L'inferenza logica, l'urna di Bernoulli e il teorema di Bayes

Un'introduzione alla statistica Bayesiana

Graziano Venanzoni

Laboratori Nazionali dell'INFN - Frascati

email: graziano.venanzoni@lnf.infn.it

Incontri di Fisica 7 Ott 2010

Indice

- Il concetto di inferenza e probabilità
- Esempi di *probabilità dirette*: L'urna di Bernoulli e le catene di Markov
- Stima dei parametri (e test delle ipotesi)
- Esempi di *probabilità inverse*: Test medici, inversione dell'urna, conteggi di particelle

L'inferenza in teoria delle probabilità

- In teoria delle probabilità si indica con inferenza il ragionamento induttivo o **plausibile**, che si verifica quando non si hanno informazioni complete
- Situazione tipica: "Dall'osservazione dei dati sperimentali, cosa possiamo dire per l'ipotesi H_0 in questione?"
- Ad ogni ipotesi H_0 possiamo associare un **grado** di plausibilità (*quanto* l'ipotesi è vera?). Lo stesso vale per una certa *proposizione*
- Il grado di plausibilità è un concetto quantitativo (numero), che dipende dalle esperienze precedenti, dall'evidenza (dati) che abbiamo e può cambiare non appena riceviamo nuove evidenze (o informazioni) (questo è *in sostanza* il metodo bayesiano)

Un esempio di ragionamento *plausibile*

- Supponiamo di osservare una persona uscire a mezzanotte da una vetrina infranta di una gioielleria con una maschera agli occhi e dei gioielli in mano. Credo nessuno di voi avrebbe il minimo dubbio a associare questa situazione ad un ladro che esce dopo una rapina. Qual è il *ragionamento* che ci porta a questa conclusione?
- In questo caso la mancanza di informazioni a nostra disposizione non ci permette una conclusione certa (logica deduttiva). Per esempio potrebbe esserci una valida spiegazione di innocenza (il proprietario del negozio che uscito da un festa in maschera e accortosi della vetrina infranta porta i gioielli al sicuro) . Tuttavia i dati a disposizione ci fanno "sentire" la nostra conclusione (il ladro) estremamente plausibile.

Ragionamento deduttivo verso ragionamento plausibile

- Date due proposizioni A e B, il ragionamento deduttivo ($apodexis$) si basa sulla premessa $A \Rightarrow B$, che implica i due sillogismi $forti$:
  - Se A è vero $allora$ B è vero
  - Se B è falso $allora$ A è falso
- In assenza di informazioni complete, la premessa $A \Rightarrow B$, fornisce due sillogismi $deboli$ ($epagoge$):
  - se B è vero $allora$ A diventa più plausibile
  - se A è falso $allora$ B diventa meno plausibile

  L'evidenza B non prova che A sia vero, ma il verificarsi di una delle sue conseguenze ci dà più confidenza in A.

- esempio:
  - A = "Pioverà entro le dieci di sera" (Ipotesi, in seguito H_0)
  - B = "Il cielo si annuvolerà prima delle dieci di sera" (Dato, in seguito D)

  Osservando un cielo $molto$ nuvoloso alle 21:45 non dà la certezza di pioggia, ma rende questa possibilità $molto$ plausibile.
- B è conseguenza logica, **non causale o fisica** di A.
- La plausibilità per A, dipende **fortemente** da quanto sono scure le nuvole prima delle 22:00.
- La plausibilità per A, dipende dalle informazioni passate, per esempio, se si sono viste o no le previsioni del tempo in mattinata ($prior$).
- Tornando all'esempio del "potenziale" ladro: supponiamo che queste situazioni accadono molto spesso e in tutti questi casi la persona è innocente. Col tempo questo tipo di situazioni verrebbero ignorate.

Probabilità e le regole quantitative dell' inferenza

Nel 1946 Richard Cox formulò le regole quantitative dell'inferenza. Il grado di plausibiltà di una proposizione X, condizionata alle informazioni in nostro possesso (*background*) I, è rappresentata dal numero reale $P(X|I)$, che d'ora in avanti chiameremo "probabilità". Valgono le seguenti regole:

$$P(X|I) + P(\overline{X}|I) = 1 \quad (1)$$

$$P(XY|I) = P(X|YI)P(Y|I) = P(Y|XI)P(X|I) \quad (2)$$

La prima equazione ci dice che la probabilità che una affermazione sia vera *più* la probabilità che sia falsa è uguale a 1. La seconda equazione ci dice che la probablità che *entrambe* le affermazioni X e Y siano vere è uguale alla probabilità che X sia vera dato Y vera *per* la probabilità che Y sia vera.

Teorema di Bayes e marginalizzazione

Dall'equazione (2) segue la relazione nota come teorema di Bayes:

$$P(Y|XI) = \frac{P(X|YI)P(Y|I)}{P(X|I)} \tag{3}$$

L'importanza di questa relazione è evidente se sostituiamo Y e X con ipotesi (H) e dati (D):

$$P(H|DI) \propto P(D|HI) \times P(H|I) \tag{4}$$

- $P(H|DI)$ è la probabilità a posteriori per H: essa rappresenta la probabilità che H sia vera (il nostro stato di conoscenza al riguardo della verità (o falsità) dell'ipotesi H), alla luce dei dati analizzati;
- $P(D|HI)$ è la probabilità di ottenere i dati D sotto l'ipotesi H. È chiamata verosimiglianza (**likelihood**) (nella sua dipendenza da H);
- $P(H|I)$ è chiamata probabilità a priori (o $prior$) per l'ipotesi H. Essa rappresenta la probabilità che H sia vera (la nostra conoscenza di H), prima di analizzare i dati attuali.

Consideriamo ora un set di ipotesi (proposizioni) $H_1, ..., H_n$ *mutualmente esclusive* e *esaustive*, ossia una e una sola di esse è vera:

$$\sum_{i=1}^{n} P(H_i|I) = 1 \tag{5}$$

Allora:

$$\begin{aligned} D &= D(H_1 + H_2 + ... + H_n) \\ P(D|I) &= \sum_{i=1}^{n} P(DH_i|I) = \sum_{i=1}^{n} P(D|H_iI)P(H_i|I) \end{aligned}$$

che è la proprietà di marginalizzazione. Il teorema di Bayes diventa quindi:

$$P(H|DI) = \frac{P(D|HI)P(H|I)}{P(D|I)} = \frac{P(D|HI)P(H|I)}{\sum_{i=1}^{n} P(D|H_iI)P(H_i|I)} \tag{6}$$

Nel caso di due ipotesi alternative (per es. vera o falsa):

$$P(H|DI) = \frac{P(D|HI)P(H|I)}{P(D|HI)P(H|I) + P(D|\overline{H}I)P(\overline{H}|I)} \quad (7)$$

$$P(\overline{H}|DI) = \frac{P(D|\overline{H}I)P(\overline{H}|I)}{P(D|HI)P(H|I) + P(D|\overline{H}I)P(\overline{H}|I)} \quad (8)$$

$$(9)$$

da cui si ottiene il rapporto ($odds$):

$$\frac{P(H|DI)}{P(\overline{H}|DI)} = \frac{P(D|HI)}{P(D|\overline{H}I)} \frac{P(H|I)}{P(\overline{H}|I)} \quad (10)$$

Assumiamo di non avere nessuna idea ("a priori") per dire che l'ipotesi H sia vera o falsa: $P(H|I) = P(\overline{H}|I) = 1/2$. Allora:

$$\frac{P(H|DI)}{P(\overline{H}|DI)} = \frac{P(D|HI)}{P(D|\overline{H}I)} \quad (11)$$

ed è chiaro che $P(H|D)$ è massima laddove la probabilità di osservare i dati secondo l'ipotesi che stiamo testando è molto più grande di os-

servare i dati attuali secondo l'ipotesi alternativa (metodo di maximum likelihood, M.L.). Il rapporto delle likelihood è detto anche *fattore di Bayes*

.

