Capitolo 3

L'ENTROPIA NELLA STATISTICA OGGETTIVA E SOGGETTIVA

I. ENTROPIA E GRANDI NUMERI

L'"enigmatica" ubiquità dell'entropia è dovuta soprattutto al suo significato aritmetico fondamentale (e dunque "elementare", demitizzante): l'entropia serve per contare un gran numero di oggetti inquadrabili in pochi tipi distinti.

1. Coefficienti binomiali e coefficienti multinomiali

Il coefficiente binomiale

$$\binom{n}{w} = \frac{n!}{w!(n-w)!} = \frac{n^{\underline{w}}}{w!}$$

ha un'importanza enorme in tutti i contesti in cui serve contare, e dunque nei contesti più svariati che immaginar si possa. Oltre a dire quante siano le n-uple binarie di *peso* w (quelle in cui l'1 compare esattamente w volte), esso ci dice anche quanti siano i w-sottinsiemi tratti da un insieme di n elementi (i sottinsiemi S di numerosità w in esso "scelti"); è lo stesso problema di prima, basta interpretare le n-uple binarie come funzioni caratteristiche: $x_i = 1$ se $a_i \in S$, altrimenti $x_i = 0$.

Notazione. $n^{\underline{w}}$ indica la *potenza discendente*, ossia il prodotto di w fattori $n(n-1)(n-2)(n-3)$ ... fino a $n-w+1$; in modo analogo si definisce la *potenza ascendente* $n^{\overline{w}}$. Con questa notazione: $n! = n^{\underline{w}} = 1^{\overline{w}}$.

Esercizi. Il lettore riveda le formule elementari sui coefficienti binomiali, che ne valorizzano il significato. Uno studio della funzione $f(w) = \binom{n}{w}$ (cenno: confronta con 1 i rapporti $\frac{f(w+1)}{f(w)}$) mostra in particolare che il massimo di $\binom{n}{w}$ è $\binom{n}{\lfloor \frac{n}{2} \rfloor}$.

Se n è troppo grande sorge una difficoltà, legata all'esito incerto dello scontro fra il "gigante" (il fattoriale) che sta al numeratore e i due "giganti" che stanno al denominatore: chi la spunta, qual è l'andamento asintotico del coefficiente binomiale? Più che una formula esatta ma mal decifrabile (e soggetta comunque a insidiosi errori di calcolo, ossia esatta solo in linea di principio), per $n \gg 1$ sarebbe assai meglio avere una formula approssimata, ma a misura d'uomo. Uscire da questa difficoltà ci prenderà un po' di tempo, ma è una fatica che val la pena di fare.

Il coefficiente multinomiale

$$\binom{n}{n_1, n_2, \ldots, n_K} = \frac{n!}{n_1! n_2! \ldots n_K!}$$

generalizza quello binomiale e risponde alla domanda ($\Box$!):

Assegnato un insieme $\mathcal{A}$ di K elementi, $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$, quante sono le n-uple costruite su $\mathcal{A}$ in cui ciascun elemento a_i compare esattamente n_i volte? Nel linguaggio dei it tipi, definiti veso la fine dell'ultimo capitolo (cui al lettore converrà ritornare), ciò significa che il coefficiente multinomiale è la numerosità del tipo di composizione $(n_1, n_2, \ldots, n_K)$.

2. Entropia e coefficienti multinomiali asintotici

Lo spazio $\mathcal{A}^n$ genera γ_n probabilità "nascoste" che si chiamano *empiriche*, tante quanti sono i tipi (le classi di composizione). Esse nascono da conti sulle n-uple e dunque hanno una *natura combinatoria*, non c'è nulla di aleatorio.

Sia x un'n-upla di $\mathcal{A}^n$ e sia $\{n_1, n_2, \ldots, n_K\}$ la sua composizione. Poichè le K frequenze relative di presenza delle lettere a_i in x, $f_i = \frac{n_i}{n}$, sono numeri fra 0 e 1 che sommano a 1, basta porre $P(a_i) = \text{Prob}(a_i) = f_i = \frac{n_i}{n}$, per ottenere una probabilità $P = P_x$ su $\mathcal{A}$, e dunque, tramite estensione moltiplicativa, bernoulliana, una probabilità P^n su $\mathcal{A}^n$. Ci si può divertire costringendo x a "mangiarsi la coda", ossia calcolando la probabilità di x tramite la $P^n = P_x^n$ che la x stessa ha generato:

$$P^n(x) = P(a_1)^{n_1} P(a_2)^{n_2} \ldots P(a_K)^{n_K} = \prod_{1 \le i \le K} P(a_i)^{nf_i} = 2^{-nH(P_x)}$$

Fra l'entropia e i coefficienti multinomiali (in particolare binomiali) c'è un legame molto profondo, che ora porteremo alla luce. Sia x un'n-upla, $P = P_x$ la sua probabilità empirica, e $T = T_x$ il suo tipo. Poichè tutte le $|T|$ n-uple di T hanno la stessa probabilità $2^{-nH(P)}$:

$$P^n(T) = |T| 2^{-nH(P)}$$

e dunque, qualunque probabilità essendo ≤ 1:

$$|T| = \binom{n}{n_1, n_2, \ldots, n_K} \le 2^{nH(P)}$$

Ciò basta per garantire che la crescita del coefficiente multinomiale è al più esponenziale in n di *tasso*, ossia di esponente normalizzato, $H(P)$ (è "sotto-fattoriale"). La maggiorazione è stretta? Vediamo di trovare una minorazione.

Ripartiamo $\mathcal{A}^n$ nei $\gamma_n \le (n+1)^K$ tipi che lo costituiscono, $T_1, T_2, \ldots, T_{\gamma_n}$ (fra questi c'è T, il tipo cui appartiene x). Si ha, tenendo conto che i tipi sono disgiunti ed esauriscono lo spazio:

$$1 = P^n(\mathcal{A}^n) = \sum_{1 \le j \le \gamma_n} P^n(T_j) \le \gamma_n P^n(T^*)$$

dove T^* è il tipo di probabilità massima: ebbene, dimostreremo in appendice che tale tipo è proprio quello che con le sue n-uple ha "partorito" la P: $T^* = T = T_x$; c'è insomma una corrispondenza amorosa fra il tipo-padre e la probabilità -figlia. Ma allora:

$$1 \le \gamma_n P^n(T) = \gamma_n \binom{n}{n_1, n_2, \ldots, n_K} 2^{-nH(P)}$$

e dunque

$$\binom{n}{n_1, n_2, \ldots, n_K} \geq \gamma_n^{-1} 2^{nH(P)}$$

che è la limitazione inferiore cercata. Il confronto con la maggiorazione diventa più espressivo se "trasciniamo" γ_n in alto, all'esponente; si trova:

