}\big).$$

Analogamente, una regione di stima con livello di confidenza approssimato $1 - \alpha$ corrisponde a

$$\hat{\Theta}(Y) = \left\{ \vartheta : W(\vartheta) \leq \chi^2_{p;1-\alpha} \right\}$$

Lezione 30

30.1 Statistiche legate alla verosimiglianza

Ci sono tre statistiche che possono essere usate per saggiare ipotesi su ϑ e costruire regioni di stima:

$$\left.\begin{aligned} &\textit{Wilks: } W(\vartheta) = 2\big\{\ell(\hat\vartheta) - \ell(\vartheta)\big\} \\ &\textit{Wald: } W_e(\vartheta) = (\hat\vartheta_n - \vartheta)^{\mathsf{T}} i(\vartheta)(\hat\vartheta_n - \vartheta) \\ &\textit{Rao: } W_u(\vartheta) = \ell_*(\vartheta)^{\mathsf{T}} i(\vartheta)^{-1} \ell_*(\vartheta) \end{aligned}\right\} \dot\sim \chi_p$$

Le tre statistiche sono asintoticamente equivalenti e considerano diversi aspetti della verosimiglianza

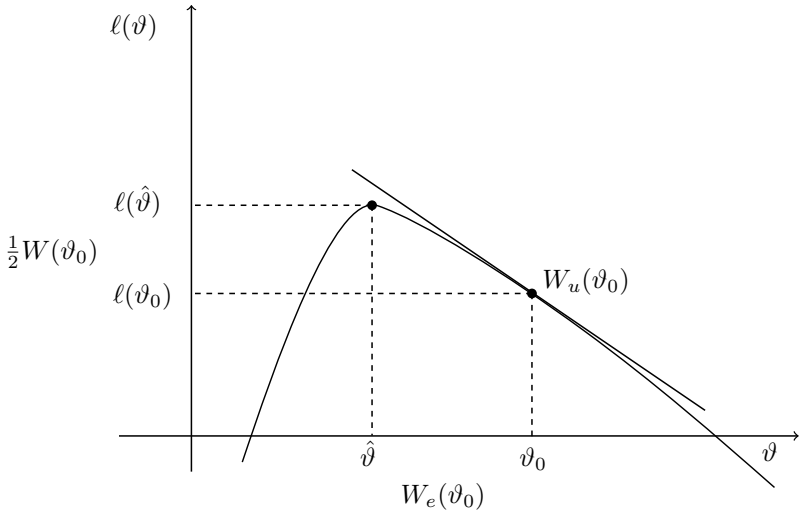


Figura 21: Comportamento delle tre statistiche di verosimiglianza.

Commenti

- $W(\vartheta)$ e $W_u(\vartheta)$ sono invarianti rispetto a riparametrizzazioni, mentre $W_e(\vartheta)$ non lo è, a causa di $i(\vartheta)$.
- $W_e(\vartheta)$ non rispetta necessariamente i vincoli imposti sullo spazio parametrico, perché approssima la log-verosimiglianza con una parabola centrata in $\hat\vartheta$.
- $W_u(\vartheta)$ coinvolge solo quantità calcolate in ϑ e non richiede $\hat\vartheta$, anche se tipicamente la smv è un passo precedente al test di ipotesi.
- $W_u(\vartheta)$ è instabile se ci si allontana da $\hat\vartheta$, se usato per costruire regioni di stima.
- $W(\vartheta)$ fornisce in generale risultati più accurati degli altri due.
- $W_u(\vartheta)$ è spesso accurato in modelli per dati discreti.

– In $W_e(\vartheta)$ è consuetudine sostituire $j(\hat{\vartheta})$ o $i(\hat{\vartheta})$ al posto di $i(\vartheta)$, in quanto

$$j(\vartheta) = i(\vartheta) + o_p(n)$$

$$\hat{\vartheta} = \vartheta + o_p(1)$$

30.1.1 Versioni unilaterali

Nel caso scalare, si possono considerare le versioni unilaterali delle tre statistiche, con distribuzione approssimata $\mathcal{N}(0,1)$, date da

$$\textit{Wilks: } R(\vartheta) = \operatorname{sgn}(\hat{\vartheta} - \vartheta_0)\sqrt{W(\vartheta_0)} = (\hat{\vartheta}_n - \vartheta_0)\sqrt{i(\vartheta_0)} + o_p(1)$$

$$\textit{Wald: } R_e(\vartheta) = (\hat{\vartheta}_n - \vartheta_0)\sqrt{i(\vartheta_0)}$$

$$\textit{Rao: } R_u(\vartheta) = \frac{\ell_*(\vartheta)}{\sqrt{i(\vartheta_0)}}$$

Le regioni di stima sono analoghe, se basate sulle versioni unilaterali, come anche le regioni di rifiuto dei test di ipotesi. Il vantaggio delle versioni unilaterali è che si possono utilizzare per verificare ipotesi unilaterali del tipo $H_1 : \vartheta \underset{(<)}{>} \vartheta_0$.

