UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche, aziendali, matematiche e statistiche "Bruno de Finetti"

Statistica (c.p.)

4. Sufficienza, principi di verosimiglianza e condizionamento

Francesco Pauli

DEAMS

Università di Trieste

A.A. 2016/2017

Indice

Sufficienza

Definizione: statistica

Una **statistica** è una funzione dei dati (e non del parametro)

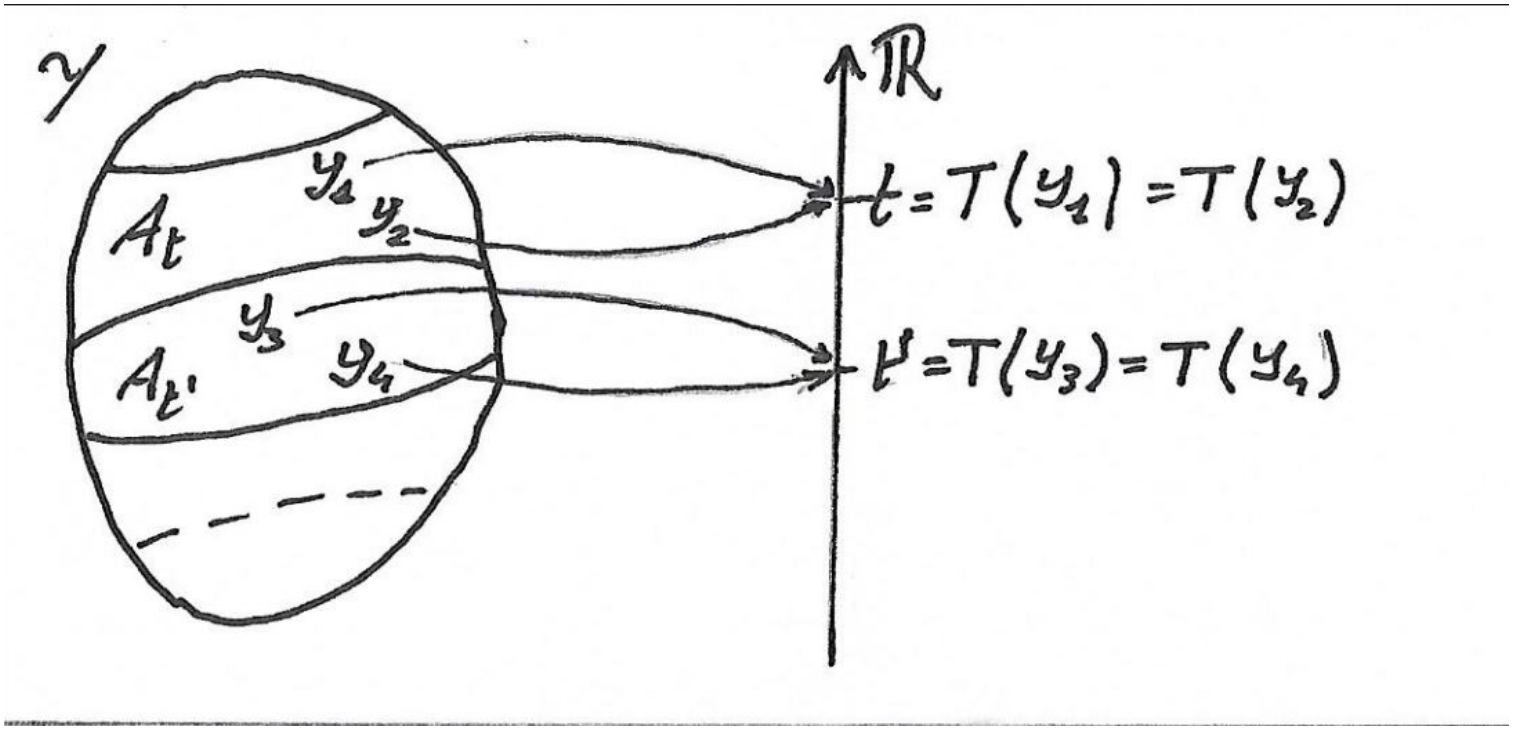
$$T : \mathcal{Y} \to \mathcal{T} \subset \mathbb{R}^k$$

A una statistica T si associa una partizione dello spazio campionario (detta partizione indotta) con generico elemento

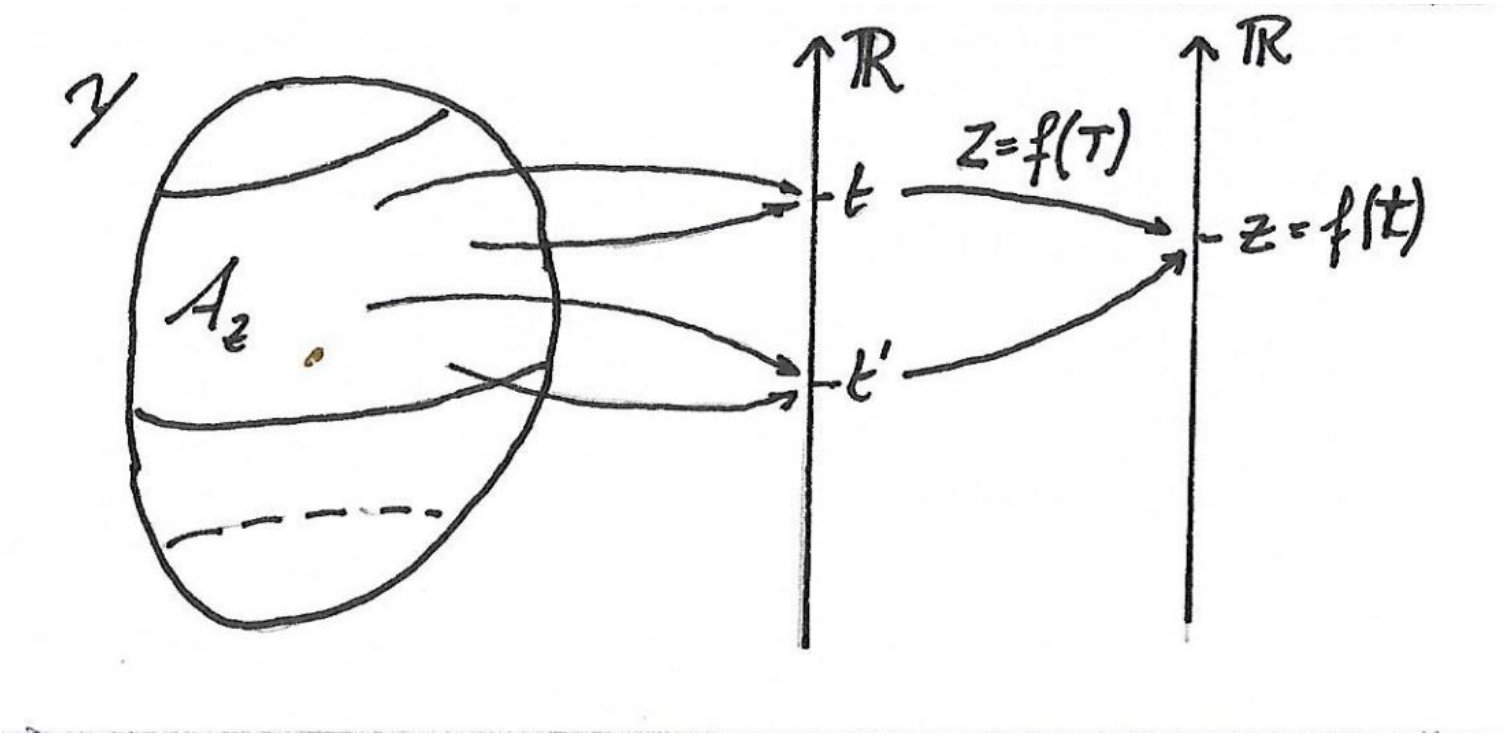
$$\mathcal{Y}_t = \{y \in \mathcal{Y} | \, T(y) = t\}, \quad t \in \mathcal{T}$$

- una qualunque trasformazione biunivoca della statistica T ha la stessa partizione associata.
- una trasformazione non biunivoca avrà una partizione associata meno fine (cioè i cui elementi sono unione di elementi della partizione di T). In questo senso essa riassume maggiormente il campione, ossia conserva, di esso, meno informazione.

