13 Statistiche sufficienti

Una domanda naturale che possiamo porci, specialmente in un momento in cui masse enormi di dati sono a disposizione, è la seguente: posso comprimere un campione di dati osservati senza perdere informazione sul parametro che mi interessa? Questa domanda ci porta alla nozione di statistiche sufficienti.

> **Spiegazione Semplice**
>
> **Cos'è una "Statistica Sufficiente"?**
> L'idea è molto intuitiva: una statistica (come la media campionaria, $\overline{X}$) è "sufficiente" se **contiene tutta l'informazione** sul parametro θ (come la media vera, μ) che era contenuta nell'intero campione originale $(X_1, \ldots, X_n)$.
> È un metodo di "compressione dati" perfetto: ci permette di buttare via l'intero dataset (che possono essere Terabyte di dati) e tenere solo un numero (o un piccolo set di numeri), con la garanzia di non aver perso *nessuna informazione* utile per stimare il nostro parametro.
> Ad esempio, se voglio stimare la media μ di una popolazione Gaussiana, non mi serve sapere i 1000 valori individuali; mi "è sufficiente" conoscere la loro somma $\sum X_i$ (o la loro media $\overline{X}$).

Definition 116 (Statistiche sufficienti)**.** *Una statistica $T_n = T(X_1, ..., X_n)$ si dice sufficiente per il parametro θ se la distribuzione del campione condizionata a T_n non dipende da θ. In altre parole, la legge $p(X_1, ..., X_n|T_n)$ non dipende dal parametro. In questa sezione, supponiamo sempre per semplicità di avere a che fare con variabili aleatorie discrete; in questo ambito, è immediato verificare che conoscendo solo la statistica sufficiente possiamo generare un nuovo campione uguale in distribuzione all'originale. Abbiamo infatti, per qualsiasi scelta di valori $(x_1, x_2, ..., x_n)$*

$$\begin{aligned}
&Pr(X_1 = x_1, ..., X_n = x_n) \\
&= Pr(X_1 = x_1, ..., X_n = x_n \cap T(X_1, ..., X_n) = T(x_1, ..., x_n)) \\
&= Pr(X_1 = x_1, ..., X_n = x_n|T(X_1, ..., X_n) = T(x_1, ..., x_n))Pr(T(X_1, ..., X_n) = T(x_1, ..., x_n)) \\
&= Pr(Y_1 = x_1, ..., Y_n = x_n|T(Y_1, ..., Y_n) = T(x_1, ..., x_n))Pr(T(X_1, ..., X_n) = T(x_1, ..., x_n)) \\
&= Pr(Y_1 = x_1, ..., Y_n = x_n)
\end{aligned}$$

dove le Y_i sono generate dalla distribuzione condizionata delle X_i a T; poichè l'uguaglianza vale per tutte le scelte di $(x_1, ..., x_n)$ abbiamo quindi dimostrato che

$$(Y_1, ..., Y_n)\underline{d}(X_1, ..., X_n).$$

Si noti che nella dimostrazione (primo passaggio) abbiamo anche utilizzato il fatto che l'evento $X_1 = x_1, ..., X_n = x_n$ è contenuto nell'evento $T(X_1, ..., X_n) = T(x_1, ..., x_n)$, e pertanto la probabilità della loro intersezione è uguale alla probabilità di $X_1 = x_1, ..., X_n = x_n$. Più in generale, denotiamo $q(T(X_1, ..., X_n); \theta)$ la funzione di probabilità della statistica sufficiente T; dalla definizione di probabilità condizionata è immediato verificare che T è sufficiente se e solo se $p(X_1, ..., X_n; \theta)/q(T(X_1, ..., X_n); \theta)$ non dipende da θ.

> **Spiegazione Semplice**
>
> **Cosa significa la Definizione Formale?**
> La definizione $p(X_1, \ldots, X_n|T_n)$ "non dipende da θ" è un modo molto tecnico per dire: "Una volta che ti ho detto il valore della statistica T_n (es. la somma è 7), il campione originale $(X_1, \ldots, X_n)$ non contiene più nessuna informazione *aggiuntiva* sul parametro θ."
> Tutta l'informazione su θ è stata "aspirata" e si trova ora in T_n. Sapere l'ordine esatto in cui sono usciti i dati (es. T-C-T-C... vs T-T-C-C...) non ci aiuta in alcun modo a stimare θ, una volta che sappiamo già che la somma è T_n.

