Brunero Liseo

Introduzione alla statistica bayesiana

Settembre 2008

Springer

Molte persone credono che il loro modo di agire e pensare sia l'unico corretto, non solo per loro stessi ma per chiunque altro. Questa ingiustificata estensione di un principio di utilità genera, di solito, una gran confusione; ma può generare tragedie se quel modo diventa lo stesso per troppi.

Anonimo, XXI secolo

Prefazione

L'approccio bayesiano all'inferenza sta acquisendo un ruolo sempre più importante nella letteratura statistica: è infatti in continuo aumento il numero di elaborazioni di dati in campo medico-sanitario, economico-finanziario, socio-politico e forse ancor di più nelle scienze sperimentali, dove si fa utilizzo più o meno esplicito di tecniche riconducibili al paradigma bayesiano dell'inferenza.

Le ragioni di questa improvvisa accelerazione, iniziata più o meno negli anni '90 del secolo scorso, della diffusione dei metodi bayesiani nella statistica applicata sono molteplici ma riconducibili a tre categorie essenziali: *i*) ragioni epistemologiche, *ii*) ragioni pragmatiche e, soprattutto, *iii*) ragioni di natura computazionali.

Da un punto di vista epistemologico, le motivazioni più cogenti per l'affermarsi del metodo bayesiano sono quelle di carattere fondazionale: l'impostazione bayesiana dell'inferenza statistica formalizza in modo semplice e diretto il ragionamento induttivo di un essere razionale che, in base alla informazioni disponibili su un certo insieme di fenomeni, in un certo istante della sua vita, vuole calcolare la probabilità di eventi futuri o, più in generale, di eventi per i quali non è noto se si siano verificati o meno. La logica bayesiana è coerente, dotata di solide basi logiche ed esente dal rischio di controesempi, sempre in agguato quando ci si muove nel campo dell'induzione, ed è necessario produrre affermazioni di natura probabilistica su eventi che non sappiamo se si verificheranno o meno.

Esistono poi motivazioni più pragmatiche: nel corso degli anni sono via via aumentate le applicazioni statistiche in cui l'esigenza di tener conto di informazioni extra-sperimentali, aspetto caratterizzante - sebbene non il più importante - dell'impostazione bayesiana, emergeva con chiarezza. In ambito epidemiologico, ad esempio, quando si valuta la probabilità che un paziente sia affetto da una certa patologia sulla base di un test diagnostico, quelle che sono le probabilità a priori sono nient'altro che le informazioni sulla prevalenza di quella malattia in quel contesto specifico e non sono meno oggettive delle informazioni sulla sensibilità e specificità del test adottato, che invece hanno una interpretazione nell'ambito della statistica classica.

In ambito economico-sociale, è sempre più importante per gli istituti nazionali di statistica e per altri enti di ricerca disporre di informazioni ad un livello di disaggregazione sufficientemente elevato: è certamente più utile, ad esempio, conoscere i livelli di disoccupazione o di natalità a livello comunale piuttosto che a livello provinciale. Questa esigenza è oggi così pressante che ha prodotto lo sviluppo di un nuovo tema di ricerca che va sotto il nome di "stima per piccole aree", dove spesso la difficoltà principale è quella di produrre informazioni anche per aree per le quali non si hanno a disposizione informazioni campionarie dirette. Una caratteristica intrinseca del

metodo bayesiano è proprio quella di poter assumere, in modo semplice e naturale, diversi livelli di associazione tra le unità campionarie, consentendo così quel fenomeno di "borrowing strength" che consente la produzione di stime sufficientemente stabili anche per quelle aree poco o per nulla coperte dall'indagine campionaria.

La solidità fondazionale del metodo bayesiano e la possibilità di integrare, attraverso il teorema di Bayes, le informazioni fornite dall'esperimento statistico con le ulteriori conoscenze "a priori" relative al problema in esame sono tuttavia cose ben note da molti decenni e non bastano da sole a giustificare l'enorme sviluppo degli ultimi anni. Ciò che ha causato la recente inversione di tendenza culturale nei confronti dei metodi bayesiani è stato senza dubbio l'enorme sviluppo di nuove metodologie computazionali che consentono ormai di analizzare, all'interno di questa impostazione, modelli statistici estremamente complessi. I cosiddetti metodi di Monte Carlo, basati o meno sulle proprietà delle catene di Markov (metodi MC e MCMC), permettono oggi di generare un campione, di dimensione qualsivoglia, di realizzazioni che possiamo considerare, almeno approssimativamente, indipendenti e somiglianti generate dalla distribuzione a posteriori dei parametri d'interesse del modello. Questo, oggi, è praticamente possibile per ogni modello statistico non importa quanto complesso. Questa potenzialità ha avuto un impatto fondamentale, soprattutto in campo applicato. Prima dell'era MCMC, l'impostazione bayesiana rappresentava un elegante modello teorico del paradigma inferenziale, insegnato soprattutto, sia in Italia che altrove, nei corsi di laurea con forte orientamento matematico. La pratica statistica era saldamente nelle mani della scuola frequentista, oggi rappresentata dalla fusione, non sempre armoniosa, di due correnti di pensiero, quella di Neyman, Pearson e Wald da un lato e quella Fisher e Cox dall'altra. Ciò che rendeva impraticabile il metodo bayesiano erano soprattutto i problemi di calcolo: aldilà di semplici modelli parametrici, infatti, non è possibile ottenere espressioni esplicite delle distribuzioni a posteriori delle quantità di interesse. Questo difficoltà ha fatto in modo che l'evoluzione della "modellistica" statistica avvenisse perlopiù in ambito frequentista.

Oggi la situazione è notevolmente diversa, a volte ribaltata. In un numero sempre crescente di ambiti applicativi, l'approccio bayesiano consente una flessibilità del modello difficilmente ottenibile mediante metodi classici.

Quanto appena descritto potrebbe lasciare intendere che il futuro sviluppo della scienza statistica sia orientato verso l'affermazione della logica bayesiana. Questo non è affatto certo: molti aspetti vanno ancora considerati e ancora oggi, ad esempio, molti studiosi sono contrari all'introduzione di informazioni extra-sperimentali nel procedimento inferenziale, intravedendo in questo la perdita di qualsiasi tipo di possibile "oggettività" delle inferenze. Questa dialettica scientifica tra diverse scuole di pensiero rende costantemente attuali due particolari capitoli del metodo bayesiano:

- lo studio delle proprietà delle distribuzioni cosiddette "convenzionali", costruite per minimizzare il contenuto informativo introdotto nella procedura e non direttamente relativo all'esperimento programmato;
- lo studio della sensibilità delle inferenze prodotte al variare degli input, con particolare riguardo alla distribuzione iniziale.

Di questi aspetti ci occuperemo, rispettivamente, nella §5.2 e nella §5.3.

Questo testo va considerato di livello introduttivo, concepito per un corso di statistica impartito nell'ambito di una laurea magistrale presso le facoltà di Economia, Scienze statistiche oppure

per studenti di Matematica. I prerequisiti necessari per la lettura del testo si limitano ad un corso di matematica generale e ad un'esposizione almeno introduttiva, al calcolo delle probabilità. Argomenti di teoria della misura, che in alcune parti renderebbero il testo più snello ed elegante sono stati volutamente evitati. La conoscenza dell'impostazione frequentista dell'inferenza non è considerata un prerequisito, ma certamente rende la lettura del testo più utile.

Dopo aver introdotto il lessico probabilistico necessario per una corretta interpretazione della logica bayesiana (capitolo 1), e una breve ma necessaria rassegna sulle tecniche di inferenza classiche basate sulla funzione di verosimiglianza (capitolo 2), i capitoli 3, 4 e 6 sono dedicati all'introduzione del metodo bayesiano e ad una rivisitazione in ottica bayesiana delle più consolidate tecniche inferenziali. Il capitolo 5 affronta invece il tema della scelta della distribuzione a priori, per molto tempo considerato il vero aspetto discriminante tra metodi bayesiani e non. Il capitolo 7 è dedicato all'illustrazione dei metodi computazionali oggi più importanti nella pratica bayesiana. Questi argomenti sono tra l'altro a tutt'oggi al centro di una frenetica attività di ricerca, e questo rende ancora difficile una loro trattazione sistematica. Prima di affrontare, nei capitoli successivi, la modellistica lineare e le sue evoluzioni, si è voluto dedicare il capitolo 8 alla discussione del tema del confronto tra modelli alternativi. Questo è uno dei settori dove le discrepanze tra metodi classici e bayesiani è più evidente e molto difficile appare una riconciliazione teorica tra le impostazioni.

Nel testo non compaiono alcuni argomenti, oggi centrali nella ricerca, come le interconnessioni fra la statistica classica e quella bayesiana in un contesto non parametrico, oppure il ruolo centrale del teorema di Bayes nelle tecniche di "machine learning". Tali argomenti, oggi essenziali per un uso efficace delle potenzialità che la statistica consente, sono tuttavia ancora troppo avanzati dal punto di vista matematico per essere trattati in modo comprensibile senza alterare la struttura del testo.

Il testo ha avuto una gestazione molto lunga, e nasce come note di un corso di statistica matematica da me tenuto per alcuni anni presso il corso di laurea in Matematica dell'università Roma Tre. A tal proposito mi fa piacere ringraziare tutti gli studenti che, leggendo e studiando le versioni precedenti, hanno segnalato diverse inesattezze. Ringrazio inoltre Alessandra Salvan, Gianfranco Adimari, Marilena Barbieri che hanno utilizzato versioni preliminari di questo testo nei loro corsi e Ludovico Piccinato che ha letto tutto con la consueta attenzione e profondità.

Roma, settembre 2008 *Brunero Liseo*

Indice

Parte I Titolo della parte

1 Teorema di Bayes e probabilità soggettiva . . . 3
1.1 Il teorema di Bayes. . . . 3
1.2 Probabilità a priori e verosimiglianze . . . 6
1.3 L'impostazione soggettiva della probabilità . . . 7
1.4 Definizione e condizione di coerenza . . . 8
Problemi . . . 10

2 Modello statistico e funzione di verosimiglianza . . . 13
2.1 Gli ingredienti di un modello statistico . . . 13
2.2 La funzione di verosimiglianza . . . 15
2.3 Uso inferenziale di $L(\theta)$ . . . 18
2.3.1 Stime di massima verosimiglianza . . . 18
2.3.2 Stima per intervalli . . . 20
2.4 Sufficienza . . . 23
2.5 Informazione di Fisher . . . 26
2.6 La divergenza di Kullback-Leibler . . . 31
2.7 Un'approssimazione della funzione di verosimiglianza . . . 32
2.8 Proprietà frequentiste delle procedure basate su $L(\theta)$ . . . 33
2.8.1 Lo stimatore di massima verosimiglianza . . . 33
2.8.2 Intervalli di confidenza . . . 34
2.8.3 Verifica di ipotesi . . . 35
2.9 Il principio di verosimiglianza . . . 37
2.10 Eliminazione dei parametri di disturbo . . . 39
2.11 La famiglia esponenziale . . . 41
2.12 Anomalie della funzione di verosimiglianza . . . 43
2.13 Esercizi . . . 44

3 Inferenza statistica da un punto di vista bayesiano . . . 47
3.1 Il teorema di Bayes e il processo induttivo . . . 47
3.2 La soggettività delle conclusioni . . . 50
3.2.1 La distribuzione a posteriori è il riassunto dell'inferenza . . . 51

3.3 La logica dell'induzione: evidenza, inferenza, decisioni ... 51
3.4 Alcune note tecniche ... 53
3.4.1 La costante di marginalizzazione ... 53
3.4.2 Alcuni aspetti matematici ... 54
3.5 Esercizi ... 54

4 Analisi di semplici modelli statistici ... 55
4.1 Dati dicotomici ... 55
4.2 Dati uniformi ... 56
4.3 La distribuzione gaussiana ... 59
4.3.1 Varianza nota ... 59
4.3.2 Media e varianza incognite ... 61
4.4 Modello di Poisson ... 63
4.5 Altri esempi notevoli ... 64
4.5.1 Confronto fra due proporzioni ... 64
4.5.2 Confronto fra due medie ... 66
4.6 La normale multivariata ... 68
4.7 Consistenza del metodo bayesiano ... 71
4.8 Esercizi ... 72