Proprietà qualitative

Vediamo ora come le regole di consistenza del ragionamento plausibile (che definiscono l'algebra della teoria delle probabilità) diano una spiegazione formale del "senso comune" Sia I la premessa $A \Rightarrow B$ ($\equiv A = AB$). Allora

$$P(B|AI) = \frac{P(AB|I)}{P(A|I)} = 1 \tag{12}$$

$$P(A|\overline{B}I) = \frac{P(A\overline{B}|I)}{P(\overline{B}|I)} = 0 \tag{13}$$

$$P(A|BI) = \frac{P(B|AI)P(A|I)}{P(B|I)} \geq P(A|I) \tag{14}$$

$$P(B|\overline{A}I) = \frac{P(\overline{A}|BI)P(B|I)}{P(\overline{A}|I)} \leq P(B|I) \tag{15}$$

Principio di indifferenza

Se la nostra conoscenza (informazioni di background I) è tale per cui $(H_1, ..., H_N)$ sono ipotesi (proposizioni) mutualmente esclusive ed esaustive (ossia una e una sola è vera, $\sum_{i=1,N} P(H_i|I) = 1$) e I non favorisce nessuna di essa, allora

$$P(H_i|I) = \frac{1}{N} \tag{16}$$

Se A è vera su un sottoinsieme M delle H_i, allora:

$$P(A|I) = \frac{M}{N}, \tag{17}$$

Questa è la definizione orginale di probabilità data da James Bernoulli (1713):La probabilità per un evento come rapporto tra casi favorevoli e possibili, quando niente porta ad aspettarci che qualche caso sia favorito rispetto agli altri. Questo principio viene anche chiamato di *ragione non sufficiente* (Laplace), in contrapposizione alla *ragione sufficiente di*

Leibniz: non abbiamo nessuna ragione per preferire un'ipotesi rispetto ad un'altra.

Piccola digressione: Questa è la situazione in cui spesso ci troviamo prima dell'osservazione dei dati: siamo in uno stato di ignoranza e basandoci sul principio di *ragione non sufficiente* assumiamo una distribuzione uniforme per la probabilità a priori. Come vedremo in seguito, questo punto è considerato dai freqentisti (ortodossi) una debolezza della statistica bayesiana.

L'urna di Bernoulli

Consideriamo un'urna contenente N palle identiche di cui M colorate di rosso (e le altre di bianco). Dobbiamo estrarre n palline una alla volta senza rimpiazzarle. Secondo il nostro background (che d'ora in avanti chiameremo B), non abbiamo nessuna informazione per estrarre una palla piuttosto che un'altra (B potrebbe rappresentare un $modello$ di un'estrazione cieca in cui l'urna viene "shakerata" ad ogni estrazione). Indichiamo con R_i la (proposizione) "palla rossa all'i-esima estrazione", e con W_i "palla bianca all'i-esima estrazione". Innanzitutto:

$$P(R_i|B) + P(W_i|B) = 1 \quad W_i = \overline{R_i};\ R_i = \overline{W_i}. \tag{18}$$

Consideriamo la prima estrazione:

$$P(R_1|B) = \frac{M}{N} \tag{19}$$

$$P(W_1|B) = \frac{N-M}{N} \tag{20}$$

- Definizione Frequentista: l'eq.(19) rappresenta la **frequenza** limite nel caso in cui l'esperimento di estrazione della prima pallina dall'urna venga ripetuta infinite volte nelle medesime condizioni. (Questo può essere ottenuto "shakerando" a "sufficienza" l'urna, prima di ogni estrazione) $\Rightarrow$ $randomizzazione$ del processo (R_1 è una variabile (evento) $random$).
- Caso Bayesiano: l'eq.(19) si riferisce solo al nostro stato di conoscenza prima dell'estrazione, determinate dal nostro background "B". (Se noi sapessimo esattamente dove è la pallina rossa (e quindi un diverso B), o ignorassimo M e N, la probabilità sarebbe diversa).

Caratteristiche dell'urna

- Probabilità di due rossi nelle prime due estrazioni:

$$P(R_2R_1|B) = P(R_2|R_1B)P(R_1|B) = \frac{M-1}{N-1}\frac{M}{N}. \qquad (21)$$

Generalizzando la probabilità di rossi nelle prime r estrazioni:

$$P(R_1R_2\cdots R_r|B) = \frac{M!(N-r)!}{(M-r)!N!}. \qquad (22)$$

Analogamente la probabilità di bianchi nelle prime w estrazioni si ottiene dallo scambio di M con (N-M):

$$P(W_1W_2\cdots W_w|B) = \frac{(N-M)!(N-w)!}{(N-M-w)!N!}. \qquad (23)$$

Consideriamo ora la probabilità di di ottenere in n estrazioni r rossi seguiti da $w = n - r$ bianchi:

$$P(R_1 \cdots R_r W_{r+1} \cdots W_n | B) = \frac{M!(N-M)!(N-n)!}{(M-r)!(N-M-w)!N!} \quad (24)$$

Si può verificare che l'eq.(24) rappresenta la probabilità di estrarre r rossi in qualsiasi ordine, o in altre parole la probabilità di estrarre r rossi in n estrazioni è la stessa(24) per ciascuna sequenza (indipendentemente dall'ordine con cui vengono estratti i rossi). Possiamo quindi chiederci qual è la probabilità di estrarre r rossi in n estrazioni, indipendentemente dall'ordine? Siccome ciascuna sequenza è esclusiva, basta moltiplicare la probabilità di ciascuna sequenza (eq.(24)) per la molteplicità di sequenza $\binom{n}{r}$:

$$P(r|N, M, n) = \frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}} \quad (25)$$

che è chiamata distribuzione ipergeometrica.

- Probabilità di un rosso alla seconda estrazione (R_2). Nella prima estrazione deve essere uscito un rosso o un bianco (eventi esclusivi e esaustivi), quindi:

$$R_2 = R_2(R_1 + W_1) = R_2R_1 + R_2W_1 \tag{26}$$

$$P(R_2|B) = \frac{M}{N} \tag{27}$$

Si può verificare in generale che:

$$P(R_i|B) = P(R_j|B) = \frac{M}{N} \quad \forall i, j \tag{28}$$

ossia (secondo il nostro modello di urna B) la probabilità di estrarre un rosso è la stessa per qualsiasi estrazione (ignorando il risultato dell'estrazione precedente). Naturalmente la conoscenza del risultato di un estrazione precedente "modifica" la probabilità per R_i (in quando cambia il mio background B). Questo è vero anche per un'estrazione successiva a R_i? In altre parole il risultato di un estrazione successiva influenza la probabilità di un'estrazione precedente?

L'inferenza logica si propaga in tutte le direzioni temporali

- La conoscenza di un risultato di un'estrazione successiva, influenza la probabilità per una estrazione precedente (R_i)?
- Risposta: Si. Es: due palline di cui una rossa, la probabilità di estrarre la prima pallina rossa diventa certezza (1 o 0) se conosciamo l'esito della seconda estrazione. L'estrazione della seconda pallina non influenza "fisicamente" l'estrazione della prima, ma cambia il nostro stato di conoscenza. (L'inferenza si basa su connessioni logiche, che non necessariamente sono influenze fisiche (causali)).
- In generale si può verificare che

$$P(R_j|R_iB) = P(R_i|R_jB) \quad \forall i,j \tag{29}$$

($P(R_iR_j|B) = P(R_i|R_jB)P(R_j|B) = P(R_j|R_iB)P(R_i|B)$) e il risultato di un estrazione successiva ha la stessa importanza di

un'estrazione precedente (per il nostro stato di conoscenza)

- Le connessioni fisiche sono causali (ossia si possono propagare solo in avanti), mentre le connessioni logiche (su cui si basa l'inferenza e quindi la probabilità) si propagano in tutte le direzioni temporali. È la nostra conoscenza (B) che viene aggiornata (secondo un'unica direzione temporale). (Es: Le scoperte da parte di paleontologi di eventi successivi all'epoca degli animali preistorici hanno effetti $logici$ su di essi in quanto modificano la nostra conoscenza al riguardo).
- Quanto vale $P(R_j R_k|I)$ (la probabilità di un rosso alla j-esima e alla k-esima estrazione)? Sia $j < k$:

$$R_j R_k = (R_1 + W_1) \cdots R_j \cdots (R_{k-1} + W_{k-1}) R_k \qquad (30)$$

che si può espandere nella somma logica di 2^{k-2} sequenze:

$$W_1 R_2 W_3 \cdots R_j \cdots R_k \qquad (31)$$

Sappiamo che la probabilità di ogni sequenza è indipendente dall'ordine in cui il rosso o il bianco appare. Quindi possiamo riordinare ciascuna

sequenza, spostando R_j e R_k all'inizio della sequenza e ottenere:

$$R_1 R_2 (R_3 + W_3) \cdots (R_k + W_k) = R_1 R_2 \tag{32}$$

e quindi:

$$P(R_j R_k|B) = P(R_1 R_2|B) = \frac{M-1}{N-1}\frac{M}{N} \tag{33}$$

$$P(W_j R_k|B) = P(W_1 R_2|B) = \frac{N-M}{N}\frac{M}{N-1} \tag{34}$$

da cui:

$$P(R_k|R_j B) = \frac{P(R_j R_k|B)}{P(R_j|B)} = \frac{M-1}{N-1} \tag{35}$$

$$P(R_k|W_j B) = \frac{P(W_j R_k|B)}{P(W_j|B)} = \frac{M}{N-1} \quad \forall j < k \tag{36}$$

Per l'equazione (29) i risultati sopra sono validi per tutti i $j \neq k$. La conoscenza dell'estrazione di una pallina rossa, modifica la conoscenza dell'urna N,M in N-1, M-1. Analogamente per la pallina

bianca. Non importa se questa informazione riguarda l'estrazione di una pallina precedente o successiva all'estrazione in questione. Non c'è connessione causale ma "logica".