Example 117. *Si consideri un campione IID di variabili Bernoulliane $X_i \sim Ber(\theta)$. La loro legge congiunta è data da*

$$p(X_1, ..., X_n; \theta) = \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}.$$

Verifichiamo che $T_n := \sum_{i=1}^n X_i$ sia una statistica sufficiente; la sua funzione di probabilità è una binomiale, data quindi da

$$q(T_n(X_1, ..., X_n); \theta) = \binom{n}{\sum_{i=1}^n X_i} \theta^{\sum_{i=1}^n X_i} (1-\theta)^{n-\sum_{i=1}^n X_i}$$

e quindi

$$\frac{p(X_1, ..., X_n; \theta)}{q(T_n(X_1, ..., X_n); \theta)} = \frac{1}{\binom{n}{\sum_{i=1}^n X_i}}$$

che è indipendente da θ.

> **Spiegazione Semplice**
>
> **Esempio 1: Lancio di Monete (Bernoulli)**
>
> - **Campione $(X_1, \ldots, X_n)$:** L'elenco esatto dei risultati dei lanci (es. $X_1 = 1, X_2 = 0, X_3 = 1, \ldots$).
> - **Parametro (θ):** La probabilità (ignota) che esca Testa (p).
> - **Statistica (T_n):** La somma $\sum X_i$, cioè il numero totale di Teste (es. $k = 7$).
>
> Per stimare θ, ci interessa l'ordine in cui sono uscite Teste e Croci? No. L'unica cosa che ci serve è il *numero totale* di Teste (la somma T_n).
> La formula $p(X|T) = 1/\binom{n}{k}$ (che non dipende da θ) lo dimostra: una volta che ti ho detto che "su n lanci, k sono Teste", la probabilità di una specifica sequenza (es. T-T-C) è 1 diviso il numero di tutte le sequenze possibili con k Teste. Il parametro θ è sparito.

Example 118. *Si consideri un campione IID di variabili Gaussiane $X_i \sim N(\theta, 1)$. La loro legge congiunta è data da*

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

consideriamo ora la statistica $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ con legge $N(\mu, \frac{1}{n})$ cioè

$$q(\overline{X}_n; \mu) = \frac{\sqrt{n}}{(2\pi)^{1/2}} exp(-\frac{n}{2}(\overline{X}_n - \mu)^2).$$

E' facile verificare che

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

da cui

$$\begin{aligned} \frac{p(X_1, ..., X_n; \mu)}{q(\overline{X}_n; \mu)} &= Const \times \frac{exp(-\frac{1}{2}\sum_{i=1}^n (X_i - \mu)^2)}{exp(-\frac{n}{2}(\overline{X}_n - \mu)^2)} \\ &= Const \times exp(-\frac{1}{2}\sum_{i=1}^n (X_i - \overline{X}_n)^2), \end{aligned}$$

e la dimostrazione della sufficienza è completata.

> **Spiegazione Semplice**
>
> **Esempio 2: Media Gaussiana**
> Allo stesso modo, se abbiamo n misurazioni da una distribuzione Gaussiana $N(\mu, 1)$ e vogliamo stimare la media vera μ.
> La media campionaria $\overline{X}_n$ (o la somma, sono equivalenti) è una statistica sufficiente.

L'algebra (che scompone $\sum(X_i-\mu)^2$ in $\sum(X_i-\overline{X}_n)^2+n(\overline{X}_n-\mu)^2$) mostra che il rapporto $p(X|\overline{X})$ dipende solo da $\sum(X_i-\overline{X}_n)^2$, che è la varianza campionaria. Questo termine non contiene μ (il parametro θ).
Ancora una volta, $\overline{X}_n$ ha "catturato" tutta l'informazione su μ contenuta nel campione.

A metà del ventesimo secolo Halmos e Savage hanno dato una caratterizzazione della suffiicenza leggermente più generale, come segue.

Theorem 119 (Criterio di Fattorizzazione di Halmos e Savage)**.** *Una statistica è sufficiente se e solo se la funzione di probabilità (o di densità) di $X_1,...,X_n$ si fattorizza come*

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

dove la funzione $h(.)$ non dipende da θ.