5 Scelta della distribuzione iniziale ... 73
5.1 Distribuzioni coniugate ... 76
5.2 Distribuzioni non informative ... 77
5.2.1 Notazione e motivazioni ... 78
5.2.2 La distribuzione uniforme ... 79
5.2.3 Il metodo di Jeffreys ... 80
5.2.4 Il metodo delle *reference priors* ... 84
5.3 La sensibilità delle conclusioni rispetto alla distribuzione a priori ... 90
5.3.1 Cenni al problema della robustezza ... 90
5.3.2 Il ruolo della dimensione campionaria ... 90
5.4 Esercizi ... 90

6 Procedure inferenziali bayesiane ... 91
6.1 Stima puntuale ... 91
6.2 Stima per intervallo ... 95
6.3 Verifica di ipotesi ... 98
6.3.1 Il caso di due ipotesi semplici ... 100
6.3.2 Il caso dell'ipotesi alternativa composta ... 102
6.3.3 Uso di distribuzioni improprie nei problemi di test ... 107
6.4 L'impostazione predittiva ... 109
6.4.1 Il concetto di sufficienza nell'impostazione predittiva ... 113
6.4.2 Calcoli predittivi ... 114
6.5 La modellizzazione gerarchica ... 116
6.5.1 L'approccio bayesiano empirico ... 117

6.6 Cenni alla teoria delle decisioni . . . 120
6.7 Esercizi . . . 124

7 Metodi computazionali . . . 125
7.1 Introduzione . . . 125
7.2 Approssimazioni analitiche . . . 127
7.2.1 Comportamento asintotico della distribuzione finale . . . 127
7.2.2 Metodo di Laplace . . . 129
7.2.3 Altri tipi di approssimazioni . . . 131
7.3 Simulazione a posteriori . . . 131
7.4 Data Augmentation . . . 135
7.5 Metodi Monte Carlo . . . 135
7.5.1 Campionamento per importanza . . . 135
7.5.2 Metodi accettazione-rifiuto . . . 140
7.5.3 Distribuzioni log-concave . . . 143
7.6 Algoritmi adattivi . . . 143
7.7 Metodi MCMC . . . 143
7.7.1 Aspetti matematici . . . 144
7.7.2 Gli algoritmi di tipo Metropolis-Hastings . . . 144
7.7.3 L'algoritmo di Gibbs . . . 148
7.7.4 Altri algoritmi . . . 152
7.7.5 Convergenza degli algoritmi MCMC . . . 153
7.8 Esercizi . . . 153

8 Scelta del modello statistico . . . 155
8.1 Introduzione . . . 155
8.2 Impostazione formale del problema . . . 158
8.3 Il fattore di Bayes . . . 159
8.3.1 Approssimazioni del fattore di Bayes . . . 161
8.3.2 Uso di distribuzioni non informative . . . 163
8.4 Metodi MC e MCMC . . . 167
8.4.1 Stima diretta della distribuzione marginale . . . 167
8.4.2 Il meta-modello . . . 168
8.4.3 L'algoritmo Reversible Jump . . . 169
8.5 Altre impostazioni . . . 170
8.5.1 Cross Validation . . . 170
8.6 Esercizi . . . 170

9 Il modello lineare . . . 171
9.1 Analisi bayesiana coniugata . . . 172
9.2 Il caso non informativo . . . 173
9.3 Regioni di credibilità. . . . 175
9.4 Regressione lineare attraverso metodi di simulazione . . . 176
9.4.1 Regressione lineare con errori a code pesanti . . . 177

9.5 Confronto tra modelli di regressione alternativi . . . 179
9.5.1 Il fattore di Bayes per modelli lineari . . . 179
9.5.2 Il calcolo della marginale di $\mathbf{y}$ . . . 179
9.5.3 Uso delle g-priors . . . 180
9.6 Previsioni . . . 181
9.7 Esercizi . . . 182

10 Modelli lineari generalizzati . . . 183
10.1 Introduzione ed esempi . . . 183
10.2 Distribuzioni a priori . . . 183
10.3 Tecniche di calcolo . . . 183
10.4 Alcune esemplificazioni . . . 183
10.4.1 Dati dicotomici . . . 183
10.4.2 Dati di conteggio . . . 183
10.4.3 sopravvivenza . . . 183
10.5 Esercizi . . . 183

11 I modelli gerarchici . . . 185
11.1 Introduzione . . . 185
11.2 Modelli gerarchici . . . 186
11.2.1 Strategie per l'analisi dei modelli gerarchici . . . 186
11.3 Il modello gerarchico gaussiano . . . 188
11.3.1 Il caso EB . . . 188
11.3.2 L'approccio HB . . . 189
11.3.3 Sulla scelta della distribuzione a priori di τ^2 . . . 191
11.4 Il calcolo dei momenti a posteriori . . . 192
11.4.1 Media e varianza dei θ_j . . . 192
11.5 Le stime finali . . . 193
11.5.1 La Strategia EB . . . 193
11.6 Approccio basato sulla simulazione . . . 198
11.7 Conclusioni . . . 198
11.8 Esercizi . . . 198

12 Approfondimenti . . . 199
12.1 Modelli a struttura latente . . . 199
12.1.1 Mistura finita di distribuzioni gaussiane . . . 199
12.1.2 Frontiera stocastica . . . 199
12.2 Il problema della stima della numerosità di una popolazione . . . 199
12.3 Scelta della numerosità campionaria . . . 199
12.4 Esercizi . . . 199

A Alcune nozioni di algebra lineare . . . 201

B Nozioni di probabilità . . . 203
B.1 Funzione generatrice dei momenti . . . 208
B.2 Convergenza di variabili aleatorie . . . 208

C Alcuni risultati e dimostrazioni . . . 209
C.1 Statistiche d'ordine . . . 209
C.2 Alcuni approfondimenti matematici . . . 210
C.2.1 Derivazione della distribuzione di Jeffreys . . . 210
C.3 Sulla scambiabilità . . . 210
C.3.1 Dimostrazione del Teorema 6.1 . . . 210
C.4 Sulle forme quadratiche . . . 212
C.4.1 Combinazione di due forme quadratiche . . . 212
C.5 Sul calcolo delle distribuzioni non informative nel modello lineare . . . 213
C.6 Sul calcolo della marginale per un modello lineare . . . 213

D Catene di Markov . . . 215
D.1 Catene in tempo discreto . . . 215
D.1.1 Distribuzione del processo ad un tempo prefissato . . . 216
D.1.2 Probabilità di assorbimento . . . 217
D.1.3 Tempi di arresto e proprietà forte di Markov . . . 218
D.1.4 Classificazioni degli stati . . . 218
D.1.5 Distribuzioni invarianti . . . 220
D.1.6 Equilibrio di una catena . . . 221
D.1.7 Reversibilità . . . 222
D.2 Catene continue . . . 222

E Le principali distribuzioni di probabilità . . . 223
E.1 Distribuzioni discrete . . . 224
E.2 Distribuzioni assolutamente continue . . . 226
E.3 Distribuzioni multivariate . . . 230

Soluzioni . . . 235

Riferimenti bibliografici . . . 237

Indice analitico . . . 241

Parte I

Titolo della parte

1

Teorema di Bayes e probabilità soggettiva

1.1 Il teorema di Bayes.

E' noto che, dati due eventi qualsiasi F e E, la probabilità dell'intersezione $F \cap E$ si può scrivere

$$P(F \cap E) = P(F|E)P(E), \tag{1.1}$$

oppure

$$P(F \cap E) = P(E|F)P(F). \tag{1.2}$$

Uguagliando la (1.1) con la (1.2) ed esplicitando rispetto a $P(F \mid E)$ si può scrivere, quando $P(E) > 0$,

$$P(F \mid E) = \frac{P(F)P(E \mid F)}{P(E)}, \tag{1.3}$$

La formula (1.3) rappresenta la forma più semplice del cosiddetto **teorema di Bayes**, dal nome di colui che, apparentemente per primo [5], utilizzò una versione leggermente più complessa dell'espressione stessa: essa insegna che la probabilità di un evento F non è una caratteristica intrinseca dell'evento, ma va calcolata sulla base delle informazioni a disposizione: il verificarsi di E, ad esempio, modifica la probabilità di F, e la trasforma in $P(F|E)$, secondo la (1.3).

Esempio 1.1

La mia collezione di CD è costituita da un 70% di dischi tradizionali e da un 30% di dischi contenenti file MP3. Tra i dischi tradizionali il 30% contiene musica rock mentre il restante 70% contiene brani di musica classica. Tra i dischi contenenti files MP3, il 10% contiene musica classica e il 90% musica rock. Scegliamo a caso un disco e sia A è l'evento { *il disco estratto è di tipo tradizionale*}, mentre R rappresenta l'evento { *il disco estratto contiene musica rock* }. Ovviamente si avrà $\Pr(A) = 0.7$; ma se dopo alcuni secondi mi rendo conto che si tratta di un disco rock, la probabilità che si tratti di un disco tradizionale diventa

$$\begin{aligned}\Pr(A \mid R) &= \frac{\Pr(A)\Pr(R \mid A)}{\Pr(R)} = \frac{\Pr(A)\Pr(R \mid A)}{\Pr\left(\bar{A}\right)\Pr\left(R \mid \bar{A}\right) + \Pr(A)\Pr(R \mid A)} \\ &= \frac{0.7 \times 0.3}{0.7 \times 0.3 + 0.3 \times 0.9} = \frac{21}{48}.\end{aligned}$$

♢

Esempio 1.2

♢

Esempio 1.3

♢

Esempio 1.4

Da un mazzo di 52 carte se ne estrae una a caso senza osservarla; se ne estrae poi una seconda che risulta essere un *Asso*. Qual è la probabilità che la prima carta estratta fosse un *Re*?
Soluzione. In questo caso identifichiamo F con l'evento $\{$*La prima carta è un Re*$\}$ e con E l'evento $\{$*La seconda carta è un Asso*$\}$. Poich $P(F) = 4/52$, $P(E) = 4/52$ (non conoscendo l'esito della prima estrazione, tutte le carte hanno la stessa probabilità di comparire come seconda carta) e $P(E|F) = 4/51$, si ha in conclusione

$$P(F \mid E) = \frac{4}{52}\frac{4}{51}/\frac{4}{52} = \frac{4}{51}.$$

Potrebbe risultare contro intuitivo il fatto che $P(E) = 4/52$ o, più in generale, che le probabilità relative alla seconda estrazione risultino uguali a quelle relative alla prima; ma quello che conta non è tanto il susseguirsi temporale degli eventi quanto l'informazione che si ha su di essi: se non conosciamo l'esito della prima estrazione al momento di calcolare la probabilità di eventi relativi alla seconda estrazione, è come se la prima se non si fosse mai verificata. Dal punto di vista matematico si può arrivare facilmente al risultato osservando che, chiamando A l'evento $\{$*La prima carta è un Asso*$\}$

$$\begin{aligned}\Pr(E) &= \Pr(E \cap A) + \Pr(E \cap A^c) = \Pr(A)\ \Pr(E \mid A) + \Pr(A^c)\ \Pr(E \mid A^c) \\ &= \frac{4}{52}\frac{3}{51} + \frac{48}{52}\frac{4}{51} = \frac{4}{52}\end{aligned}$$

♢

Esempio 1.5
Sugli aerei esiste una spia luminosa che si accende in fase di atterraggio quando il carrello non fuoriesce regolarmente. Può succedere però che la spia si illumini anche se il carrello non ha avuto alcun problema. Sia A l'evento $\{$*Carrello in ordine* $\}$ e sia B l'evento $\{$*Spia accesa*$\}$. È noto, da indagini di laboratorio, che

$$\Pr(B \mid A) = 0.005, \quad \Pr(B \mid A^c) = 0.999;$$

in altri termini la spia si accende erroneamente solo cinque volte su 1000 mentre non si accende quando dovrebbe soltanto una volta su 1000. Infine le statistiche di bordo riportano che la frequenza relativa di volte in cui il carrello non ha funzionato correttamente è pari al 3%. Calcolare la probabilità che, in caso di spia accesa, si tratti di un falso allarme.
Soluzione: Dalle informazioni di bordo sappiamo che $P(A) = .97$; si tratta di calcolare $P(A \mid B)$:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(A^c)P(B \mid A^c)} = \frac{0.97 \times 0.005}{0.97 \times 0.005 + 0.03 \times 0.999} = 0.139.$$

♢

Un modo efficace di interpretare la formula di Bayes è quello di considerare l'evento E come un insieme di sintomi (effetti) e l'evento F come una possibile malattia (causa) associata a tali sintomi.