$$\binom{n}{n_1, n_2, \ldots, n_K} = 2^{n(H(P)-\epsilon_n)}$$

con $0 \leq \epsilon_n \leq \frac{1}{n}\log\gamma_n \leq \frac{K}{n}\log(n+1)$; l'infinitesimo positivo ϵ_n è assai ben controllato: la maggiorazione di ϵ_n è infatti uniforme, non c'è traccia di P. L'esito dello scontro fra giganti è ora perfettamente chiaro, abbiamo trovato l'approssimazione che volevamo:

$$\binom{n}{n_1, n_2, \ldots, n_K} \approx 2^{nH(P)}$$

Ossia: il coefficiente multinomiale esplode esponenzialmente, il suo "ordine di grandezza" (il logaritmo) cresce linearmente con n, mentre il suo tasso è mansueto, ed è circa uguale all'entropia $H(P)$ (la differenza fra l'entropia e il tasso del coefficiente multinomiale è infinitesima). In sintesi: *l'entropia di Shannon è il surrogato asintotico del coefficiente multinomiale*; in particolare *la funzione entropica è il surrogato asintotico del coefficiente binomiale*:

$$\binom{n}{w} \approx 2^{nh(\frac{w}{n})}$$

Di solito i fattoriali si approssimano asintoticamente con la formula di Stirling. La strada alternativa che abbiamo seguito noi non solo è rapida, ma è anche simpatica: usare la probabilità come strumento al servizio della combinatorica implica infatti rovesciare i ruoli che sono abituali. Il "metodo probabilistico" ha acquistato una grande importanza nella ricerca matematica a partire dai lavori di Pál Erdös negli anni 50.

Esercizio in appendice.

Lemmino aritmetico. $\frac{n!}{m!} \leq n^{n-m}$ (distingui i due casi $n \geq m$ e $n \leq m$).

Lemma promesso. Se T è un tipo generico e T_x è il tipo cui x appartiene: $P_x^n(T) \leq P_x^n(T_x)$

Cenno: cercando di vincere la noia, scrivi il rapporto delle probabilità per esteso e applica il lemmino ai K rapporti $\frac{n(a_i|y)!}{n(a_i|x)!}$, dove $y \in T$, e dove $n(.|.)$ indica il numero di presenze della lettera al primo argomento nell'n-upla al secondo (rammenta che $P_x(a_i) = \frac{n(a_i|x)}{n}$):

$$\frac{P_x^n(T)}{P_x^n(T_x)} = \frac{|T|}{|T_x|} \times \frac{P_x^n(x)}{P_x^n(y)} = \prod_i \frac{n(a_i|y)!}{n(a_i|x)!} \times \prod_i \frac{n(a_i|x)^{n(a_i|y)} n^{n(a_i|x)}}{n(a_i|x)^{n(a_i|x)} n^{n(a_i|y)}}$$

$$\leq \prod_i n^{n(a_i|x)-n(a_i|y)} = n^{n-n} = 1$$

Esercizio: Se x appartiene al tipo T di composizione nF:

$$-n^{-1}\log P^n(x) = H(F\|P) = H(F) + D(F\|P),$$
$$n^{-1}\log|T| = H(F) - \delta \text{ con } 0 \leq \delta \leq n^{-1}\log\gamma_n$$
$$\Rightarrow -n^{-1}\log P^n(T) = D(F\|P) + \delta \text{ (detto alla buona: } P^n(T) \approx 2^{-nD(F\|P)})$$

3. Probabilità oggettive e grandi numeri

Nel seguito ci serve per ragioni "tecniche" una successione $\{\epsilon_n\}_{n\geq 1}$ infinitesima "ma non troppo", nel senso che $\lim \epsilon_n = 0$, ma già $\lim \sqrt{n}\epsilon_n = +\infty$. Ad esempio, nella seconda dimostrazione della legge dei grandi numeri riportata sotto prenderemo proprio $\epsilon_n = \frac{1}{\sqrt[4]{n}}$.

Definizione: Un'n-upla $x \in \mathcal{A}^n$ di composizione $nF = n(f_1, \ldots, f_K)$ di dice tipica secondo P se in essa non compaiono eventuali lettere di P-probabilità nulla, e se $\forall i \; |f_i - p_i| \leq \epsilon_n$.

Dunque, se un n-upla è tipica, $d_M(F,P) \leq K\epsilon_n \to 0$. La tipicità dipendendo solo dalla composizione, per cui un'n-upla è o non è tipica assieme a tutte quelle del suo tipo. Se non crea ambiguità diremo solo che la x è *tipica*, e dimenticheremo P. (Una precisazione: se $p_i = 0$, abbiamo rafforzato la richiesta $|f_i - p_i| \leq \epsilon_n$, ciò che diminuisce il numero delle sequenze tipiche, ma non altera la loro probabilità globale.)

> Per dirla alla buona, in un'n-upla tipica probabilità e frequenze relative di presenza delle singole lettere sono pressocchè uguali.
> Il teorema che segue dice che *quasi sicuramente* in un'n-upla lunga frequenze relative di presenza e probabilità sono *quasi uguali* (la tipicità dell'uscita è insomma quasi sicura). Questo risultato cardine del calcolo delle probabilità fu dimostrato da Bernoulli, e viene invocato a sostegno della loro ideologia dai partigiani della probabilità *statistica* (frequentista, oggettiva), secondo i quali la probabilità è una frequenza di successo asintotica.
> Di esso daremo due dimostrazioni; la prima è più "professionale"; la seconda non è la più pregnante sul mercato, ma in cambio è "teorico-informazionale"; il lettore potrà notare che in essa non abbiamo usato la disuguaglianza di Čebyšev e neppure la nozione di varianza.

Teorema (legge dei grandi numeri). Se n diverge, tende a 1 la probabilità P^n delle n-uple tipiche.
Dimostrazione. Se X è l'n-upla aleatoria emessa dalla sorgente, fissa i e applica la disuguaglianza di Čebyšev al numero aleatorio delle presenze di a_i in X, $N(a_i|X)$, che è *binomiale* di speranza matematica np_i e varianza $np_i(1-p_i) \leq \frac{n}{4}$. Usa la limitazione dell'unione (*union bound*), che dà:

$$P^n\{X \text{non tipica}\} \leq \frac{K}{4n\epsilon_n^2} \to 0$$

La legge dei grandi numeri come l'abbiamo enunciata è chiamata "debole" dai probabilisti. Il lettore non rimpianga la legge forte, che si occupa solo di valori al limite (in questo caso la probabilità limite = 1 delle sequenze tipiche). Ciò che interessa è soprattutto la limitazione uniforme della probabilità di errore: essa può essere usata non solo per passare al limite, in quella regione asintotica del mondo in cui "tutte le vacche sono ugualmente grigie", ma anche, e più concretamente, per n "molto grande".
Vedremo ora che le sequenze tipiche, a meno che P non sia uniforme, in percentuale sono davvero poche.

