ENTROPIA E INFORMAZIONE

Strumenti e idee teorico-informazionali nella compressione dei dati, nella crittografia e nella gestione dell'incertezza

Andrea Sgarro

1. **L'entropia di Shannon**

2. **L'entropia nella compressione dei dati**

3. **L'entropia nella statistica oggettiva e soggettiva**

4. **L'entropia e l'equivocazione nella segretezza e nell'autenticazione**

5. **Le entropie eterodosse nella gestione dell'incertezza**

ENTROPY AND INFORMATION

Information-theoretic tools and ideas in data compression, cryptography and uncertainty management

1. **Shannon entropy**

2. **Entropy in data compression**

3. **Entropy in objective and subjective statistics**

4. **Entropy and equivocation in secrecy and authentication**

5. **Unorthodox entropies in uncertainty management**

Capitolo 1

L'ENTROPIA DI SHANNON

Il concetto di *entropia* ha un fascino irresistibile: esso interviene nei contesti più disparati, termodinamica, ecologia, informazione, talvolta con tinte catastrofiche (morte dell'universo, caos informazionale). Non è neppure chiaro *quante* siano le entropie, se una o due (entropia termodinamica, entropia informazionale). Protagonista di questo libro è l'entropia informazionale di Claude Shannon (1916-2001).

1. Entropia e probabilità

Sia pure in sordina - in sordina nel senso che la *semantica* dell'entropia, il suo significato, ancora non si vedrà - entriamo subito nel vivo, e introduciamo l'*entropia di Shannon*. Diamo per scontato che il lettore sappia, almeno per sommi capi, che cos'è la probabilità, cui ad ogni buon conto abbiamo dedicato l'ultima sezione del capitolo.

L'*universo del* nostro *discorso* è molto semplicemente un insieme finito $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$ di K oggetti, che spesso troveremo suggestivo chiamare *lettere* o *simboli* dell'*alfabeto* $\mathcal{A}$; salvo specifica contraria $K \geq 2$. Ogni lettera a_i ha la probabilità $p_i = \text{Prob}\{a_i\}$ di venir scelta; $\forall i \; p_i \geq 0$, $\sum_i p_i = 1$, per cui $p_i \leq 1$.

Nel seguito specifiche come la $\forall i$ verranno di norma sottintese quando sono ovvie nel contesto; sommatorie tacite si intendono estese a tutti i valori dell'indice; qui l'indice va da 1 a K. Un vettore di probabilità basta ad individuare una distribuzione (una legge) di probabilità su uno spazio finito; è per questo che i termini vettore e distribuzione verranno usati in maniera interscambiabile.

Variare il *vettore di probabilità* P, che è un vettore reale a K componenti, significa variare la legge probabilistica che regola la scelta aleatoria delle lettere di $\mathcal{A}$. L'insieme di *tutti* i vettori di probabilità (di tutte le leggi aleatorie possibili su $\mathcal{A}$) è il dominio di definizione dell'entropia, il suo *ambiente*. Si tratta di un sottospazio *compatto* e *convesso* di $\Re^K$, che per $K = 2$ è un segmento e per $K = 3$ è un triangolo (□ ! Compatto, lo rammentiamo, significa chiuso e limitato; per la convessità: al variare del parametro λ fra 0 e 1 il "punto" $\lambda P + \overline{\lambda} Q$ percorre il segmento di estremi P e Q ed è esso stesso un vettore di probabilità).

Sia P un vettore di probabilità su $\mathcal{A}$; si definisce la sua *entropia* come:

$$H(P) \doteq -\sum_i p_i \log_2 p_i \equiv \sum_i p_i \log_2 \frac{1}{p_i}$$

con la convenzione di continuità $0 \log_2 0 = 0$ (e infatti $\lim_{x \to 0^+} x \log x = 0$). D'ora in poi ometteremo l'indicazione della base 2 dei logaritmi (sulla scelta della base ritorneremo fra poco; certo si è che in un contesto così spesso binario i logaritmi binari si impongono da sé). In particolare, per $K = 2$, $\mathcal{A} = \{0, 1\}$, $P = (1 - x, x)$, $0 \leq x \leq 1$:

$$H(P) = h(x) = -x \log_2 x - (1 - x) \log_2 (1 - x)$$

La funzione $h(x)$ si chiama anche *funzione entropia*; essa è simmetrica, concava, ha pendenza infinita in 0 e 1, dove dunque non è derivabile. Inoltre: $0 \leq h(x) \leq 1$; i minimi corrispondono alle due probabilità *deterministiche*, o *degeneri*, il massimo alla probabilità *uniforme* (□ ! disegna il grafico).