Ad esempio, per verificare $H_0 : \vartheta = \vartheta_0$ contro $H_1 : \vartheta > \vartheta_0$, una regione di rifiuto di livello approssimato α è data da

$$R = \{y \in \mathcal{Y} : r(\vartheta_0) > z_{1-\alpha}\}$$

e il livello di significatività osservato è

$$\alpha^{\text{oss}} = P_{\vartheta_0}\big(r(\vartheta_0) > r^{\text{oss}}\big) \approx 1 - \Phi(r^{\text{oss}}).$$

Lezione 31

31.1 Esempi di test di ipotesi approssimati

Esempio (Modello binomiale)
Sia $Y_i \sim \text{Bin}(n, \vartheta)$, la log-verosimiglianza è

$$\ell(\vartheta) = y \log \vartheta + (n - y) \log(1 - \vartheta).$$

La funzione score e la derivata seconda sono

$$\ell_*(\vartheta) = \frac{y}{\vartheta} - \frac{n-y}{1-\vartheta},$$

$$\ell_{**}(\vartheta) = -\frac{y}{\vartheta^2} - \frac{n-y}{(1-\vartheta)^2};$$

L'equazione di verosimiglianza ha soluzione $\hat{\vartheta} = y/n$, che è massimo globale di $\ell(\vartheta)$. L'informazione osservata e attesa sono rispettivamente

$$j(\vartheta) = \frac{y}{\vartheta^2} + \frac{n-y}{(1-\vartheta)^2};$$

$$i(\vartheta) = \frac{n}{\vartheta(1-\vartheta)}.$$

Infine, si noti che $j(\hat{\vartheta}) = \frac{n}{\hat{\vartheta}(1-\hat{\vartheta})} = i(\hat{\vartheta})$.

Considerando i dati sul comportamento aggressivo, si vuole verificare l'ipotesi nulla $H_0 : \vartheta \leq 0.2$ contro $H_1 : \vartheta > 0.2$, utilizzando le tre statistiche test $r(\vartheta), r_e(\vartheta), r_u(\vartheta)$. Si ha

$$W(\vartheta) = 2n\Big\{\hat{\vartheta} \log \frac{\hat{\vartheta}}{\vartheta} + (1 - \hat{\vartheta}) \log \frac{1-\hat{\vartheta}}{1-\vartheta}\Big\}.$$

Inoltre, si verifica che

$$r_e(\vartheta) = (\hat{\vartheta} - \vartheta)\sqrt{j(\hat{\vartheta})} = \frac{\hat{\vartheta} - \vartheta}{\sqrt{\frac{\hat{\vartheta}(1-\hat{\vartheta})}{n}}}$$

$$r_u(\vartheta) = \frac{\ell_*(\vartheta)}{\sqrt{i(\vartheta)}} = \frac{\hat{\vartheta} - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}}.$$

I corrispettivi livelli di significatività approssimati sono

$$r(\vartheta): \quad \alpha^{\text{oss}} = 0.051$$

$$r_e(\vartheta): \quad \alpha^{\text{oss}} = 0.056$$

$$r_u(\vartheta): \quad \alpha^{\text{oss}} = 0.069$$

Si possono usare queste statistiche per trovare intervalli di confidenza approssimati per ϑ.

Per la statistica $W(\vartheta)$, quasi sempre è necessario risolvere numericamente l'equazione

$$W(\vartheta) - \chi^2_{1;1-\alpha} = 0,$$

Esempio (GLM esponenziale)
(?) TODO

Lezione 32

32.1 Test localmente più potente

A volte si possono costruire test uniformemente più potenti (UMP) nel caso di verifica di ipotesi unilaterali per parametri scalari, se il modello ha rapporto di verosimiglianza monotono.

Altre volte, è possibile costruire test con potenza più alta possibile per $H_0 : \vartheta = \vartheta_0$ contro $H_1 : \vartheta \underset{(<)}{>} \vartheta_0$ almeno per alternative locali, cioè quando ϑ è vicino a ϑ_0: il test così trovato è detto *localmente più potente* (LMP).

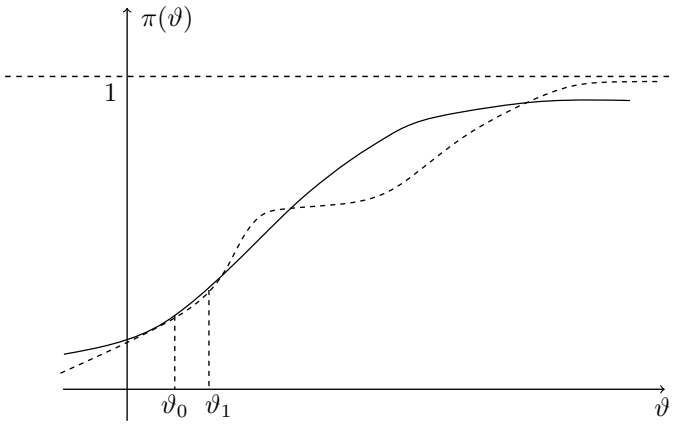


Figura 22: LMP

Sia $\vartheta_1 = \vartheta_0 + \varepsilon$, allora per discriminare tra due valori si ha

$$\begin{aligned}
\frac{p(y;\vartheta_1)}{p(y;\vartheta_0)} &= \frac{p(y;\vartheta_0+\varepsilon)}{p(y;\vartheta_0)} \\
&= \frac{1}{p(y;\vartheta_0)}\left(p(y;\vartheta_0) + \varepsilon\frac{\partial}{\partial\vartheta}p(y;\vartheta)|_{\vartheta=\vartheta_0} + \dots\right) \\
&= 1 + \varepsilon\cdot\frac{\frac{\partial}{\partial\vartheta}p(y;\vartheta_0))}{p(y;\vartheta_0)} \\
&= 1 + \varepsilon\cdot\ell_*(\vartheta_0)
\end{aligned}$$

Quando $\varepsilon \to 0$, il rapporto di verosimiglianza e la funzione punteggio sono equivalenti. Dunque, il test localmente più potente rifiuta H_0 per valori grandi della funzione punteggio.