Esempio 1.6 [*Possibili cause di un sintomo*]

Tizio si reca dal medico perch ha notato alcuni strani puntini rossi sulla sua cute (E=*insorgenza di puntini rossi*). Tizio non sa a quali cause far risalire tali sintomi. Il medico sostiene che le possibili cause sono tre: un banale fungo della pelle (F_1), varicella (F_2), una grave malattia (F_3). Per semplicità assumiamo che una e una sola delle tre cause possa aver effettivamente agito. Il medico sa anche quanto è verosimile osservare E quando si è malati di F_1, F_2, oppure F_3. Infatti studi precedenti indicano che $P(E \mid F_1) = 0.5$, $P(E \mid F_2) = 0.7$, mentre $P(E \mid F_3) = 0.99$. In pratica, in presenza del fungo, si ha una probabilità su due di osservare i puntini rossi, mentre, nel caso della grave malattia (F_3) l'insorgenza dei puntini è pressoch certa. E' il caso che Tizio si preoccupi? ♢

Soluzione. Prima di iniziare a preoccuparsi, è bene che Tizio calcoli, secondo la formula di Bayes, le probabilità a posteriori delle tre possibili malattie. Per fare questo però occorrono le probabilità a priori che Tizio, non essendo un esperto del settore, non conosce: il medico, che assumiamo esperto, sostiene che, nella città di Tizio l'insorgenza di F_1, soprattutto in quella stagione, è molto comune mentre le altre due malattie hanno una scarsa diffusione: egli quantifica tali valutazioni nelle seguenti probabilità:

$$P(F_1) = 0.7 \quad P(F_2) = 0.2 \quad P(F_3) = 0.1$$

Va notato che la somma delle tre probabilità sopra assegnate è 1: infatti stiamo assumendo che una e una sola causa abbia veramente agito. Non vi è invece alcun motivo per cui le tre probabilità condizionate assegnate precedentemente (le $P(E|F_i), i = 1, 2, 3$) sommino a 1. Alla luce di questi dati la probabilità che Tizio sia affetto da F_3 è

$$P(F_3|E) = \frac{P(F_3)P(E|F_3)}{P(E)} = \frac{0.1 \times 0.99}{P(E)} = \frac{0.099}{P(E)}. \tag{1.4}$$

Allo stesso modo

$$P(F_2|E) = \frac{P(F_2)P(E|F_2)}{P(E)} = \frac{0.7 \times 0.2}{P(E)} = \frac{0.14}{P(E)}, \tag{1.5}$$

$$P(F_1|E) = \frac{P(F_1)P(E|F_1)}{P(E)} = \frac{0.5 \times 0.7}{P(E)} = \frac{0.35}{P(E)}. \tag{1.6}$$

Pur senza calcolare $P(E)$, siamo in grado di tranquillizzare Tizio. Infatti,

$$\frac{P(F_1|E)}{P(F_3|E)} = \frac{0.35}{0.099} = 3.\overline{53}$$

e

$$\frac{P(F_1|E)}{P(F_2|E)} = \frac{0.35}{0.14} = 2.5.$$

In pratica la presenza del fungo è 3 volte e mezzo più probabile della malattia F_3 e 2 volte e mezzo più probabile della varicella. Se poi vogliamo calcolare le effettive probabilità a posteriori occorre calcolare $P(E)$. Questo si può fare in due modi, ovviamente equivalenti.

(a) Metodo formale: perch E si verifichi, deve verificarsi uno tra i tre eventi F_i; quindi

$$E = (E \cap F_1) \cup (E \cap F_2) \cup (E \cap F_3);$$

essendo poi le cause incompatibili,

$$\begin{aligned} P(E) &= P(E \cap F_1) + P(E \cap F_2) + P(E \cap F_3) \\ &= P(F_1)P(E|F_1) + P(F_2)P(E|F_2) + P(F_3)P(E|F_3) \\ &= 0.589 \end{aligned} \tag{1.7}$$

(b) Metodo più semplice: dalle formule (1.4),(1.5) e (1.6) si evince che $P(E)$ non è altro che un fattore di normalizzazione delle tre quantità suddette, necessario affinch la loro somma sia 1. Basta quindi sommare le tre quantità, uguagliare il risultato a 1 ed esplicitare rispetto a $P(E)$.

Per concludere, viene fornita una versione più formale del teorema di Bayes.

Teorema 1.1 (Teorema di Bayes) . *Sia E un evento contenuto in $F_1 \cup F_2 \cup \cdots \cup F_k$, dove gli F_j, $j = 1, \cdots, k$ sono eventi a due a due incompatibili (il verificarsi di uno di essi esclude la possibilità che se ne possa verificare un altro). Allora, per ognuno dei suddetti F_j vale la seguente formula*

$$P(F_j|E) = \frac{P(F_j)P(E|F_j)}{\sum_{i=1}^{k} P(F_i)P(E|F_i)}. \tag{1.8}$$

Dimostrazione 1.1 *Lasciata per esercizio*

La dimostrazione del teorema è molto semplice nel caso in cui il numero di eventi incompatibili $F_1, \cdots, F_k$ risulti finito. Qualora essi rappresentino un'infinità numerabile, occorre un momento di zelo, e specificare che, nell'impostazione comune del calcolo delle probabilità, quella sistematizzata da Kolmogorov nel 1933, il teorema continua ad essere ugualmente valido; al contrario, nell'impostazione di de Finetti [32], la (1.7) non è più garantita e occorre assumere tale uguaglianza o condizioni che la implichino. Nel seguito, salvo avviso contrario, ci muoveremo nell'ambito dell'impostazione di Kolmogorov.

1.2 Probabilità a priori e verosimiglianze

Nella formula (1.8) il membro di sinistra prende il nome di *probabilità finale (o a posteriori)* dell'evento F_j: il termine finale sta a significare *dopo che è noto che si è verificato E*. Come già osservato, il denominatore del membro di destra della (1.8) è un semplice fattore di normalizzazione; nel numeratore, invece, compaiono due quantità: la $P(F_j)$ è la probabilità a priori dell'evento F_j (nell'esempio medico, rappresenta la probabilità che qualcuno sia affetto dalla malattia F_j indipendentemente dall'aver riscontrato o meno i sintomi E); la $P(E \mid F_j)$ rappresenta invece la *verosimiglianza* di F_j, ovvero la probabilità che si manifestino i sintomi E quando si è affetti dalla malattia F_j. La formula (1.8) fornisce così un modo sintetico di valutare il grado di incertezza che abbiamo sul verificarsi di un evento, basandoci sia sulle informazioni a priori che abbiamo riguardo l'evento stesso, sia su ulteriori conoscenze sopraggiunte, magari mediante un apposito test, come nell'esempio precedente.

Volendo confrontare le probabilità a posteriori di due tra le k possibili cause, ad esempio F_h e F_j si ha

$$\frac{P(F_h|E)}{P(F_j|E)} = \frac{P(F_h)}{P(F_j)} \frac{P(E|F_h)}{P(E|F_j)}.$$

A conferma di quanto osservato in precedenza, si vede che il rapporto delle probabilità a posteriori di due eventi è pari al prodotto dei due rapporti: $P(F_h)/P(F_j)$ è il rapporto *a priori* mentre il rapporto delle verosimiglianze $P(E|F_h)/P(E|F_j)$ viene spesso indicato con B e prende il nome di *fattore di Bayes*: esso rappresenta un indicatore di evidenza relativa per una possibile ipotesi F_h rispetto ad un' altra ipotesi F_j, basato esclusivamente sui fatti osservati (l'evento E) e non su valutazioni soggettive sul verificarsi degli eventi F_j, $j = 1, \cdots, k$. Un valore di B pari a 1 corrisponde al caso di eguale evidenza per le due ipotesi a confronto.

1.3 L'impostazione soggettiva della probabilità

E' bene chiarire subito un aspetto essenziale: la probabilità non è una caratteristica intrinseca degli eventi per i quali viene calcolata bensì può dipendere dalla percezione che l'individuo ha degli eventi stessi. Quando si lancia una moneta presa a caso da un salvadanaio, siamo tutti pronti a sostenere che la probabilità che la moneta dia testa (T) sia pari a 0.5: in realtà, a voler essere pignoli, avremmo dovuto verificare che la moneta fosse regolare (che, ad esempio, non fosse una moneta con due teste!) e che non presentasse vistose alterazioni.

Allo stesso modo ci appare naturale, estraendo a caso una pallina da un'urna che ne contiene 10 rosse (R) e cinque blu (B), che la probabilità che la pallina estratta sia B sia posta pari a 1/3. Ma se chiediamo ad un gruppo di persone di valutare la probabilità che la squadra di calcio A superi la squadra B nella prossima partita di campionato, è verosimile aspettarci tante differenti risposte e nessuno trova da ridire sul fatto che un tifoso della squadra A reputi più probabile l'evento $\{$*vittoria della squadra A*$\}$ rispetto, ad esempio, ad un tifoso della squadra B.

E' giustificabile tutto ciò? Esistono casi in cui la probabilità è soggettiva (variabile da individuo a individuo) ed altri in cui è invece uguale per tutti? Certamente no.

La probabilità che un individuo associa ad un evento è *sempre* soggettiva: essa rappresenta il grado di fiducia che l'individuo pone nel verificarsi dell'evento. Essa si colloca dunque, non già all'interno dell'evento bensì tra l'individuo e il mondo esterno: è dall'interazione che scaturisce tra l'individuo e l'evento, dall'interesse che per l'individuo suscita l'evento che nasce la valutazione della probabilità (si veda [25]).

Risulta allora del tutto normale che individui differenti, di fronte al lancio di una moneta, in assenza di particolari informazioni sulla moneta stessa, concordino nel sostenere che, non foss'altro per ragioni di simmetria, la probabilità che la moneta dia T è uguale alla probabilità che la moneta dia C e quindi entrambe valgano 0.5. Ma la partita di calcio è un qualcosa di ben più complesso e ciascun individuo, con le sue informazioni e le sue distorsioni (tifo, pregiudizi, superstizioni, etc..) finirà con l'associare all'evento *vince la squadra A* una probabilità differente dagli altri.

Una prima conseguenza della soggettività della probabilità è che non esiste una probabilità *corretta*, se non forse in alcuni casi speciali. Anche se, come abbiamo visto, un gran numero di persone concorda nell'assegnare probabilità 0.5 all'evento $\{$*la moneta dà T*$\}$, non esiste alcun meccanismo fisico per "verificare" tale valutazione e non servirebbero nemmeno un gran numero di prove ripetute per eliminare il dubbio che la probabilità di T sia 0.5001 e non 0.5.

Il fatto che non esista una probabilità *corretta* per un dato evento, non ci autorizza però ad associare agli eventi probabilità scelte a caso: pur nella soggettività delle valutazioni, le probabilità debbono soddisfare alcune condizioni di *coerenza*.

Negli anni '20 e '30, B. de Finetti, con una serie di scritti (si vedano, ad esempio [32], e [36]), gettò le basi per la costruzione della teoria soggettiva della probabilità: a tal fine egli utilizzo lo schema teorico, e il linguaggio, delle scommesse. Nel prossimo paragrafo verrà illustrata tale impostazione arrivando così alla definizione soggettiva di probabilità: inoltre, attraverso la condizione di coerenza, verranno riottenuti quei postulati che altre teorie della probabilità introducono in modo esogeno. La profonda influenza che la figura di Bruno de Finetti tuttora esercita nella probabilità e nella statistica possono essere apprezzati appieno mediante la lettura dei suoi due volumi, [33], apparsi poi in lingua inglese in [34] e [35].