Probabilità di R_k supponendo che un rosso esca in almeno una delle estrazioni successive

$$R_{later} \equiv R_{k+1} + R_{k+2} + ... + R_N \tag{37}$$

$$P(R_k|R_{later}B) = \frac{P(R_{later}|R_kB)P(R_K|B)}{P(R_{later}|B)} \tag{38}$$

Indichiamo con $\overline{R}_{later} = W_{k+1}W_{k+2}\cdots W_N$ "nessun rosso per $i \geq$ k" Allora $P(\overline{R}_{later}|B)$ è (per la proprietà di scambiabilità) uguale alla probabilità di estrarre le prime n-k palle bianche:

$$P(\overline{R}_{later}|B) = \frac{\binom{N-M}{n-k}}{\binom{N}{n-k}} = C/D \tag{39}$$

Analogamente $P(\overline{R}_{later}|R_kB)$ è lo stesso risultato per il caso di N-1 palle di cui M-1 rosse:

$$P(\overline{R}_{later}|R_kB) = \frac{\binom{N-M}{n-k}}{\binom{N-1}{n-k}} = C/B \tag{40}$$

e quindi:

$$P(R_k|R_{later}B) = \frac{M}{N}\frac{1-C/B}{1-C/D} = \tag{41}$$

$$= \frac{M}{N-n+k}\frac{\binom{N-1}{n-k}-\binom{N-M}{n-k}}{\binom{N}{n-k}-\binom{N-M}{n-k}} \tag{42}$$

Caso particolare: estraiamo 3 palline (n=3) da un'urna di 4 palle due delle quali rosse (M=2). Quanto vale $P(R_1|R_2+R_3)$?

$$P(R_1|R_2+R_3B) = \frac{6-2}{12-2} = \frac{2}{5} \tag{43}$$

Per confronto:

$$P(R_1|B) = \frac{1}{2} \quad P(R_1|R_2B) = P(R_2|R_1B) = \frac{1}{3} \tag{44}$$

$$P(R_1|R_2 + R_3B) > P(R_1|R_2B). \ \textit{Antiintuitivo?} \tag{45}$$

(La conoscenza che il rosso occorra almeno una volta nelle due estrazioni successive dovrebbe "diminuire" la probabilità che esso accada alla prima estrazione più della conoscenza che occorra solo una volta (R_2), in contrasto con la (45)).

In realtà R_2 riduce per il calcolo della probabilità di R_1 le palline M e N di un'unità; $R_1 + R_2$ riduce N di 2 e M$\geq$1. Intuitivamente:

$$P(R_1|R_2 + R_3B) = \frac{M_{eff}}{N-2} \tag{46}$$

$$P(R_2R_3|R_2 + R_3B) = \frac{P(R_2R_3R_{later}|B)}{P(R_{later}|B)} = \frac{P(R_2R_3|B)}{P(R_{later}|B)} \tag{47}$$

$$= \frac{1/2 \times 1/3}{5/6} = \frac{1}{5} \tag{48}$$

Intuitivamente dato R_{later} c'è una probabilità di $1/5$ che due palle rosse siano rimosse, così il numero effettivo è $1+1/5 = 6/5$. Cosi il numero M_{eff} di palle rosse rimanenti per la prima estrazione è $4/5$, da cui:

$$P(R_1|R_2 + R_3B) = \frac{4/5}{2} = \frac{2}{5} \tag{49}$$

in accordo con quanto ottenuto con il calcolo precedente più rigoroso ma meno intuitivo.

Valore di aspettazione e varianza

Se una variabile X assume i valori $(x_1, ..., x_n)$ in n situazioni mutualmente esclusive e esaustive, con probabilità $(p_1, ..., p_n)$, la quantità

$$< X >= E(X) = \sum_{i=1}^{n} p_i x_i \tag{50}$$

è l'aspettazione o valore atteso di X (= la media pesata su tutti i possibili valori, con peso dato dalla probabilità di ciascun valore)

Analogamente definiamo varianza di X:

$$\sigma^2(X) =< X^2 > - < X >^2 \tag{51}$$

che rappresenta l'incertezza sulla variabile X ($\sigma(X)$ è chiamata anche deviazione standard).

Dato R_{later}, qual'è il valore atteso del numero di palline rosse per la prima estrazione?

Ci sono tre possibilità mutualmente esclusive compatibili con R_{later}: R_2W_3, W_2R_3, R_2R_3, per i quali M è (1,1,0). Allora:

$$P(R_2W_3|R_{later}B) = \frac{P(R_2W_3|B)}{P(R_{later}|B)} = \frac{1/2 \times 2/3}{5/6} = \frac{2}{5} \quad (52)$$

$$P(W_2R_3|R_{later}B) = \frac{2}{5} \quad (53)$$

$$P(R_2R_3|R_{later}B) = \frac{1}{5} \quad (54)$$

Da cui:

$$< M >= 1 \times \frac{2}{5} + 1 \times \frac{2}{5} + 0 \times \frac{1}{5} = \frac{4}{5} \quad (55)$$

che è quello che abbiamo indicato nell'esempio precedente come valore di aspettazione: $M_{eff} =< M >$

Più in generale possiamo dire che quando la frazione $F = M/N$ di palline rosse è conosciuta, $P(R_1|B) = F$; quando F è sconosciuta: $P(R_1|B) =< F >$

Campionamento con reinserimento

Supponiamo ora che dopo ogni estrazione *blindly* rimettiamo la palla nell'urna (definisce la nostra conoscenza B') La probabilità di estrarre due palle rosse in successione è:

$$P(R_1R_2|B') = P(R_2|R_1B')P(R_1|B') \tag{56}$$

$P(R_1|B')$ è evidentemente M/N, ma cosa dire per $P(R_2|R_1B')$? Essa dipenderà dalle assunzioni che facciamo per descrivere l'operazione di re-inserimento della palla nell'urna. Questa operazione è "funzione" delle proprietà geometriche (dimensioni) e fisiche (elasticità, attrito) delle palle e dell'urna. La descrizione completa è troppo complicata! Rendiamo la descrizione fisica ancora più *estrema*. Shakeriamo l'urna dopo avere reinserito la palla e *assumiamo* che dopo lo scuotimento dell'urna l'operazione precedente di "estrazione e inserimento" di R_1 non abbia nessun effetto per l'estrazione successiva. Tutto questo può essere tradotto nel termine di "randomizzazione", in cui l'estrazione e

inserimento di ciascuna palla non ha influenza per l'estrazione successiva (vedremo successivamente un caso diverso). Allora $P(R_2|R_1B') = P(R_2|B') = M/N$, ossia la probabilità di estrarre una palla rossa è M/N per qualsiasi estrazione.

La probabilità di estrarre r palline rosse in n estrazioni, indipendentemente dall'ordine è:

$$\binom{n}{r}\left(\frac{M}{N}\right)^r\left(\frac{N-M}{N}\right)^{n-r} \tag{57}$$

che è la distribuzione binomiale con probabilità di singolo successo = M/N. La distribuzione binomiale è anche il limite $N \to \infty$ di un'urna senza reinserimento.







Come si tiene conto delle correlazioni?