Partizione associata a una statistica



Partizione associata a una statistica



Trasformazione non biunivoca

Media campionaria e massimo

Dato un campione $(y_1, y_2) \in \mathcal{Y} = \mathbb{R}^2$ la media campionaria

$$T(y_1, y_2) = (y_1 + y_2)/2$$

è una statistica la cui partizione indotta è

$$A_t = \{(y_1, y_2) | y_1 + y_2 = t\}$$

cioè è l'insieme delle rette parallele di inclinazione -1.
Il trasformato non biunivoco $T' = |T|$ ha partizione associata

$$A_t = \{(y_1, y_2) | y_1 + y_2 = t \text{ o } y_1 + y_2 = -t\}.$$

Anche $T = \max\{y_1, y_2\}$ è una statistica, la cui partizione associata è

$$A_t = \{(y_1, y_2) | y_2 < y_1 = t \text{ o } y_1 < y_2 = t\}.$$

Statistica sufficiente

In generale una statistica riassume le osservazioni; una statistica sufficiente per un parametro θ è

- tale per cui le informazioni che il campione y porta su θ sono contenute anche nella statistica $T(y)$
- in altre parole, nota la statistica $T(y)$ il fatto di conoscere y non aggiunge informazioni sul parametro.

Definizione: statistica sufficiente

Nell'ambito del modello $(\mathcal{Y}, p_\theta, \Theta)$ una statistica $T(\cdot) : \mathcal{Y} \to \mathbb{R}^k$ è sufficiente per θ se e solo se

$$p(y|T = t) \text{ non dipende da } \theta$$

(La definizione fa riferimento a un modello $p_\theta(y)$, non ha senso un'affermazione del tipo 'la media campionaria è statistica sufficiente per la media della popolazione' se non si è precisata la distribuzione della popolazione.)

Statistica sufficiente

Una definizione equivalente centrata sulla verosimiglianza recita che

Definizione: statistica sufficiente

Nell'ambito del modello $(\mathcal{Y}, p_\theta, \Theta)$ una statistica $T(\cdot) : \mathcal{Y} \to \mathbb{R}^k$ è sufficiente per θ se e solo se

$$T(y_1) = T(y_2) \quad \Rightarrow \quad L(\theta; y_1) \propto L(\theta; y_2)$$

Statistiche sufficiente per ogni problema sono: l'intero campione $\mathbf{y}$; la funzione di varosimiglianza $L(\theta)$ (intesa come funzione di θ, quindi calcolabile conoscendo $\mathbf{y}$).

Caratterizzazione della statistica sufficiente

Teorema: caratterizzazione della statistica sufficiente

Sia $(\mathcal{Y}, p_\theta, \Theta)$ un modello statistico e sia $T : \mathcal{Y} \to \mathbb{R}^k$ una statistica.
Sono equivalenti le seguenti affermazioni:

(i) $T(y) = T(z) \Rightarrow L(\theta, y) \propto L(\theta, z)$;

(ii) (teorema di Neyman)
esistono due funzioni h e g tali che
$p_\theta(y) = h(y)g(T(y), \theta)$;

(iii) T è sufficiente
$p_\theta(y|T = t) = p(y|T = t) \ \ \forall\theta$;

Dimostrazione, (i)⇒(ii)

La (i) implica che la verosimiglianza è funzione di y solo attraverso $T(y)$[1], ossia che

$$L(\theta) \propto g(T(y); \theta),$$

essendo al contempo $L(\theta) \propto p_\theta(y)$ si ha che

$$\frac{p_\theta(y)}{g(T(y); \theta)} = h(y)$$

pertanto

$$p_\theta(y) = h(y) g(T(y); \theta).$$

[1] Si può costruire una particolare funzione $g(T(y); \theta)$:

- si consideri la partizione di $\mathcal{Y}$ indotta da T: $A_t = \{y | T(y) = t\}$
- per ciascun valore t si selezioni un elemento $y_t^{(0)} \in A_t$ e si definisca $g(T(y); \theta) = p_\theta(T(y_t^{(0)}))$

si ha allora, se $y \in A_t$, $L(\theta; y) = p_\theta(y) \propto p_\theta(y_t^{(0)}) = g(T(y); \theta)$.

Dimostrazione, (ii)⇒(iii)

Si ha

$$p_\theta(t) = \int_{y|T(y)=t} p_\theta(y)dy = \int_{y|T(y)=t} h(y)g(T(y),\theta)dy =$$
$$= g(T(y),\theta)\int_{y|T(y)=t} h(y)dy = g(T(y),\theta)h^*(y)$$

e quindi

$$p(y|T=t) = \frac{g(T(y),\theta)h(y)}{g(T(y),\theta)\int_{y|T(y)=t} h(y)dy} = \frac{h(y)}{\int_{y|T(y)=t} h(y)dy}$$

che non dipende da θ. (Nota: la dimostrazione non è assolutamente rigorosa, renderla tale richiede nozioni di teoria della misura per definire la distribuzione condizionata.).

Dimostrazione, (iii)→(i)

Si ha

$$\begin{aligned} L(\theta, y) &\propto p_\theta(y) \\ &= p_\theta(t)p(y|T=t) \text{ teorema di Bayes} \\ &\propto p_\theta(t)p(z|T=t) \text{ per la (iii)} \\ &= p_\theta(z) \propto L(\theta, z) \end{aligned}$$

Statistica sufficiente e partizione associata

Consideriamo la partizione dello spazio campionario $\mathcal{Y}$ associata alla statistica sufficiente: $\{\mathcal{Y}_t\}$: l'informazione rilevante per il parametro non è l'esatto campione osservato ma solo in quale costituente della partizione esso si trovi.

Questo traspare dalla caratterizzazione **(iii)**, in base alla quale possiamo pensare al meccanismo generatore dei dati come un meccanismo a due stadi, al primo livello si determina il valore di $T(y)$, al secondo, condizionatamente al valore di $T(y)$, si determina y: solo la prima fase è regolata da una legge di probabilità che dipende da θ.

È trasparente a questo punto l'affermazione per cui non tutta l'informazione portata dal campione è rilevante per il parametro.

Esempio binomiale

Consideriamo un campione $Y_1, Y_2, Y_3 \sim IID(\text{Bernoulli}(\theta))$, per il quale la statistica sufficiente è $S = \sum_{i=1}^{3} Y_i$.