1.4 Definizione e condizione di coerenza

Prima di addentrarci nel linguaggio delle scommesse, è bene chiarire che cosa si intende per evento.

Definizione 1.1 *Un evento è un ente logico che può assumere solo due valori: vero (V) o falso (F). Inoltre la situazione sperimentale deve essere tale per cui, una volta effettuata la prova, si è in grado di verificare se l'evento si sia manifestato come V oppure come F.*

Ad esempio, la proposizione {*La squadra A vincerà il campionato nel 2010*} è un evento, che potrà essere dichiarato vero o falso nel mese di giugno del 2010. Al contrario, la proposizione {*La tal moneta dà Testa con probabilità 0.5* } non rappresenta un evento perch non siamo in grado di verificarne la verità o meno: È un evento invece il seguente {*Nei prossimi dieci lanci, la tal moneta fornirà 3 T e 7 C*}:

Possiamo ora dare la definizione di probabilità [25]:

Definizione 1.2 *La probabilità di un evento E, per un dato individuo, in un certo momento della sua vita, è il prezzo $P(E) = p$ che egli ritiene giusto pagare (o ricevere da uno scommettitore) per partecipare ad una scommessa in cui vincerà (o pagherà)* 0 *se E non si verifica oppure* 1*, qualora E si verifichi.*

È importante sottolineare che l'individuo deve produrre lo stesso valore di p sia nelle vesti di scommettitore che nel ruolo del *Banco*. Se ad esempio l'evento su cui scommettiamo è $A = \{$*vince la squadra A*$\}$ e Tizio ritiene che $p = P(A) = 0.4$ allora Tizio deve essere disposto a

- pagare 0.4 per ricevere 1 in caso di vittoria di A (e 0 altrimenti)
  oppure
- pagare 0.6 per ricevere 1 in caso di mancata vittoria di A (e 0 altrimenti)

C'è da notare che in questo modo la valutazione della probabilità non dipende dall'entità della posta in palio in quanto tutti ragionamenti fin qui esposti funzionano ugualmente se le poste vengono moltiplicate per una somma S. Abbiamo già detto che la probabilità è soggettiva ma deve rispettare una condizione di "coerenza".

Definizione 1.3 *. Una valutazione di probabilità sugli n eventi $E_1, E_2, \cdots, E_n$ si dice coerente se nessuna combinazione di scommesse sugli eventi consente una vincita certa (indipendentemente dagli eventi E_i, $i = 1, \cdots, n$, che si verificano effettivamente).*

Esempio 1.7
Consideriamo il caso di una corsa a cui partecipano n cavalli, e siano $p_1, p_2, \cdots, p_n$ le probabilità di vittoria assegnate agli n cavalli. Consideriamo il caso in cui

$$p_1 + p_2 + \cdots + p_n = C < 1;$$

Allora è sufficiente scommettere una posta S su ogni cavallo partecipante alla gara per garantirsi una vincita certa. Infatti la quota pagata per partecipare alle scommesse sarà

$$p_1 S + p_2 S + \cdots + p_n S = CS < S$$

a fronte di una vincita certa pari a S (un cavallo vincerà certamente). ♢
Nella definizione di probabilità non è espressamente richiesto che la probabilità di un evento debba essere un numero compreso tra 0 e 1. Questo vincolo emerge naturalmente se però vogliamo che la nostra probabilità sia coerente. Infatti

Teorema 1.2 *Condizione necessaria e sufficiente affinch $P(E)$ sia coerente è che*

$$0 \leq P(E) \leq 1$$

In particolare, se $P(E) = 0$, l'evento è impossibile, se $P(E) = 1$, l'evento si dice certo.

Dimostrazione 1.2 *Sia $p = P(E)$ e assumiamo di scommettere una posta S sul verificarsi di E. Quando E si verifica il guadagno ottenuto dalla scommessa è $W(E) = S - pS = S(1-p)$. Quando E non si verifica si ha invece $W(\bar{E}) = -pS$. Se prendiamo $p < 0$, allora basta scommettere una quantità S positiva per garantirci una vincita sicura. Se invece prendiamo $p > 1$, sarà sufficiente prendere una posta S negativa (ovvero, invertire i termini della scommessa) per garantirci una vincita certa.*

Ne segue che $0 \leq P(E) \leq 1$. Inoltre, se l'evento E è certo si avrà certamente $W(E) = (1-p)S$ e, per non avere vincite certe, deve per forza essere $W(E) = 0$, da cui $p = 1$; allo stesso modo si verifica che p deve essere 0 nel caso di eventi impossibili.

È possibile derivare, attraverso la condizione di coerenza tutte le più familiari regole del calcolo delle probabilità, come ad esempio il teorema delle probabilità totali.

Meritano un discorso a parte le probabilità condizionate che, nell'impostazione soggettiva, sono considerate vere e proprie probabilità ma riferite ad eventi subordinati (del tipo $E_1 \mid E_2$): in termini di scommesse la probabilità condizionata $P(\cdot \mid \cdot)$ si definisce esattamente come nel caso precedente quando E_2 si verifica, mentre non si procede alla scommessa (non si valuta la probabilità) se, al contrario, non si verifica E_2.

Esempio 1.8
In una sala scommesse si accettano scommesse sull'esito dell'incontro di calcio tra la squadra A e la squadra B. Gli esperti sostengono che il giocatore *Pallino* è molto importante per la squadra A, le cui probabilità di vittoria sono molto diverse con Pallino in campo o meno. Siano E_1 l'evento

{Vince la squadra A} e E_2 l'evento { Pallino gioca}. Uno scommettitore può decidere di pagare un prezzo p per partecipare ad una scommessa relativa all'evento $E_1 \mid E_2$. In questo caso gli esiti possibili della scommessa sono:

- Gioca Pallino e la squadra A vince: Tizio incassa 1;
- Gioca Pallino e la squadra A perde: Tizio incassa 0;
- Non gioca Pallino: la scommessa è annullata e a Tizio viene restituita la posta p

♢

Dalla precedente definizione di probabilità condizionata discendono direttamente, attraverso la condizione di coerenza, la legge delle probabilità composte così come il Teorema di Bayes.

Problemi

1.1. Ogni giorno Mario tenta di comprare il quotidiano. Egli prova di mattina (M) con probabilità 1/3, di sera (S) con probabilità 1/2 oppure si dimentica del tutto (D) con probabilità 1/6. La probabilità di trovare effettivamente il giornale (G) è pari a 0.9 se va di mattina, 0.2 se va di sera e, ovviamente 0 se non va affatto.

Una sera torna a casa e la moglie vede che Mario ha effettivamente comprato il giornale. Qual è la probabilità che lo abbia comprato di mattina?

1.2. Una certa specie di criceti può nascere con il manto nero o marrone a seconda dell'associazione tra due geni ognuno dei quali può assumere il valore A oppure B. Se i due geni sono simili (AA oppure BB) il criceto è omozigote, altrimenti è detto eterozigote. Il criceto nasce marrone solo se è omozigote di tipo AA. Il figlio di una coppia di criceti porta con s i due geni, uno da ogni genitore: se il genitore è eterozigote il gene ereditato è A o B con la stessa probabilità; se il parente è omozigote, con probabilità pari a 1, trasmette il suo unico gene. Supponiamo che un criceto nero sia nato da una coppia di due eterozigoti.

(a) Qual è la probabilità che questo criceto sia omozigote?

Supponiamo ora che tale criceto sia poi accoppiato ad una cricetina marrone e che tale accoppiamento produca 7 figli, tutti neri

(b) Usa il teorema di Bayes per determinare la nuova probabilità che lo stesso criceto risulti omozigote.

1.3. Ogni mattina il lattaio ci lascia sulla porta di casa una bottiglia di latte. Egli riceve forniture in eguale misura dalle centrali di Roma e Latina ed ogni mattina sceglie a caso la bottiglia che ci lascia. Il latte di Roma raggiunge l'ebollizione in un tempo in minuti che può considerarsi una v.a. $N(2, 3)$ mentre quello di Latina ha un tempo di ebollizione pari ad una v.a. $N(2.5, 4)$. Una certa mattina cronometriamo il tempo necessario all'ebollizione del latte appena ricevuto e registriamo 2 minuti e 18 secondi. Qual è la probabilità che si tratti di latte di Roma?

1.4. Dimostrare il Teorema 1.1.

1.5. Ogni individuo appartiene ad uno dei quattro gruppi sanguigni O (si legge "zero"); A; B; AB. In una popolazione le frequenze dei quattro gruppi sono rispettivamente $\pi_O; \pi_A; \pi_B; \pi_{AB}$. Per poter eseguire una trasfusione di sangue da un donatore a un ricevente occorre seguire regole specifiche: O può ricevere solo da O; A può ricevere da O e da A; B può ricevere da O e da B; AB può ricevere da O, da A, da B e da AB. Si dice anche che il gruppo O è donatore universale e il gruppo AB è ricevente universale. Si estraggono a caso un donatore e un ricevente. Calcolare

(a) la probabilità che la trasfusione sia possibile;
(b) la probabilità che il ricevente sia di gruppo AB sapendo che la trasfusione è possibile.

1.6. Un test radiologico per la tubercolosi ha esito incerto: la probabilità che il test risulti positivo su un malato è $1-\beta$; la probabilità che il test risulti positivo su un non malato è invece pari ad α. La frequenza relativa o *prevalenza* di malati nella popolazione è pari a γ. Un individuo, selezionato a caso nella popolazione e sottoposto a test, risulta positivo. Qual è la probabilità che egli sia sano?

1.7. L'urna U_1 contiene 1 pallina bianca e n_1 palline nere; l'urna U_2 contiene n_2 palline bianche e 1 nera. Si estrae a caso una pallina dall'urna U_1 e la si mette nell'urna U_2; poi si estrae a caso una pallina dall'urna U_2 e la si mette nell'urna U_1. Trovare la distribuzione di probabilità del numero finale di palline bianche nell'urna U_1.

1.8. Durante un intero anno, il numero di raffreddori che un individuo contrae può essere considerato una v.a. X con distribuzione di Poisson di parametro 5. Viene immessa sul mercato una nuova medicina: essa risulta efficace sul 75% della popolazione, e per tali persone il numero di raffreddori contratti in un anno, condizionatamente all'uso della medicina, è una v.a. di Poisson di parametro 3. Sul restante 25% della popolazione la medicina è inefficace. Se un individuo a caso prende la medicina e in un anno ha due raffreddori, qual è la probabilità che appartenga alla categoria di persone su cui la medicina ha effetto?

1.9. Il 10% della popolazione soffre di una seria malattia. Ad un individuo estratto a caso vengono somministrati due test diagnostici indipendenti. Ciascuno dei due test fornisce una diagnosi corretta nel 90% dei casi. Calcolare la probabilità che l'individuo sia effettivamente malato nelle due ipotesi alternative:

(a) entrambi i test siano positivi;
(b) un solo test sia positivo.

1.10. In una fabbrica di bibite, le bottiglie che essa stessa produce vengono sottoposte a un controllo prima di essere riempite. Il 30% delle bottiglie prodotte sono difettose. La probabilità che l'ispettore si accorga che una bottiglia è difettosa, e quindi la scarti, è 0.9. Mentre la probabilità che l'ispettore giudichi erroneamente difettosa una bottiglia buona è 0.2. Qual è la probabilità che una bottiglia scartata sia difettosa? E la probabilità che una bottiglia giudicata buona sia invece difettosa?

1.11. La moneta M_1 dà testa con probabilità 0.3, la moneta M_2 con probabilità 0.5 e la moneta M_3 con probabilità 0.7. Viene scelta a caso una moneta e lanciata finché non si ottiene testa per la seconda volta. Sapendo che la seconda testa si è avuta al quinto tentativo, stabilire quale delle monete ha la probabilità più alta di essere stata lanciata.

1.12. Si sappia che le donne in una specifica famiglia possono essere portatrici di emofilia con probabilità 0.5. Se la madre è portatrice, allora i suoi figli maschi, indipendentemente l' uno dall' altro, possono essere emofiliaci, ciascuno con probabilità 0.5. Se la madre non è portatrice, allora i figli maschi non sono emofiliaci.