Teorema ("completamento aritmetico" della legge dei grandi numeri). Il tasso dell'insieme di tutte le sequenze tipiche rispetto a P tende all'entropia $H(P)$.

Dimostrazione: Se T è un tipo "tipico" (di n-uple tipiche di composizione $nF \approx nP$) il suo tasso, a meno di un infinitesimo, è $H(F)$, che a sua volta differisce da $H(P)$ per un'infinitesimo, l'entropia essendo una funzione continua. La numerosità delle tipiche è la somma delle numerosità dei "tipi tipici", ossia è la somma di un "piccolo" numero di addendi esponenziali della forma $2^{n[H(P)+\cdots]}$, dove i puntini indicano una correzione infinitesima. Lasciamo al lettore i dettagli della dimostrazione; l'analista numerico, usando il lemma sulla continuità uniforme dell'entropia, potrà perfino dare limitazioni uniformi agli infinitesimi coinvolti. □

Corollario. Se P non è uniforme il rapporto fra la numerosità delle sequenze tipiche e quella di tutte le sequenze tende almeno esponenzialmente a 0.

Esercizio: Sbirciando se è il caso in un testo elementare di probabilità, il lettore ridimostri la disuguaglianza di Čebyšev e ricalcoli la media (la speranza matematica, il valore atteso) e la varianza della distribuzione binomiale.

4. Codici a blocco con segnale d'allarme

I codici con elementi a lunghezza variabile, a causa dei problemi di sincronizzazione, introducono di fatto errori nello schema di codifica/decodifica. La sincronizzazione fra il flusso di lettere primarie che entrano nel codificatore e il flusso di simboli secondari che ne escono viene resa possibile da una *memoria tampone*, o *buffer*, entro cui vengono messe in attesa le lettere in entrata quando il codificatore è troppo impegnato a causa dell'addensarsi di informazioni improbabili (ma non impossibili). Tuttavia, se si è molto sfortunati, il buffer trabocca ed è costretto a tagliare l'informazione in arrivo. I codici a blocco non hanno problemi di sincronizzazione, ma non possono comprimere i dati: per ragioni di equità in questa sezione concederemo anche a loro una probabilità d'errore "trascurabile", allentando il criterio di attendibilità fin qui imposto.

Esercizio. Codifica le n-uple di una sorgente bernoulliana con parole di codice distinte di lunghezza fissa $l = l_n$. Dimostra che, in queste condizioni, il tasso del codice ottimo tende a $\log_\beta K$. (Le parole devono essere distinte perchè il codice sia UD; allora $\beta^l = \beta^{l_n} \geq K^n$, e dunque il tasso ottimo è $R_n = \frac{\lceil n \log_\beta K \rceil}{n}$.)

Dunque i codici B-B (*codici a blocco*) non consentono la compressione dei dati, ma solo la loro mera *trascrizione*. Se tuttavia allentiamo il criterio di attendibilità che ci siamo imposti, la compressione è possibile; rifà capolino l'entropia.

Codifica. Numera le n-uple tipiche. La parola di codice di un'n-upla tipica è il suo numero d'ordine trascritto in binario aggiungendo se serve zeri in testa; la parola di codice comune a tutte le n-uple non tipiche è il blocco di soli zero. La lunghezza delle parole di codice (che sono blocchi binari) è $\lceil \log(\tau_n + 1) \rceil$, dove τ_n è la numerosità delle n-uple tipiche.

Decodifica. Inverti la codifica se la parola di codice non è il blocco nullo; altrimenti fa' scattare un segnale di allarme.

Prestazioni. La probabilità d'errore (o meglio d'allarme, di mancata decodifica) tende a 0, il tasso del codice tende a $H(P)$ ($\Box$!).

Nei termini del capitolo II il decodificatore proposto è *flessibile* e l'errore è *rilevato* (*detected error*).
Si può dimostrare che, anche con il nuovo criterio di attendibilità allentato, non è possibile "battere l'entropia", in accordo con quanto affermato dalla "legge generale dell'entropia". Ciò è dimostrato nella parte finale di questa sezione, che è riservata ai lettori più impegnati.

Esercizio. Se $\underline{x}$ è una sequenza tipica di lunghezza n e p^- è la più piccola probabilità positiva in P:

$$\exp\{-n[H(P) - K\delta_n \log p^-]\} \leq P^n(\underline{x}) \leq \exp\{-n[H(P) + K\delta_n \log p^-]\}$$

Si rammenti che $P^n(\underline{x}) = \exp\{-nH(F\|P)\}$; stante la tipicità di $\underline{x}$, si ha $F \approx P$, o meglio $\sum_i |f_i - p_i| \leq K\delta_n$, per cui $|H(F\|P) - H(P)| \leq -K\delta_n \log p^-$.

Esercizio: disuguaglianza della somma logaritmica, o *log-sum inequality*. Dati i due K-vettori non negativi $\{a_1, a_2, \ldots, a_K\}$ e $\{b_1, b_2, \ldots, b_K\}$ a somma $\sum a_i \doteq A$ e $\sum b_i \doteq B$ si ha:

$$\sum_{1 \leq i \leq K} a_i \log \frac{a_i}{b_i} \geq A \log \frac{A}{B}$$

con uguaglianza sse i due vettori sono proporzionali. Cenno: "forza" le due K-uple a diventare vettori di probabilità dividendo le loro componenti per A e per B. Controlla a parte i casi $A = 0$, $B = 0$.

Teorema di codifica inverso. Se $\mathcal{C}_n$ è un insieme di n-uple la cui probabilità tende a 1 con n:

$$\liminf_n R_n \doteq \liminf_n \frac{\log |\mathcal{C}_n|}{n} \geq H(P)$$

Dimostrazione. Se $\mathcal{C}_n$ è un insieme di n-uple, tipiche o non tipiche, la cui probabilità tende a 1 con n, anche la probabilità di $\mathcal{C}_n \cap \mathcal{T}_n$ tende a 1 ($\Box$! $\mathcal{T}_n$ è l'insieme di tutte le

n-uple tipiche). Usiamo la disuguaglianza *log-sum* con il numero degli addendi s uguale alla numerosità di $\mathcal{C}_n \cap \mathcal{T}_n$, $q_i = \frac{1}{s}$, e $p_i = P^n(\underline{x}_i)$, $x_i \in \mathcal{C}_n \cap \mathcal{T}_n$:

