

Statistiche sufficienze

La nozione di statistica sufficiente sorge nel momento in cui ci si chiede se sia possibile comprimere un campione di dati osservabili senza perdere informazioni sul parametro cercato

Def

Una statistica $\bar{T}_n(x_1, \dots, x_n)$ si dice sufficiente per il parametro θ se e solo se la distribuzione del campione condizionato a \bar{T}_n non dipende da θ

Prendendo il caso discreto, denotando la funzione di probabilità di \bar{T}_n come

$$q(\bar{T}_n(x_1, \dots, x_n); \theta)$$

O chiamando la funzione di probabilità congiunta del campione

$$p(x_1, \dots, x_n; \theta) \sim \text{il corrispettivo della densità nel discreto}$$

Allora \bar{T} è una statistica sufficiente se

$$\frac{p(x_1, \dots, x_n; \theta)}{q(\bar{T}(x_1, \dots, x_n); \theta)} \text{ non dipende da } \theta$$

Ese

Se $x_1, \dots, x_n \sim \text{iid. Ber}(\theta)$

Sapendo che la loro funzione di prob. congiunta è della forma

$$p(x_1, \dots, x_n; \theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}$$

Proviamo a vedere se $\bar{T}_n = \sum_{i=1}^n x_i$ sia una statistica sufficiente

Scriviamo ora la sua funzione di probabilità

$$q(\bar{T}(x_1, \dots, x_n); \theta) = \binom{n}{\sum_{i=1}^n x_i} \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}$$

Calcoliamo ora

Se è possibile vedere come non dipenda da θ

↓

$$\frac{p(x_1, \dots, x_n; \theta)}{q(\bar{T}(x_1, \dots, x_n); \theta)} = \frac{1}{\binom{n}{\sum_{i=1}^n x_i}}$$

\bar{T} è una statistica sufficiente

Teo Sufficientezza di Halmos-Savage

Una statistica è sufficiente se e solo se la funzione di probabilità condizionata o densità di X_1, \dots, X_n

si s'abborra come

$$p(X_1, \dots, X_n; \theta) = g(T(X_1, \dots, X_n; \theta)) h(X_1, \dots, X_n)$$

Dove $h(\cdot)$ non dipende da θ

Dim

① Sufficientezza \Rightarrow Fattorizzazione

Questa implicazione è ovvia, basta prendere

$$g(T(X_1, \dots, X_n; \theta)) = q(T(X_1, \dots, X_n; \theta))$$

$h(X_1, \dots, X_n) =$ la legge delle X_i condizionata a T

② Fattorizzazione \Rightarrow Sufficientezza

Prendiamo il caso discreto e partizioniamo il dominio di X_1, \dots, X_n in classi di equivalenza

corrispondenti ai valori della statistica sufficiente

Definiamo allora i sottosinsiemi di \mathbb{R}^n come

$$A_{\bar{T}(X_1, \dots, X_n)} \subseteq \mathbb{R}^n = \{Y = Y_1, \dots, Y_n : \bar{T}(Y) = \bar{T}(X)\}$$

Prendiamo ora il rapporto

$$\frac{p(X_1, \dots, X_n; \theta)}{S_{\bar{T}}(T(X_1, \dots, X_n; \theta))}$$

È dimostrato come questo non dipenda da θ

$$\frac{p(X_1, \dots, X_n; \theta)}{S_{\bar{T}}(T(X_1, \dots, X_n; \theta))} = \frac{g(T(X_1, \dots, X_n; \theta)) h(X_1, \dots, X_n)}{\sum_{(Y_1, \dots, Y_n) \in A_{\bar{T}(X)}} p(Y_1, \dots, Y_n; \theta)} = \frac{g(T(X_1, \dots, X_n; \theta)) h(X_1, \dots, X_n)}{\sum_{(Y_1, \dots, Y_n) \in A_{\bar{T}(X)}} g(T(Y_1, \dots, Y_n; \theta)) h(Y_1, \dots, Y_n)}$$

Non dipende

$$= \frac{g(T(X_1, \dots, X_n; \theta)) h(X_1, \dots, X_n)}{g(T(X_1, \dots, X_n; \theta)) \sum_{(Y_1, \dots, Y_n) \in A_{\bar{T}(X)}} h(Y_1, \dots, Y_n)} = \frac{h(X_1, \dots, X_n)}{\sum_{(Y_1, \dots, Y_n) \in A_{\bar{T}(X)}} h(Y_1, \dots, Y_n)}$$

$\cancel{\text{d}} \theta$

$\cancel{\text{d}} \theta$

Es

Sia $X_1, \dots, X_n \sim \text{iid } N(\mu, 1)$

La loro legge di prob. condizionata e' del tipo

$$p(X_1, \dots, X_n; \mu) = \frac{1}{(2\pi)^n} e^{-\frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2}$$