(a) Se il primo figlio maschio di una donna nella famiglia non è emofiliaco, qual è la probabilità che anche il secondo non sia emofiliaco?

(b) Se i primi due figli maschi di una donna della famiglia non sono emofiliaci, qual è la probabilità che la madre sia portatrice di emofilia?

2

Modello statistico e funzione di verosimiglianza

Questo capitolo va considerato come un breve compendio di inferenza classica che si prefigge due obiettivi principali:

- introdurre i concetti e gli strumenti matematici, con relativa notazione, che costituiscono la base del metodo inferenziale e che vengono utilizzati sia in ambito classico che in ambito bayesiano;
- rendere la lettura di questo testo il più possibile indipendente da nozioni di inferenza statistica preliminari.

È evidente però che quanto segue in questo capitolo non può considerarsi esauriente per una competenza nelle discipline del calcolo di probabilità e della statistica classica. Il lettore interessato può consultare, ad esempio, [30] oppure [26] per una trattazione esauriente dei fondamenti del calcolo delle probabilità e [2] o [66] per quanto concerne l'inferenza non bayesiana.

2.1 Gli ingredienti di un modello statistico

Nel linguaggio comune un esperimento statistico viene percepito come l'osservazione parziale di un fenomeno quantitativo, effettuata in modo da poter trarre informazioni anche sulla parte non osservata. Tale percezione, troppo vaga, deve essere formalizzata in modo chiaro e privo di ambiguità. Cominciamo allora a definire lo spazio dei possibili risultati di un esperimento, ovvero l'insieme di tutte le possibili realizzazioni numeriche relative alla misurazione di un determinato fenomeno.

Definizione 2.1 *Si definisce $\mathcal{X}$ l'insieme di tutti i possibili risultati osservabili in un esperimento.*

Esempio 2.1 [*Lancio di una moneta*]
Se l'esperimento consiste nel lancio di una moneta, i cui risultati possibili sono Testa (T) e Croce (C), si avrà $\mathcal{X} = \{T, C\}$; in genere si preferisce codificare i possibili risultati in modo numerico: ad esempio si potrebbe porre $T = 1$ e $C = 0$, cosicch $\mathcal{X} = \{1, 0\}$. Se la stessa moneta viene lanciata un numero $n \geq 1$ di volte, allora lo spazio $\mathcal{X}$ sarà formato da tutte le $n-$ple i cui elementi possono essere 0 oppure 1, ovvero

$$\mathcal{X} = \left\{ (\overbrace{0, 0, \cdots, 0, 0}^{n \text{ volte}}), (\overbrace{0, 0, \cdots, 0}^{n\text{-}1 \text{ volte}}, 1), \cdots, (\overbrace{1, 1, \cdots, 1, 1}^{n \text{ volte}}) \right\}.$$

In forma sintetica si può esprimere $\mathcal{X}$ come il prodotto cartesiano dell'insieme $\{0, 1\}$ per s stesso ripetuto n volte, ovvero $\mathcal{X} = \{0, 1\}^n$. ♢

Esempio 2.2 [*Tempo di attesa*]
Se invece l'esperimento consiste nel misurare, in minuti, il tempo di attesa che trascorriamo una certa mattina in banca prima che arrivi il nostro turno allo sportello, il risultato dell'esperimento potrà essere, in linea teorica qualunque valore reale positivo, cosicch $\mathcal{X} = \mathbb{R}^+$. ♢
Una volta definito l'insieme $\mathcal{X}$, consideriamo la variabile aleatoria X il cui supporto, l'insieme dei valori che può assumere, coincide con $\mathcal{X}$. Per definire un modello statistico occorre selezionare un insieme di leggi di probabilità, una delle quali si assume che sia la vera legge di probabilità di X.

Definizione 2.2 *Si definisce $\mathcal{P}$ la famiglia di tutte le possibili leggi di probabilità associabili alla variabile aleatoria X.*

Esempio 2.1 (continua). In questo caso X può assumere solo i valori 0 e 1. È ragionevole assumere allora che, fissato un valore $\theta \in [0,1]$, si abbia $P(X = 1;\theta) = \theta$ e, di conseguenza, $P(X = 0;\theta) = 1 - \theta$. In questa formalizzazione, θ gioca il ruolo di parametro incognito. In questo caso si avrà

$$\mathcal{P} = \{P(\cdot;\theta) : P(X = 1;\theta) = \theta, \theta \in [0,1]\};$$

in altri termini si assume per X un modello di tipo Bernoulliano, la cui distribuzione generica verrà indicata col simbolo $Be(\theta)$. ♢
Può accadere che il risultato d'interesse dell'esperimento non sia quello della variabile aleatoria X, bensì quello di una sua funzione $t(X)$.

Definizione 2.3 *Con riferimento allo spazio dei risultati $\mathcal{X}$ si chiama statistica ogni funzione*

$$t : \mathcal{X} \to \mathbb{R}^k, \quad k \geq 1,$$

che associa ad ogni punto $x \in \mathcal{X}$, una funzione a k valori

$$t(x) = (t_1(x), t_2(x), \cdots, t_k(x)).$$

Esempio 2.1 (continua). Supponiamo ora che la stessa moneta venga lanciata n volte e i lanci, condizionatamente al valore di θ, siano indipendenti. Questo schema è tra i più frequenti nella pratica statistica: può essere utilizzato tutte le volte in cui si effettuano prove ripetute di un esperimento che fornisce risposte dicotomiche (successo o insuccesso, favorevole o contrario, sopra o sotto una determinata soglia, etc.); quasi sempre, in questo tipo di esperimenti, la variabile aleatoria osservabile d'interesse è rappresentata da $Y = \{$*numero totale di successi*$\}$ o, se vogliamo, $Y = \{$*numero di* 1 *nella* $n-$*pla osservata*$\}$. Il modello naturale di riferimento è allora quello Binomiale, che indicheremo col simbolo $\text{Bin}(n,\theta)$: assumeremo cioè che lo spazio dei possibili risultati sia relativo alla variabile aleatoria Y, ovvero

$$\mathcal{Y} = \{0, 1, 2, \cdots, n\},$$

mentre la famiglia $\mathcal{P}$ è costituita da tutte le leggi di probabilità binomiali $Bin(n,\theta)$, con n fissato pari al numero di prove ripetute e $\theta \in [0,1]$,

$$\mathcal{P} = \left\{ p(\cdot;\theta) : P(Y = y;n,\theta) = \binom{n}{y} \theta^y \ (1-\theta)^{n-y}, \quad \theta \in [0,1] \right\}.$$

Riprenderemo queste idee nella §2.4 quando si introdurrà il concetto di sufficienza. ♢

Negli esempi precedenti il numero di leggi di probabilità in $\mathcal{P}$ è pari al numero dei punti che formano l'intervallo $[0,1]$; esiste cioè una corrispondenza biunivoca tra l'insieme $\mathcal{P}$ e l'intervallo chiuso $[0,1]$ che prende il nome di spazio parametrico.

Definizione 2.4 *Si definisce* **spazio parametrico***, e verrà indicato con il simbolo Ω, l'insieme dei valori assumibili dal parametro θ.*

Definizione 2.5 *Si definisce* **modello statistico** *e si indica col simbolo $\mathcal{E}$, la terna*

$$\mathcal{E} = (\mathcal{X}, \mathcal{P}, \Omega)\,. \tag{2.1}$$

Ogni volta che faremo riferimento ad un modello statistico, assumeremo implicitamente che il modello in questione sia *identificabile*.

Definizione 2.6 *Un modello statistico si dice* **identificabile** *se comunque consideriamo due misure di probabilità della famiglia $\mathcal{P}$, $P(\cdot, \boldsymbol{\Omega}_1)$ e $P(\cdot, \boldsymbol{\Omega}_2)$, con $\theta_1 \neq \theta_2$, è possibile individuare almeno un sottoinsieme $E \subset \mathcal{X}$ per il quale*

$$\Pr(E; \theta_1) \neq \Pr(E; \boldsymbol{\Omega}_2)\,. \tag{2.2}$$

Tutte le volte che Ω è rappresentabile come un sottoinsieme dello spazio euclideo $\mathbb{R}^k$, per qualche k intero, parleremo di modello *parametrico*; altrimenti si dice che il modello è *non parametrico*.

Esempio 2.3 [*Modello non parametrico*]
Sia X il tempo di durata di una certa lampadina e consideriamo, come possibili leggi di probabilità su $\mathcal{X} = (0, \infty)$, tutte quelle dotate di densità di probabilità decrescente in $\mathcal{X}$. In questo caso non è possibile individuare la singola legge di probabilità in $\mathcal{P}$ attraverso un numero finito di parametri: si tratta dunque di un problema di inferenza non parametrica. ♢

In questo testo ci occuperemo quasi esclusivamente di modelli parametrici: alcuni esempi di inferenza non parametrica secondo un approccio bayesiano verranno discussi nella §**??**. ?.?.

Una volta definito il modello statistico, viene concretamente effettuato l'esperimento statistico e la realizzazione $(X = x_0)$ viene utilizzata per estrarre informazioni su quale, tra le possibili leggi in $\mathcal{P}$, abbia realmente operato nel generare x_0.

2.2 La funzione di verosimiglianza

La trattazione che segue dovrebbe soffermarsi su alcuni aspetti matematici non del tutto trascurabili. Tuttavia per perseguire l'obiettivo di mantenere una certa agilità del testo, faremo delle assunzioni semplificatrici. Assumeremo allora che la famiglia $\mathcal{P}$ di leggi di probabilità che costituisce il modello statistico possa essere di due tipi:

- Tutte le leggi in $\mathcal{P}$ sono assolutamente continue, ovvero dotate di una funzione di densità $f(\cdot; \theta)$, non negativa su $\mathcal{X} \subset \mathbb{R}^n$, per qualche n e per ogni possibile valore di $\theta \in \Omega$.
- Tutte le leggi in $\mathcal{P}$ sono di tipo discreto, ovvero, per ogni $\theta \in \Omega$, i valori che la variabile aleatoria X assume con probabilità positiva sono al più un insieme numerabile. In questo caso la generica distribuzione di probabilità di X si indica col simbolo $p(\cdot; \theta)$.

Per ulteriori approfondimenti su tali aspetti si possono consultare diversi testi che approfondiscono a diversi livelli l'argomento. Suggeriamo [30] per gli aspetti probabilistici e [2] o [68] per le implicazioni inferenziali.

Assumere un modello statistico corrisponde a limitare la scelta fra le possibili leggi aleatorie che descrivono il fenomeno osservabile alla famiglia di distribuzioni $\mathcal{P}$ o, analogamente, all'insieme delle "etichette" Ω. Occorre ora stabilire in che modo il risultato osservato dell'esperimento ($X = x_0$) fornisca supporto ai diversi elementi di Ω. Consideriamo il seguente esempio binomiale.

Esempio 2.4 [*Verosimiglianza binomiale*]
Una moneta dà Testa (oppure il valore 1) con probabilità incognita θ; essa viene lanciata $n = 10$ volte e i lanci possono essere considerati indipendenti condizionatamente al valore di θ. Per semplicità di esposizione supponiamo che θ possa assumere i soli valori $\Omega = \{0, 0.1, 0.2, \cdots, 0.9, 1\}$. Il risultato dell'esperimento, ovvero il numero di Teste ottenute in dieci lanci, è allora, per ogni valore fissato di θ, una variabile aleatoria di tipo $Bin(10, \theta)$. *Prima* di osservare il risultato dell'esperimento è possibile elencare, per ogni $\theta \in \Omega$, la legge di probabilità di Y. La righe della Tabella 2.1 mostrano tutte le possibili leggi di probabilità della variabile aleatoria Y secondo i diversi valori di θ.