$$\sum_{1 \leq i \leq s} P^n(\underline{x}_i) \log[|\mathcal{C}_n \cap \mathcal{T}_n| P^n(\underline{x}_i)] \geq P^n(\mathcal{C}_n \cap \mathcal{T}_n) \log P^n(\mathcal{C}_n \cap \mathcal{T}_n)$$

per cui, *a fortiori*:

$$\log |\mathcal{C}_n \cap \mathcal{T}_n| \geq -\sum_i P^n(\underline{x}_i) \log P^n(\underline{x}_i) - \tau_n$$

dove $\tau_n \doteq -P^n(\mathcal{C}_n \cap \mathcal{T}_n) \log P^n(\mathcal{C}_n \cap \mathcal{T}_n)$ tende a 0 con n. Confrontiamo la somma su $\underline{x}_i$ con $nH(P)$. Se $\underline{x}_i$ è tipica di lunghezza n, la limitazione sulla sua probabilità (cfr. l'esercizio) dà :

$$P^n(\underline{x}_i) \leq \exp\{-n[H(P) + K\delta_n \log p^-]\} \leq \exp\{-n[H(P) - \rho_n]\}$$

dove $\rho_n \doteq -K\delta_{n^-} \log p^-$ tende a 0 con n, per cui:

$$\log |\mathcal{C}_n \cap \mathcal{T}_n| \geq n[H(P) - \rho_n] P^n(\mathcal{C}_n \cap \mathcal{T}_n) - \tau_n \geq n[H(P) - \rho_n][1 - \epsilon_n] - \tau_n$$

dove anche ϵ_n tende a zero. Dopo aver diviso tutto per n, la parte destra tende a $H(P)$ al crescere di n, quella sinistra è superiormente limitata da $R_n = \frac{\log |\mathcal{C}_n|}{n}$. $\square$

Il teorema inverso appena dimostrato è un *inverso debole*, come usano dire i teorici dell'informazione. Vale anche l'inverso *forte*: nel suo enunciato la richiesta che la probabilità delle n-uple tenda a 1 viene allentata, e si richiede solo che essa sia "discosta da zero", anche se molto bassa. Si è dimostrato che neppure in questo caso non si riesce a battere asintoticamente l'entropia.

II. DIVERGENZA E DISCRIMINABILITÀ STATISTICA

5. Probabilità oggettive: discriminazione statistica a massima verosimiglianza

Teorema (comportamento asintotico del rapporto di verosimiglianza). Se P e Q sono due vettori di probabilità su $\mathcal{A}$ con $P \ll Q$, e se x è una sequenza tipica per la "legge casuale" P, il *rapporto di verosimiglianza* $\frac{P^n(x)}{Q^n(x)}$ tende esponenzialmente a $+\infty$.

Dimostrazione. Infatti:

$$\frac{P^n(x)}{Q^n(x)} = \frac{2^{-nH(F\|P)}}{2^{-nH(F\|Q)}} = 2^{n(D(F\|Q) - D(F\|P))}$$

dove nF è la composizione di x. Poichè l' n-upla x è P-tipica ($F \approx P$), si ha $F \ll P$ (le lettere di P-probabilità 0 non possono comparire in x), e anche, per la transitività della relazione $\ll$, $F \ll Q$. Dunque $D(F\|Q) - D(F\|P)$ è uguale a $D(P\|Q)$ a meno di un infinitesimo $\delta_n \to 0$:

$$\frac{P^n(x)}{Q^n(x)} = 2^{n[D(P\|Q) + \delta_n]}$$

Il teorema è alla base di una nota ricetta statistica: se devi *discriminare* fra un certo numero (finito) di vettori di probabilità, uno e uno solo dei quali descrive la "vera legge di probabilità " che regola il fenomeno aleatorio che stai studiando, esegui un "gran numero" di esperimenti e poi opta per il vettore di probabilità che dà all'ennupla delle osservazioni probabilità massima (premia la "spiegazione" di ciò che hai osservato che lo rende più probabile). La legge dei grandi numeri, infatti, garantisce o quasi la tipicità delle osservazioni rispetto al vettore di probabilità vero.

Se $P \not\ll Q$ il rapporto di verosimiglianza "salta" a $+\infty$; un valore così estremo è ben interpretabile, poichè in questo caso esiste una lettera possibile per la P ma impossibile per la Q: quando essa si presenterà (e la legge dei grandi numeri garantisce un tanto), lo sperimentatore scoprirà la verità senza neppure dover aspettare di raggiungere il "mondo asintotico".

Il metodo di massima verosimiglianza è descritto da Galileo, ma venne usato già dagli Arabi nella decifrazione delle scritture segrete: nel IX secolo il "filosofo degli Arabi" Al-Kindi compilò accurati *istogrammi*, come diremmo noi oggi, con le frequenza delle lettere arabe. E' un paradosso della storia della statistica che il contributo arabo alla nascita della statistica sia pressocché ignorato.

6. Discriminazione statistica a massima probabilità a posteriori

La probabilità *soggettiva*, o (neo-)*bayesiana*, misura su una scala che va da zero (= impossibilità) a uno (= certezza) il grado di fiducia *a priori* che lo sperimentatore ha nel verificarsi di un evento aleatorio. A dir il vero, non ha sempre senso parlare di "sperimentatore", perchè l'evento di cui si valuta la probabilità potrebbe non essere ripetibile (e dunque non essere suscettibile di verifica sperimentale). Le probabilità soggettive (bayesiane) verificano le stesse proprietà di calcolo delle probabilità oggettive (statistiche) dalle quali sono dunque *formalmente* indistinguibili. Preannunciamo tuttavia che nell'ultimo capitolo, dedicato alle probabilità non ortodosse, questa comoda parentela formale verrà contestata.

Esempio. Supponiamo che ci siano in circolazione due tipi di monete, quelle eque in cui testa e croce escono con la stessa frequenza, $P = (1/2, 1/2)$, e quelle truccate in cui tende a uscire testa, $Q = (3/4, 1/4)$. Abbiamo in mano una moneta di cui non sappiamo nulla, ammessso ciò sia possibile. La lanciamo $n = 4$ volte, esce tre volte testa e una volta croce. Il principio di massima verosimiglianza ci raccomanda di decidere: "la moneta è truccata, la vera distribuzione di probabilità è la Q"; tuttavia, stante il basso numero di esperimenti, la correttezza solo asintotica del principio è allarmante. A dire il vero la raccomandazione è del tutto assurda se sappiamo a priori che le monete eque sono molto più numerose di quelle truccate: il buon senso che ci avverte che anche dopo le tre teste è ragionevole "puntare" sull'equità della moneta.