In generale:

$$0 \leq H(P) \leq \log K$$

con $H(P) = 0$ sse P è deterministica, $H(P) = \log K$ sse P è uniforme (le dimostrazioni sono relegate agli esercizi in calce); $H(P)$ è una funzione concava di P (la dimostrazione è rinviata a quando parleremo di *entropie condizionate*).

Esercizio. Se $K = 3$ l'entropia ricorda un *wigwam* concavo che i pellerossa hanno attaccato a terra nei tre vertici del triangolo che fa da pavimento; il wigwam ha tre porte verticali spalancate a forma di funzione entropia, ciò che fa un gran giro d'aria. Fa il disegno ricordando che in questo caso l'entropia si puo scrivere come una funzione $h(x, y)$ di *due* variabili libere.

Entropia e funzione entropica sono nomi impegnativi, di sapore termodinamico, che a suo tempo cercheremo di giustificare. Consigliamo fin d'ora, per i vantaggi euristici che comporta, di vedere l'entropia di P come una *misura dell'incertezza* associata all'esito dell'esperimento aleatorio X di legge probabilistica P che lo sperimentatore sta per effettuare. Il lettore apprezzerà il fatto che sotto questa luce le proprietà già dimostrate acquistano un significato perspicuo. In particolare è giusto che l'incertezza sia nulla in corrispondenza a situazioni in cui il risultato dell'esperimento è noto deterministicamente già a priori, e sia massima quando gli esiti a priori possibili sono equiprobabili, e non si sa davvero "che pesci pigliare".

Claude Shannon pubblicò nel 1948 un articolo, *A mathematical theory of communication*, che fonda la *teoria dell'informazione*, e prefigura la teoria dei codici. La lettera H (acca) era in origine una η greca maiuscola, iniziale di Entropia.

Esercizio. Se l'esperimento consiste nel lancio di una moneta equa, $H(P) = 1$; se invece consiste nel lancio di n monete eque $H(P) = n$ (a priori ci sono 2^n esiti possibili ed equiprobabili).

L'esercizio precedente dà una chiara idea di quale sia l'ordine di grandezza della nostra *unità di misura* per l'incertezza: essa si chiama *bit* (da *binary unit*). Se nella formula dell'entropia si usano i logaritmi in base $\beta > 1$, si dice che l'entropia è misurata in β-it, in particolare in *nat* (= *natural units*) nel caso dei logaritmi naturali.

Esercizio. Se $H(P)$ e $H_\beta(P)$ sono l'entropia di P in bit e in β-it, $H(P) = \alpha H_\beta(P)$, con α costante positiva indipendente da P (calcolala!). E' la tipica legge di trasformazione che si ha quando si cambia l'unità di misura.

Esercizio. Dimostra la disuguaglianza $H(P) \geq 0$, e stabilisci le condizioni di uguaglianza. Cenno: una somma di K addendi non negativi è nulla sse lo sono tutti i K addendi.

Esercizio-lemma 1. Dimostra la disuguaglianza $\ln x \leq x - 1$, con uguaglianza sse $x = 1$ (il logaritmo naturale è una funzione strettamente concava che sta sotto la sua tangente nel punto di ascissa 1 e la tocca solo nel punto di tangenza).

Esercizio-lemma 2. Dimostra la disuguaglianza $\ln x \geq 1 - \frac{1}{x}$, con uguaglianza sse $x = 1$ (nel lemma precedente metti $\frac{1}{x}$ al posto di x).

Esercizio. Dimostra la disuguaglianza $H(P) \leq \log K$, e stabilisci le condizioni di uguaglianza. Cenno: pensando dapprima all'entropia misurata in nat, applica K volte il lemma 2 ai K fattori di $\ln K - H_e(P) = \sum p_i \ln(Kp_i)$.

Complementi

Partiamo dalla *formula di diramazione*, che consente di ridurre il calcolo dell'entropia di un vettore a K componenti al calcolo delle entropie di due vettori con meno componenti.

Se A è un sottinsieme proprio di $\mathcal{A}$ di probabilità P positiva (almeno una lettera a di A ha probabilità positiva; $A \subset \mathcal{A}$), possiamo costruire un nuovo vettore di probabilità P_A sull'alfabeto *parziale* A ponendo $P_A(a) = P(a)/P(A)$ (e $P_A(a) = 0$ per $a \notin A$).