Supponiamo che l'estrazione e il reinserimento di una palla rossa da un'urna aumenti la probabilità di un rosso all'estrazione successiva di $\epsilon > 0$, mentre l'estrazione di un bianco diminuisca la probabilità di un rosso di $\delta > 0$, e che l'influenza di estrazioni precedenti sia trascurabile rispetto a ϵ e δ. ϵ e δ possono modellizzare l'influenza (questa volta fisica, causale) che un'estrazione ha per quella successiva. Sia "C" il nostro "Background". Allora:

$$P(R_k|R_{k-1}C) = p + \epsilon, \quad P(R_k|W_{k-1}C) = p - \delta \tag{58}$$

$$P(W_k|R_{k-1}C) = 1 - p - \epsilon, \quad P(W_k|W_{k-1}C) = 1 - p + \delta \tag{59}$$

dove $p = M/N$. La probabilità di estrarre r rosse e (n-r) bianche in un determinato ordine è:

$$p(p+\epsilon)^c(p-\delta)^{c'}(1-p+\delta)^w(1-p-\epsilon)^{w'} \tag{60}$$

se la prima estrazione dà rosso, altrimenti il primo fattore deve essere (1-p).c è il numero delle coppie RR,c' è il numero delle coppie WR, w è il numero delle coppie WW e w' è il numero delle coppie RW (il colore a sinistra indica l'estrazione *precedente* a quella di destra).

Evidentemente:

$$c + c' = \begin{bmatrix} r-1 \\ r \end{bmatrix} \qquad w + w' = \begin{bmatrix} n-r \\ n-r-1 \end{bmatrix}$$

a seconda che a prima estrazione dia rosso (up) o bianco (down). Ora, essendo $\epsilon, \delta \ll p < 1$, possiamo usare $(1 + \frac{\epsilon}{p})^c \sim exp(\frac{\epsilon c}{p})$:

$$p^r (1-p)^{n-r} exp\Big(\frac{\epsilon c - \delta c'}{p} + \frac{\delta w - \epsilon w'}{1-p}\Big) \qquad (61)$$

La probabilità di estrarre r rossi e $n-r$ bianchi dipende (questa volta) dalla sequenza con cui i rossi e bianchi appaiono e p uò essere molto diversa dalla predizione binomiale. Supponiamo $N = 2M$, $p = 1/2$,

$\epsilon = \delta$. L'esponenziale in (61) diventa

$$exp\Big(2\epsilon[(c - c') + (w - w')]\Big) \tag{62}$$

Allora possiamo calcolarci la probabilità di avere lunghe sequenze di rossi (o bianchi). In particolare se $n \to \infty$, $c \gg c'$ $(w \gg w')$, allora $exp \gg 1$, e non possiamo credere all'ipotesi di randomizzazione. È un dato di fatto che nella ripetizione di esperimenti in condizioni $identiche$, queste lunghe sequenze si osservano.

Calcoliamoci ora alcune probabilità tenendo conto delle correlazioni

(58):

$$
\begin{aligned}
P(R_1|C) &= \frac{M}{N} \quad q = 1 - p = P(W_1|C) = \frac{N-M}{M} \\
P(R_2|C) &= P(R_2R_1|C) + P(R_2W_1|C) \\
&= P(R_2|R_1C)P(R_1|C) + P(R_2|W_1C)P(W_1|C) \\
&= (p+\epsilon)p + (p-\delta)q = p + (p\epsilon - q\delta) \\
P(R_3|C) &= P(R_3|R_2C)P(R_2|C) + P(R_3|W_2C)P(W_2|C) \\
&= (p+\epsilon)(p + p\epsilon - q\delta) + (p-\delta)(q - p\epsilon + q\delta) \\
&= p + (1+\epsilon+\delta)(p\epsilon - q\delta)
\end{aligned}
$$

Come aspettato $P(R_k|C)$ dipende da k.

Ma quanto vale $P(R_k|C)$ per $k \to \infty$? Utilizziamo un altro metodo (*catene di Markov*). Calcoliamo la probabilità per la k-esima estrazione come un vettore

$$
V_k = \begin{bmatrix} P(R_k|C) \\ P(W_k|C) \end{bmatrix} \qquad V_1 = \begin{bmatrix} p \\ q \end{bmatrix}
$$

e l'equazione (58) si può esprimere come

$$V_k = MV_{k-1} \tag{63}$$

dove

$$M = \begin{pmatrix} p+\epsilon & p-\delta \\ q-\epsilon & q+\delta \end{pmatrix}$$

Questo definisce una catena di Markov delle probabilità e M è chiamata matrice di transizione. Si può verificare che:

$$V_k = M^{k-1}V_1 \tag{64}$$

la cui soluzione generale si ottiene diagonalizzando la matrice:

$$M^{k-1} = \frac{1}{1-\epsilon-\delta}\begin{pmatrix} p-\delta+(\epsilon+\delta)^{k-1}(q-\epsilon) & p-\delta(1-(\epsilon+\delta)^{k-1}) \\ (q-\epsilon)(1-(\epsilon+\delta)^{k-1}) & q-\epsilon+(\epsilon+\delta)^{k-1}(p-\delta) \end{pmatrix}$$

E la soluzione generale è:

$$P(R_k|C) = \frac{(p-\delta)-(\epsilon+\delta)^{k-1}(p\epsilon-q\delta)}{1-\epsilon-\delta} \tag{65}$$

Si può verificare che questa soluzione è corretta (casi particolari/assenza di correlazioni).

Consideriamo ora il limite per $k \to \infty$:

$$P(R_k|C) \to \frac{p-\delta}{1-\epsilon-\delta} \tag{66}$$

Consideriamo ora $P(R_k|R_jC)$ (per $j < k$), che si può ottenere seguendo la (63)

$$V_k = M^{k-j} V_j \tag{67}$$

dove (rosso nella j-esima estrazione)

$$V_j = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

da cui:

$$P(R_k|R_jC) = \frac{(p-\delta) + (\epsilon+\delta)^{k-j}(q-\epsilon)}{1-\epsilon-\delta} \tag{68}$$

che tende allo stesso limite (66) per $k \gg j$ (il termine esponenziale che decade con la distanza).

È importante notare che (in generale) dal momento che $P(R_k|C) \neq P(R_j|C)$:

$$P(R_j|R_kC) \neq P(R_k|R_jC) \tag{69}$$

ossia la presenza di correlazioni ("fisiche") rende l'inferenza "all'indietro" diversa da quella "in avanti". Questo è un esempio di processo *Markoviano irreversibile* che accade molto spesso in fisica.

Notiamo che esiste (almeno) un caso in cui la (69) non è vera: $p\epsilon = q\delta$. Allora

$$\frac{p-\delta}{1-\epsilon-\delta} = p \quad \frac{q-\epsilon}{1-\epsilon-\delta} = q \quad \epsilon + \delta = \frac{\epsilon}{q} \tag{70}$$

e:

$$\begin{aligned} P(R_k|C) &= p \; \forall k \\ P(R_k|R_jC) &= P(R_j|R_kC) = p + q\left(\frac{\epsilon}{q}\right)^{|k-j|} \; \forall k, j \end{aligned}$$

In questo caso la simmetria avanti-indietro è ristabilita, sebbene le correlazioni ϵ, δ operino solo in avanti.

Stima elementare dei parametri (e Test delle ipotesi)

"Gioco a carte con una persona che ritengo perfettamente onesta. Qual e' la probabilità che egli abbia un Re? È 1/8. Questo è un problema delle probabilità degli effetti. Gioco a carte con una persona che non conosco. Su dieci mani ha estratto il Re 6 volte. Qual é la probabilità che sia un baro? Questo è un problema delle probabilità delle cause. Possiamo dire che è un problema essenziale del metodo sperimentale.....le leggi sono conosciute attraverso gli effetti osservati. Il dedurre dagli effetti le leggi che ne sono le cause corrisponde a risolvere un problema delle probabilità delle cause."

Poincaré

Nei paragrafi precedenti abbiamo analizzato le caratteristiche (e le conseguenze) delle distribuzioni di probabilità o *probabilità dirette*:

Da un'ipotesi H (nel nostro caso il contenuto (M,N) dell'urna), qual è la probabilità di ottenere un certo dato (nel nostro caso un certa sequenza di palle rosse e bianche) ?

- Probabilità diretta (o *degli effetti*): Causa $\rightarrow$ Effetto ($H \rightarrow D$)

Nel mondo reale, ci troviamo molto più spesso nella situazione opposta: abbiamo a disposizione dei dati (D), ma non conosciamo l'ipotesi (H) corretta. Ci troviamo così ad affrontare il problema inverso o delle *probabilità delle cause*:

Da i dati D, qual è la probabilità che una certa specificata ipotesi H sia vera? E naturalmente...cosa sapevamo su H, prima di osservare i dati?