$\mathcal{Y}$	$p_\theta(y)$	$\mathcal{Y}_s$	$p_\theta(s)$	$p(y\|S)$
$(0,0,0)$	$(1-\theta)^3$	$\mathcal{Y}_0$	$(1-\theta)^3$	1
$(1,0,0)$	$\theta(1-\theta)^2$			$1/3$
$(0,1,0)$	$\theta(1-\theta)^2$	$\mathcal{Y}_1$	$3\theta(1-\theta)^2$	$1/3$
$(0,0,1)$	$\theta(1-\theta)^2$			$1/3$
$(1,1,0)$	$\theta^2(1-\theta)$			$1/3$
$(1,0,1)$	$\theta^2(1-\theta)$	$\mathcal{Y}_2$	$3\theta^2(1-\theta)$	$1/3$
$(0,1,1)$	$\theta^2(1-\theta)$			$1/3$
$(1,1,1)$	θ^3	$\mathcal{Y}_3$	θ^3	1

Esempio binomiale

Consideriamo un campione $Y_1, Y_2, Y_3 \sim IID(\text{Bernoulli}(\theta))$, per il quale la statistica sufficiente è $S = \sum_{i=1}^{3} Y_i$.

$\mathcal{Y}$	$p_\theta(y)$	$\mathcal{Y}_s$	$p_\theta(s)$	$p(y\|S)$
$(0,0,0)$	$(1-\theta)^3$	$\mathcal{Y}_0$	$(1-\theta)^3$	1
$(1,0,0)$	$\theta(1-\theta)^2$			1/3
$(0,1,0)$	$\theta(1-\theta)^2$	$\mathcal{Y}_1$	$3\theta(1-\theta)^2$	1/3
$(0,0,1)$	$\theta(1-\theta)^2$			1/3
$(1,1,0)$	$\theta^2(1-\theta)$			1/3
$(1,0,1)$	$\theta^2(1-\theta)$	$\mathcal{Y}_2$	$3\theta^2(1-\theta)$	1/3
$(0,1,1)$	$\theta^2(1-\theta)$			1/3
$(1,1,1)$	θ^3	$\mathcal{Y}_3$	θ^3	1

Meccanismo generatore per (Y_1, Y_2, Y_3)

(1) Genero S secondo

$$P(S = s) = \binom{3}{s} \theta^s (1-\theta)^{3-s}$$

(2) Dato S scelgo una terna $y = (Y_1, Y_2, Y_3)$ usando le condizionate.
Ad es. se $S = 1$ genero secondo

$$p(y|S=1) = \begin{cases} 1/3 & \text{se } y = (1,0,0) \\ 1/3 & \text{se } y = (0,1,0) \\ 1/3 & \text{se } y = (0,0,1) \end{cases}$$

Il passo (2) non dipende da θ.

Statistica sufficiente minimale

Definizione: statistica sufficiente minimale

Nell'ambito del modello $(\mathcal{Y}, p_\theta, \Theta)$ una statistica $T(\cdot) : \mathcal{Y} \to \mathbb{R}^k$ è **sufficiente minimale** per θ se e solo se

$$T(y_1) = T(y_2) \Leftrightarrow L(\theta; y_1) \propto L(\theta; y_2)$$

per ogni $\theta \in \Theta$.

Caratterizzazioni della statistica sufficiente minimale

Equivalenti alla definizione sono le affermazioni

- Una statistica è sufficiente minimale se e solo se ogni altra statistica sufficiente è funzione di essa.
- Una statistica sufficiente minimale è determinata dalla partizione di verosimiglianza, dove quest'ultima è la partizione dello spazio campionario tale per cui due elementi stanno nello stesso costituente se e solo se portano a verosimiglianze proporzionali (cioè uguali).
- Una statistica T è sufficiente minimale se e solo se vale la

$$\frac{L(\theta; y)}{L(\theta; z)} = c(y, z) \Leftrightarrow T(y) = T(z)$$

quest'ultima caratterizzazione è immediatamente utilizzabile per mostrare che una statistica è sufficiente minimale.

Proprietà della statistica sufficiente minimale

Risulta, immediatamente, che una statistica sufficiente minimale è la funzione di verosimiglianza (nel senso precisato sopra in relazione ad essa come statistica sufficiente).

Si ha inoltre che se una statistica è sufficiente (minimale) ogni sua trasformazione biunivoca è sufficiente (minimale).

Statistiche sufficienti per il modello normale

Consideriamo un campione $(Y_1, \ldots, Y_n)$ IID da una $\mathcal{N}(\mu, \sigma^2)$, il parametro è $\theta = \mu$ (σ^2 è noto), la funzione di densità del campione è quindi

$$p_\theta(\mathbf{y}) = \underbrace{(2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right\}}_{h(y)} \underbrace{\exp\left\{-\frac{n}{2\sigma^2}(\mu^2 - 2\mu\bar{y})\right\}}_{g(T(y);\theta)},$$

da cui emerge che $\bar{y}$ è sufficiente in virtù del criterio (ii).

D'altra parte notiamo che anche $(\bar{y}, \sum_{i=1}^{n} y^2)$ o, se è per questo, $(\bar{y}, \sum_{i=1}^{n}(y_i - \bar{y})^2)$ sono statistiche sufficienti.

Statistiche sufficienti per il modello normale

Con riferimento alla caratterizzazione (iii) la densità condizionata di $\mathbf{y}|T = t$ è 0 se $\sum_i y_i \neq nt$ ed è altrimenti pari a

$$
\begin{aligned}
p(y|T=t) &= \frac{p_\theta(y)}{p_\theta(t)} \\
&= \frac{(2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(\sum_i y_i^2 + n\mu^2 - 2\mu n\bar{y})\right\}}{(2\pi)^{-1/2}(\sigma^2/n)^{-1/2}\exp\left\{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2\right\}} \\
&= (2\pi)^{-(n-1)/2}(\sigma^2)^{-(n-1)/2}n^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(\sum_i y_i^2 - n\bar{y}^2)\right\} \\
&= (2\pi)^{-(n-1)/2}(\sigma^2)^{-(n-1)/2}n^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i-\bar{y})^2\right\}
\end{aligned}
$$

Statistica sufficiente minimale per il modello normale

Per individuare una statistica sufficiente minimale consideriamo la verosimiglianza

$$L(\theta, \mathbf{y}) = \exp\left\{-\frac{n}{2\sigma^2}(\mu^2 - 2\mu\bar{y})\right\}$$

calcolata in due punti arbitrari $\mathbf{y}_1$ e $\mathbf{y}_2$, e calcoliamo il rapporto

$$\begin{aligned}\frac{L(\theta, \mathbf{y}_1)}{L(\theta, \mathbf{y}_2)} &= \exp\left\{-\frac{n}{2\sigma^2}(\mu^2 - 2\mu\bar{y}_1 - \mu^2 + 2\mu\bar{y}_2)\right\} \\ &= \exp\left\{-\frac{\mu n}{\sigma^2}(\bar{y}_2 - \bar{y}_1)\right\}\end{aligned}$$

questo non dipende dal parametro, μ, se e solo se $\bar{y}_2 = \bar{y}_1$, quindi $\bar{y}$ è statistica sufficiente minimale, come ogni sua trasformazione biunivoca.