Il lettore avrà riconosciuto in P_A una *probabilità condizionata*, quando si ipotizza il verificarsi dell'evento A. L'agnizione è essenziale, ma per il momento i nostri interessi vanno solo all'uso delle tabelle.

Oltre all'alfabeto parziale, ci serviremo dell'alfabeto *ingrossato* $\mathcal{A}^*$ di $K - |A| + 1$ lettere, in cui le $|A|$ lettere di A vengono considerate "indistinguibili" (vengono "fuse" in un'unica "letterona", la A appunto); anche la probabilità P può venir ingrossata, ponendo $P^*(a) = P(a)$ per $a \notin A$, $P^*(A) = P(A) = \sum_{a \in A} P(a)$.

$$\textit{Formula di diramazione}: H(P) = H(P^*) + P(A)H(P_A)$$

La formula è un'immediata conseguenza delle proprietà elementari dei logaritmi (□ !) e consente di calcolare l'entropia di un vettore a K componenti usando $K - 1$ volte la tabella 2. Il lettore rifletta al significato euristico della formula in termini di incertezze.

Esercizio 1. Tabula i valori di $-x \log_2 x$ da $x = 0$ a $x = 1$ con passo Δ.

Esercizio 2. Tabula i valori di $h(x)$ da $x = 0$ a $x = \frac{1}{2}$ con passo Δ.

L'entropia $H(P)$ è una funzione continua di P definita su uno spazio compatto, e dunque è *uniformemente* continua. A soddisfazione degli analisti numerici, l'uniformità viene precisata e concretizzata nel risultato che segue; la limitazione superiore a destra della disuguaglianza non dipende né da P né da Q e tende a 0 insieme con la distanza fra P e Q.

Faremo riferimento alla distanza di Manhattan $d_M(P,Q) = \sum_i |p_i - q_i|$ (o distanza del taxi; a Trieste diremmo distanza del Borgo Teresiano: le ragioni del nome sono chiare non appena si faccia uno schizzo per $K = 2$). In rapporto alla comune distanza euclidea fra vettori di probabilità $d_E(P,Q) = \sqrt{\sum_i (p_i - q_i)^2}$ si ha $d_E \leq d_M \leq K d_E$, per cui le due *metriche* $d_E(P,Q)$ e $d_M(P,Q)$ definiscono la stessa *topologia* (□ ! cenni: da una parte $d_E^2 \leq d_M^2$, dall'altra $d_E \geq \max_i |p_i - q_i| \geq \frac{d_M}{K}$).

Disuguaglianza: uniforme continuità dell'entropia. Se $d_M(P,Q) \leq \frac{1}{2}$, si ha

$$|H(P) - H(Q)| \leq -d_M(P,Q) \log \frac{d_M(P,Q)}{K}$$

Dimostrazione. Questa dimostrazione è "tecnica" e noiosa, per cui il lettore può comprare il risultato "a scatola chiusa". Partiamo da una premessa analitica che il lettore di buon gusto scorrerà di malanimo. Anzi, prima della premessa, per poi non interromperci, facciamo notare che $r \leq s \leq t \Rightarrow |s| \leq \max[-r, t]$ (□ ! distingui i due casi $s \geq 0$, $s \leq 0$). Se ora $f(x)$ è una funzione concava del numero reale x, e se $[a, a + \Delta]$ e $[x, x + \Delta]$, con $a \leq x$ e $\Delta \geq 0$, sono due intervalli dell'insieme (intervallo esso stesso, semiretta o retta) su cui la funzione f è definita, si ha $f(x + \Delta) - f(x) \leq f(a + \Delta) - f(a)$ (essenzialmente questa è la definizione di funzione concava). Passando a *tre* intervalli di ampiezza Δ e punti di partenza $a \leq x \leq b$, si trova $f(b+\Delta) - f(b) \leq f(x+\Delta) - f(x) \leq f(a+\Delta) - f(a)$. Dunque, per il valore assoluto $|f(x + \Delta) - f(x)|$:

$$|f(x + \Delta) - f(x)| \leq \max[f(a + \Delta) - f(a), f(b) - f(b + \Delta)]$$

Applichiamo la disuguaglianza appena ottenuta alla funzione concava $f(x) = -x \log x$ con $a = 0$, $b + \Delta = 1$ $\big($al solito $f(0) \doteq 0$ per continuità; calcola la derivata seconda per dimostrare la concavità di $f(x)\big)$. Notato che $f(0) = f(1) = 0$, la maggiorazione precedente si scrive:

$$|f(x + \Delta) - f(x)| \leq \max[f(\Delta), f(1 - \Delta)]$$

La lotta per il massimo è vinta da $f(\Delta)$ sull'intervallo $[0, \frac{1}{2}]$, da $f(1 - \Delta)$ sull'intervallo $[\frac{1}{2}, 1]$ (□ !; confronta i grafici delle funzioni $f(x)$ e $f(1 - x)$). Dunque, e ciò conclude la parte noiosa:

$$|f(x + \Delta) - f(x)| \leq f(\Delta)\ , \quad 0 \leq \Delta \leq \frac{1}{2}$$

Posto $\delta_i = |p_i - q_i|$, useremo questa disuguaglianza K volte; $\sum_i \delta_i \doteq d_M(P,Q) \equiv d_M \leq \frac{1}{2}$ garantisce $\delta_i \leq \frac{1}{2}$. Si ha: $|H(P) - H(Q)| \leq \sum_i |f(p_i) - f(q_i)|$, e dunque:

$$|H(P) - H(Q)| \leq \sum_i f(\delta_i) \equiv -\sum_i \delta_i \log \delta_i$$

Il membro di destra si può scrivere in maniera elegante, e in apparenza stravagante, dopo aver introdotto il vettore di probabilità (*sic* !) $R \doteq (\frac{\delta_1}{d_M}, \frac{\delta_2}{d_M}, \ldots, \frac{\delta_K}{d_M})$:

$$|H(P) - H(Q)| \leq d_M\big(H(R) - \log d_M\big) \leq d_M(\log K - \log d_M) \equiv -d_M \log \frac{d_M}{K}$$

L'ultima disuguaglianza è la limitazione superiore naturale di qualunque entropia, anche di un entropia "artificiosa" come $H(R)$. (A esser pignoli, nella dimostrazione dobbiamo supporre $d_M \neq 0$, ma se $d_M = 0$ non c'è nulla da dimostrare, visto che in questo caso $P = Q$.) QED

Un semplice studio di funzione e la $d_M \leq K d_E$ danno subito i due corollari:

Corollario 1. Se $d_M(P,Q) \leq \epsilon \leq \frac{1}{e} \approx 0,37$, si ha $|H(P) - H(Q)| \leq -\epsilon \log \frac{\epsilon}{K}$

Corollario 2 (distanza euclidea).
Se $d_E(P,Q) \leq \tau \leq \frac{1}{Ke}$, si ha $|H(P) - H(Q)| \leq -K\tau \log \tau$

2. Richiami sulla probabilità.

La parola probabilità si incontra in contesti che appaiono poco conciliabili: in effetti ci sono almeno *tre* probabilità (tre concezioni distinte della probabilità).

`Probabilità combinatoria, o classica.`
La probabilità che esca un numero pari sul dado che sto per lanciare è di tre sesti, cioè di un mezzo, perchè gli eventi possibili sono sei, per ragioni di simmetria sono equiprobabili, e quelli favorevoli sono tre. Il soggetto (lo sperimentatore, l'esperto) deve assumersi la responsabilità di saper distinguere i dadi costruiti "a regola d'arte" da quelli truccati.

`Probabilità empirica, o oggettiva, frequentista, statistica.`
La probabilità che il pezzo prodotto da questa macchina sia difettoso è del 3%, come ha mostrato la sperimentazione ripetuta e registrata un altissimo numero di volte. Si noti che non siamo di fronte al "valore esatto" della probabilità, ma solo a una sua *stima statistica*, sperabilmente accurata; il valore esatto è asintotico, raggiungibile solo dopo aver effettuato ... infiniti esperimenti. Di solito (con una certa *nonchalance* filosofica), la concezione empirica "divora" quella combinatoria: si suppone che nel caso di dadi costruiti "a regola d'arte", su cui si sono fatti esperimenti per centinaia d'anni in tutti gli angoli del globo, i valori limite possano considerarsi noti. Questa popolarissima "scappatoia filosofica" circola col nome pomposo di *legge empirica del caso.*