Spiegazione Semplice

Un "Test" Pratico per la Sufficienza
La definizione formale (Def. 116) che usa la probabilità condizionata $p(X|T)$ è un incubo da usare per le dimostrazioni.
Questo teorema (il Criterio di Fattorizzazione) ci dà un "test" molto più semplice. Per dimostrare che T è sufficiente, ti basta fare un po' di algebra:

1. Prendi la funzione di verosimiglianza $p(X_1,\ldots,X_n;\theta)$.
2. Prova a "spezzarla" (fattorizzarla) in due pezzi moltiplicati tra loro:
   - **Pezzo g:** Una funzione g che dipende dal parametro θ, ma che "vede" i dati X *solo* attraverso la statistica $T(X)$.
   - **Pezzo h:** Una funzione h che può dipendere da tutti i dati X in modo complicato, ma *non deve* contenere il parametro θ.

Se riesci a fare questa separazione, hai dimostrato che $T(X)$ è sufficiente.

Dimostrazione. L'implicazione sufficienza $\Rightarrow$ fattorizzazione è ovvia: basta prendere per $g(T;\theta)=q(T;\theta)$ (la legge di q) e h la legge delle X_i condizionata a T, che sappiamo essere indipendente da θ. L'interesse del Teorema è quindi nella seconda parte, cioè nel caso in cui la fattorizzazione non corrisponda necessariamente alla coppia legge di T/ legge di $X_1,...,X_n$ condizionata a T. Come sempre in questa sezione ci limitiamo il caso discreto, e partizioniamo il dominio di $X_1,...,X_n$ in classi di equivalenza corrispondenti ai valori della statistica sufficiente; definiamo cioè i sottoinsiemi di $\mathbb{R}^n$

$$A_{T(x_1,...,x_n)}=\{(x_1',...,x_n'):T(x_1',...,x_n')=T(x_1,...,x_n)\}.$$

Andiamo a calcolare, per $x_1,...,x_n$ qualsiasi ma fissati

$$\begin{aligned}\frac{p(x_1,...,x_n;\theta)}{q(T(x_1,...,x_n);\theta)} &= \frac{g(T(x_1,...,x_n;\theta))h(x_1,...,x_n)}{q(T(x_1,...,x_n);\theta)}\\ &= \frac{g(T(x_1,...,x_n;\theta))h(x_1,...,x_n)}{\sum_{(x_1',...,x_n')\in A_{T(x_1,...,x_n)}} p(x_1',...,x_n';\theta)}\\ &= \frac{g(T(x_1,...,x_n;\theta))h(x_1,...,x_n)}{\sum_{(x_1',...,x_n')\in A_{T(x_1,...,x_n)}} g(T(x_1,...,x_n;\theta))h(x_1',...,x_n')}\\ &= \frac{h(x_1,...,x_n)}{\sum_{(x_1',...,x_n')\in A_{T(x_1,...,x_n)}} h(x_1',...,x_n')};\end{aligned}$$

l'ultima frazione non dipende da θ, il che è sufficiente a concludere la dimostrazione della sufficienza. □

13.1 Il Teorema di Rao-Blackwell

L'idea intutiva dietro le statistiche sufficienti è quella di condensare l'informazione del campione aleatorio conservando quello che è utile per risalire al valore "vero" del parametro. Sembra quindi inuitivo che gli stimatori efficienti debbano essere costruiti solo a partire da statistiche sufficienti. Questa idea è resa rigorosa dal teorema di Rao-Blackwell; prima di enunciarlo, dobbiamo ricordare alcune nozioni sul valor medio condizionato, restringendoci solamente al caso discreto. Siano X Y valori aleatorie discrete con momento secondo finito e funzione di probabilità congiunta $P_{XY}(x, y)$; la densità di X condizionata a $Y = y$ è data da

$$p_{X|Y}(x|y) = \begin{cases} \frac{p_{XY}(x,y)}{p_Y(y)}, & se\ p_Y(y) > 0 \\ 0 & altrimenti \end{cases}$$

(Nota: il testo originale riporta $p_X(x)$ al denominatore e $p_{X|Y}(x, y)$ a sinistra, ma $p_Y(y)$ e $p_{X|Y}(x|y)$ sono corretti)

ed il valor medio condizionato a $Y = y$ naturalmente

$$E[X|Y = y] = \sum_{x \in Dom(X)} x p_{X|Y}(x|y)$$

(Nota: il testo originale presenta una formula parziale e diversa $E[X|Y = y] = \sum x p_{XY}(x, y)/p_Y$)