Prima di eseguire gli esperimenti le due *ipotesi* P e Q potrebbero avere, come nell'esempio, una loro *probabilità a priori*

$$\eta_P = \text{Prob}\{P \text{ vera}\} \ , \ \eta_Q = \text{Prob}\{Q \text{ vera}\}$$

con $\eta_P + \eta_Q = 1$. Una volta eseguiti gli esperimenti aleatori $\underline{X}$ (il vettore aleatorio $\underline{X} = X_1 X_2 \ldots X_n$ rappresnta gli esiti aleatori degli n esperimenti, che nell'esempio si sono concretati nel vettore *osservato* $\underline{x}$, formato da tre teste e una croce, diciamo in quest'ordine), le probabilità a priori vanno *aggiornate* a due *probabilità a posteriori*

$$\eta_P(n) = \text{Prob}\{P \;\text{ vera } |\underline{X} = \underline{x}\} \;\; , \;\; \eta_Q(n) = \text{Prob}\{Q \;\text{ vera } |\underline{X} = \underline{x}\}$$

Il meccanismo di aggiornamento è la celebre *regola di Bayes*, proposta a metà Settecento, la quale recita:
la probabilità a posteriori di un'ipotesi è proporzionale alla sua probabilità a priori $\times$ *la verosimiglianza delle osservazioni sotto quell'ipotesi*
In formule:

$$\text{Prob}\{P \;\text{ vera } |\underline{X} = \underline{x}\} \;\propto \eta_P P^n(\underline{x}) \;\; , \;\; \text{Prob}\{Q \;\text{ vera } |\underline{X} = \underline{x}\} \;\propto \eta_Q Q^n(\underline{x})$$

La costante di proporzionalità si determina subito tenendo conto che anche le probabilità a posteriori devono sommare a uno. La regola di Bayes è dimostrata a fine sezione; per inciso: la generalizzazione a tre o più ipotesi è immediata. Il principio di discriminazione statistica a massima probabilità a posteriori è di saggezza palese, poiché raccomanda di scegliere l'ipotesi cui compete la probabilità a posteriori massima. Esso è dunque basato sul confronto con 1 del rapporto

$$\frac{\eta_P \; P^n(\underline{x})}{\eta_Q \; Q^n(\underline{x})}$$

e coincide col principio di massima verosimiglinza nel caso in cui le due probabilità a priori siano uguali. Come si vede, anche il nuovo principio è corretto asintoticanente, ma può dare risultati molto diversi dal principio di massima verosimiglianza su un basso numero di esperimenti, nel caso in cui le probabilità a priori siano molto asimmetriche. Le probabilità a priori hanno carattere *soggettivo*, nel senso che sono basate su un certo *stato di conoscenze*, e variano al variare di tale stato, ad esempio perché si sono eseguiti degli esperimenti. Nulla vieta che anche le conoscenze a priori abbiano un carattere empirico, e dunque siano in questo senso "oggettive" ("soggettivo" non vuol dire affatto "arbitrario"). Nell'esempio le due probabilità a priori potrebbero ben essere uguali alle percentuali empiricamente accertate delle monete eque e false che si sanno essere in circolazione.

Esercizio. Siano $E_1, E_2, \ldots, E_K$ eventi incompatibili ed esaurienti le cui K probabilità $\text{Prob}(E_i)$ sommano a uno (nel nostro caso si tratta delle $K = 2$ probabilità a priori η_P ed η_Q). Sia F un ulteriore evento (di probabilità positiva; nel nostro caso è l'evento $\underline{X} = \underline{x}$). Allora le probabilità condizionate $\text{Prob}(E_i|F)$ sono proporzionali al prodotto $\text{Prob}(E_i)\text{Prob}(F|E_i)$ (la costante di proporzionalità non dipende da i). Cenno: sviluppa in due modi le probabilità congiunte $\text{Prob}(E_i \wedge F)$.

III. LA MISURA DELL'INCERTEZZA

7. Gli assiomi dell'entropia e l'incertezza statistica

Una misura di incertezza *adeguata* deve soddisfare a certe proprietà (a certi *desiderata*) che noi riteniamo "imprescindibili": gli *assiomi*, si usa dire un po' a sproposito. Se siamo molto fortunati il sistema degli assiomi ha un'unica soluzione. Di seguito proporremo un sistema di assiomi che individua univocamente l'entropia di Shannon come misura adeguata dell'incertezza statistica (fra i sistemi di assiomi disponibili nella letteratura spiccano quelli di Hinčin e quelli di Faddev, degli anni 50). Poichè il nostro interesse si accentra sui lati statistici dell'incertezza, presupporremo un "assioma tacito" su cui dovremo ritornare, perchè è molto impegnativo: l'incertezza dipende dalle probabilità e *da nient'altro*!

Per ogni $K \geq 1$ e per ogni vettore di probabilità $P = (p_1, p_2, \ldots, p_K)$ a K componenti cerchiamo una funzione $\Xi_K(P)$ (la nostra *incognita funzionale*) tale da soddisfare gli:

Assiomi di adeguatezza:

1. $\Xi_K(P)$ è una funzione continua di P (continuità)

2. $\Xi_K(P)$ è una funzione simmetrica di P (simmetria)

3. $\Xi_K(P) = \Xi_{K+1}(P, 0)$ (espansibilità)

4. $\Xi_2(\frac{1}{2}, \frac{1}{2}) = 1$ (normalizzazione)

5. $\Xi_K(P) \leq \Xi_K(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$ (massima incertezza)

6. $\Xi_{K\times H}(p_1 q_1, \ldots, p_i q_j, \ldots, p_K q_H) = \Xi_K(p_1, \ldots, p_K) + \Xi_H(q_1, \ldots, q_H)$ (additività sotto indipendenza; $H \geq 1$)

7. $\Xi_K(p_1, \ldots, p_H, p_{H+1}, \ldots, p_K) =$
   $= \Xi_{K-H+1}(p_H^*, p_{H+1}, \ldots, p_K) + p_H^* \Xi_H(\frac{p_1}{p_H^*}, \ldots, \frac{p_H}{p_H^*})$
   (assioma dei due passi o di diramazione; si è posto $p_H^* = p_1 + \ldots + p_H$; $K \geq H \geq 1$)

Commentiamo gli assiomi:

1. *Continuità*: variando "di poco" P varia "di poco" la sua incertezza.

2. *Simmetria*: l'incertezza è insensibile al "nome" degli esiti elementari.

3. *Espansibilità*: l'incertezza è insensibile all'aggiunta e/o all'eliminazione di eventi comunque impossibili.

4. *Normalizzazione*: è una banale scelta dell'unità di misura; chi non predilige i *bit* può senza sforzo adattare l'assioma ai propri gusti.