`Probabilità epistemica, o soggettiva.`
La probabilità che Israeliani e Palestinesi firmino l'accordo entro Pasqua è del 60%. In quest'ultimo caso non ci sono simmetrie né c'è la possibilità di ripetere gli esperimenti, non esistendo nessuna fantascientifica macchina del tempo che ci faccia ritornare centinaia di volte a Natale, diciamo, per ricontrollare ogni volta come vada poi a finire: qui c'è solo l'opinione di un esperto di politica disposto a scommettere sulla pace 6 contro 4 - l'obiezione casomai è come sia giunto a una valutazione numerica cosìprecisa. La probabilità epistemica ha la pretesa di essere "onnivora": in effetti se lo stato di conoscenza del

soggetto è determinato "essenzialmente" dalla sua esperienza statistica (dagli esperimenti che ha svolto) la valutazione numerica "soggettiva" di una probabilità finisce col coincidere con quella "oggettiva" dell'empirista (nell'esempio precedente il soggettivista potrebbe ben puntare 3 contro 97 sull'uscita di uno scarto).

Dal punto di vista della "matematica astratta" la lacerazione filosofica è irrilevante, poichè, come si può mostrare, il calcolo delle probabilità combinatorie, oggettive e soggettive rimane lo stesso - i teoremi non cambiano.

Il calcolo delle probabilità, che è lo strumento più tradizionale di gestione dell'incertezza (ma non l'unico, come vedremo), è nato ufficialmente nel 1654 con un memorabile scambio di sette lettere fra Blaise Pascal e Pierre de Fermat; in esse è risolto un problema di gioco d'azzardo futile solo per chi non sappia vedere la fertile metafora in esso celata. E' fin dai primordi che la probabilità si presenta scissa fra due concezioni che solo in un primo tempo i pionieri del nuovo calcolo riuscirono ecletticamente e fruttuosamente a contemperare: da una parte la probabilità *oggettiva*, legata a proprietà fisiche (a "dati di fatto"), come la simmetria di un dado o la regolarità statistica registrata in una serie di esperimenti, dall'altra la probabilità *soggettiva*, che misura il "grado di fiducia" (soggettivo ma razionale) del soggetto giudicante circa la possibile realizzazione di un evento, grado di fiducia che dipende dallo stato di conoscenze nel quale il soggetto giudicante si trova in quel momento.
Entrambe le concezioni, quella oggettiva e quella soggettiva, rivendicano la stessa paternità: l'*Ars Conjectandi* di Jacob Bernoulli, che comparve postuma a Basilea nel 1713. Posta a buon uso nell'Ars Conjectandi, l'ambiguità col tempo divenne pesante, tant'è che agli inizi dell'Ottocento, ma senza gran successo, Siméon-Denis Poisson, anche lui illustre fisico oltre che illustre matematico, propose di usare due parole distinte, *probabilité* e *chance*, e di riservare la prima alla concezione epistemica, soggettiva, proprio come vuole la filologia. Una divagazione filologica non è in effetti fuori luogo: è interessante capire il perchè di un'ambiguità che all'inizio suggerıilluminanti analogie, ma che poi portò a fraintendimenti e ad abusi. Grosso modo fino al Seicento, il termine "probabile", affine in maniera ovvia a parole come "provabile" o "approvabile", era univoco e significava "approvato dall'autorità religiosa", "attestato dalle sacre Scritture". Non sorprende dunque che, senza contraddizione, i miracoli venissero considerati eventi probabili e insieme impossibili (*sic*). Come mai alla fine del Seicento il significato originale venne rapidamente obliterato e sostituito da due significati distinti e conflittuali? Secondo quanto affermano gli storici, ciò sarebbe dovuto a una ragione profonda, legata al ribaltamento delle fonti di autorità tradizionali, quelle religiose, che vennero soppiantate dall'autorità della "natura", natura che l'uomo interroga tramite il metodo sperimentale, e dall'autorità della "mente illuminata". Al tramonto dell'età dei lumi l'accordo fra natura (probabilità oggettiva) e mente illuminata (probabilità epistemica) si incrinò definitivamente. Tutto ciò può lasciar perplessi, tanto più che la probabilità, e con lei la statistica che le è ancella, hanno registrato e continuano a registrare successi pratici e applicativi che sono sotto gli occhi di tutti.

Anche se probabilità, casualità e determinismo sono parole di significato largamente extramatematico, in questa fine di sezione prenderemo un punto di vista formale, da teoria della misura, e parleremo di "pesi" e di "misure". A chi conosce un po' di probabilità sarà chiaro che lo schema proposto è (per ora!) bernoulliano, per cui le componenti dell'n-upla x possono esser viste come le prime n uscite di un processo stocastico stazionario e senza memoria (in termini familiari allo scienziato naturale: come i risultati osservati in n esperimenti replicati "nelle stesse condizioni").