	$\Pr(Y = y)$										
θ	0	1	2	3	4	5	6	7	8	9	10
0	1	0	0	0	0	0	0	0	0	0	0
0.1	0.348	0.387	0.193	0.057	0.011	0.001	0	0	0	0	0
0.2	0.107	0.268	0.302	0.201	0.088	0.026	0.005	0.0007	0	0	0
0.3	0.028	0.121	0.233	0.267	0.200	0.103	0.037	0.009	0.001	0	0
0.4	0.006	0.040	0.121	0.215	0.251	0.201	0.111	0.042	0.010	0.001	0
0.5	0.0009	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.0009
0.6	0	0.001	0.010	0.042	0.111	0.201	0.251	0.215	0.121	0.040	0.006
0.7	0	0	0.001	0.009	0.037	0.103	0.200	0.267	0.233	0.121	0.028
0.8	0	0	0	0.0007	0.005	0.026	0.088	0.201	0.302	0.268	0.107
0.9	0	0	0	0	0	0.001	0.011	0.057	0.193	0.387	0.348
1	0	0	0	0	0	0	0	0	0	0	1

Tabella 2.1. Distribuzioni di probabilità di Y per diversi valori di θ

Supponiamo ora che l'esperimento fornisca il risultato $\{Y = y_0 = 7\}$. È ragionevole allora considerare, nella tabella, solo i valori della colonna corrispondente all'evento osservato[1] $\{Y = 7\}$ e interpretare come misure dell'evidenza che $\{Y = 7\}$ fornisce ai diversi valori di θ, le probabilità che aveva l'evento $Y = 7$ di verificarsi secondo i vari θ. In altre parole i valori della colonna della tabella relativa a $\{Y = 7\}$ ci dicono quanto sono *verosimili* i valori di θ alla luce del risultato osservato. Così, ad esempio, quando si osservano 7 Teste su 10 lanci, il fatto che la moneta sia regolare ($\theta = 0.5$) ha una verosimiglianza pari a 0.117, mentre l'ipotesi che la moneta sia distorta e fornisca testa nel 60% dei casi viene valutata con una verosimiglianza superiore, pari a 0.215.

[1] Alcune scuole inferenziali, prima fra tutte quella classica, basate sulla teoria di Neyman e Pearson, propongono metodi inferenziali che sono in chiaro contrasto con tale "ragionevole" considerazione; non approfondiremo qui tali aspetti fondazionali: il lettore interessato può consultare [68]

Detto in altro modo equivalente, il valore ($\theta = 0.6$) è

$$\frac{\Pr(Y = 7; 0.6)}{\Pr(Y = 7; 0.5)} = \frac{0.215}{0.117} = 1.838$$

volte più verosimile del valore $\theta = .0.5$. ♢

Tuttavia, nella pratica statistica, l'insieme Ω non è composto da un numero finito di possibili valori di θ e un approccio tabellare non è più possibile: la naturale estensione del ragionamento precedente conduce alla definizione della cosiddetta funzione di verosimiglianza [2].

Definizione 2.7 *Con riferimento al modello statistico (2.1), si chiama funzione di verosimiglianza associata al risultato $X = x_0$ la funzione $L : \Omega \to [0, \infty)$ che associa, ad ogni valore di $\theta \in \Omega$, la probabilità $p(X = x_0; \theta)$ (nel caso discreto) oppure la densità di probabilità $f(x_0; \theta)$ (nel caso assolutamente continuo).*

Esempio 2.4 (continua). Consideriamo ora il caso in cui Ω è l'intervallo chiuso $[0, 1]$. P er $n = 10$ e $y_0 = 7$ la funzione di verosimiglianza vale

$$L(\theta) = \Pr(Y = 7; \theta) = \binom{10}{7} \theta^7 (1 - \theta)^3, \tag{2.3}$$

e viene rappresentata nella Figura 2.1(b); nella Figura 2.1(d) viene considerato il caso con $n = 50$ e $y_0 = 35$.

Esempio 2.5 [*Verosimiglianza normale*]

Si osservano n replicazioni $(X_1, X_2, \cdots, X_n)$ di una variabile aleatoria $X \sim N(\mu, \sigma_0^2)$, che, per μ fissato, risultano indipendenti; il valore di σ_0^2 va considerato, per semplicità, noto. La realizzazione dell'esperimento consiste in un vettore di osservazioni $z_0 = (x_1, x_2, \cdots, x_n)$. La funzione di verosimiglianza è allora definita come

$$L(\mu) = f(z_0; \mu) = \prod_{j=1}^{n} f(x_j; \mu) = \prod_{j=1}^{n} \frac{1}{\sigma_0 \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma_0^2} (x_j - \mu)^2 \right\}.$$

Attraverso semplici elaborazioni algebriche si può scrivere, denotando con $\bar{x}$ la media campionaria osservata e con $s^2 = \sum_{j=1}^{n} (x_j - \bar{x})^2 / n$ la varianza campionaria osservata,

$$\begin{aligned} L(\mu) =& \frac{1}{\sigma_0^n (2\pi)^{n/2}} \exp\left\{ -\frac{1}{2\sigma_0^2} \sum_{j=1}^{n} (x_j - \mu)^2 \right\} \\ =& \frac{1}{\sigma_0^n (2\pi)^{n/2}} \exp\left\{ -\frac{1}{2\sigma_0^2} \sum_{j=1}^{n} (x_j - \bar{x} + \bar{x} - \mu)^2 \right\} \\ =& \frac{1}{\sigma_0^n (2\pi)^{n/2}} \exp\left\{ -\frac{1}{2\sigma_0^2} \left[\sum_{j=1}^{n} (x_j - \bar{x})^2 + n(\bar{x} - \mu)^2 \right] \right\}, \end{aligned}$$

da cui finalmente,

$$L(\mu) = \frac{1}{\sigma_0^n (2\pi)^{n/2}} \exp\left\{ -\frac{n}{2\sigma_0^2} \left[s^2 + (\bar{x} - \mu)^2 \right] \right\}. \tag{2.4}$$

La figura 2.1, nel riquadro (a) mostra il grafico della $L(\mu)$ nel caso particolare in cui $n = 10$, $\sigma_0^2 = 4$ e le osservazioni sono

$$z_0 = (2.71, 3.53, 3.76, 3.24, 2.73, 2.36, 1.66, 3.97, 2.89, 1.52),$$

con $\bar{x} = 2.84$ e $s^2 = 0.61$. Nel riquadro (c) è invece proposta la funzione di verosimiglianza per lo stesso contesto, ma ottenuta da un campione di $n = 50$ osservazioni che hanno fornito una media campionaria pari a $\bar{x} = 3.15$. ♢

L'esempio precedente suggerisce alcune considerazioni, di natura generake.

1. La funzione di verosimiglianza è definita a meno di una costante.

La funzione di verosimiglianza stabilisce un sistema di pesi *relativi* con cui viene misurata l'evidenza a favore dei vari valori di θ. Se la $L(\theta)$ viene moltiplicata per un termine $c(x_0)$ dipendente dal campione osservato, ma non dal parametro θ, il contributo informativo relativo resta inalterato. Ad esempio, nel precedente esempio, dove il parametro d'intereesse era la media μ, possono essere eliminati dalla (2.4) tutti i fattori che non coinvolgono μ e scrivere semplicemente

$$L(\mu) \propto \exp\left\{-\frac{n}{2\sigma_0^2}(\bar{x} - \mu)^2\right\}. \tag{2.5}$$

In alcuni casi per risolvere questa indeterminazione e, allo stesso tempo, avere a disposizione un indicatore di evidenza che assuma un ben preciso *range* di valori si preferisce utilizzare la versione relativa della funzione di verosimiglianza, $L_R(\theta)$, che si ottiene semplicemente dividendo $L(\theta)$ per il suo valore massimo, a patto che questo risulti finito: si ottiene così

$$L_R(\theta) = \frac{L(\theta)}{\sup_{\theta\in\Omega} L(\theta)}. \tag{2.6}$$

In questo modo si ottiene che $0 \leq L_R(\theta) \leq 1$, per ogni $\theta \in \Omega$, e $L_R(\theta)$ può a ben diritto essere considerata come un indice di evidenza sperimentale a favore di θ, basato sull'osservazione campionaria.

2. La funzione di verosimiglianza non è una distribuzione di probabilità.

Il sistema di pesi relativo costituito da $L(\theta)$, oppure da $L_R(\theta)$ non rappresenta una distribuzione di probabilità su Ω. Va sottolineato che, in una impostazione classica dell'inferenza, è il risultato sperimentale X e non il parametro θ ad essere considerato aleatorio. Se riconsideriamo la Tabella 2.1 si può notare che, mentre le righe rappresentano le distribuzioni di probabilità della variabile aleatoria Y sotto i diversi valori di θ (e, come tali, sommano a 1), le colonne rappresentano le possibili funzioni di verosimiglianza associate ai possibili risultati dell'esperimento, e nulla le vincola ad avere somma unitaria.

2.3 Uso inferenziale di $L(\theta)$

La funzione di verosimiglianza è lo strumento attraverso cui vengono soppesati i diversi valori dei parametri. Attraverso di essa è possibile produrre sintesi inferenziali di diverso tipo. Ad esempio è naturale considerare come stima puntuale del parametro incognito θ, l'argomento che massimizza la funzione $L(\theta)$.

2.3.1 Stime di massima verosimiglianza

Definizione 2.8 *Si chiama stima di massima verosimiglianza il valore $\hat{\theta} \in \Omega$ tale che*

$$L(\hat{\theta}) \geq L(\theta), \qquad \forall\theta \neq \hat{\theta}.$$

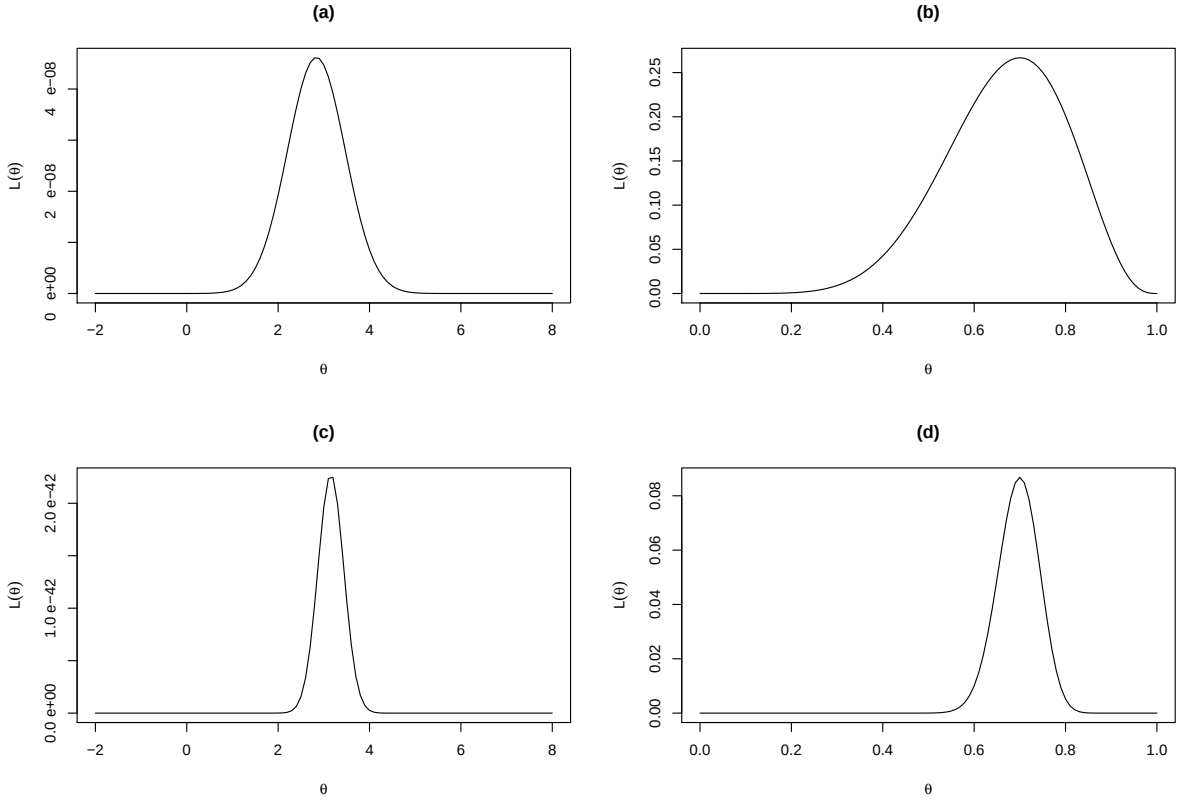

Figura 2.1. Funzioni di verosimiglianza per gli esempi normale (casi (a) e (c)) e binomiale ((b) e (d))

Va detto che il valore $\hat{\theta}$ non necessariamente esiste n tanto meno è unico. È facile costruire esempi in cui, ad esempio, la funzione di verosimiglianza risulta illimitata: si veda [2]. Nei modelli più frequentemente usati, è facile ottenere il valore $\hat{\theta}$, attraverso la massimizzazione analitica della funzione di log-verosimiglianza definita come il logaritmo della funzione di verosimiglianza. Nel caso frequente di un campione di osservazioni $(x_1, x_2, \cdots, x_n)$, realizzazioni indipendenti e somiglianti di una variabile aleatoria X con funzione di (densità di) probabilità $f(\cdot;\theta)$ si avrà

$$\ell(\theta) = \log L(\theta) = \sum_{j=1}^{n} \log f(x_j;\theta). \tag{2.7}$$

Esempio 2.6 [*Modello esponenziale*]
La durate delle telefonate che il centralino dell'Università di Roma "La Sapienza" riceve quotidianamente possono essere considerate variabili aleatorie indipendenti con distribuzione Esp(θ). Per acquisire informazioni sul parametro incognito θ si registra la durata di $n = 10$ telefonate ricevute in un certo intervallo di tempo.