Statistica sufficiente minimale per il modello normale

Generalizziamo l'esempio abbandonando l'ipotesi σ^2 noto

$$L(\theta, \mathbf{y}) = (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 + n\mu^2 - 2\mu n\bar{y}\right)\right\}$$

andiamo alla ricerca di una statistica sufficiente minimale, si ha

$$\begin{aligned}\frac{L(\theta, \mathbf{y}_1)}{L(\theta, \mathbf{y}_2)} &= \frac{(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_{1,i}^2 + n\mu^2 - 2\mu n\bar{y}_1\right)\right\}}{(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_{2,i}^2 + n\mu^2 - 2\mu n\bar{y}_2\right)\right\}} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_{1,i}^2 - \sum_{i=1}^{n} y_{2,i}^2 + 2\mu n(\bar{y}_2 - \bar{y}_1)\right)\right\}\end{aligned}$$

che non dipende dal parametro se e solo se $\sum_{i=1}^{n} y_{1,i}^2 = \sum_{i=1}^{n} y_{2,i}^2$ e $\bar{y}_2 - \bar{y}_1$, quindi $(\bar{y}, \sum_i y_i^2)$ è statistica sufficiente minimale.
Notiamo che questo implica che anche $(\bar{y}, S^2)$ con $S^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$, che è trasformazione biunivoca di $(\bar{y}, \sum_i y_i^2)$, è sufficiente minimale.

modello per una proporzione

Nel modello $(y_1, \ldots, y_n)$ IID da una Bernoulli di parametro θ ci si chiede se le seguenti statistiche sono sufficienti/ sufficienti minimali

$$
\begin{aligned}
(i) \quad & T_1(\mathbf{y}) = (y_1, \ldots, y_n) \\
(ii) \quad & T_2(\mathbf{y}) = (y_{(1)}, \ldots, y_{(n)}) \\
(iii) \quad & T_3(\mathbf{y}) = (y_1, y_2) \\
(iv) \quad & T_4(\mathbf{y}) = (\sum_{i=1}^{n} y_i)
\end{aligned}
$$

Statistiche sufficienti e campioni casuali semplici

Se si è di fronte a un c.c.s. $y_i \sim f(y;\theta)$, $i = 1, \ldots, n$, essendo la distribuzione congiunta

$$p_\theta(y) = \prod_{i=1}^{n} f(y_i;\theta)$$

e quindi indipendente dall'ordine delle osservazioni, la statistica d'ordine $(y_{(1)}, \ldots, y_{(n)})$ è sufficiente.

Esempio: distribuzione di Cauchy

Non è scontato che esistano statistiche sufficienti al di là di quelle banali (intero campione, statistica d'ordine), ad esempio se si ha un campione iid da una Cauchy di parametro θ, ossia

$$f(y_i;\theta) = \frac{1}{\pi(1+(y_i-\theta)^2)},$$

dove $y_i \in \mathbb{R}$ e $\theta \in \mathbb{R}$, si ha

$$p_\theta(y) = \prod_{i=1}^{n} \frac{1}{\pi(1+(y_i-\theta)^2)} = \pi^{-n} \prod_{i=1}^{n} (1+(y_i-\theta)^2)^{-1}$$

che non consente fattorizzazioni.

Esempio: distribuzione uniforme

Si considera un campione iid da $y_i \sim \text{Unif}(\theta, \theta + 1)$, allora

$$p_\theta(y) = \prod_{i=1}^{n} I_{[\theta,\theta+1]}(y_i) = I_{]-\infty,y_{(1)}]}(\theta) I_{[y_{(n)}-1,+\infty[}(\theta)$$

La statistica $(y_{(1)}, y_{(n)})$ è sufficiente minimale.

È tipico che la statistica sufficiente minimale abbia la stessa dimensione del parametro, ciò però non è scontato.

Esempio: distribuzione normale con CV noto

Si considera un campione iid da $y_i \sim \mathcal{N}\left(\theta, \theta^2\right)$, allora

$$\begin{aligned} p_\theta(y) &= (2\pi)^{-n/2}(\theta^2)^{-n/2} \exp\left\{-\frac{1}{2\theta^2}\left(\sum_{i=1}^{n} y_i^2 - 2\theta n\bar{y} + n\theta^2\right)\right\} \\ &= (2\pi)^{-n/2}(\theta^2)^{-n/2} \exp\left\{-\frac{\sum_{i=1}^{n} y_i^2}{2\theta^2} - \frac{n\bar{y}}{\theta}\right\} \exp\left\{-\frac{n}{2}\right\} \end{aligned}$$

La statistica $(\bar{y}, S^2)$ è sufficiente minimale.

SMV e sufficienza

È semplice mostrare che lo SMV $\hat{\theta}$ è funzione della statistica sufficiente minimale.
Infatti se T è una statistica sufficiente, si ha

$$L(\theta) \propto h(y)g(T(y), \theta) \propto g(T(y), \theta)$$

quindi

$$\hat{\theta} = \underset{\theta \in \Theta}{\text{argmax}} g(T(y); \theta)$$

cioè θ può essere espresso come funzione di qualunque statistica sufficiente, e quindi in particolare anche di qualunque statistica sufficiente minimale.

Indice

Introduzione

I principi di verosimiglianza e quello di condizionamento fanno parte dei fondamenti della statistica, rispondono o vorrebbero rispondere alla domanda su come si possa o si debba inferire sulla distribuzione che genera i dati (sul parametro) a partire dalle osservazioni (dal campione).

In particolare, il principio di verosimiglianza e di condizionamento individuano (se li si accetta) quali siano le informazioni rilevanti ai fini dell'inferenza e quali invece debbano essere ignorate.

Essi vanno visti in parallelo (in parte in contrasto, come verrà chiarito) con il principio del campionamento ripetuto che qui ricordiamo e che è il fondamento del cosiddetto approccio frequentista.

Principio del campionamento ripetuto

Principio del campionamento ripetuto

Il **principio del campionamento ripetuto** prescrive di valutare le procedure di inferenza sulla base del loro risultato atteso in un gran numero di ripetizioni dell'esperimento.

Il PCR prevede, pragmaticamente, di scegliere una procedura se questa garantisce di ottenere buoni risultati con elevata frequenza, ovvero di ottenere difficilmente risultati "sballati" in ipotetiche ripetizioni dell'esperimento.

Il PCR non è prescrittivo, non dice come si ottengono le procedure ed è possibile, anche in esempi semplici, che procedure che portano a risultati diversi siano parimenti giustificate secondo il PCR.

Indice

Enunciato del principio debole di verosimiglianza

L'enunciazione del Principio di Verosimiglianza non è del tutto scontata, si trovano in letteratura diversi enunciati con sottili differenze che però, essendo sottili anche le interpretazioni, portano a affermare cose diverse, soprattutto in relazione alle conseguenze del principio stesso.

Principio debole di verosimiglianza (PDV)

Con riferimento a un modello $(\mathcal{Y}, p_\theta, \Theta)$ se

$$y, z \in \mathcal{Y} \text{ tali che } L(\theta; y) \propto L(\theta; z)$$

allora y e z devono portare alle medesime conclusioni inferenziali.

Enunciato del principio debole di verosimiglianza

L'enunciazione del Principio di Verosimiglianza non è del tutto scontata, si trovano in letteratura diversi enunciati con sottili differenze che però, essendo sottili anche le interpretazioni, portano a affermare cose diverse, soprattutto in relazione alle conseguenze del principio stesso.