In particolare, visto che per n grande, $\ell_*(\vartheta) \dot{\sim} \mathcal{N}(0, i(\vartheta))$, la regione di rifiuto di un test localmente

più potente sarà

$$R_\alpha = \left\{ y \in \mathcal{Y} : \ell_*(\vartheta_0; y) \geq i(\vartheta_0)^{\frac{1}{2}} z_{1-\alpha} \right\}$$
$$= \{ y \in \mathcal{Y} : r_u(\vartheta_0) \geq z_{1-\alpha} \}.$$

Sotto l'ipotesi alternativa, si ha che

$$\mathbb{E}_{\vartheta_1}\left[\ell_*(\vartheta_0)\right] = \int \ell_*(\vartheta_0) p(y; \vartheta_0 + \varepsilon)\, dy$$
$$= \int \ell_*(\vartheta_0) \left(p(y; \vartheta_0) + \varepsilon \ell_*(\vartheta_0) p(y; \vartheta_0) + \ldots \right) dy$$
$$\approx \varepsilon \cdot i(\vartheta_0).$$

In modo simile, anche $\mathbb{V}_{\vartheta_1}\left[\ell_*(\vartheta_0)\right] = i(\vartheta_0) + O(\varepsilon \cdot n)$. Dunque, la potenza del test LMP è

$$P_{\vartheta_1}\left(\ell_*(\vartheta_0) > i(\vartheta_0)^{\frac{1}{2}} z_{1-\alpha}\right) = P_{\vartheta_1}\left(\frac{\ell_*(\vartheta_0) - \varepsilon i(\vartheta_0)}{i(\vartheta_0)^{\frac{1}{2}}} > \frac{i(\vartheta_0)^{\frac{1}{2}} z_{1-\alpha} - \varepsilon i(\vartheta_0)}{i(\vartheta_0)^{\frac{1}{2}}} \right)$$
$$\approx 1 - \Phi(z_{1-\alpha} - \varepsilon \cdot i(\vartheta_0)^{\frac{1}{2}})$$
$$= \Phi(z_\alpha + \varepsilon \cdot i(\vartheta_0)^{\frac{1}{2}}),$$

dove $\varepsilon = \vartheta_1 - \vartheta_0$ e $i(\vartheta_0) = n \cdot i_1(\vartheta_0)$. Dipende dunque da quanto grande è n, dall'informazione in ϑ_0 e da quanto distante è ϑ_1 da ϑ_0.

Esempio (Cauchy(ϑ, 1))

Sia

$$p(y_i; \vartheta) = \left(\pi(1 + (y_i - \vartheta)^2)\right)^{-1},$$

la verosimiglianza è

$$\ell(\vartheta) = -\sum_{i=1}^{n} \log\left(1 + (y_i - \vartheta)^2\right).$$

Qui non esiste un test UMP per ipotesi unilaterali, ma si può usare il test localmente più potente basato sullo score:

$$\ell_*(\vartheta) = \sum_{i=1}^{n} \frac{2(y_i - \vartheta)}{1 + (y_i - \vartheta)^2}.$$

L'informazione attesa è

$$i(\vartheta) = \mathbb{V}_\vartheta\left[\ell_*(\vartheta)\right]$$
$$= n\, \mathbb{E}_\vartheta\left[\left(\frac{2(Y_i) - \vartheta}{1 + (Y_i - \vartheta)^2} \right)^2 \right]$$
$$= \ldots \text{(?) guardare passaggi con t-Student}$$
$$= \frac{n}{2}$$

Quindi, il test LMP rifiuta H_0 per valori grandi di

$$r_u(\vartheta_0) = \sqrt{\frac{2}{n}} \sum_{i=1}^{n} \frac{2(y_i - \vartheta_0)}{1 + (y_i - \vartheta_0)^2},$$

che ha distribuzione approssimata nulla $\mathcal{N}(0, 1)$.

32.2 Parametri di disturbo e verosimiglianza profilo

Si consideri il caso $\vartheta = (\psi, \lambda)$, con ψ parametro di interesse. La log-verosimiglianza profilo si comporta asintoticamente come una verosimiglianza propria, per cui è possibile usarla per verificare test di ipotesi e costruire intervalli di confidenza.

Si può definire

$$\begin{aligned} W_p(\psi) &= 2\big\{\ell_p(\hat{\psi}) - \ell_p(\psi)\big\} \\ &= 2\big\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\big\} \end{aligned}$$

Si può vedere che, sotto ϑ_0,

$$W_p(\psi_0) \dot{\sim} \chi^2_k,$$

cioè asintoticamente $\ell_p(\psi)$ si comporta come una log-verosimiglianza per ψ.

Dim.

$$\begin{aligned} W_p(\psi_0) &= 2\big\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi_0, \hat{\lambda}_{\psi_0})\big\} \\ &= 2\big\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi_0, \lambda_0) + \ell(\psi_0, \lambda_0) - \ell(\psi_0, \hat{\lambda}_{\psi_0})\big\} \\ &= 2\big\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi_0, \lambda_0)\big\} - 2\big\{\ell(\psi_0, \hat{\lambda}_{\psi_0}) - \ell(\psi_0, \lambda_0)\big\}, \end{aligned}$$

dove

- $2\big\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi_0, \lambda_0)\big\}$ è la statistica log-rapporto di verosimiglianza per l'ipotesi nulla completa e ha la distribuzione asintotica nulla χ^2_p, uguale a quella di

$$\ell_*(\psi_0, \lambda_0)^\mathsf{T} i(\psi_0, \lambda_0)^{-1} \ell_*(\psi_0, \lambda_0).$$