5. *Incertezza massima*: l'incertezza è massima di fronte all'equiprobabilità , quando nessun esito è favorito, neppure su un piano meramente statistico, e non si sa "che pesci pigliare"

6. *Additività*: l'incertezza sul risultato di due esperimenti indipendenti (che non c'entrano per niente l'uno con l'altro) è uguale alla *somma* delle incertezza sui singoli esperimenti.

7. *Assioma dei due passi* (*fusione*, o, visto a rovescio, *diramazione*).
L'incertezza sull'esperimento viene rimossa in due passi: in un primo tempo si è bonari e i primi H esiti vengono considerati indistinguibili; il secondo passo, in cui si rimuove l'incertezza residua, è *condizionato* e si presenta con probabilità p_H^* (le probablità $\frac{p_1}{p_H^*}, \ldots, \frac{p_H}{p_H^*}$ sono appunto le probabilità condizionate quando l'uscita è una delle lettere fuse).

Teorema. L'unica soluzione dei sette assiomi di adeguatezza è l'entropia di Shannon $H(P)$ misurata in bit.

Dimostrazione. Che l'entropia di Shannon verifichi gli assiomi di adeguatezza è ovvio. Supponiamo viceversa di aver in mano una soluzione $\Xi_K(P)$, e dimostriamo che essa coincide con l'entropia di Shannon. Procederemo per passi. Osserviamo subito che la simmetria e l'assioma degli eventi impossibili ci consentono di aggiungere e togliere zeri a piacimento; sempre grazie alla simmetria non occorre fondere proprio nelle posizioni iniziali. Poniamo $\xi_K = \Xi_K(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$; per l'ultimo assioma con $H = 1$ si ha $\xi_1 = 0$, com'è giusto; rammenta che $\xi_2 = 1$ per l'assioma di normalizzazione.

I. ξ_K è una successione in K non decrescente.
E infatti $\xi_K = \Xi_{K+1}(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}, 0) \leq \xi_{K+1}$ per gli assiomi 3 e 5.

II. $\xi_{K^n} = n\xi_K$ (è una proprietà di tipo logaritmico)
Per induzione su n. La tesi è vera per $n = 0$ e per $n = 1$. Assunto $\xi_{K^{n-1}} = (n-1)\xi_K$, usa l'additività pensando alla scomposizione $K^n = K^{n-1} \times K$ (le probabilità coinvolte sono tutte uniformi).

III. $\xi_K = \log K$
Sia $n \geq 1$ sia $t = \lfloor n \log K \rfloor$, per cui $2^t \leq K^n < 2^{t+1}$. Sia la ξ che il logaritmo conservano la crescenza (la ξ forse in senso debole) e verificano la proprietà $f(r^s) = rf(s)$. Ma allora: $t \leq n\xi_K \leq t + 1$ (rammenta che $\xi_2 = 1$) e anche $t \leq n \log K < t + 1$. Dunque sia ξ_K che $\log K$ stanno nell'intervallo $[\frac{t}{n}, \frac{t+1}{n}]$ di ampiezza $\frac{1}{n}$ "arbitrariamente piccola". Non c'è altra via d'uscita: ξ_K e $\log K$ coincidono.

IV. $\Xi_2(P) = H(P)$
Per la continuità di $\Xi_K(P)$ basta dimostrare l'assunto per P a componenti razionali, $P = (\frac{r}{n}, \frac{n-r}{n})$. Applichiamo due volte l'assioma dei due passi al vettore di probabilità uniforme a n componenti, fondendo i primi r esiti e gli ultimi $n - r$ esiti. Si trova ciò che serve: $\xi_n = \Xi_2(P) + \frac{r}{n}\xi_r + \frac{n-r}{n}\xi_{n-r}$, ossia proprio $\Xi_2(P) = H(P)$.

V. $\Xi_K(P) = H(P)$ per ogni K; già sappiamo che ciò è vero per $K = 1, 2$. Usiamo un'induzione su K: basta fondere due esiti e applicare l'assioma dei due passi. □

Osservazione: L'uso dell'assioma dell'additività è stato piuttosto parco e si è limitato al caso delle probabilità uniformi.

Dunque, *se esiste* una misura adeguata dell'incertezza, questa è l'entropia di Shannon. Rimane il dubbio che l'entropia di Shannon violi altri assiomi che abbiamo dimenticato nella lista, ma che sono altrettanto cogenti di quelli di cui ci siamo ricordati; se così fosse non esisterebbe nessuna misura di incertezza come Dio comanda! Il dubbio viene reso più tormentoso dalle obiezioni che seguono. Esse serviranno a capir meglio che cosa abbiamo in mano, e che cosa non dobbiamo illuderci di avere.

Prima obiezione. Primo gioco: lanci una moneta equa, se esce testa vinci un euro, se esce croce lo perdi. Secondo gioco: lanci un rublo equo, se esce testa lo passi al giocatore successivo, se esce croce ... be', è una variante della *roulette* russa, sei spacciato.

Commento. In entrambi i casi la tua incertezza prima di lanciare la moneta è la stessa, di un bit. Prima di scandalizzarti, ricorda l'"assioma tacito": l'entropia di Shannon misura solo gli aspetti statistici dell'incertezza, e ignora gli altri aspetti (semantico, pragmatico) che pure nella vita reale sono essenziali. Per ribadire: l'entropia di Shannon è una misura dell'*incertezza statistica* (si è detto anche: dell'*incertezza selettiva*). Usare l'entropia di Shannon per valutare i lati non meramente statistici dell'incertezza è un abuso che pure gli entusiasti, pronti a confrontare l'entropia in bit di *T'amo o pio bove* con quella del *Cinque maggio* per capirne le virtù poetiche, hanno commesso.

Seconda obiezione. Sei un pignolo e hai in mano una moneta di cui non sai nulla, la tua incertezza è massima. In termini numerici essa è di *circa* un bit (essendo pignolo ti sei ricordato che c'è una probabilità piccolissima, ma non nulla, che la moneta non mostri nè testa nè croce, ma rimanga in bilico: gli esiti elementari sono dunque tre, testa, croce e in piedi). Ora su tre esiti elementari l'incertezza è resa massima da una moneta truffaldina che rimane in piedi con la probabilità di un terzo: una comportamento ben strano, che solo il baro che se ne serve conosce, perchè solo lui su quella moneta la sa lunga (*non* è ignorante).

Commento. Non scandalizzarti, c'è un equivoco in cui tanti sono caduti: una cosa è l'*incertezza*, altra cosa è l'*ignoranza*. L'incertezza sull'esito è davvero massima nel caso della moneta truffaldina. Nell'ultimo capitolo vedremo che per gestir bene l'ignoranza bisogna sfondare i confini delle probabilità "ortodosse".