La formulazione matematico-statistica del contesto descritto è allora:

$$X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} \text{Esp}(\theta),$$

ovvero ogni singola X_j ha funzione di densità

$$f(x;\theta) = \theta \exp\{-\theta x\}\, \mathbf{1}_{(0,\infty)}(x), \qquad \theta > 0,$$

dove il simbolo $\mathbf{1}_A(x)$ rappresenta la funzione indicatrice d'insieme, che vale 1 per ogni $x \in A$ e 0 altrove. La funzione di verosimiglianza associata all'esperimento è

$$L(\theta) = \prod_{j=1}^{n} f(x_j;\theta) = \theta^n \exp\left\{-\theta \sum_{j=1}^{n} x_j\right\}, \qquad \theta > 0,$$

e la conseguente funzione di log-verosimiglianza risulta pari a

$$\ell(\theta) = n\log(\theta) - \theta\sum_{j=1}^{n} x_j = n\log(\theta) - n\theta\bar{x},$$

dove $\bar{x}$ è la media campionaria; è facile ora massimizzare $\ell(\theta)$:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{n}{\theta} - n\bar{x} = 0,$$

da cui risulta che il valore $\hat{\theta} = 1/\bar{x}$ è uno zero della della derivata prima di $\ell(\theta)$. Che $\hat{\theta}$ sia effettivamente un punto di massimo lo si deduce dal fatto che la derivata seconda di $\ell(\theta)$ è negativa per ogni valore di $\theta > 0$. Dunque la stima di massima verosimiglianza per il parametro di una distribuzione esponenziale è pari al reciproco della media campionaria. Nella §2.8 discuteremo brevemente le proprietà frequentiste delle procedure basate sulla massimizzazione di $L(\theta)$. ♢

2.3.2 Stima per intervalli

Quando esiste, la versione relativa della funzione di verosimiglianza, $L_R(\theta)$, rappresenta un sistema di pesi compresi tra 0 e 1. Un intervallo di verosimiglianza può allora essere costituito da tutti i valori di θ per i quali la funzione di verosimiglianza relativa è non inferiore ad una certa soglia. Possiamo così definire *intervallo di verosimiglianza di livello* k, con $k \in [0,1]$, l'insieme

$$\mathcal{L}_k = \{\theta \in \Omega : L_R(\theta) \geq k\}.$$

Non esiste un criterio oggettivo per la scelta di k. Fisher (???) propose l'utilizzo delle soglie 1/20, 1/100. Tuttavia, queste scelte convenzionali non hanno riscosso lo stesso successo di altri valori altrettanto convenzionali che vengono quotidianamente utilizzati nella pratica statistica come il livello di significatività del 5% nella verifica di ipotesi (vedi oltre)????

Esempio.[Verosimiglianza normale] (continua). In questo contesto la verosimiglianza (2.5) calcolata in $\hat{\theta}$ vale 1 cosicch $L_R(\theta) = L(\theta)$. L'insieme $\mathcal{L}_k$ è

$$\mathcal{L}_k = \left\{\theta \in \mathbb{R} : \quad \exp\left\{-\frac{n}{2\sigma_0^2}(\bar{x}-\theta)^2\right\} \geq k\right\},$$

che può analogamente essere scritto come

$$\left\{\theta \in \mathbb{R} : \quad \frac{n(\bar{x}-\theta)^2}{\sigma_0^2} \leq k'\right\},$$

con $k' = -2\log k$; ne consegue facilmente allora che

$$\mathcal{L}_k = \left(\bar{x} - \sqrt{\frac{2\log k}{n}}\sigma_0, \bar{x} + \sqrt{\frac{2\log k}{n}}\sigma_0\right).$$

Nell'ambito della statistica classica esiste comunque una teoria alternativa alla costruzione di stime intervallari, che non si basa sulla espressione della funzione di verosimiglianza osservata bensì sulla distribuzione campionaria degli stimatori puntuali di θ e che prende il nome di "regioni di confidenza". Torneremo su questi aspetti nella §2.8.2.

La struttura di $\mathcal{L}_k$ è così identica a quella di un intervallo di confidenza: è possibile associare ad ogni livello k il corrispondente livello di confidenza $1-\alpha$ [68]. Va da s che questa completa coincidenza operativa tra le soluzioni classiche e quelle basate sulla funzione di verosimiglianza si verifica solo in pochi casi, soprattutto quando si adotta il modello normale. Ritorneremo su questi aspetti nella §6.2 a proposito degli intervalli di stima di tipo bayesiano.

Alcuni esempi

Concludiamo questa sezione con alcuni esempi di utilizzo della funzione di verosimiglianza in contesti leggermente più complessi.

Esempio 2.7 [*Modelli cattura-ricattura*]
Dopo un'indagine censuaria nella città XXX, il cui obiettivo specifico è di rilevare tutte le unità della popolazione di riferimento, una specifica circoscrizione della città, diciamo *yy*, viene analizzata nuovamente e con maggior impegno, per rilevare tutte le unità abitanti in quella zona: l'obiettivo della seconda indagine è di produrre una stima dell'efficacia dell'indagine censuaria, attraverso la stima del suo livello di copertura, ovvero la stima della percentuale degli individui "catturati" nella prima indagine. Sia N il numero incognito di unità che vivono nella circoscrizione *yy*, e sia n_1 il numero di persone rilevate dall'indagine censuaria nella circoscrizione stessa. Nella seconda rilevazione vengono "catturati" n_2 individui, dei quali m erano già stati osservati nella prima occasione, mentre gli altri n_2-m risultano nuove "catture". Per semplicità di esposizione assumiamo che ogni individuo abbia la stessa probabilità p di essere catturato in ogni occasione[2] e che tale probabilità sia uguale per tutti gli individui.

Consideriamo allora come realizzazione dell'esperimento la terna (N_1, N_2, M). La loro distribuzione congiunta, per un valore fissato di N e P, è data da

$$p(n_1, n_2, m; N, p) = p(n_1; N, p)p(n_2; n_1, N, p)p(m; n_1, n_2, N, p);$$

il primo fattore, la legge di n_1, è di tipo $\text{Bin}(N, p)$ (ogni tentativo di cattura degli N individui è una prova bernoulliana con probabilità di successo pari a p); il secondo fattore, per l'indipendenza delle due occasioni di cattura, non dipende da n_1 ed è ancora di tipo $\text{Bin}(n, p)$; infine la legge di m condizionata ad (n_1, n_2) non dipende da p ed ha distribuzione ipergeometrica, ovvero

$$\Pr(M = m \mid n_1, n_2, N, p) = \binom{N-n_1}{n_2-m}\binom{n_1}{m}/\binom{N}{n_2}.$$

Ne segue che, dopo facili semplificazioni,

$$\begin{aligned} L(N,p) &\propto \binom{N}{n_1}\binom{N-n_1}{n_2-m}p^{n_1+n_2}(1-p)^{2N-n_1-n_2} \\ &\propto \frac{N!}{(N+m-n_1-n_2)!}p^{n_1+n_2}(1-p)^{2N-n_1-n_2} \end{aligned} \tag{2.8}$$

[2] questa assunzione è chiaramente poco realistica; ad esempio, quando si applicano modelli del genere al problema della stima di popolazioni animali, è ragionevole supporre che individui più deboli siano più facilmente catturabili.

Per ottenere una stima di N si può ad esempio considerare la funzione di verosimiglianza calcolata in $p = \hat{p}_N$, ovvero sostituendo a p la sua stima di massima verosimiglianza assumendo N noto. Si vede facilmente che che $\hat{p}_N = (n_1 + n_2)/(2N)$. Questo modo di agire conduce alla cosiddetta funzione di verosimiglianza profilo, sulla quale torneremo più avanti, che in questo esempio è pari a

$$\hat{L}(N) = \frac{N!}{(N + m - n_1 - n_2)!} \frac{(2N - n_1 - n_2)^{2N - n_1 - n_2}}{(2N)^{2N}}$$

che può essere massimizzata numericamente. Consideriamo un esempio in cui, la prima cattura conduce ad identificare $n_1 = 40$ individui, la seconda cattura conduce a $n_2 = 30$, dei quali $m = 25$ già osservati nella prima occasione. Si ha dunque $\hat{p}_N = 55/(2N)$, mentre la funzione di verosimiglianza profilo è raffigurata nella figura 2.2. Il valore più verosimile appare chiaramente $N = 48$, che produce una stima del livello di copertura pari a $\hat{p} = 55/96 = 0.572$.

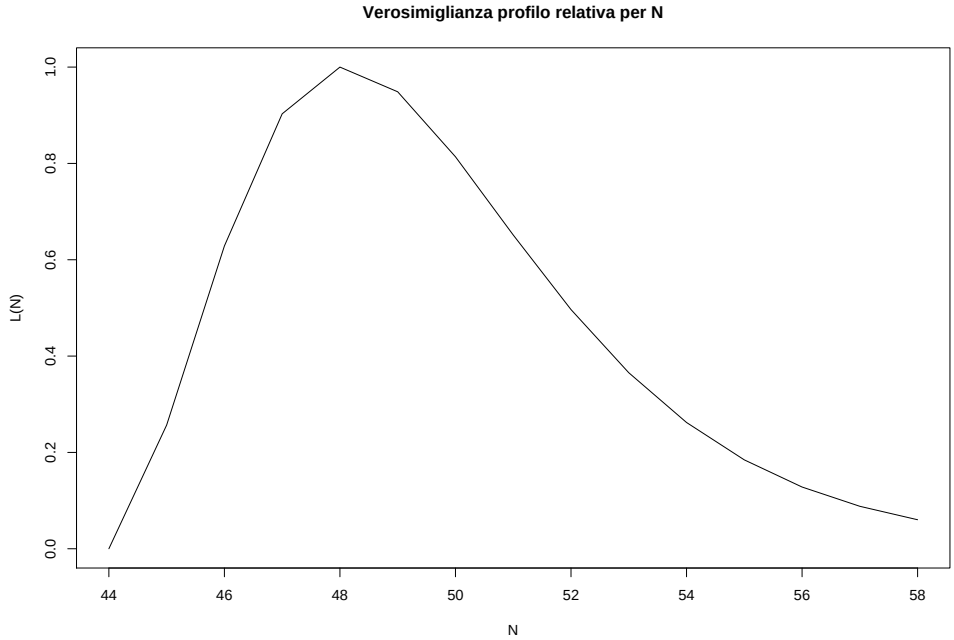

Figura 2.2. Verosimiglianza profilo e curve di livello per la funzione di verosimiglianza bivariata per l'Esempio 2.7; dal secondo grafico si può notare l'informazione sui due parametri sia difficilmente separabile

♢

Esempio 2.8 [*Osservazioni a informazione limitata*].