Consideriamo ora la statistica $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ con legge

$$q(\bar{X}_n; \mu) = \frac{1}{(2\pi)^2} e^{-\frac{1}{2} (\bar{X}_n - \mu)^2}$$

E' facile vedere che

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X}_n + \bar{X}_n - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{i=1}^n (\bar{X}_n - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu)^2$$

Quindi possiamo riscrivere

$$p(X_1, \dots, X_n; \mu) = \frac{1}{(2\pi)^n} e^{-\frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2} = \frac{1}{(2\pi)^n} e^{-\frac{1}{2} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu)^2 \right)} = \frac{1}{(2\pi)^n} \left[e^{-\frac{1}{2} \sum_{i=1}^n (X_i - \bar{X}_n)^2} e^{-\frac{n}{2} (\bar{X}_n - \mu)^2} \right]$$

Possiamo notare che la legge di \bar{X}_n $q(\bar{X}_n; \mu)$ e' proprio q

Per questo appunto la statistica \bar{X}_n e' una statistica sufficiente

Questo e' verificabile anche se facciamo

$$\frac{p(X_1, \dots, X_n; \mu)}{q(\bar{X}_n; \mu)} = h = \frac{1}{(2\pi)^n} e^{-\frac{1}{2} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

Che e' indipendente da μ

Teo di Rao-Blackwell

L'idea intuitiva dietro alle statistiche sufficienti e' quella di considerare l'informazione del campione aleatorio

Conservando quello che e' utile per risalire al "vero" valore del parametro

Sembra quindi intuitivo che gli stimatori efficienti debbano essere costruiti a partire da statistiche sufficienti

Prima di enunciare il Teo bisogna ricordare alcune nozioni sul voto medio condizionato

Prendiamo 2 v.a. continue X, Y con densità congiunta

$f_{X,Y}(x,y)$ \rightsquigarrow indica come distribuita la prob. sul piano (x,y)

Da qui ricorriamo le densità marginale

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dx$$

Calcoliamo ora i valori attesi di X, Y

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx$$

$$E[Y] = \int_{-\infty}^{+\infty} y f_Y(y) dy$$

La densità condizionata di X dato $Y=y$ è

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Di conseguenza

$$E[X|Y=y] = \int x f_{X|Y}(x|y) dx \rightsquigarrow \text{c'è una funzione di } y, \text{ ma essendo fissato il risultato è un numero reale}$$

Ponendo però y con la v.a. Y abbiamo che

$$E[X|Y] = \int x f_{X|Y}(x|Y) dx = \psi(Y) \rightsquigarrow \text{In funzione di } Y$$

\rightsquigarrow indica la migliore stima di X usando l'informazione contenuta in Y

Lemma

Siano X, Y due v.a. di quadrato integrabile definite sullo stesso spazio di probabilità

Allora

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y])$$

Dim

$$\begin{aligned}
 \text{Var}(x) &= E[(x - E[x])^2] = E[(x - E[X|Y] + E[X|Y] - E[X])^2] = \\
 &= E_Y \left[E[(x - E[X|Y] + E[X|Y] - E[X])^2] \middle| Y \right] = \\
 &= E_Y \left[E[(x - E[X|Y])^2] \middle| Y \right] + E_Y \left[(E[X|Y] - E[X])^2 \middle| Y \right] \\
 &= E \left[\text{Var}(x|Y) \right] + \text{Var}(E[X|Y]) \quad \text{??} \Rightarrow \text{D} \rightarrow \text{chiedere a Leo perde il senso}
 \end{aligned}$$

Teo di Rao-Blackwell

Sia $W = W(X_1, \dots, X_n)$ uno stimatore con le seguenti caratteristiche

a) $E[W] = \theta_0 \rightsquigarrow$ non distorto

b) $E[W^2] < \infty \rightsquigarrow$ Var Simba

Preso T una statistica sufficiente

Allora è possibile creare un nuovo stimatore W' b.c.

$$W' = E[W|T] \rightsquigarrow \text{p}(\bar{T})$$

Con le seguenti caratteristiche

a) $E[W|T] = \theta_0 \rightsquigarrow$ non distorto

b) $\text{Var}(W') \leq \text{Var}(W) \rightsquigarrow$ efficiente

Dim

a) $E[W'] = E[W|T] = E[W] = \theta_0$

b) $\text{Var}[W] = E[\text{Var}(W|T)] + \text{Var}(E[W|T]) \geq \text{Var}(E[W|T]) = \text{Var}(W')$

Oss

In questa dimostrazione non compare minimamente il concetto di statistica sufficiente

Ma in realtà usarla ci garantisce che il valor medio condizionato non dipenda dal parametro e che quindi

$\psi(\bar{x})$ sia effettivamente uno stimatore