Se evitiamo di fissare il valore della variabile aleatoria Y, avremo ora una nuova variabile aleatoria (funzione di Y) della forma $\psi(Y) = E[X|Y]$, che ad ogni valore di Y associa il valore del valor medio condizionato corrispondente a quel valore

$$E[X|Y] = \sum_{x \in Dom(X)} x \frac{p_{XY}(x, Y)}{p_Y(Y)} = \sum_{x \in Dom(X)} x p_{X|Y}(x|Y);$$

in altre parole, $\psi(Y)$ assume il valore $\psi(y) = E[X|Y = y]$ con probabilità $p_Y(y)$ (la legge marginale di Y. E' immediato verificare che (la cosiddetta legge del valor medio iterato)

$$E[E[X|Y]] = E[X];$$

infatti

$$\begin{aligned} E[E[X|Y]] &= \sum_y E[X|Y = y] p_Y(y) = \sum_y \sum_{x \in Dom(X)} x \frac{p_{XY}(x, y)}{p_Y(y)} p_Y(y) \\ &= \sum_{x \in Dom(X)} x \sum_y p_{XY}(x, y) \\ &= \sum_{x \in Dom(X)} x p_X(x) = E[X]. \end{aligned}$$

Allo stesso modo possiamo definire la variabile aleatoria (funzione di Y)

$$Var[X|Y] = \sum_{x \in Dom(X)} (x - E[X|Y])^2 p_{X|Y}(x|Y),$$

che corrisponde alla varianza condizionata di X, in funzione del valore aleatorio di Y. Vale il seguete risultato.

Lemma 120. *Siano X, Y variabili aleatorie di quadrato integrabile definite sullo stesso spazio di probabilità. Abbiamo*

$$Var[X] = E[Var[X|Y]] + Var[E[X|Y]].$$

Spiegazione Semplice

Scomposizione della Varianza
Questo lemma è un risultato fondamentale della probabilità, noto come **Legge della Varianza Totale**. Dice che la variabilità (Varianza) totale di una variabile X può essere scomposta in due parti:

1. $E[Var[X|Y]]$**:** La **varianza interna media**. È la variabilità *residua* di X che *non* è spiegata da Y. (Quanto "balla" ancora X *dopo* che Y è noto, in media).

2. $Var[E[X|Y]]$**:** La **varianza esterna**. È la variabilità di X che *è spiegata* da Y. (Quanto "balla" la media di X al variare di Y).

Varianza Totale = Varianza non Spiegata + Varianza Spiegata.

Dimostrazione. Abbiamo che

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= E[(X - E[X|Y] + E[X|Y] - E[X])^2] \\ &= E[(X - E[X|Y])^2] + E[(E[X|Y] - E[X])^2] \\ &\quad + 2E[(X - E[X|Y])(E[X|Y] - E[X])]. \end{aligned}$$

Ora per definizione abbiamo

$$E[(E[X|Y] - E[X])^2] = E_Y[(E[X|Y] - E[X])^2] = Var[E[X|Y]];$$

inoltre

$$\begin{aligned} E[(X - E[X|Y])(E[X|Y] - E[X])] &= E_Y[E[(X - E[X|Y])(E[X|Y] - E[X])|Y]] \\ &= E_Y[(E[X|Y] - E[X])E[(X - E[X|Y])|Y]] \\ &= 0, \end{aligned}$$

perché $E[(X - E[X|Y])|Y] = 0$. Infine

$$E[(X - E[X|Y])^2] = E[E[(X - E[X|Y])^2|Y]] = E[Var[X|Y]],$$

e il Lemma è dimostrato. □

Possiamo finalmente arrivare al risultato principale, che ci garantisce che condizionando uno stimatore non distorto su una statistica sufficiente se ne ottiene un altro ancora non distorto e con varianza non maggiore.

Theorem 121 (Rao-Blackwell)**.** *Sia $W_n = W_n(X_1, ..., X_n)$ uno stimatore non distorto e con varianza finita del parametro θ, e sia $T_n = T(X_1, ..., X_n)$ una statistica sufficente per θ. Definiamo $\psi(T) = E[W_n|T]$; allora abbiamo*

$$E[\psi(T)] = \theta \text{ e } Var[\psi(T)] \leq Var[W_n]$$

Spiegazione Semplice

Perché le Statistiche Sufficienti sono così importanti?
Questo teorema è la risposta. Ci fornisce una **ricetta per migliorare uno stimatore**.

1. **Partenza:** Hai uno stimatore W_n che è "onesto" (non distorto, $E[W_n] = \theta$) ma magari non molto "preciso" (ha una varianza $Var[W_n]$ alta).