- $2\big\{\ell(\psi_0, \hat{\lambda}_{\psi_0}) - \ell(\psi_0, \lambda_0)\big\}$ è la statistica per verificare l'ipotesi nulla $H_0 : \lambda = \lambda_0$ nel sottomodello con ψ_0 fissato, ed ha la distribuzione χ_{p-k}, uguale a quella di

$$\ell_\lambda(\psi_0, \lambda_0)^\mathsf{T} i_{\lambda\lambda}(\psi_0, \lambda_0)^{-1} \ell_\lambda(\psi_0, \lambda_0)$$

Si ha quindi

$$W_p(\psi_0) = \ell_*(\psi_0, \lambda_0)^\mathsf{T} i(\psi_0, \lambda_0)^{-1} \ell_*(\psi_0, \lambda_0) - \ell_\lambda(\psi_0, \lambda_0)^\mathsf{T} i_{\lambda\lambda}(\psi_0, \lambda_0)^{-1} \ell_\lambda(\psi_0, \lambda_0) + o_p(1).$$

Nel seguito si omette l'argomento delle funzioni, per semplificare la notazione. Considerato ψ, λ entrambi scalari (no inversioni di blocchi che son pesanti), si ha

$$\begin{aligned} W_p &= (\ell_\psi, \ell_\lambda) \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}^{-1} \begin{pmatrix} \ell_\psi \\ \ell_\lambda \end{pmatrix} - \ell_\lambda i_{\lambda\lambda}^{-1} \ell_\lambda \\ &= \frac{1}{i_{\psi\psi} i_{\lambda\lambda} - i_{\psi\lambda}^2} (i_{\lambda\lambda} \ell_\psi^2 - 2 i_{\psi\lambda} \ell_\psi \ell_\lambda + i_{\psi\psi} \ell_\lambda^2) - \frac{\ell_\lambda^2}{i_{\lambda\lambda}} \\ &= \dots \\ &= \left(\frac{\ell_\psi - \frac{i_{\psi\lambda}}{i_{\lambda\lambda}} \ell_\lambda}{\sqrt{i_{\psi\psi} - \frac{i_{\psi\lambda}^2}{i_{\lambda\lambda}}}} \right)^2, \end{aligned}$$

con

$$\mathbb{E}_{\psi_0, \lambda_0} \left[\ell_\psi - \frac{i_{\psi\lambda}}{i_{\lambda\lambda}} \ell_\lambda \right] = \mathbb{E}_{\psi_0 \lambda_0} \left[\ell_\psi \right] - \frac{i_{\psi\lambda}}{i_{\lambda\lambda}} \mathbb{E}_{\psi_0, \lambda_0} \left[\ell_\psi \right] = 0 - 0.$$

$$\mathbb{V}_{\lambda_0, \psi_0} \left[\ell_\psi - \frac{i_{\psi\lambda}}{i_{\lambda\lambda}} \ell_\lambda \right] = (?) FINIRE \dots = i_{\psi\psi} - \frac{i_{\psi\lambda}^2}{i_{\lambda\lambda}}.$$

Dunque, per il teorema limite centrale, $W_p(\psi_0) \xrightarrow{d} \chi_1^2$ e di conseguenza

$$r_p(\psi_0) = \operatorname{sgn}(\hat{\psi} - \psi_0) \sqrt{W_p(\psi_0)} \xrightarrow{d} \mathcal{N}(0, 1).$$

Se si considerano valori generici di k, p si possono ripetere gli stessi passaggi tramite algebra matriciale e risultati su inversioni di matrici a blocchi, per ottenere (conti)

$$W_p(\psi_0) \dot{\sim} \chi_k^2.$$

□

Si possono definire analogamente le quantità di Wald e Rao tramite

$$\hat{\vartheta} \dot{\sim} \mathcal{N}_p \left(\binom{\psi}{\lambda}, i(\psi, \lambda)^{-1} \right) \implies \hat{\psi} \dot{\sim} \mathcal{N}_k \big(\psi, i^{\psi\psi}(\psi, \lambda) \big),$$

da cui si ottiene

$$W_{p_e}(\psi) = (\hat{\psi} - \psi)^\intercal \left(j^{\psi\psi}(\hat{\vartheta}) \right)^{-1} (\hat{\psi} - \psi) \dot{\sim} \chi_k^2.$$

La statistica punteggio, infine, si ottiene direttamente sostituendo a λ la sua stima $\hat{\lambda}_\psi$ nella forma quadratica, ottenendo

$$W_{p_u}(\psi) = \frac{\partial}{\partial \psi} \ell_p(\psi)^\intercal i^{\psi\psi}(\psi, \hat{\lambda}_\psi) \frac{\partial}{\partial \psi} \ell_p(\psi) \dot{\sim} \chi_k^2.$$

In questa statistica, si può sostituire l'informazione osservata, mentre senza parametro di disturbo si usa quasi sempre l'informazione attesa. Per $k = 1$, si possono definire anche le corrispondenti versioni unilaterali delle statistiche, con distribuzioni approssimate normali standard, sotto ψ.

Commenti

- I risultati asintotici sono stati ottenuti sotto condizioni di regolarità, per cui la qualità delle approssimazioni dipende anche dalla dimensione del parametro.
- L'invarianza rispetto a riparametrizzazioni si intende rispetto a quelle che *non alterano l'interesse*, e sono invarianti le statistiche $W_p(\psi)$ e $W_{p_u}(\psi)$ (come anche le versioni unilaterali). Come nel caso senza parametri di disturbo, $W_{p_e}(\psi)$ non è invariante per riparametrizzazioni.