- Probabilità inversa (o *delle cause*): Effetto $\longrightarrow$ Causa ($D \longrightarrow H$)

Di questo problema (detto dell'inferenza statistica), ci occuperemo ora

Il problema è trattato diversamente dai bayesiani e dai frequentisti, come vedremo da alcuni esempi nel seguito. Tuttavia spesso le predizioni coincidono (o si discostano di poco), mentre si possono avere delle differenze sostanziali in casi particolari, laddove per esempio si ha a che fare con pochi eventi (e quindi l'informazioni a priori diventa importante).

Test delle ipotesi

Il problema della stima delle ipotesi viene risolto dai probabilisti classici (soggettivisti o bayesiani) usando il teorema di Bayes, (3):

$$P(H|DI) \propto P(D|HI) \times P(H|I) \tag{71}$$

ossia la probabilità per H (probabilità *a posteriori*) viene ottenuta *aggiornando* la probabilità *a priori* alla luce dei risultati sperimentali (verosimiglianza o *likelihood*).

Nel caso discreto:

$$P(H|DI) = \frac{P(D|HI)P(H|I)}{P(D|I)} = \frac{P(D|HI)P(H|I)}{\sum_{i=1}^{n} P(D|H_iI)P(H_i|I)} \tag{72}$$

mentre al continuo ($H_i \to \theta$):

$$p(\theta|DI) = \frac{P(D|HI)P(H|I)}{P(D|I)} = \frac{P(D|HI)P(H|I)}{\int d\theta P(D|\theta I)P(\theta|I)} \tag{73}$$

Attenzione a non confondere $P(H|DI)$ con $P(D|HI)$ perchè si possono ottenere risultati errati!







Test dell'AIDS

Supponiamo di sottoporci ad un test dell'AIDS. Assumiamo che il test sia molto efficiente sui malati (99.9%), mentre dia un 0.2% di falso positivo. Supponiamo che il risultato sia positivo. Posso concludere di essere malato?

Dati:

Sia H_0 l'ipotesi "sono malato" ($\bar{H}_0$ l'ipotesi "sono sano"), e sia D il dato "sono risultato positivo al test". Allora:

$$P(D|H_0) = 0.99 \quad (74)$$

$$P(D|\bar{H}_0) = 0.002 \quad (75)$$

$$(76)$$

Posso concludere che siccome $P(D|\bar{H}_0) = 0.2\% \Rightarrow P(\bar{H}_0|D) = 0.2\%$ ossia che alla luce del risultato positivo al test sono malato al 99.8% (Questo è quello che si sente spesso dire nelle diagnosi mediche)?

Vedremo successivamente che ciò non è vero.

Es. 2: "Eventi strani ad Hera". Nel 1997 due esperimenti di fisica delle particelle elementari presso il laboratorio Desy in Germania hanno osservato un eccesso anomalo di eventi, che hanno una probabilità inferiore all' 1% di provenire dal Modello Standard.

Quindi dal momento che $P(D|SM) \leq 1\%$ possiamo concludere che $P(SM|D) \leq 1\%$ ossia che questi dati indichino al 99% la presenza di nuova fisica? ($\Leftrightarrow P(\bar{SM}|D) > 99\%$) Anche in questo caso la risposta è no.

Es. 3: Gioco onestamente alla lotteria (ipotesi H_0). Vinco. Dal momento che $P(vinto|onesto) \sim 0\%$ $allora$ $P(onesto|vinto) \sim 0\%$, ossia ho giocato in maniera disonesta?

La risposta a questi esempi è chiaramente no, e il modo opportuno per vederlo è di utilizzare (correttamente) il teorema di Bayes (Laplace).

Risolviamo il problema del malato di AIDS. La probabilità che io sia malato alla luce del risultato positivo del test è

$$P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0) + P(D|\overline{H_0})P(\overline{H_0})} \tag{77}$$

Dobbiamo conoscere la P *a priori* di essere malato di AIDS (ossia prima di aver effettuato il test). In generale ogni informazione a nostra disposizione (sesso, età, tipo di vita, etc....) è utile a valutarla. In generale possiamo assumere una probabilità dello 0.1% in base alle statistiche sui malati di AIDS. In questo caso:

$$P(H_0|D) = \frac{0.99 \times 10^{-3}}{0.99 \times 10^{-3} + 0.02 \times (1 - 10^{-3})} = 0.047 \tag{78}$$

ossia la probabilità di essere malato di AIDS essendo risultato positivo al test è minore del 5% (e non del 99.8% come erroneamente si sarebbe supposto). Quindi quando si va a fare un test diagnostico bisogna sempre valutare l'incidenza della malattia tra la popolazione (nel caso in cui essa sia rara il falso positivo non vuol dire assolutamente che il soggetto

sia malato). Supponiamo di ripetere il test e di risultare di nuovo positivo. In questo caso (esercizio lasciato per casa)

$$P(H_0|D_2D_1) = \frac{P(D_2D_1|H_0)P(H_0)}{P(D_2D_1)} \quad (79)$$

$$= \frac{P(D_2|D_1H_0)P(D_1|H_0)P(H_0)}{P(D_2|D_1)P(D_1)} \quad (80)$$

$$= \frac{P(D_2|H_0)P(H_0|D_1)}{P(D_2|H_0)P(H_0|D_1) + P(D_2|\overline{H_0})P(\overline{H_0}|D_1)} \quad (81)$$

$$= \frac{0.99 \times 0.047}{0.99 \times 0.047 + 0.02 \times (1 - 0.047)} = 0.71 \quad (82)$$

ossia la probabilità di essere malato sale al 71%.

Superamento del metodo di falsificazione

Ricordiamo ancora una volta lo schema *Bayesiano*

Probabilità finale $\propto$ verosimiglianza $\times$ probabilità iniziale.

Ne segue che:

- affinché la probabilità finale si annulli è sufficiente che sia nulla la verosimiglianza relativa ad una qualsiasi osservazione.
- affinché si raggiunga l'assoluta certezza (probabilità finale uguale a 1) è necessario invece che sia diversa da zero la verosimiglianza e che sia uguale a 1 la probabilità iniziale.

Quindi, se delle osservazioni sono assolutamente incompatibili con una teoria, questa è falsificata:

$$P(osservazioni|teoria) = 0 \Rightarrow P(teoria|osservazioni) = 0 \quad (83)$$

Al contrario, affinché una teoria sia dichiarata certa è necessario, oltre che spieghi i dati sperimentali, che essa sia ritenuta certa fin dall'inizio.

Questa è una posizione dogmatica inaccettabile.
Si noti come, al contrario del metodo usuale di falsificazione, l'induzione probabilistica basata sull'aggiornamento bayesiano, permette di classificare in ordine di credibilità le teorie non falsificate. Questo è in accordo con il modo di procedere $de\ facto$ della ricerca scientifica.







Critiche alla stima bayesiana

La stima probabilistica basata sul metodo bayesiano è rigettata dai statistici frequentisti (*ortodossi*), come Fisher etc..., con l'assunto che la stima deve essere basata *unicamente* sui dati e quindi non deve fare riferimento a probabilità iniziali o finali. Si utilizza quindi solo la verosimiglianza.

Anche il concetto di probabilità cambia (per i frequentisti): è definita come la frequenza relativa in una popolazione potenzialmente infinita (R. von Mises, R. A. Fisher)

Notiamo comunque che all'aumentare delle osservazioni, la dipendenza dai priori tende ad essere sempre meno significativa, e quindi la stima Bayesiana è guidata dalla verosimiglianza (G. D'agostini). Quindi i due approcci tendono a dare stessi risultati, anche se le due scuole divergono sul significato da dare alle stime (valor medi, incertezze, livelli di confidenza, etc...).

In particolare nell'approccio frequentista non ha senso parlare di $P(H_0|D)$, in quanto H_0 non è una variabile randomica. Di conseguenza la bontà di un ipotesi viene stabilita attraverso opportuni test che confrontano i valori attesi sotto una determinata ipotesi (che usa la verosimiglianza $P(D|H_0)$) con i i valori osservati e definendo una $soglia$ per accettare o meno l'ipotesi stessa.

Problema dell'inversione dell'urna

Assumiamo di conoscere il risultato di un'estrazione ($D \equiv (n, r)$) Cosa possiamo dire (*inferire*) sul contenuto dell'urna (N,R)?