2. **Ingrediente Chiave:** Trovi una statistica *sufficiente* T_n (che "aspira" tutta l'informazione su θ dal campione).

3. **Ricetta:** Crei un nuovo stimatore $\psi(T)$ calcolando la media del tuo vecchio stimatore W_n *dato* T_n. (In pratica, $\psi(T) = E[W_n|T_n]$).

Risultato: Il tuo nuovo stimatore $\psi(T)$ è **sempre migliore** (o al peggio uguale) del vecchio W_n:

- È ancora "onesto" (non distorto): $E[\psi(T)] = \theta$.
- È più "preciso" (ha una varianza minore): $Var[\psi(T)] \leq Var[W_n]$.

Questo teorema ci dice che se stiamo cercando lo stimatore "migliore" (con varianza minima), dobbiamo cercarlo *solo* tra le funzioni di statistiche sufficienti.

Dimostrazione. La dimostrazione è una applicazione immediata dei risultati precedenti sul valor medio condizionato. In particolare

$$E[\psi(T)] = E[E[W_n|T]] = E[W_n] = \theta,$$
$$Var[W_n] = E[Var[W_n|T]] + Var[E[W_n|T]]$$
$$\geq Var[E[W_n|T]] = Var[\psi(T)].$$

□

Spiegazione Semplice

La Dimostrazione (spiegata dal Lemma 120)
La dimostrazione è incredibilmente elegante e usa direttamente il Lemma 120 (Legge della Varianza Totale):
$Var[W_n] = E[Var[W_n|T]] + Var[E[W_n|T]]$
Guardiamo i due pezzi:

- $Var[E[W_n|T]]$: Questa è la varianza del nostro *nuovo* stimatore $\psi(T)$.
- $E[Var[W_n|T]]$: Questa è la "varianza interna media". Poiché una varianza non può mai essere negativa, questo termine è sempre ≥ 0.

Quindi, la formula ci dice: Var(Vecchio Stimatore) = (Termine ≥ 0) + Var(Nuovo Stimatore)
Questo implica matematicamente che Var(Vecchio Stimatore) $\geq Var$(Nuovo Stimatore).

Remark 122. *Dove abbiamo usato nella dimostrazione precedente la sufficienza? Apparentemente non gioca alcun ruolo, ma in realtà ci garantisce che il valor medio condizionato non dipenda dal parametro, e quindi che $\psi(T)$ sia effettivamente uno stimatore. Ad esempio consideriamo un campione di sole due osservazioni X_1, X_2 Gaussiane indipendenti con valor medio μ e varianza 1, e consideriamo lo stimatore $\overline{X}_2 = \frac{1}{2}(X_1 + X_2)$. Consideriamo*

$$\psi(X_1) := E[\overline{X_2}|X_1] = \frac{X_1 + \mu}{2}$$

Questa variabile aleatoria ha valor medio μ e varianza $\frac{1}{4} < Var(\overline{X}_2) = \frac{1}{2}$; non si tratta però di uno stimatore, perchè il suo valore non può essere determinato senza conoscere il valore del parametro μ.

Spiegazione Semplice

Perché la "Sufficienza" è la chiave? (Un punto sottile)
Questo remark è cruciale. La dimostrazione $Var(W_n) \geq Var(\psi(T))$ funziona sempre, ma *non serve a nulla* se $\psi(T) = E[W_n|T]$ non è un vero stimatore.
Uno "stimatore" (Def. 80) deve essere una formula calcolabile *solo* dai dati $(X_1, \ldots, X_n)$, senza conoscere il parametro θ (altrimenti che stima è?).

- **L'esempio fallimentare:** Il testo usa $T = X_1$ (che *non* è sufficiente per μ). Se proviamo ad applicare Rao-Blackwell, il "nuovo stimatore" è $\psi(X_1) = E[\overline{X}_2|X_1] = \frac{X_1+\mu}{2}$. Questa formula è *inutilizzabile*: per calcolarla serve μ, che è proprio il valore che non conosciamo!
- **Perché la sufficienza salva tutto:** Se T è *sufficiente* (Def. 116), la distribuzione $p(X|T)$ *non dipende da θ*. Di conseguenza, anche la media $E[W_n|T]$ (che si calcola usando $p(X|T)$) non dipenderà da θ. Sarà una vera formula calcolabile solo dai dati, e quindi un vero stimatore.