Lezione 33

Esempio ($\mathcal{N}(\mu, \sigma^2)$)
Se σ^2 è il parametro di interesse, data

$$\ell(\mu, \sigma^2) = -\frac{n}{\sigma^2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2,$$

la stima di μ per σ^2 fissato è soluzione di

$$\ell_\mu(\mu, \sigma^2) = 0 \iff \hat{\mu}_{\sigma^2} = \hat{\mu} = \bar{y}.$$

La verosimiglianza profilo per σ^2 è quindi

$$\ell_p(\sigma^2) = \ell(\bar{y}, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

che si massimizza in $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$. La funzione punteggio profilo è

$$\begin{aligned}\frac{\partial \ell_p(\sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{n}{2(\sigma^2)^2}\hat{\sigma}^2 \\ &= \frac{n}{2(\sigma^2)^2}(\hat{\sigma}^2 - \sigma^2)\end{aligned}$$

e l'informazione profilo osservata è

$$j_p(\hat{\sigma}^2) = \frac{n}{2(\hat{\sigma}^2)^2}.$$

Infine, l'informazione attesa è

$$i(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2(\sigma^2)^2} \end{pmatrix} \implies i(\mu, \sigma^2)^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2(\sigma^2)^2}{n} \end{pmatrix}$$

Le tre statistiche, con distribuzione approssimata χ^2_1, sono quindi

$$W_p(\sigma^2) = n\Big(\frac{\hat{\sigma}^2 - \sigma^2}{\sigma^2} - \log\left(\hat{\sigma}^2/\sigma^2\right)\Big);$$

$$W_{p_e}(\sigma^2) = (\hat{\sigma}^2 - \sigma^2)^2\frac{n}{2(\hat{\sigma}^2)^2};$$

$$W_{p_u}(\sigma^2) = (\hat{\sigma}^2 - \sigma^2)^2\frac{n}{2(\sigma^2)^2}.$$

Le corrispondenti versioni unilaterali si ottengono in modo automatico prendendo la radice (o radice con segno) delle quantità.

Esercizio: Trovare la condizione per cui l'estremo inferiore dell'intervallo basato su $W_{p_e}(\sigma^2)$

risulta negativo e trovare quando l'intervallo basato su $W_{p_u}(\sigma^2)$ risulta della forma $(a, +\infty)$. Mostrare che le due condizioni sono le stesse.

Esempio ($\mathbf{Exp(\alpha + \beta x_i)}$)
Dato un modello $Y_i \sim \text{Exp}(\lambda_i)$, con $\lambda_i = \alpha + \beta x_i$ e x_i dicotomica, si vuole studiare il parametro di interesse β in presenza del parametro di disturbo α.

$$\ell_\alpha(\alpha, \beta) = -n_0\bar{y}_0 - n_1\bar{y}_1 + \frac{n_0}{\alpha} + \frac{n_1}{\alpha + \beta} = 0,$$

che porta a un'equazione

$$(n_0\bar{y}_0 + n_1\bar{y}_1)\alpha^2 + \big(\beta(n_0\bar{y}_0 + n_1\bar{y}_1) - n_0 - n_1\big)\alpha - \beta n_0 = 0.$$

Si ha inoltre uno spazio parametrico della forma $\alpha + \beta > 0 \implies \alpha > \max(0, -\beta)$. Quindi, delle due soluzioni è solo ammissibile quella con segno positivo

$$\hat{\alpha}_\beta = \frac{-b + \sqrt{b^2 - 4ac}}{2a}.$$

Si ottiene la log-verosimiglianza profilo per β

$$\ell_p(\beta) = -\hat{\alpha}_\beta \left(\frac{n_0}{\hat{\alpha}_\beta} + \frac{n_1}{\hat{\alpha}_\beta + \beta} \right) - \beta \frac{n_1}{\hat{\alpha}_\beta + \beta} + n_0 \log \hat{\alpha}_\beta + n_1 \log(\hat{\alpha}_\beta + \beta)$$

e

$$W_p(\beta) = 2 \left\{ \ell_p(\hat{\beta}) - \ell_p(\beta) \right\}.$$

Ricordando che

$$i^{\beta\beta} = \big(i_{\beta\beta} - i_{\alpha\beta} i_{\alpha\alpha}^{-1} i_{\beta\alpha}\big)^{-1},$$

si ottiene il valore dell'informazione attesa

$$i^{\beta\beta} = \frac{\alpha^2}{n_0} + \frac{(\alpha + \beta)^2}{n_1} = j^{\beta\beta},$$

da cui la statistica di Wald

$$W_{p_e}(\beta) = (\hat{\beta} - \beta)^2 j^{\beta\beta}(\hat{\alpha}, \hat{\beta})^{-1} \underset{\beta}{\dot{\sim}} \chi^2_1.$$

Infine, la statistica di Rao è

$$W_{p_u}(\beta) = \ell_\beta(\beta, \hat{\alpha}_\beta)^2 i^{\beta\beta}(\beta, \hat{\alpha}_\beta).$$

Sotto l'ipotesi nulla, $\hat{\alpha}_0 = \frac{1}{\bar{y}}$, per cui si ottengono facilmente

33.1 Verifica di ipotesi su parametri definiti implicitamente

A volte, l'ipotesi nulla $H_0 : \vartheta = \Theta_0$ si può esprimere con k vincoli indipendenti e regolari sul parametro ϑ, del tipo

$$\begin{cases} g_1(\vartheta) = 0 \\ g_2(\vartheta) = 0 \\ \vdots \\ g_k(\vartheta) = 0 \end{cases}$$

In alcuni casi, si può riparametrizzare il modello come $\vartheta = (\kappa, \psi)$, in modo che i vincoli corrispondano alla verifica dell'ipotesi $\psi = \psi_0$ tramite log-verosimiglianza profilo.