Principio di sufficienza (PS)

Con riferimento a un modello $(\mathcal{Y}, p_\theta, \Theta)$ per il quale T è una statistica sufficiente, se

$$y, z \in \mathcal{Y} \text{ tali che } T(y) = T(z)$$

allora y e z devono portare alle medesime conclusioni inferenziali.

Enunciato del principio debole di verosimiglianza

L'enunciazione del Principio di Verosimiglianza non è del tutto scontata, si trovano in letteratura diversi enunciati con sottili differenze che però, essendo sottili anche le interpretazioni, portano a affermare cose diverse, soprattutto in relazione alle conseguenze del principio stesso.

Discorsivamente, e imprecisamente, il PV dice che *la verosimiglianza contiene tutte le informazioni che il campione fornisce sul parametro.*

Esempio: implicazioni del PDV

Navighiamo in un laghetto sinché osserviamo 20 pesci dei quali interessa il colore (rosso o non rosso), abbiamo quindi il modello

$$\left(\mathcal{Y} = \{0,1\}^{20}, f_\theta = \theta^{\sum_i y_i}(1-\theta)^{3-\sum_i y_i}, \Theta = [0,1]\right)$$

consideriamo i due campioni

$$\begin{aligned} y &= (1,1,0,1,0,1,0,0,1,0,1,1,0,0,0,0,0,0,0,0) \quad \textstyle\sum_i y_i = 7 \\ z &= (0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1) \quad \textstyle\sum_i z_i = 7 \end{aligned}$$

in base al PDV devono portare alle medesime conclusioni in quanto

$$L(\theta; y) = \theta^7(1-\theta)^{13} = L(\theta; z)$$

In sostanza l'implicazione è che possiamo usare come modello

$$\left(\mathcal{S} = \{0,\ldots,20\}, f_\theta = \binom{n}{s}\theta^s(1-\theta)^{20-s}, \Theta = [0,1]\right)$$

dove S è il numero di successi (pesci rossi).

Enunciato del principio forte di verosimiglianza

Il rafforzamento del PDV consiste nella sua applicazione a esperimenti diversi.

Principio forte di verosimiglianza (PFV)

Con riferimento a due modelli

$$(\mathcal{Y}_f, f_\theta, \Theta) \text{ e } (\mathcal{Y}_g, g_\theta, \Theta)$$

che condividono il parametro θ, se

$$y \in \mathcal{Y}_f, z \in \mathcal{Y}_g \text{ tali che } L_f(\theta; y) \propto L_g(\theta; z)$$

allora y e z devono portare alle medesime conclusioni inferenziali.

Significato del PFV

Con questo, si afferma non solo che conta unicamente la verosimiglianza, ma che conta solo la verosimiglianza relativa al campione.

Per dirla con Jeffreys, i risultati sperimentali che non si sono verificati non possono avere alcuna rilevanza.

Sebbene anche questo possa, di primo acchito, apparire ragionevole, esso è ben lungi dall'essere accettato e molte procedure statistiche non vi si attengono.

Tra le implicazioni del PV che destano maggiori critiche vi è il fatto che, come mostra anche il prossimo esempio, se aderiamo al PV dobbiamo ignorare lo schema di campionamento.

Campo di applicazione del PFV

Il PV si applica a un modello completamente specificato (p_θ, Θ), cioè possiamo confrontare e discernere solo tra possibilità all'interno di (p_θ, Θ).

Ad esempio, in un modello lineare con ipotesi gaussiana il PV conduce alla stima dei parametri, le procedure di analisi dei residui, ad esempio il confronto dei quantili empirici con i quantili teorici della normale sono fuori dal suo campo di applicazione perché attengono alla verifica delle assunzioni all'interno delle quali esso opera.

In pratica, quindi, si ammette una fase dell'inferenza al di fuori della logica del PFV per verificare il modello.

Esempio: implicazioni del PFV

Consideriamo, accanto all'esperimento

$$\left(\mathcal{S} = \{0, \ldots, 20\}, f_\theta = \binom{n}{s}\theta^s(1-\theta)^{20-s}, \Theta = [0,1]\right)$$

un secondo esperimento in cui continuiamo a osservare sino a vedere 7 pesci rossi, sia Z il numero di pesci non rossi visti sinché osservo 7 rossi

$$\left(\mathcal{Z} = \mathbb{N}, g_\theta = \binom{z+7-1}{z-1}\theta^7(1-\theta)^z, \Theta = [0,1]\right)$$

I due esperimento condividono il parametro θ, è diverso lo spazio campionario: nel primo esperimento esso è $\{0,1\}^{20}$, nel secondo è

$$\bigcup_{n \leq 7} \{v \in \{0,1\}^n \text{ t.c. } \sum v_i = 7 \text{ e } v_n = 1\}$$

Esempio: implicazioni del PFV

Consideriamo, accanto all'esperimento

$$\left(\mathcal{S} = \{0, \ldots, 20\}, f_\theta = \binom{n}{s}\theta^s(1-\theta)^{20-s}, \Theta = [0,1]\right)$$

un secondo esperimento in cui continuiamo a osservare sino a vedere 7 pesci rossi, sia Z il numero di pesci non rossi visti sinché osservo 7 rossi

$$\left(\mathcal{Z} = \mathbb{N}, g_\theta = \binom{z+7-1}{z-1}\theta^7(1-\theta)^z, \Theta = [0,1]\right)$$

Osserviamo $z = 7$ nel primo esperimento e $z = 13$ nel secondo, si ha allora

$$L_f(\theta; y) = \theta^7(1-\theta)^{13} = L_g(\theta; z),$$

in base al PFV dovrei giungere alle medesime conclusioni del primo esperimento.

In effetti questo ha un *appeal* intuitivo, poiché in entrambi i casi ho osservato 20 pesci di cui 7 rossi.

Esempio: implicazioni del PFV

Consideriamo il problema di verifica d'ipotesi

$$H_0 : \theta \geq 0.5 \quad H_1 : \theta < 0.5$$

nei due esperimenti si hanno le regioni di rifiuto e i valori p

$$\{Y \leq k_\alpha\} \qquad \{Z \geq k_\alpha\}$$

$$P(Y \leq 7; \theta = 0.5) = 0.131588 \qquad P(Z \geq 13; \theta = 0.5) = 0.08753$$

quindi il valore p non rispetta il PFV

Il punto è che nelle procedure di verifica d'ipotesi non conta solo ciò che si è osservato, ma anche il resto dello spazio campionario (il valore p è la probabilità di qualcosa che non è stato osservato).

Le procedure frequentiste non rispettano il PFV

Principio di verosimiglianza e regola d'arresto

Una delle più rilevanti (e discusse) conseguenze del principio di verosimiglianza è che la regola d'arresto di un esperimento, purché sia non informativa per θ ossia non dipenda dal parametro θ, non è rilevante ai fini delle conclusioni da trarre.

Che la regola d'arresto sia rilevante nell'impostazione frequentista è naturale: dalla regola d'arresto dipende lo spazio campionario e quindi i possibili risultati nelle future ripetizioni dell'esperimento.

Va tenuto presente che il PFV prescrive di non considerare la regola d'arresto nell'inferenza ma non autorizza a usare procedure frequentiste (tipicamente test) sul campione così costruito.