8. Due principi: massima entropia e minima entropia.

Abbiamo già argomentato che l'entropia è una misura adeguata di incertezza, mentre *non* è una misura adeguata di ignoranza: la confusione fra incertezza e ignoranza è alla base di una serie di fraintendimenti “storici” nell’uso del principio di massima entropia, cui sono dedicate le due sezioni che seguono.

Bernoulli e poi Laplace caldeggiarono l’uso del seguente principio di scelta delle probabilità a priori: *in caso di totale ignoranza scegli la probabilità uniforme*, ossia, in termini equivalenti: *in caso di totale ignoranza scegli la probabilità a entropia massima* (principio *di indifferenza*, o *della ragione insufficiente*). Il secondo enunciato che coinvolge l’entropia è filologicamente incongruo, ma ha il vantaggio che regge anche quando sono noti dei vincoli sulla probabilità da scegliere che mettono fuori gioco quella uniforme (a dire il vero regge bene solo se il massimo dell’entropia è unico, ciò che comunque succede in almeno un caso tanto importante quanto fortunato, come vedremo nel capitolo che segue). Visto così, il principio di massima entropia è epistemologicamente fragile: e infatti render massima l’entropia porta a render massima l’incertezza e non l’ignoranza. I successi pratici del principio nella termodinamica sono legati a un’altra “faccia” dell’entropia, quella aritmetico-asintotica; ciò che si massimizza è un coefficiente multinomiale (= numero dei microstati che danno origine al medesimo macrostato), poi approssimato con l’entropia, piuttosto che una misura di informazione. Sia come sia, il principio è spesso visto come la scelta “meno impegnativa” di un vettore di probabilità all’interno di una famiglia di vettori ammissibili.

Molto meno popolare è il principio di minima entropia, che raccomanda, fra i “modelli probabilistici” disponibili, di scegliere quello cui compete l’incertezza minima; lo si è fatto risalire a una massima di Lao Tzu del sesto secolo a.C.:

conoscere l’ignoranza è forza
ignorare la conoscenza è debolezza

Se si vuole, il principio di massima entropia è un principio di massima modestia, mentre quello di minima entropia è un principio di massima arroganza.

9. Entropia massima a media vincolata.

Nelle sezioni 2 e 3 come unità di misura sceglieremo i *nat* (la base “naturale” e).

Sia data una funzione f che alle lettere dell’alfabeto $\mathcal{A}$ associa dei numeri, $f : \mathcal{A} \to \Re$, $f_i \doteq f(a_i)$; per esempio f potrebbe essere la lunghezza delle parole di un codice di Huffman. Se X è la lettera aleatoria emessa dalla sorgente, la *variabile aleatoria* (il *numero aleatorio*) $f(X)$ ha un suo *valor medio* (una sua *speranza matematica*), ovviamente compresa fra $\min_i f_i$ e $\max_i f_i$:

$$Ef(X) = \sum_i p_i f_i$$

(invece di $Ef(X)$ scriveremo anche Ef, o perfino $E_P f$ per mettere in evidenza rispetto a quale probabilità si calcola la media). Dato un numero $\overline{f}$ compreso fra $\min_i f_i$ e $\max_i f_i$, di frequente si impone alla P, ossia alla distribuzione di probabilità della X, il vincolo $Ef(X) = \overline{f}$; ci porremo il problema di render massima l’entropia di P sotto quest’unico vincolo (e poi di renderla minima nel capitolo seguente). I due esercizi seguenti mostrano che il problema è interessante solo per $K \geq 3$ e solo se $\overline{f}$ è compreso strettamente fra $\min_i f_i$ e $\max_i f_i$.

Esercizio 1. Sia $K = 2$, $f_2 > f_1$, $\overline{f} \in [f_1, f_2]$. L'*unica* P che verifica il vincolo è la $p_1 = \frac{f_2 - \overline{f}}{f_2 - f_1}$, $p_2 = \frac{\overline{f} - f_1}{f_2 - f_1}$, la cui entropia vale $h(\frac{\overline{f} - f_1}{f_2 - f_1})$.

Esercizio 2. Se $\overline{f} = \max f_i$ (o, rispettivamente, $\overline{f} = \min f_i$) e le lettere a_i per cui $f(a_i) = \overline{f}$ sono in tutto s, l'entropia massima vale $\ln s$ ed è raggiunta in corrispondenza a una P uniforme su quelle lettere ($p_i = \frac{1}{s}$) e nulla altrove. Se $\min f_i = \max f_i = \overline{f}$ il vincolo è "vuoto" (è verificato da tutte le P).

Quando il problema è invece interessante, ci soccorre ancora una volta la non negatività della divergenza $D(P\|Q)$ (la disuguaglianza di Gibbs). "Inventiamo" una nuova probabilità Q ponendo $q_i \propto e^{-\beta f_i}$, dove β è una costante qualunque, che più avanti sceglieremo con astuzia; il denominatore delle probabilità q_i vale $\sigma \equiv \sigma_\beta = \sum_i e^{-\beta f_i}$. Per inciso: una distribuzione di probabilità le cui componenti sono proporzionali a $e^{-\beta f_i}$ si chiama la *distribuzione canonica* (o *distribuzione di Gibbs*) di parametro β. Si ha

$$\sum_i p_i \ln \frac{p_i}{e^{-\beta f_i}} \geq -\ln \sigma$$

ossia:

Lemma. $\forall \beta$, $\forall P$, $H(P) \leq \ln \sigma + \beta E_P f$; l'uguaglianza vale sse $p_i \propto e^{-\beta f_i}$.

Vediamo di scegliere β in modo che il vettore P che realizza l'uguale nel lemma dia alla funzione f una speranza matematica uguale proprio al numero $\overline{f}$. Dev'essere:

$$E_P f = \sum_i \frac{e^{-\beta f_i}}{\sigma} f_i = \overline{f}$$

Il lemma che segue è un noioso intermezzo di calcolo, che mostra come questo desiderio sia realizzato da una e una sola scelta di β; il numero β^* si può trovare fino alla precisione desiderata con metodi standard di analisi numerica, basati sul dimezzamento progressivo, e dunque di velocità esponenziale.

Lemma. Se $\min_i f_i < \overline{f} < \max_i f_i$, l'equazione

$$\sum_i f_i e^{-\beta f_i} = \overline{f} \sum_i e^{-\beta f_i} \tag{$*$}$$

ha un'unica soluzione β^*.