Sappiamo che la funzione di likelihood per l'ipotesi (N,R) è la distribuzione ipergeometrica:

$$p(D|NRI) = \binom{N}{n}^{-1} \binom{R}{r} \binom{N-R}{n-r} \tag{84}$$

In generale il caso in cui N e R sono entrambi sconosciuti non è interessante, in quanto le probabilità a priori non sono mai troppo complesse per estrarre informazioni su N (si può solo concludere che $N \geq n$). Ossia i *prior* (in generale) non sono *informativi* per N. È più interessante il caso quando N è conosciuto e vogliamo inferire il valore di R:

$$p(R|DNI) = p(R|NI)\frac{p(D|NRI)}{p(D|NI)} \tag{85}$$

Cosa possiamo dire per la *prior* $p(R|NI)$?

Prior uniforme

Supponiamo di non sapere nulla su R prima dell'estrazione dell'urna. In questo caso possiamo assumere per R una distribuzione di probabilità uniforme su tutti i possibili valori di R ($0 \leq R \leq N$):

$$p(R|NI_0) = \frac{1}{N+1} \quad 0 \leq R \leq N \tag{86}$$

Notiamo che prima dell'estrazione:

$$< R > \; = \sum_{R=0}^{N} p(R|NI_0)R = \frac{1}{(N+1)} \frac{N(N+1)}{2} = \frac{N}{2} \tag{87}$$

$$\sigma^2(R) \; = \; < R^2 > - < R >^2 = \frac{N(2N+1)}{6} - \frac{N^2}{4} = \frac{N(N+2)}{12} \tag{88}$$

Priors non informativi

La scelta della distribuzione uniforme per la *prior*, è una conseguenza del *principio di indifferenza (o di ragione non sufficiente)*. Essa vorrebbe descrivere uno stato di ignoranza in cui non abbiamo nessuna ragione per prediligere una certa ipotesi.

Sembrerebbe quindi che in assenza di informazioni (prima dell'estrazione), la distribuzione uniforme sia quella che garantisca il minimo di assunzioni sui valori dell'esperimento. Questa assunzione è estremamente delicata e fortemente criticata dai frequentisti: "se non sappiamo nulla, non sappiamo nulla. Ossia non siamo in grado di valutare l'adeguatezza di alcuna distribuzione e quindi neppure dell'uniforme". Ricordiamo che i **frequentisti** rifiutavano qualunque assunzione che non trovasse l'immediato riscontro sperimentale (Fisher).

A questa critica **Jeffreys** (uno dei padri del metodo bayesiano), risponde che la probabilità è soggettiva e la scelta di una prior (per esempio quella uniforme) non è cruciale, ma può avvenire per convenzione (attraverso delle regole formali).

Ritorniamo al problema dell'inversione dell'urna: dalla (85):

$$P(R|DNI_0) = \frac{P(D|RNI_0)P(R|NI_0)}{\sum_{R=0}^{N} P(D|RNI_0)P(R|NI_0)} = \frac{\binom{R}{r}\binom{N-R}{n-r}}{S} \qquad (89)$$

dove S è la costante di normalizzazione:

$$S = \sum_{R=0}^{N} \binom{R}{r}\binom{N-R}{n-r} = \binom{N+1}{n+1} \qquad (90)$$

e qundi la prob a posteriori per R diventa:

$$p(R|DNI_0) = \binom{N+1}{n+1}^{-1}\binom{R}{r}\binom{N-R}{n-r} \qquad (91)$$

Osserviamo che a parte una costante di normalizzazione la (91) coincide con la verosimiglianza. Questa è una caratteristica generale, ossia nel caso di prior uniformi, la stima bayesiana o frequentista (dove si usa-massimizza la verosimiglianza) danno identici risultati.

Assumiamo di non aver nessun dato $(n = r = 0) \Rightarrow p(R|DNI_0) = 1/(N+1) = p(R|NI_0)$.

Se invece estraiamo una sola pallina rossa, n=r=1

$$p(R|DNI_0) = \frac{2R}{N(N+1)} \tag{92}$$

Come si vede, e come ci si aspetta, P(R=0)=0.

Il valore più probabile di R (per cui è massima la prob a posteriori) si ottiene ponendo $p(\hat{R}) = p(\hat{R} - 1)$ da cui:

$$\hat{R} = (N+1)\frac{r}{n} \tag{93}$$

ossia la frazione di palline rosse nell'urna è circa uguale alla frazione di rosse nell'estrazione (così come aspettato). Calcoliamo ora il valore di aspettazione di R:

$$< R >= E(R|DNI_0) = \sum_{R=0}^{N} Rp(R|DNI_0). \tag{94}$$

Notiamo che:

$$< R > +1 = \frac{(N+2)(r+1)}{(n+2)} \tag{95}$$

e quindi il valore di aspettazione della frazione di palline rosse lasciate nell'urna dopo l'estrazione diventa:

$$< F >= \frac{< R > -r}{N-n} = \frac{r+1}{n+2} \tag{96}$$

La regola di successione di Laplace

Vogliamo calcolare ore la probabilità di estrarre una pallina rossa dopo aver estratto r rosse in n estrazioni. Definiamo $R_{n+1} \equiv rosso\ nella\ n{+}1\text{-}esima\ estrazione$. Allora, usando la proprietà di marginalizzazione e usando la (91):

$$p(R_{n+1}|DNI_0) = \sum_{R=0}^{N} p(R_{n+1}|RDNI_0)p(R|DNI_0) = \sum_{R=0}^{N} \frac{R-r}{N-n}\binom{N+1}{n+1}^{-1}\binom{R}{r}\binom{N-R}{n-r}$$

che diventa (esercizio per casa)

$$p(R_{n+1}|DNI_0) = \frac{r+1}{n+2} \tag{97}$$

come l'espressione (96).

L'equazione (96) è conosciuta anche con il nome di *regola di successione di Laplace* (Laplace la derivò in un modo diverso). Notare che è indipendente da N.

Torniamo al problema di prima e calcoliamo ora la varianza di R:

$$\sigma^2(R) =< R^2 > - < R >^2 = \frac{p(1-p)}{n+3}(N+2)(n+2) \qquad (98)$$

dove $p =< F >= (r+1)/(n+2)$.

Notate che in assenza di dati (n=r=0) $\sigma^2(R) = N(N+2)/12$, in accordo con (88). A questo punto possiamo calcolare stima ed errore di F (=frazione delle palline rosse rimaste nell'urna dopo l'estrazione):

$$F_{est} = p \pm \sqrt{\frac{p(1-p)}{n+3}\frac{N+2}{N-n}} \qquad (99)$$

e nel limite di infinite palle ($N \to \infty$)

$$F_{est} = p \pm \sqrt{\frac{p(1-p)}{n+3}}, \qquad (100)$$

corrispondente al risultato della distribuzione binomiale.

Qusto risultato si applica molto bene ai sondaggi di opinione: assumiamo che in un campione di n persone intervistate, il 41% favorisce un certo candidato, e che l'errore sia del 3%. Allora:

$$F_{est} =< F > (1 \pm 0.03) \rightarrow n + 3 = \frac{1-p}{p}\frac{1}{0.03^2} = 1599 \qquad (101)$$

ossia sono state intervistate 1600 persone.

Atri prior interessanti

Come cambiano i risultati precedenti con i seguenti prior?

1. Prior uniforme troncata (almeno una palla rossa e bianca nell'urna):

$$p(R|I_0) = \frac{1}{N-1}, \quad 1 \leq R \leq N-1 \tag{102}$$

2. Prior concava:

$$p(R|I_C) = \frac{A}{R(N-R)}, \quad 1 \leq R \leq N-1 \tag{103}$$

3. Prior binomiale:

$$p(R|I_B) = \binom{N}{R} g^R (1-g)^{N-R}, \quad 0 \leq R \leq N \tag{104}$$

(g è la probabilità per una pallina di essere rossa).

Proprietà interessanti. Per casa!

Esempio di inferenza bayesiana applicata ad un problema di fisica

Supponiamo di disporre di una sorgente radioattiva che emette delle particelle (ad esempio positroni) ad un $rate\ p$ ($\sharp$ particelle/sec). Supponiamo ora che il nostro rivelatore registri dei conteggi ad un $rate$ θ. Dalle misure di conteggi $c_1, c_2, ...$ cosa possiamo dire del $rate$ di particelle emesse $n_1, n_2, ...$, e quindi delle caratteristiche della sorgente?