PFV e procedure bayesiane

Le procedure bayesiane sono tutte basate sulla distribuzione a posteriori

$$\pi(\theta|\mathbf{y}) \propto \pi(\theta)l(\theta;\mathbf{y})$$

Questo è sufficiente per alcuni per affermare che il PV è compatibile con l'approccio Bayesiano e anzi che il PV è una conseguenza del paradigma bayesiano.

La differenza che invece altri mettono in evidenza è la distribuzione a priori. Questa dipende da altro che non è la verosimiglianza (e potenzialmente dal disegno sperimentale, ad esempio se si utilizzano le a priori non informative di Jeffrey si perviene a scelte diverse a seconda che la f.d.p. del campione sia binomiale o binomiale negativa).

Un importante elemento comune dei due approcci è il fatto di ignorare la regola d'arresto.

Rigetto del PFV

Se si ritiene che il PFV implichi la non validità delle procedure frequentiste, volendo usare queste ultime si rigetta il PFV.

Si può argomentare (Pawitan) che il PFV si ferma al fatto che l'informazione dei due campioni è la stessa se tutta l'informazione è portata dalla verosimiglianza. Il rifiuto deriverebbe dal fatto che l'informazione che si usa per trarre le conclusioni non si limita ai dati-verosimiglianza.

Va anche ricordato che i metodi frequentisti portano sotto condizioni abbastanza generali a risultati approssimativamente uguali a quelli Bayesiani, per cui la validità dei secondi porterebbe comunque ad affermare la validità, almeno approssimativamente, dei primi.

Indice

Statistica ancillare

L'altro elemento del discorso richiede preliminarmente la definizione di statistica ancillare.

Definizione: statistica ancillare

Considerato un modello $(\mathcal{Y}, p_\theta, \Theta)$, una statistica A è **ancillare** se

(i) $S = (T, A)$ è una statistica sufficiente minimale;

(ii) la distribuzione di A non dipende da θ.

Principio di condizionamento

Sul come trattare situazioni di questo tipo è stato proposto il

Definizione: principio di condizionamento (PC)

Con riferimento a un modello $(\mathcal{Y}, p_\theta, \Theta)$ per il quale $T = (S, A)$ è una statistica sufficiente e A è ancillare, l'inferenza va fatta condizionatamente al valore assunto dalla statistica ancillare A, cioè si opera con

$$L_a(\theta) = f(a) f_{T|A=a}(t|a, \theta)$$

Il condizionamento appare una scelta obbligata in alcune situazioni, come quelle dei prossimi esempi.

Statistiche ancillari e PC, esperimento mistura

Un manufatto ritrovato durante uno scavo archeologico viene spedito a un laboratorio per la datazione. È noto che il laboratorio dispone di due macchinari per la datazione $M = 1, 2$, l'errore di misura è distribuito normalmente con una varianza di 49 anni per il primo strumento e di 196 per il secondo. È noto che la probabilità di usare ciascuno strumento per una particolare misura è 0.5.

Detta θ l'età dell'oggetto, dunque, la misura che ottengo è

$$Y = \theta + Z$$

dove Z è la v.a. mistura delle due normali, cioè con densità

$$f(z) = \frac{1}{2}\phi(z; 0, 7) + \frac{1}{2}\phi(z; 0, 14)$$

Statistiche ancillari e PC, esperimento mistura

Se con M indichiamo il macchinario usato si ha

$$E(Y) = E(Y|M=1) = E(Y|M=2) = \theta$$

$$V(Y) = E(V(Y|M)) + V(E(Y|M)) = \frac{1}{2}49 + \frac{1}{2}196 = 105$$

Sicché se ottengo una misura pari a 1000, l'i.c. al 95% per θ è

$$1000 \pm 1 - 96\sqrt{105} \quad \rightarrow \quad [990, 1010]$$

Supponiamo però che il laboratorio ci dica anche che è stato usato il macchinario 1, il campione è allora $(M = 1, Y = 1000)$, l'i.c. sensato per θ si basa sulla distirbuzione di $Y|M = 1$ e quindi

$$1000 \pm 1 - 96\sqrt{49} \quad \rightarrow \quad [993, 1007]$$

dove M è ancillare.

Statistiche ancillari e PC, numerosità campionaria

Si consideri il seguente esperimento in due fasi

1. $A = 10^X$, con X distribuito secondo una Bernoulli(0.5);
2. $Y_1, \ldots, Y_a \sim \text{IID}(\mathcal{N}(\theta, 1))$.

allora $(\bar{Y}, A)$ è una statistica sufficiente minimale per θ e A è ancillare.

Lo SMV è $\bar{Y}$, e per esso si ha

$$V(\bar{Y}|A = a) = 1/a$$

mentre

$$V(\bar{Y}) = E(V(\bar{Y}|A)) + V(E(\bar{Y}|A)) = \frac{1}{2}1 + \frac{1}{2}\frac{1}{10} = \frac{11}{20}$$

(Si noti che $E(\bar{Y}|A) = \theta$ non dipende da A.)

Significato del PC

Negli esempi fatti notiamo che la statistica ancillare ha effetto sulla valutazione della precisione della stima, che è basata sul principio del campionamento ripetuto.

Altri esempi comuni di statistiche ancillari: numerosità campionaria quando è fissa, le covariate di un modello lineare.

In questi esempi "giocattolo" il ricorso al PC è naturale, in generale non è scontato, bisogna ad esempio considerare il fatto che di statistiche ancillari ve ne può essere più d'una nel qual caso non è chiaro a quale ci si debba condizionare (v. anche `https://normaldeviate.wordpress.com/2012/07/28/statistical-principles/`).

Statistica ancillare per la distribuzione uniforme

Siano $(Y_1, \ldots, Y_n) \sim IID(U(\theta, \theta+1))$, con $\theta \in \Theta = \mathbb{R}$.

$$p_\theta(y_i) = \begin{cases} 1 & \text{se } \theta \leq y_i \leq \theta + 1 \\ 0 & \text{altrimenti} \end{cases} \quad \Rightarrow \quad p_\theta(\mathbf{y}) = \begin{cases} 1 & \text{se } y_{(n)} - 1 \leq \theta \leq y_{(1)} \\ 0 & \text{altrimenti} \end{cases}$$

ne segue che $(Y_{(1)}, Y_{(n)})$ è statistica sufficiente minimale.