Come regola generale, si utilizza $W(\vartheta)$, facendo la differenza tra massimo globale e massimo vincolato sotto l'ipotesi nulla

$$W_p^{H_0} = 2\big\{\ell(\hat{\vartheta}) - \ell(\hat{\vartheta}_0)\big\}.$$

Quando i k vincoli sono indipendenti, si ha che

$$W_p^{H_0} \xrightarrow[H_0]{d} \chi^2_k.$$

Esempio ($\mathbf{Exp(\alpha + \beta x_i)}$)
Nel caso precedente, si poteva riparametrizzare il modello tramite $(\alpha, \beta) \to (\mu_0, \mu_1)$, dove

$$\mu_i = \mathbb{E}\left[Y_i | x_i = j\right], \quad j = 0, 1.$$

La log-verosimiglianza è

$$\ell(\mu_0, \mu_1) = -n_0 \log \mu_0 + \frac{n_0}{\mu_0}\bar{y}_0 - n_1 \log \mu_1 - \frac{n_1}{\mu_1}\bar{y}_1$$

e l'ipotesi nulla $H_0 : \beta = 0$ nella nuova parametrizzazione corrisponde all'ipotesi nulla $H_0 : \mu_0 = \mu_1$. La log-verosimiglianza è pari a

$$\ell(\mu, \mu) = \ldots = -(n_0 + n_1) \log \mu - \frac{1}{\mu}(n_0\bar{y}_0 + n_1\bar{y}_1),$$

con stima di massima verosimiglianza $\hat{\mu} = \bar{y}$.

Lezione 34

34.1 Test W e W_p per bontà di adattamento

Esempio (Genetic linkage)
Dall'esempio del genetic linkage, se $Y = (125, 18, 20, 34)$, si assume che $Y \sim \mathrm{Mn}_k(n, \vartheta)$, dove $\vartheta = (\vartheta_1, \dots, \vartheta_k) \in \Delta_k$, Δ_k simplesso k-dimensionale.

La log-verosimiglianza è

$$\ell(\vartheta) = \sum_{j=1}^{k} y_j \log \vartheta_j.$$

Dal momento che c'è il vincolo sui ϑ_j, si può scrivere $\vartheta_k = 1 - \sum_{j=1}^{k-1} \vartheta$ e

$$\ell(\vartheta) = \sum_{j=1}^{k-1} y_j \log \vartheta_j + y_k \log \big(1 - \sum_{j=1}^{k-1} \vartheta_j\big).$$

Assumendo che tutti $y_j > 0$, si ha

$$\frac{\partial \ell(\vartheta)}{\partial \vartheta_j} = \frac{y_j}{\vartheta_j} - \frac{y_k}{\vartheta_k},$$

per cui, risolvendo le equazioni di verosimiglianza, si ottiene

$$\frac{y_j}{\hat\vartheta_j} = \frac{y_k}{\hat\vartheta_k} \implies y_j = \hat\vartheta_j \frac{y_k}{\hat\vartheta_k} \overset{\text{somma}}{\implies} \sum_{j=1}^{k-1} y_j = \frac{y_k}{\hat\vartheta_k} \sum_{j=1}^{k-1} \hat\vartheta_j,$$

dunque poiché $\sum_{j=1}^{k} \hat\vartheta_j = 1 - \hat\vartheta_k$, si ha

$$n - y_k = \frac{y_k}{\hat\vartheta_k}(1 - \hat\vartheta_k),$$

da cui $\hat\vartheta_k = \frac{y_k}{n}$. Sostituendo,

$$\frac{y_j}{\hat\vartheta_j} = n \implies \hat\vartheta_j = \frac{y_j}{n}.$$

La matrice di informazione osservata è $(k-1) \times (k-1)$ con generico elemento

$$j(\vartheta)_{ij} = \begin{cases} \frac{y_j}{\vartheta_j^2} + \frac{y_k}{\left(1+\sum_{j=1}^{k-1} \vartheta_j\right)^2} & \text{se } i = j \\ \frac{y_k}{\left(1+\sum_{j=1}^{k-1} \vartheta_j\right)^2} & \text{se } i \neq j \end{cases}$$

$$= \begin{cases} \frac{y_j}{\vartheta_j^2} + \frac{y_k}{\vartheta_k^2} & \text{se } i = j \\ \frac{y_k}{\vartheta_k^2} & \text{se } i \neq j \end{cases}$$

Sostituendo $\hat{\vartheta}$, si ottiene

$$j(\hat{\vartheta}) = n\left(\frac{1}{\hat{\vartheta}_k}\mathbf{1}_{k-1}\mathbf{1}_{k-1}^\top + \hat{D}\right), \quad \hat{D} = \operatorname{diag}\left(\frac{1}{\hat{\vartheta}_1}, \dots, \frac{1}{\hat{\vartheta}_{k-1}}\right).$$

La matrice di informazione attesa è l'analogo

$$i(\vartheta) = n\left(\frac{1}{\vartheta_k}\mathbf{1}_{k-1}\mathbf{1}_{k-1}^\top + D\right), \quad D = \operatorname{diag}\left(\frac{1}{\vartheta_1}, \dots, \frac{1}{\vartheta_{k-1}}\right).$$