Questo è un tipico problema a due livelli (emissione-rivelazione), in cui noi osserviamo solo il risultato finale. Da questo dobbiamo fare la nostra migliore "inferenza" sulla causa originale e sulle condizioni intermedie. Mostreremo come queste stime dipendano dalle conoscenze di background (prior). Esempi analoghi si possono fare in medicina dove, dalla percentuale di malati si stima il tasso di incidenza di una malattia.

Ciascun rivelatore (quindi anche il nostro) è caratterizzato da un parametro ϵ, detto *efficienza*, che rappresenta la probabilità che una particella incidente dia un conteggio.

La probabilità che n particelle incidenti, diano luogo a c conteggi sarà dunque una binomiale:

$$b(c|n,\epsilon) = \binom{n}{c}\epsilon^c(1-\epsilon)^{n-c} \tag{105}$$

dove abbiamo assunto l'indipendenza di ciascun evento (conteggio). Assumiamo ϵ noto (dopo cercheremo di capire come si può calcolare).

La probabilità di emissione della nostra sorgente è invece descritta da una distribuzione *poissoniana*:

$$p(n|\mu) = \frac{e^{-\mu}\mu^n}{n!} \tag{106}$$

dove n è il numero di particelle emesse al secondo, e μ il valore di aspettazione (medio) di particelle emesse al secondo (intensità della sorgente).

Assumiamo di conoscere anche μ:

$$p(c|\epsilon\mu) = \sum_{n=0}^{\infty} p(cn|\epsilon\mu) = \sum_{n=0}^{\infty} p(c|n\epsilon\mu)p(n|\epsilon\mu) \qquad (107)$$

Da cui si ricava:

$$p(c|\epsilon\mu) = \sum_{n=c}^{\infty} \left[\frac{n!}{c!(n-c)!}\epsilon^c(1-\epsilon)^{n-c}\right]\left[\frac{e^{-\mu}\mu^n}{n!}\right] = \frac{e^{-\epsilon\mu}(\epsilon\mu)^c}{c!} \qquad (108)$$

che è una *poissoniana* di valor medio $\epsilon\mu$. Ossia il valor medio dei conteggi è uguale al prodotto del valor medio delle particelle emesse per l'efficienza del rivelatore (come atteso intuitivamente).

Vogliamo ora, dai dati a nostra disposizione (c, ϵ e μ), stimare n. Questo è un caso tipico in fisica delle particelle elementari, poiché spesso si a che fare con queste situazioni per calcolare sezioni d'urto di reazioni, etc.... Il teorema di Bayes ci da:

$$p(n|c\epsilon\mu) = \frac{p(c|n\epsilon)p(n|\mu)}{p(c|\epsilon\mu)} \qquad (109)$$





essendo $p(c|n\epsilon\mu) = p(c|n\epsilon)$ e $p(n|\epsilon\mu) = p(n|\mu)$ Con un pò di algebra otteniamo:

$$p(n|c\epsilon\mu) = \frac{e^{-\mu(1-\epsilon)}[\mu(1-\epsilon)]^{n-c}}{(n-c)!} \tag{110}$$

che è di nuovo una distribuzione di Poisson di parametro $\mu(1-\epsilon)$, *shiftato* di c (perchè $n \geq c$). Il valore di aspettazione di n è:

$$< n >= \sum_n np(n|c\epsilon\mu) = c + \mu(1-\epsilon) \tag{111}$$

che coincide con $\hat{n}$ (massimo della probabilità).

Se avessimo voluto risolvere questo problema nell'approccio frequentista, avremo massimimizzato la verosimiglianza $p(c|n\epsilon) \rightarrow \hat{n}_{M.L} = \frac{c}{\epsilon}$.

Torniamo ora al nostro problema e confrontiamo due *prior* diversi:

- A: nessuna informazione sulla sorgente/i
- B: sa che è un'unica sorgente a vita media lunga.

Come cambiano le inferenze per A e B?

Per A abbiamo già visto come la distribuzione uniforme possa descrivere uno stato di completa ignoranza (principio di *indifferenza o di ragione non sufficiente*):

$$p(n|I_A) = \frac{1}{N}, \quad 0 \leq n < N \tag{112}$$

da cui, applicando il teorema di Bayes otteniamo:

$$p(n|c\epsilon I_A) = Ap(c|\epsilon n) \quad 0 \leq n < N \tag{113}$$

dove A è un fattore di normalizzazione:

$$A^{-1} = \sum_{n=0}^{N} p(c|n\epsilon) \rightarrow \frac{1}{\epsilon} \ \ (N \rightarrow \infty) \tag{114}$$

da cui otteniamo:

$$p(n|c\epsilon I_A) = \epsilon p(c|n\epsilon) = \binom{n}{c} \epsilon^{c+1}(1-\epsilon)^{n-c} \tag{115}$$

Per A quindi il valore più probabile di n è lo stesso della stima di M.L.:

$$\hat{n}_A = \frac{c}{\epsilon} \tag{116}$$

mentre il valore di aspettazione è:

$$< n >= \sum_{n=c}^{\infty} np(n|c\epsilon I_A) = \frac{c+1-\epsilon}{\epsilon} \qquad (117)$$

Cosa direbbe invece B? Egli sa che c'è una sorgente radioattiva ma non ne conosce l'intensità, che assume di probabilità uniforme:

$$p(n|I_B) = \int_0^{\infty} d\mu p(n|\mu)p(\mu|I_B) = \frac{1}{\mu_0}\int_0^{\infty} d\mu \frac{\mu^n e^{-\mu}}{n!} = \frac{1}{\mu_0} = const. \qquad (118)$$

ossia B otterrà gli stessi risultati di A.

Jeffreys' prior

Harold Jeffreys (vissuto nella metà del secolo scorso) ha proposto che il modo proprio per esprimere la *completa ignoranza* di una variabile continua positiva è di assegnare probabilità uniforme al logaritmo della

prior, ossia:

$$p(x|I_J) \propto \frac{1}{x} \quad (0 \leq x < \infty) \tag{119}$$

Si può vedere che questa prior è *formalmente* ben definita (invarianza di scala). Utilizzando questa prior, si ottiene:

$$p(n|I_J) = \frac{1}{n}, \quad p(c|I_j) = \frac{1}{c}, \quad p(n|c\epsilon I_J) = \frac{c}{n} p(c|\epsilon n) \tag{120}$$

Da cui si ottengono le seguenti stime:

$$\hat{n}_J = \frac{c - 1 + \epsilon}{\epsilon}; \quad < n >= \frac{c}{\epsilon} \tag{121}$$

Si noti come il valore di aspettazione coincida con la stima di M.L.

Supponiamo ora di avere a disposizione un secondo conteggio c_2 (chiamiamo il primo conteggio c_1) . Quale sarà la stima di A e B per n_1, n_2? A non ha nessuna conoscenza che possa collegare i due conteggi tra di loro e quindi otterrà per n_2 (vedi (116,117)):

$$(\hat{n}_2)_A = \frac{c}{\epsilon}; < n_2 >_A = \frac{c_2 + 1 - \epsilon}{\epsilon} \tag{122}$$

Quindi la conoscenza di c_2 non lo aiuta a migliorare la stima per n_1 (e viceversa), che rimane quella di prima.