Dimostrazione. Moltiplicati ambedue i membri per $e^{\beta \overline{f}}$, e posto $f_i - \overline{f} \doteq d_i$, l'equazione si può scrivere $\sum_i d_i e^{-\beta d_i} = 0$, ossia anche, chiamato $\phi(\beta)$ il primo membro, $\phi(\beta) = 0$. La derivata $\phi'(\beta)$ è negativa ($\phi'(\beta) = -\sum_i d_i^2 e^{-\beta d_i} < 0$; le d_i non sono tutte nulle), e dunque $\phi(\beta)$ è una funzione di β strettamente decrescente. Si ha $\phi(-\infty) = \sum_{d_i>0} d_i e^{+\infty} = +\infty$ e $\phi(+\infty) = \sum_{d_i<0} d_i e^{+\infty} = -\infty$ (la nostra notazione "compatta" due passaggi al limite; nota che esiste almeno una d_i positiva e almeno una negativa). □

Dunque, posto $\sigma^* = \sum_i e^{-\beta^* f_i}$ e $p_i^* = \frac{e^{-\beta^* f_i}}{\sigma}$, il vettore P^* di componenti p_i^* dà alla funzione f il valore atteso desiderato. Usiamo ora il lemma con $\beta = \beta^*$, e limitatamente alle P che soddisfano il vincolo sul valore atteso (fra esse c'è anche P^*). Ne viene:

Teorema. Se $\min_i f_i < \overline{f} < \max_i f_i$ e se β^* è l'unica soluzione dell'equazione $(*)$, l'entropia $H(P)$ sotto il vincolo $Ef(X) \equiv \sum_i f_i p_i = \overline{f}$ è resa massima dalla distribuzione canonica P^* di parametro β^*. Il valore dell'entropia massima è :

$$H(P^*) = \ln \sigma^* + \beta^* \, \overline{f}$$

dove $\sigma^* \doteq \sum_i e^{-\beta^* f_i}$.

(Nel calcolo di $H(P^*)$ rammenta che $\sum_i p_i^* f_i = \overline{f}$)

10. Entropia minima a media vincolata

Risolveremo il problema opposto a quello della sezione precedente, rendere *minima* l'entropia $H(P)$ sotto il vincolo $Ef(X) = \overline{f}$. Il punto chiave è che l'entropia $H(P)$ è una funzione *strettamente concava* di P; ne viene subito che:

Se $[P, Q]$ è il segmento che unisce due vettori di probabilità distinti, l'entropia $H(R)$ sotto il vincolo $R \in [P, Q]$ *non* raggiunge il valore minimo all'interno del segmento, ma su uno o entrambi dei suoi estremi (□ !).

Nel seguito supporremo che la funzione f assuma K valori tutti distinti; che ciò non sia restrittivo è assicurato dall'esercizio che segue:

Esercizio. Sia $\mathcal{I}$ un insieme di due o più lettere di P-probabilità positiva su cui la f è costante; sia P^* un vettore di probabilità ottenuto da P assegnando tutta la probabilità $P(\mathcal{I})$ a un'unica lettera di $\mathcal{I}$ (le altre lettere di $\mathcal{I}$ hanno dunque P^*-probabilità nulla). Allora $E_P f = E_{P^*} f$ e $H(P^*) < H(P)$. (Usa la proprietà di diramazione dell'entropia: ci si riconduce al problema di render minima l'entropia su $\mathcal{I}$ *senza alcun vincolo.*)

Grazie a quest'osservazione, se la f ha valori ripetuti è sufficiente risolvere il problema di minimo sull'alfabeto ingrossato ottenuto fondendo fra di loro le lettere su cui la f è costante; se P^* risolve il problema di minimo per l'alfabeto ingrossato, basta assegnare la probabilità di ciascuna lettera "grossa" a una qualunque delle lettere "magre" che la compongono. Ancora più ovvia è la ragione per cui possiamo escludere che il valor medio $\overline{f}$ venga effettivamente assunto dalla funzione f:

Esercizio. Se esiste una lettera a_i per la quale $f(a_i) = \overline{f}$, l'entropia vincolata minima vale 0 ed è raggiunta per P deterministica su a_i.

Assumeremo dunque, senza restrizione effettiva, che la f non abbia valori ripetuti e che non assuma il valore $\overline{f}$. Il lemma che segue scende appunto dalla concavità stretta dell'entropia; useremo anche l'esercizio 1 della sezione precedente.

Lemma. Se P è un vettore di probabilità con tre o più componenti positive, esiste un vettore di probabiltà P^* con *due* componenti positive tale che $E_{P^*}f = E_P f$ e $H(P^*) < H(P)$.

Dimostrazione. Sia $\mathcal{I}$ l'insieme delle lettere di P-probabilità positiva. Devono esistere due lettere a_j e a_i in $\mathcal{I}$ tali che $f_i > \overline{f} > f_j$ ($\square$!). Costruiamo un nuovo vettore di probabilità P^* (distinto da P) come nell'esercizio 1. A somiglianza di P e P^*, *tutti* i punti della *retta* che unisce P e P^* verificano le uguaglianze $\sum_i x_i = 1$, $\sum_i x_i f_i = \overline{f}$, $x_i = 0$ se $a_i \notin \mathcal{I}$ ($\square$!). La retta interseca il poliedro compatto e convesso dei vettori di probabilità in un segmento $[R, S]$; di là del segmento una delle componenti x_i con $a_i \in \mathcal{I}$ deve diventare negativa, per cui si azzera sul bordo: in altre parole R e S hanno al più $|\mathcal{I}| - 1$ componenti positive. Per la stretta convessità dell'entropia o R o S (o entrambi) hanno entropia $< H(P)$. $\square$

> Di fatto abbiamo dimostrato una proprietà molto più generale, e cioè che una funzione strettamente concava definita su un *poliedro* convesso nello spazio euclideo multidimensionale assume il valore minimo su uno o più vertici del poliedro; nel nostro caso il poliedro convesso è ottenuto intersecando quello di tutti i vettori di probabilità con l'*iperpiano* dei vettori reali su cui è verificato il vincolo.

Il lemma dimostra che la soluzione del problema di minimo ha esattamente *due* componenti positive. Grazie all'esercizio 1 della sezione precedente il problema si riduce al confronto di entropie binarie del tipo $h(\frac{f_i - \overline{f}}{f_i - f_j}) = h(\frac{\overline{f} - f_j}{f_i - f_j})$, $f_i > \overline{f} > f_j$.

Teorema. Sia f una funzione che non assume valori ripetuti e che non assume il valore $\overline{f}$. Sotto il vincolo $Ef = \overline{f}$, l'entropia è minima in corrispondenza a una distribuzione di probabilità concentrata su una lettera per cui $f_i > \overline{f}$ di probabilità $\frac{f_i - \overline{f}}{f_i - f_j}$ e su una lettera per cui $\overline{f} > f_j$ di probabilità $\frac{\overline{f} - f_j}{f_i - f_j}$; l'entropia minima vale:

$$\min_{f_i > \overline{f} > f_j} h\Big(\frac{f_i - \overline{f}}{f_i - f_j}\Big)$$