Se su vuole verificare ora l'ipotesi nulla $H_0 : \vartheta = \vartheta_0 = (\frac{3}{16}, \dots, \frac{1}{16})$, si ha la statistica log-rapporto di verosimiglianza

$$\begin{aligned} W(\vartheta_0) &= 2\left(\ell(\hat{\vartheta}) - \ell(\vartheta_0)\right) \\ &= 2\sum_{i=1}^{k} y_j \log\left(\hat{\vartheta}_j/\vartheta_{0j}\right) \dot{\sim} \chi^2_3 \end{aligned}$$

in quanto il parametro ha dimensione $k - 1 = 3$. Osservando che $n\hat{\vartheta}_j = y_j$ e $n\vartheta_{0j} = a_j$, si può dividere all'interno del logaritmo per n/n, si ottiene

$$W(\vartheta_0) = 2\sum_{j=1}^{k} y_j \log\frac{y_j}{n_j}.$$

Analogamente, si vede che

$$W_u(\vartheta_0) = \sum_{k=1}^{k} \frac{(y_j - a_j)^2}{a_j} \dot{\sim} \chi^2_3,$$

che coincide con la statistica X^2 di Pearson.

Si verifichi ora l'ipotesi nulla $H_0 : \vartheta_1 = \frac{2+\omega}{4}, \vartheta_2 = \frac{1-\omega}{4}, \vartheta_3 = \frac{1-\omega}{4}, \vartheta_4 = \frac{\omega}{4}$, con $\omega \in (0, 1)$.

In questo caso, il test $W_p^{H_0}$ è dato da

$$W_p^{H_0} = 2\left(\ell(\hat{\vartheta}) - \ell(\vartheta(\hat{\omega}))\right) \dot{\sim} \chi^2_{k-1-p}$$

dove $\ell(\vartheta(\hat{\omega}))$ è il massimo della verosimiglianza nel modello sotto H_0. La distribuzione ha $k - 1 - p$ gradi di libertà, dove $p = \dim(\omega) = 1$.

Commento

La verifica di ipotesi è un test di *bontà di adattamento* del modello, che confronta le frequenze osservate y_j contro quelle attese a_j sotto il modello assunto. Questo si può fare in generale per un generico modello $p(y; \omega)$, non necessariamente multinomiale, partizionando lo spazio campionario S_Y in $E_1, \dots, E_k$, ottenendo quindi una multinomiale per $f_1, f_2, \dots, f_k$, frequenze osservate in $E_1, E_2, \dots, E_k$ e $\vartheta_j(\omega) = P(Y \in E_j; \omega)$.

34.2 Modelli non regolari

I risultati asintotici generali valgono per modelli regolari, le cui condizioni sono abbastanza generali (LM, GLM, ...) e hanno contribuito alla popolarità dei metodi di verosimiglianza. Tuttavia, ci sono diverse eccezioni che non soddisfano le condizioni di regolarità, per cui bisogna verificare le condizioni caso per caso.

Esempio (Supporto dipendente da ϑ)
Nel caso Unif$(0, \vartheta)$, si ha la funzione di verosimiglianza

$$L(\vartheta) = \prod_{i=1}^{n} \frac{1}{\vartheta} \mathbb{1}_{[0,\vartheta]}(y_i) = \frac{1}{\vartheta^n} \mathbb{1}_{[y_{(n)},\infty)}(\vartheta).$$

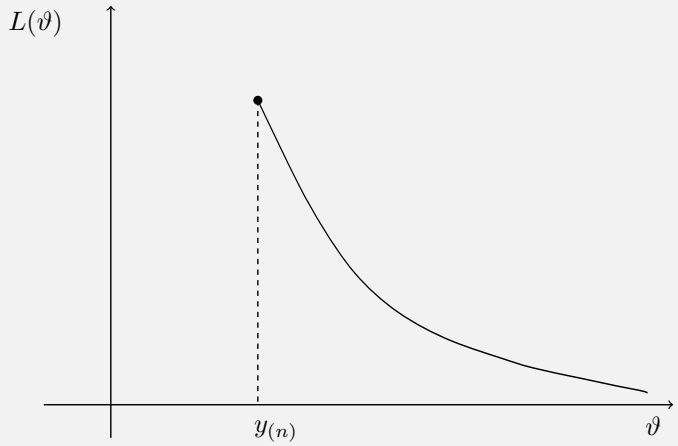


Figura 23: unif

Chiaramente, $\hat{\vartheta}$ non è soluzione dell'equazione di verosimiglianza, l'approssimazione parabolica della log-verosimiglianza relativa non è valida, $\mathbb{E}_\vartheta\left[\ell_*(\vartheta)\right] \neq 0$ e $\hat{\vartheta}$ non è asintoticamente normale. Questo non vuol però dire che lo stimatore abbia cattive proprietà, solo che non valgono i risultati asintotici generali.