Ogni trasformazione biunivoca di $(Y_{(1)}, Y_{(n)})$ è sufficiente minimale, in particolare lo è

$$(R = Y_{(n)} - Y_{(1)}, M = (Y_{(1)} + Y_{(n)})/2)$$

mostreremo ora che R è ancillare, per ricavare la distribuzione di R procediamo in tre passi

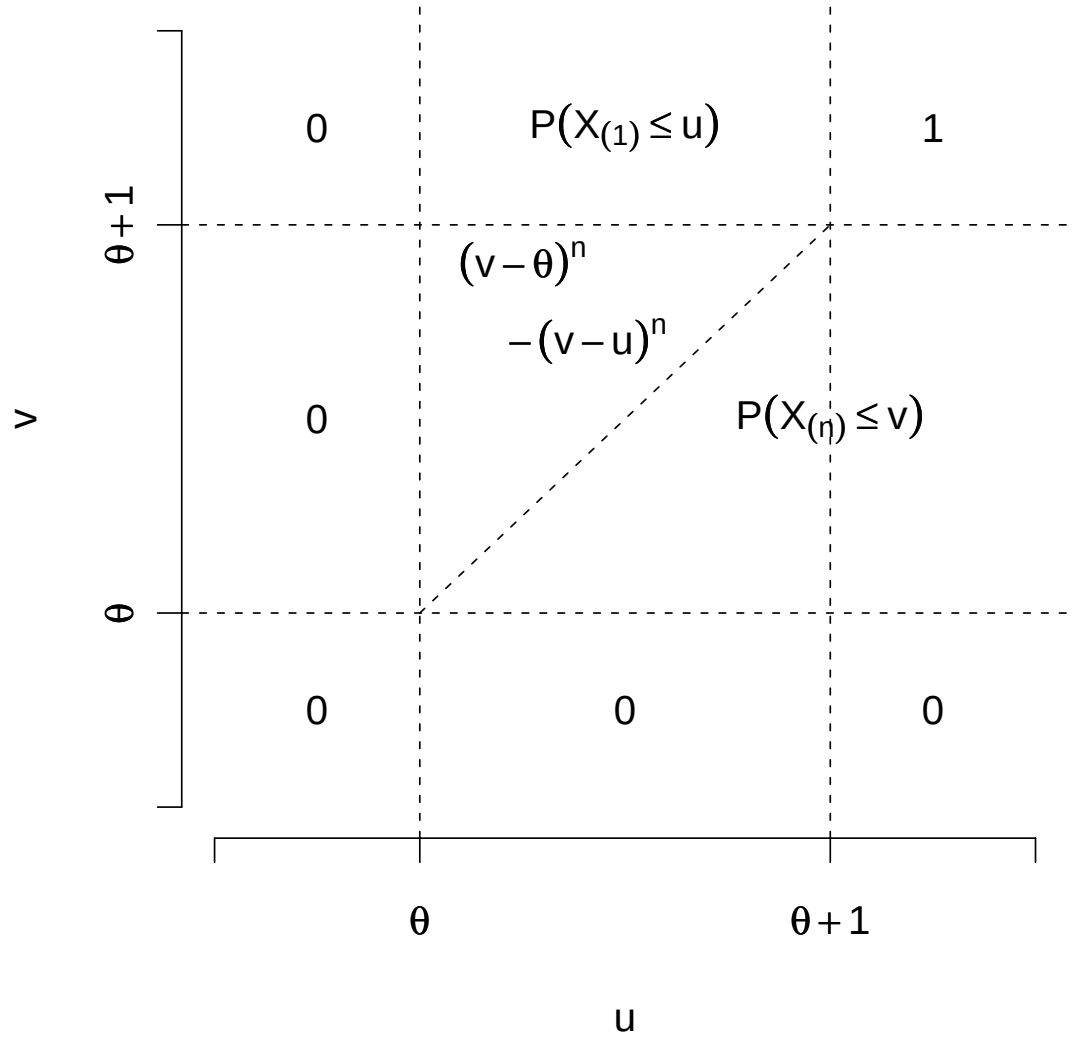
1. Ricaviamo la distribuzione di $(Y_{(1)}, Y_{(n)})$.
2. Otteniamo la distribuzione di (R, M) per trasformazione.
3. Otteniamo la distribuzione di R per marginalizzazione.

Distribuzione di $(Y_{(1)}, Y_{(n)})$

La FdR di $(Y_{(1)}, Y_{(n)})$ è

$$
\begin{aligned}
F(u,v) &= P(Y_{(1)} \leq u \cap Y_{(n)} \leq v) \\
&= P(Y_{(n)} \leq v) - P(Y_{(1)} > u \cap Y_{(n)} \leq v) \\
&= P\left(\bigcap_{i=1}^{n} Y_i \leq v\right) - P\left(\bigcap_{i=1}^{n} Y_i > u \cap \bigcap_{i=1}^{n} Y_i \leq v\right) \\
&= \prod_{i=1}^{n} P\left(Y_i \leq v\right) - \prod_{i=1}^{n} P\left(u < Y_i \leq v\right) \\
&= \begin{cases}
1 & \text{se } u, v > \theta + 1 \\
1 - P(Y_{(1)} > u) = 1 - (1 - u + \theta)^n & \text{se } u > \theta, v > \theta + 1 \\
P(Y_{(n)} \leq v) = (v - \theta)^n & \text{se } u > \theta, \theta \leq v \leq u \\
(v - \theta)^n - (v - u)^n & \text{se } \theta \leq u \leq v \leq \theta + 1 \\
0 & \text{altrove } (u < \theta \text{ o } v < \theta)
\end{cases}
\end{aligned}
$$

Distribuzione di $(Y_{(1)}, Y_{(n)})$



Distribuzione di (R, M)

La funzione di densità di $(Y_{(1)}, Y_{(n)})$ è

$$f_\theta(u, v) = \frac{\partial}{\partial v}\frac{\partial}{\partial u}F(u, v) = n(n-1)(v-u)^{n-2}$$

se $\theta \leq u, v \leq \theta + 1$ e 0 altrimenti, si ha poi

$$\begin{cases} u = m - r/2 \\ v = m + r/2 \end{cases}$$

con Jacobiano di trasformazione

$$J = \begin{bmatrix} 1 & -1/2 \\ 1 & +1/2 \end{bmatrix} \Rightarrow |J| = 1$$

si ha quindi la funzione di densità di (r, m)

$$h_\theta(r, m) = n(n-1)r^{n-2}$$

se $0 < r < 1$ e $\theta + r/2 < m < \theta + 1 - r/2$ e 0 altrimenti.

Distribuzione di R

La densità di R è dunque, per $0 < r < 1$,

$$\begin{aligned} h_\theta(r) &= \int h_\theta(r, m) dm \\ &= \int_{\theta + r/2}^{\theta + 1 - r/2} n(n-1) r^{n-2} dm \\ &= n(n-1) r^{n-2} (1 - r) \end{aligned}$$

che è una Beta di parametri $n - 1$ e 2, non dipende quindi da θ, c.v.d.

Che R sia rilevante nell'inferenza è ovvio, basti notare che se $M = 1/2$ e $R = 1$ allora $\theta = 0$ è l'unico valore di θ che assegna densità non nulla al campione, così non è se, ad esempio, $M = 1/2$ e $R = 1/2$

Statistica ancillare, uniforme discreta

Per un esempio più semplice matematicamente ma sulla falsariga del precedente, si consideri un campione Y_1, Y_2 iid dalla distribuzione

$$P_\theta(Y = \theta) = P_\theta(Y = \theta + 1) = P_\theta(Y = \theta + 2) = 1/3$$

con θ intero. Si procede come nell'esempio precedente, la statistica $(R = Y_{(2)} - Y_{(1)}, M = (Y_{(2)} + Y_{(1)})/2)$ è sufficiente minimale ed è facile verificare che R è ancillare. Supponendo di aver osservato $M = m$ dove m è un intero, si confronti l'inferenza a seconda del valore di R.

Indice

Legame tra PC e PFV

PFV e PC sono formalmente collegati, in particolare il PC e il PS implicano il PFV e viceversa. Questo risulato è stato mostrato da Birnbaum.

Preliminarmente introduciamo una notazione formale in cui con

$$E = (Y, f_\theta)$$

si indica un esperimento in cui Y è il campione e f_θ è il modello parametrico (è sottinteso nella descrizione lo spazio parametrico Θ).