La *potenza cartesiana n-esima* $\mathcal{A}^n$ dell'alfabeto $\mathcal{A}$ è costituita dalle K^n *sequenze* di *lunghezza* n, o n-uple, costruite su $\mathcal{A}$. Alle n-uple e agli insiemi di n-uple in $\mathcal{A}^n$ associeremo ora dei "pesi", in un senso un po' vago ma suggestivo di questo termine. Partiremo dai "pesi atomici" delle lettere di $\mathcal{A}$, e poi specificheremo le regole per calcolare il peso di un'n-upla a partire dai pesi delle singole lettere che la costituiscono (*piano diacronico*), e il peso di un insieme di n-uple a partire dai pesi delle singole n-uple che lo costituiscono (*piano sincronico*).

Usiamo il termine diacronico perchè troviamo suggestivo pensare che ciascuna n-upla sia generata via via da una *sorgente di informazione* dalla prima all'n-esima lettera componente; va da sé che, eccedendo in termini, stiamo preparando il terreno agli usi futuri di questi concetti. Il termine sincronico è usato più che altro per contrapposizione, come se tutte le n-uple dell'insieme A fossero state generate "in parallelo" da tante sorgenti quante sono le n-uple.

Piano istantaneo: I pesi P (le probabilità) delle K lettere sono K numeri non negativi che sommano ad 1: $p_i = P(a_i) = \text{Prob}(a_i) \geq 0$, $\sum_i p_i = 1$

Ribadiamo che sommatorie "tacite" si intendono estese a tutti i valori dell'indice.

Piano diacronico: Se dagli n "atomi" si passa all'n-upla che ne è costituita, la probabilità bernoulliana ha comportamento *moltiplicativo*, nel senso che per ogni n-upla $\underline{x} = x_1 x_2 \dots x_n$ si ha:

$$P^n(\underline{x}) = \text{Prob}(\underline{x}) = \prod_{1 \leq i \leq n} \text{Prob}(x_i)$$

Piano sincronico: Se si passa a un insieme di n-uple la probabilità bernoulliana ha comportamento additivo, nel senso che il "peso complessivo" di un insieme $A \subseteq \mathcal{A}$ di oggetti, di n-uple, è la somma dei "pesi" degli oggetti che lo compongono:

$$P^n(A) = \text{Prob}(A) = \sum_{\underline{x} \in A} \text{Prob}(\underline{x})$$

Esercizio. $H(P^n) = nH(P)$. (Per induzione su n.) La formula ben sottolinea il "carattere logaritmico" dell'entropia. Sul piano euristico: l'incertezza circa il risultato globale di n esperimenti "ripetuti nelle stesse condizioni" è la somma delle incertezze nei singoli esperimenti.

L'additività "sincronica" implica le seguenti relazioni (con $m = P^n$) che faranno la gioia di chi ama la teoria della misura, visto che al loro interno riconoscerà gli assiomi delle misure su spazi finiti (sotto non abbiamo specificato n, poichè qui la dimensione è irrilevante; $\mathcal{U}$ è l'insieme "pieno"; nel caso di P^n, $\mathcal{U} = \mathcal{A}^n$):

$$m(A) \geq 0\ , \quad m(\emptyset) = 0$$
$$m(A \cup B) = m(A) + m(B) - m(A \cap B)$$
$$m(A \cup B) \leq m(A) + m(B)$$

$A \cap B = \emptyset \Rightarrow m(A \cup B) = m(A) + m(B)$
$m(\overline{A}) = m(\mathcal{U}) - m(A)$

Esercizio. La formula seguente riflette la mancanza di memoria dei processi bernoulliani (indipendenza delle n componenti aleatorie):

$$P^{nm}(A \times B) = P^n(A)P^m(B) \ , \quad A \subseteq \mathcal{A}^n \ , \ B \subseteq \mathcal{A}^m \ , \quad A \times B \subseteq \mathcal{A}^{n+m}$$

Cenno: dimostra prima la $P^{nm}(\underline{x} \times B) = P^n(\underline{x})P^m(B)$, poi somma su $x \in \mathcal{A}$. Come corollario si trova $P^n(\mathcal{A}^n) = 1$ (per induzione).

Dal punto di vista di un teorico della misura deformato professionalmente la probabilità altro non è se non una *misura normata a* 1. Inutile aggiungere che questo punto di vista non rende affatto ragione di un concetto che è uno dei più affascinanti e coinvolgenti della scienza e della filosofia moderne.