Nella fattispecie, $\hat{\vartheta} = Y_{(n)}$ e la sua distribuzione è

$$P_\vartheta(\hat{\vartheta} \leq c) = P_\vartheta(Y_i \leq c)^n = \begin{cases} 0 & \text{se } c < 0 \\ (c/\vartheta)^n & \text{se } 0 \leq c < \vartheta \\ 1 & \text{se } c > \vartheta \end{cases}$$

Per $n \to \infty$, si consideri $n(\vartheta - \hat{\vartheta}) = U_n \in (0, n\vartheta)$, allora

$$\begin{aligned} P_\vartheta(U_n \le u) &= P_\vartheta(\vartheta - \hat{\vartheta} \le u/n) \\ &= 1 - P_\vartheta\big(\hat{\vartheta} < \vartheta - \frac{u}{n}\big) \\ &= 1 - \left(\frac{\vartheta - \frac{u}{n}}{\vartheta}\right)^n \\ &= 1 - \Big(1 - \frac{u}{n\vartheta}\Big)^n \\ &\xrightarrow{n\to\infty} 1 - e^{-u/\vartheta}, \end{aligned}$$

per cui si ha che $U_n \xrightarrow{d} \text{Exp}(1/\vartheta)$. Inoltre, questo significa che $n(\vartheta - \hat{\vartheta}) = O_p(1) \implies \hat{\vartheta} - \vartheta = O_p(n^{-1})$, per cui la convergenza di $\hat{\vartheta}$ a ϑ è più veloce di quello che si ha usualmente nei modelli regolari, in cui $\hat{\vartheta} - \vartheta = O_p(n^{-1/2})$.

Osservazione

Bisogna ricavarsi le proprietà a mano, ma a volte si possono avere stimatori anche più efficienti che nel caso regolare. Altre volte, le proprietà sono invece pessime:

Esempio ($\dim(\Theta)$ dipende da n)

Siano $y_i = (y_{1i}, y_{2i})$ coppie di osservazioni indipendenti da $\mathcal{N}(\mu_i, \sigma^2)$, con parametro $\vartheta = (\mu_1, \mu_2, \dots, \mu_n, \sigma^2) \in \Theta = \mathbb{R}^n \times \mathbb{R}^+$. La log-verosimiglianza è

$$\ell(\vartheta) = -n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \Big[\big(y_{1i} - \mu_i\big)^2 + \big(y_{2i} - \mu_i\big)^2 \Big],$$

con funzione punteggio

$$\begin{aligned} \ell_{\mu_i}(\vartheta) &= \frac{1}{\sigma^2} \left[\big(y_{1i} - \mu_i\big) + \big(y_{2i} - \mu_i\big) \right] \\ \ell_{\sigma^2}(\vartheta) &= \dots \end{aligned}$$

Le equazioni di verosimiglianza per μ_i hanno come soluzioni

$$\hat{\mu}_i = \frac{y_{1i} + y_{2i}}{2},$$

per cui

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{2n} \sum_{i=1}^{n} \Big[\big(y_{1i} - \bar{y}_i\big)^2 + \big(y_{2i} - \bar{y}_i\big)^2 \Big] \\ &= \frac{1}{4n} \sum_{i=1}^{n} (y_{1i} - y_{2i})^2. \end{aligned}$$

Si noti che $Y_{1i} - Y_{2i} \sim \mathcal{N}(0, 2\sigma^2)$, per cui

$$\mathbb{E}_\vartheta \left[\hat{\sigma}^2\right] = \frac{1}{4n} 2n\sigma^2 = \frac{\sigma^2}{2}$$

$$\mathbb{V}_\vartheta \left[\hat{\sigma}^2\right] = \frac{(\sigma^2)^2}{2}$$

per cui converge a $\sigma^2/2$ e dunque non è consistente. Accade anche che W_p ha valore atteso di ordine $O(n)$, ma non $o(n)$, e quindi non converge in distribuzione ad un χ^2_1.

Osservazione 1

Questi problemi si dicono di Neyman-Scott, quando ci sono dati stratificati con parametri specifici per strato. Se p è abbastanza grande relativamente a n, questo ci dice che potrebbero esserci problemi nelle approssimazioni asintotiche della verosimiglianza.

Osservazione 2

Altri casi di non regolarità si hanno quando il modelli non è identificabile, quando $\Theta \not\subseteq \mathbb{R}^p$, e quindi la verosimiglianza non è derivabile, quando l'informazione non cresce all'aumentare di n e quando il vero valore del parametro è sulla frontiera dello spazio parametrico.

Ad esempio, nei modelli ad effetti casuali si fa il test per $H_0 : \sigma_\alpha = 0$ per verificare che siano significativamente presenti gli effetti casuali. In molti di questi casi è ancora possibile fare inferenza basata sulla verosimiglianza, solo che i risultati asintotici non sono garantiti.

Riferimenti bibliografici

Azzalini, A. (2001). *Inferenza statistica: una presentazione basata sul concetto di verosimiglianza*. Second. Springer-Verlag.

Basu, D. (1955). «On statistics independent of a complete sufficient statistic». In: *Sankhyā: the indian journal of statistics* 15.

Casella, G. e Berger, R. L. (2001). *Statistical Inference*. 2 edizione. Australia ; Pacific Grove, CA: Duxbury Pr.

Efron, B. e Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. New York, NY: Cambridge University Press.

Evans, M. J. e Rosenthal, J. S. (2006). *Probability and Statistics: The Science of Uncertainty*. New York: W H Freeman & Co.

Liseo, B. (2010). «Introduzione alla statistica bayesiana».

Neyman, J. e Pearson, E. (1933). «On the problem of the most efficient tests of statistical hypotheses». In: *Philosophical transactions of the royal society of london series a* 231, pp. 289–337.

Pace, L. e Salvan, A. (1996). *Teoria della statistica. Metodi, modelli, approssimazioni asintotiche*. Padova: CEDAM.

Pace, L. e Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Singapore: World Scientific Pub.

Pace, L. e Salvan, A. (2001). *Introduzione alla statistica: Inferenza, verosimiglianza, modelli*. Vol. 2. CEDAM.

Welsh, A. H. (1996). *Aspects of Statistical Inference*. John Wiley & Sons.