Per B la situazione è diversa, in quanto lui assume che i conteggi c_1, c_2 e quindi n_1, n_2 provengano da un'unica sorgente. Utilizzando il teorema di Bayes:

$$p(n_1|c_2c_1\epsilon I_B) = \frac{p(c_2|n_1c_1\epsilon I_B)p(n_1|c_1\epsilon I_B)}{p(c_2|c_1\epsilon I_B)} = \frac{p(c_2|n_1\epsilon I_B)p(n_1|c_1\epsilon I_B)}{p(c_2|c_1\epsilon I_B)} \tag{123}$$

e la proprietà di marginalizzazione si ottiene:

$$p(c_2|n_1\epsilon I_B) = \int_0^\infty d\mu p(c_2|\mu n_1\epsilon I_B)p(\mu|n_1\epsilon I_B) \tag{124}$$

dove $p(c_2|\mu n_1\epsilon I_B) = p(c_2|\mu\epsilon I_B)$ (calcolata in (108)). Utilizzando la (118)

$$p(\mu|n_1\epsilon I_B) = \frac{p(n_1|\mu\epsilon I_B)p(\mu|\epsilon I_B)}{p(n_1|\epsilon I_B)} = p(n_1|\mu\epsilon) \tag{125}$$

Otteniamo quindi:

$$p(c_2|n_1\epsilon I_B) = \int_0^\infty d\mu \left[\frac{e^{-\mu\epsilon}(\mu\epsilon)^{c_2}}{c_2!}\right]\left[\frac{e^{-\mu}\mu^{n_1}}{n_1!}\right] = \binom{n_1+c_2}{c_2}\frac{\epsilon^{c_2}}{(1+\epsilon)^{n_1+c_2+1}} \tag{126}$$

da cui infine otteniamo:

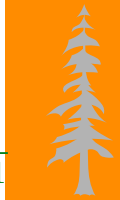
$$p(n_1|c_2c_1\epsilon I_B) = \binom{n_1+c_2}{c_1+c_2}\left(\frac{2\epsilon}{1+\epsilon}\right)^{c_1+c_2+1}\left(\frac{1-\epsilon}{1+\epsilon}\right)^{n_1-c_1} \tag{127}$$

Notate che si sarebbe potuto ottenere lo stesso risultato usando la proprietà di marginalizzazione (lasciato per casa):

$$p(n_1|c_2c_1\epsilon I_B) = \int_0^\infty d\mu p(n_1|\epsilon\mu c_1 I_B)p(\mu|\epsilon c_2 c_1 I_B) \tag{128}$$

Al solito possiamo calcolare valore più probabile e valor medio di n_1:

$$(\hat{n}_1)_B = \frac{c_1}{\epsilon} + (c_2 - c_1)\frac{1-\epsilon}{2\epsilon}; < n_1 >_B = \frac{c_1+1-\epsilon}{\epsilon} + (c_2 - c_1 - 1)\frac{1-\epsilon}{2\epsilon} \tag{129}$$



Riusciamo a capire il risultato? I conteggi c_1 e c_2 migliorano la conoscenza di μ, che a sua volta è rilevante per la stima di n_1.

Mentre per A, ogni sequenza $n_i \longrightarrow c_i$ è (logicamente) indipendente dalle altre, per B c_1 può essere ricavato a partire da c_1 secondo la sequenza $c_2 \longrightarrow n_2 \longrightarrow \mu \longrightarrow n_1$, e si può verificare che:

$$p(n_1|\epsilon c_1 c_2 I_B) \propto p(n_1|\epsilon c_1 I_B) p(n_1|\epsilon c_2 I_B) \tag{130}$$

Ossia l'informazione su n_1 è aggiornata alla luce dell'informazione c_2 (secondo il fattore $p(n_1|\epsilon c_2 I_B)$).

Va da sé che all'aumentare del numero dei conteggi migliora la stima di n_1.

L'accuratezza delle misure è quantificata dalla varianza:

$$\sigma^2(n_1|\epsilon c_1 I_A) = \frac{(c_1+1)(1-\epsilon)}{\epsilon^2}, \tag{131}$$

$$\sigma^2(n_1|\epsilon c_1 c_2 I_B) = \frac{(c_1+c_2+1)(1-\epsilon^2)}{4\epsilon^2}, \tag{132}$$

$$\sigma^2(n_1|\epsilon c_1 I_J) = \frac{c_1(1-\epsilon)}{\epsilon^2} \tag{133}$$

($\sigma^2_{n_2}$ si ricavano per simmetria).

Rappresentiamo in tabella seguente i risultati ottenuti per i valori numerici $c_1 = 10$ e $c_2 = 16$.

			Problema 1	Problemaa 2
			n_1	n_1 n_2
A	valore più prob	100	100	160
	$<> \pm\sigma$	109± 31	109± 31	169± 39
B	valore più prob	100	127	133
	$<> \pm\sigma$	109± 31	131.5± 25.9	137.5± 25.9
Jeffreys	valore più prob	91	121.5	127.5
	$<> \pm\sigma$	100± 30	127± 25.4	133± 25.4

Come si vede l'informazione aggiuntiva per I_B ha permesso una stima più accurata.

Generalizzazione a n conteggi e formule asintotiche

Supponiamo ora che B faccia $c_1, c_2, ...c_m$ conteggi. Allora:

$$p(n_k|\epsilon c_1....c_m I_B) = \int_0^\infty d\mu p(n_k \mu|\epsilon c_1....c_m I_B) = \int_0^\infty d\mu p(n_k|\mu\epsilon c_k I_B) p(\mu|\epsilon c_1....c_m I_B)$$

(se l'intensità della sorgente μ è conosciuta allora tutti i c_i con $i \neq k$ sono irrilevanti per n_k). Usiamo di nuovo il teorema di Bayes:

$$\begin{aligned} p(\mu|\epsilon c_1....c_m I_B) &= p(\mu|\epsilon I_B)\frac{p(c_1...c_m|\epsilon\mu I_B)}{p(c_1...c_m|\epsilon I_B)} \\ &= (const.) \times p(\mu|\epsilon I_B) p(c_1|\epsilon\mu I_B)...p(c_m|\epsilon\mu I_B) \end{aligned}$$

Usando la (108) otteniamo:

$$p(\mu|\epsilon c_1....c_m I_B) = s^c \frac{(m\epsilon)^{c+1} e^{-m\mu\epsilon}}{c!} \tag{134}$$

dove $c \equiv c_1 + ... + c_m$ è il numero totale di conteggi in m secondi. Otteniamo quindi le stime per μ:

$$\hat{\mu} = \frac{c}{m\epsilon}, \quad <\mu>= \frac{c+1}{m\epsilon}, \quad \sigma^2(\mu) = \frac{c+1}{m^2\epsilon^2} = \frac{<\mu>}{m\epsilon} \tag{135}$$

Come aspettato, la distribuzione $p(\mu|\epsilon c_1....c_m I_B)$ si fa sempre più stretta all'aumentare di m:

$$p(\mu|\epsilon c_1....c_m I_B) \to \delta(\mu - \mu') \tag{136}$$

dove:

$$\mu' \equiv \lim_{m\to\infty} \frac{c_1 + ... + c_m}{m\epsilon} \tag{137}$$

Si può verificare che l'andamento asintotico della (134) è una gaussiana:

$$p(\mu|\epsilon c_1....c_m I_B) \to A e^{-\frac{c(\mu-\hat{\mu})^2}{2\hat{\mu}^2}} \tag{138}$$

e nel limite $c \to \infty$:

$$(\mu)_{est} = \hat{\mu}\left(1 \pm \frac{1}{\sqrt{c}}\right) \tag{139}$$

Quindi asintoticamente B approccia la conoscenza esatta sull'intensità della sorgente. Si può verificare infine che nel limite $m \to \infty, c \to \infty, (c/m\epsilon) \to \mu' = const.$, la probabilità a posteriori per n_1 è una poissoniana:

$$p(n_k|c_1...c_m I_B) \to \frac{e^{-\mu'(1-\epsilon)}}{(n_k - c_k)!}[\mu'(1-\epsilon)]^{n_k - c_k} \tag{140}$$

che è identica alla (110). Quindi la deviazione standard di B passa da 26 a 10.8, mentre per A e' 31. La stima per n_k è:

$$< n_k >= c_k + < \mu > (1-\epsilon) \tag{141}$$

dove $< \mu >= (c+1)/m\epsilon$ e analogamente:

$$\hat{n_k} = c_k + \hat{\mu}(1-\epsilon) \tag{142}$$

dove $\hat{\mu}$ è definito in (135).

Bibliografia (essenziale)

- E.T. Jaynes "Probability Theory", Cambridge University Press (2003)
- G. D'agostini "Bayesian Reasoning in Data Analysis", World Scientific (2003) (vedere anche `http://www.roma1.infn.it/~dagos/prob+stat.html`)
- D. Costantini "I fondamenti storico-filosofici delle discipline statistico-probabilistiche", Bollati Boringhieri (2004)
- H. Jeffreys, "Theory of Probability", Oxford University Press, New York, (1961) 3rd ed.
- R. E. Kass and L. Wasserman, "The Selection of Prior Distributions by Formal Rules", Journal of the American Statistical Association 91, 1343 (1996). (`http://www.jstor.org/stable/2291752,http://lib.stat.cmu.edu/~kass/papers/rules.pdf`)
- J. M. Bernardo, " Bayesian Statistics" Encyclopedia of Life Support Systems (EOLSS). Probability and Statistics, (R. Viertl, ed). Oxford, UK: UNESCO (www.eolss.net) (`http://www.uv.es/~bernardo/BayesStat2.pdf`)