Indichiamo poi con

$$\text{Ev}(E, y)$$

l'informazione sperimentale sul parametro θ (possiamo anche immaginare che Ev siano la conclusioni inferenziali).

Con questo formalismo enunciamo nuovamente i principi di sufficienza, verosimiglianza e condizionamento

PS, PFV, PC riformulati

Principio di sufficienza (bis)

Con riferimento a un esperimento $E = (Y, \theta, f_\theta)$ per il quale T è una statistica sufficiente, se

$$y, z \in \mathcal{Y} \text{ tali che } T(y) = T(z)$$

allora

$$\text{Ev}(E, y) = \text{Ev}(E, z).$$

PS, PFV, PC riformulati

Principio forte di verosimiglianza (bis)

Con riferimento a due esperimenti $E_i = (Y_i, \theta, f_\theta^{(i)})$ che condividono il parametro θ, se

$$y \in \mathcal{Y}_1, z \in \mathcal{Y}_2 \text{ tali che } L_1(\theta; y) \propto L_2(\theta; z)$$

allora

$$\text{Ev}(E_1, y) = \text{Ev}(E_2, z).$$

PS, PFV, PC riformulati

Principio di condizionamento (bis)

Con riferimento a due esperimenti $E_i = (Y_i, \theta, f_\theta^{(i)})$ che condividono il parametro θ, si definisca E^* l'esperimento mistura per cui con probabilità $1/2$ si effettua l'esperimento E_1, altrimenti si effettua l'esperimento E_2. Allora

$$\mathrm{Ev}(E^*, (j, Y_j)) = \mathrm{Ev}(E_j, (j, Y_j)).$$

Teorema di Birnbaum

Il teorema di Birnbaum asserisce che il principio di verosimiglianza è una conseguenza dei due principi di sufficienza e di condizionamento.

Teorema di Birnbaum (1962)

Il principio forte di verosimiglianza è una conseguenza del principio di sufficienza e di quello di condizionamento e viceversa.

Nel seguito se ne dà una dimostrazione nel caso discreto. Essa può essere generalizzata al caso continuo, ma si può anche argomentare che, dato che qualunque esperimento reale è discreto, le versioni discrete dei principi e del teorema sono tutto quanto serve.

Dimostrazione, (PS $\wedge$ PC) $\Rightarrow$ PFV

Dati due esperimenti con parametro θ in comune

$$E_i = (Y_i, \theta, f_\theta^{(i)})$$

definiamo l'esperimento mistura

$$E^* = \left(Y^* = (j, Y_j), \theta, f_\theta^* = \frac{1}{2} f_\theta^{(1)} + \frac{1}{2} f_\theta^{(2)} \right)$$

Siano y_1^* e y_2^* risultati sperimentali di E_1 e E_2 tali che

$$L_2(\theta) = f_\theta^{(2)}(y_1^*) = c f_\theta^{(1)}(y_1^*) = c L_1(\theta)$$

Definiamo la statistica

$$T(J, Y_J) = \begin{cases} (1, y_1^*) & \text{se } J = 2, Y_2 = y_2^* \\ (J, Y_J) & \text{altrimenti} \end{cases}$$

Dimostrazione, (PS $\wedge$ PC) $\Rightarrow$ PFV

$$T(J, Y_J) = \begin{cases} (1, y_1^*) & \text{se } J = 2, Y_2 = y_2^* \\ (J, Y_J) & \text{altrimenti} \end{cases}$$

Il valore della statistica è lo stesso in $(1, y_1^*)$ e $(2, y_2^*)$

$$T(1, y_1^*) = T(2, y_2^*) = (1, y_1^*);$$

negli altri punti dello spazio campionario il valore della statistica coincide col campione

$$T(J, Y_J) = \begin{cases} (1, y_1) & \text{se } J = 2, Y_2 = y_2^* \\ (1, y_1^*) & \text{se } J = 1, Y_2 = y_1^* \\ (J, Y_J) & \text{in qualunque altro punto dello spazio campionario} \end{cases}$$

Dimostrazione, (PS $\wedge$ PC) $\Rightarrow$ PFV

T è sufficiente, infatti considerando la distribuzione condizionata del campione a $T = t$ si ha

$$P(Y^* = (j, y_j) | T = t \neq (1, y_1^*)) = \begin{cases} 1 & \text{se } (j, y_j) = t \\ 0 & \text{altrimenti} \end{cases}$$

e, nel caso $T = (1, y_1^*)$ si ha

$$P(Y^* = (1, y_1^*) | T = (1, y_1^*)) + P(Y^* = (2, y_2^*) | T = (1, y_1^*)) = 1$$

e

$$P(Y^* = (1, y_1) | T = (1, y_1^*)) = \frac{\frac{1}{2} f_\theta^{(1)}(y_1)}{\frac{1}{2} f_\theta^{(1)}(y_1) + \frac{1}{2} f_\theta^{(2)}(y_2)} = \frac{\frac{1}{2} f_\theta^{(1)}(y_1)}{\frac{1}{2} f_\theta^{(1)}(y_1) + \frac{1}{2} c f_\theta^{(1)}(y_1)} = \frac{1}{1 + c}$$

ossia non dipende dal parametro.

Dimostrazione, (PS $\wedge$ PC) $\Rightarrow$ PFV

Possiamo allora applicare il principio di sufficienza all'esperimento mistura E^* e otteniamo

$$\mathrm{Ev}(E^*, (1, y_1)) = \mathrm{Ev}(E^*, (2, y_2))$$

da cui la tesi.

Dimostrazione, PFV $\Rightarrow$ (PS $\wedge$ PC)

Assumiamo ora come ipotesi il principio forte di verosmiglianza. La verosimiglianza dell'esperimento E^* per (j, y_j) è

$$L^*(\theta) = 0.5 f_\theta^{(j)}(y_j) \propto f_\theta^{(j)}(y_j)$$

e quindi le conclusioni dell'esperimento E^* devono essere uguali a quelle ottenute condizionando a J

$$\text{Ev}(E^*, (j, y_j)) = \text{Ev}(E, y_j)$$

Infine, che il principio forte di verosimiglianza implichi il PS è ovvio.

Il problema

- Le procedure frequentiste non rispettano il PFV, il che porterebbe a rifiutarlo.
- Tuttavia, PS e PC sono (o almeno appaiono) del tutto naturali.
- In virtù del teorema di Birnbaum accettare PS e PC dovrebbe implicare l'accettazione del PFV.

Tutto ciò è piuttosto critico per la filosofia della statistica (e della scienza più in generale), poiché sembrerebbe di dover abbandonare la logica frequentista.
Vi è un dibattito non concluso su questi aspetti, di cui riportiamo alcuni aspetti.

Aspetti legati alla validità del teorema, delle premesse e delle implicazioni

Premesse

- il PC è naturale in contesti semplici, la validità generale è contestabile

Dimostrazione

- La dimostrazione di Birnbaum è contestata sotto vari aspetti, ne esiste però una nuova versione (più tecnica) che dovrebbe superare tali obiezioni.

Implicazioni

- la validità del PFV non implica necessariamente la non validità delle procedure basate sul PCR, queste hanno una giustificazione in termini di efficacia nel lungo periodo (Neyman-Pearson)