[67], pag.24, considera il seguente esempio di tipo bernoulliano: si lancia n volte una moneta che dà testa (T) con una certa propensione incognita θ, e i lanci possono essere considerati indipendenti. Il risultato dell'esperimento è la realizzazione della v.a. X =*numero di T in n lanci*; tuttavia tale valore non viene reso noto con precisione, e si sa soltanto che il valore osservato di X risulta minore o uguale ad m, con $m \leq n$. La funzione di verosimiglianza per θ associata a tale esperimento, o meglio al contenuto informativo dell'esperimento, cioè il valore m, è allora

$$L(\theta) = P(X \leq m; \theta) \propto \sum_{k=0}^{m} \binom{n}{k} \theta^k (1 - \theta)^{n-k}. \tag{2.9}$$

Una funzione di verosimiglianza approssimata $\tilde{L}(\theta)$ è ottenibile attraverso l'approssimazione normale alla distribuzione binomiale. Poich $X \sim$ Bin(n, θ), si vede facilmente che

$$\tilde{L}(\theta) \approx \Phi\left(\frac{m - n\theta + 0.5}{\sqrt{n\theta(1-\theta)}}\right)$$

La Figura 2.4 mostra la funzione di verosimiglianza (2.9) in due diverse situazioni, entrambe con $n = 15$; nel primo caso si ha $m = 4$ e, nel secondo, $m = 12$. Nel grafico sono incluse anche le versioni approssimate. Anche questo esempio verrà ridiscusso in ottica bayesiana nel capitolo (**??**)
♢

Esempio 2.9 [*Batteri in sospensione*].

[27], pag. 171, considera il seguente esempio. In una sospensione batteriologica, il numero di batteri per centimetro cubo (c.c) può essere considerata come una variabile aleatoria X con legge di tipo Po(θ). Per ogni campione estratto dalla sospensione è possibile stabilire solamente se essa risulti sterile (in quanto nessun batterio è presente), oppure non sterile (almeno un batterio presente). Prelevando n campioni, ognuno composto da h c.c. di sospensione, si rileva che k degli n campioni risultano sterili. Si vuole determinare la stima di massima verosimiglianza di θ.

Per ogni centimetro cubo di sospensione, la probabilità di non osservare alcun batterio è pari a $\Pr(X = 0; \theta) = e^{-\theta}$. Se il singolo prelievo si riferisce ad h c.c., allora la probabilità di non osservare alcun batterio in quel prelievo è $e^{-h\theta}$. Se gli n prelievi vengono effettuati in modo indipendente, la verosimiglianza per θ vale allora

$$L(\theta) = e^{-hk\theta}(1 - e^{-h\theta})^{n-k}. \tag{2.10}$$

Ne segue che

$$\frac{\partial \log L}{\partial \theta}(\theta) = -kh + h\frac{n-k}{1-e^{-h\theta}}e^{-h\theta};$$

uguagliando a zero la derivata, si ottiene

$$\hat{\theta} = -\frac{1}{h}\log\frac{k}{n}.$$

Ad esempio, per $h = 1$, $n = 20$ e $k = 7$ si ottiene la funzione di verosimiglianza relativa descritta nella Figura 2.5, e il valore di θ che la massimizza è pari a circa 1.05.

♢

2.4 Sufficienza

Negli esempi considerati nei paragrafi precedenti si può notare come la funzione di verosimiglianza associata ad un risultato sperimentale $x_0 = (x_1, \cdots, x_n)$ non dipenda in maniera esplicita dalle n realizzazioni osservate bensì da una loro trasformazione non biunivoca, ovvero una statistica nel senso della definizione (2.3).

Esempio 2.5 (continua). Nel caso di dati indipendenti e somiglianti con distribuzione $N(\theta, \sigma_0^2)$, abbiamo già visto che

$$L(\theta) \propto \exp\left\{-\frac{n}{2\sigma_0^2}(\bar{x} - \theta)^2\right\};$$

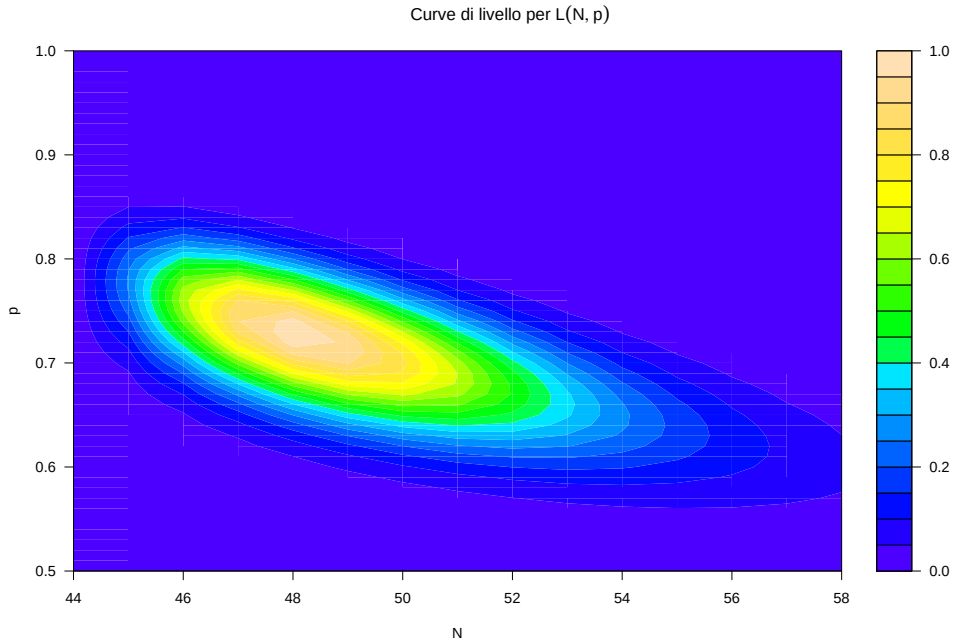

Figura 2.3. Verosimiglianza profilo e curve di livello per la funzione di verosimiglianza bivariata per l'Esempio 2.7; dal secondo grafico si può notare l'informazione sui due parametri sia difficilmente separabile

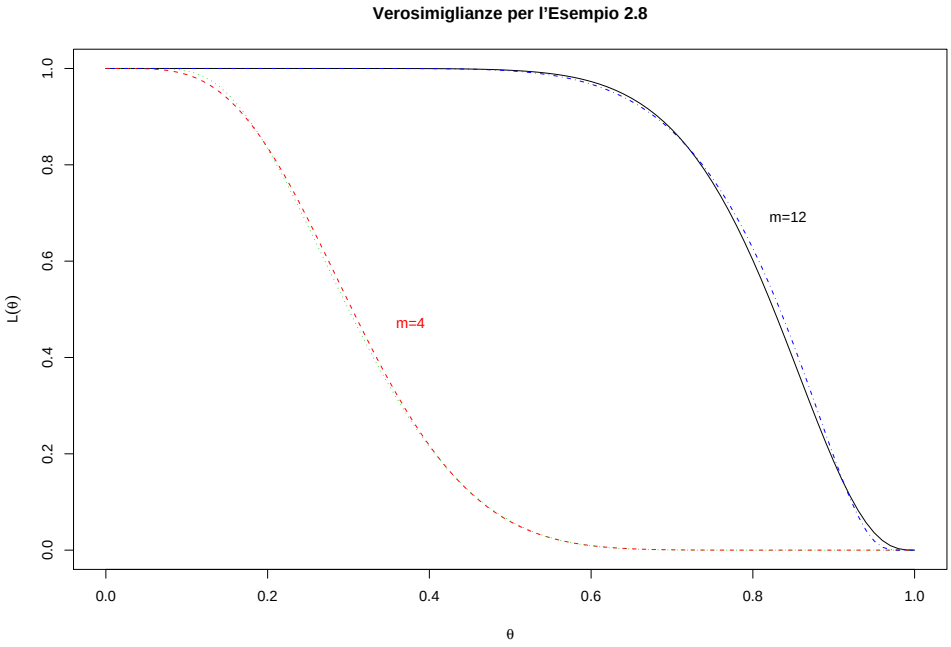

Figura 2.4. Funzioni di verosimiglianza per l'Esempio 2.8: le linee continue rappresentano le verosimiglianze esatte mentre quelle tratteggiate si riferiscono alle approssimazioni gaussiane; si può notare come l'evento $X \leq 4$, essendo molto più informativo dell'evento $X \leq 12$, produca una funzione di verosimiglianza molto più discriminante tra i diversi valori ti θ.

dunque $L(\theta)$ dipende da x_0 solo attraverso la media campionaria $\bar{x} = \sum x_i/n$. Tutte le possibili n-ple dello spazio (ovvero tutti i possibili campioni n-dimensionali) $\mathcal{X}$ che danno luogo alla stesso valore della media campionaria forniranno dunque la stessa funzione di verosimiglianza o, se vogliamo, lo stesso contributo informativo per la stima del parametro θ. Nell'esempio 2.5 dunque, $\bar{x}$ rappresenta la trasformazione dai dati non biunivoca che racchiude tutta l'evidenza sperimentale che il campione fornisce sul parametro θ; si usa dire in tal caso che la media campionaria rappresenta una *statistica sufficiente*, o *riassunto esaustivo*, del risultato sperimentale x_0 ai fini dell'inferenza sul parametro θ. Diamo ora la definizione formale di statistica sufficiente.

Definizione 2.9 *Con riferimento al modello statistico $(\mathcal{X}, \mathcal{P}, \Omega)$, si dice che la statistica t è sufficiente per il parametro θ se e solo se, qualunque sia il campione osservato $x \in \mathcal{X}$, si può scrivere la funzione di verosimiglianza come*

$$L(\theta) = c\,(x)\; g\,(t(x); \theta)\,,$$

dove $c(\cdot)$ è funzione del solo campione osservato x mentre $g(\cdot, \cdot)$ dipende sia da θ che da x ma da quest'ultimo solo attraverso la statistica $t(\cdot)$.

Esempio 2.6 (continua). Per osservazioni indipendenti da una distribuzione esponenziale con parametro θ abbiamo già visto come la funzione di verosimiglianza si possa scrivere come

$$L(\theta) = \theta^n \exp\left\{-\theta \sum x_i\right\} = \theta^n \exp\left\{-n\theta\bar{x}\right\};$$

ne segue che sia la statistica *somma delle osservazioni*, $t_1 = \sum x_i$, che la statistica *media campionaria*, $t_2 = \bar{x}$, rappresentano possibili statistiche sufficienti per θ.

Esempio 2.10 [*Normale con media e varianza incognite*]

In presenza di osservazioni indipendenti e somiglianti con distribuzione $N(\mu, \sigma^2)$ entrambi incogniti, il parametro $\theta = (\mu, \sigma^2)$ risulta bidimensionale; la funzione di verosimiglianza associata ad un campione n-dimensionale $(x_1, x_2, \cdots, x_n)$ è, ricordando la (2.4), pari a

$$L(\mu, \sigma^2) = \frac{1}{\sigma^n} \exp\left\{-\frac{n}{2\sigma^2}\left[s^2 + (\bar{x} - \mu)^2\right]\right\};$$

in questo caso la statistica sufficiente per il parametro bidimensionale θ è a sua volta bidimensionale ed è data da $t_1(x) = s^2$, $t_2(x) = \bar{x}$. ♢

La coincidenza tra le dimensioni del parametro e della statistica sufficiente non è una proprietà generale: essa si verifica solo nell'ambito della cosiddetta famiglia esponenziale; tratteremo brevemente di questo argomento nella §2.11: il lettore interessato ad approfondimenti può consultare [66] oppure [2].

È possibile comunque incontrare esempi in cui la statistica sufficiente presenta una dimensione diversa da quella del parametro.

Esempio 2.11 [*Distribuzione uniforme*]

Siano $X_1, X_2, \cdots, X_n$ variabili aleatorie indipendenti e somiglianti con distribuzione $U(\theta, 2\theta)$, con $\theta > 0$. La densità vale dunque

$$f(x; \theta) = \frac{1}{\theta}\, \mathbf{1}_{(\theta, 2\theta)}(x).$$

Considerando che il supporto delle X_j dipende, per ampiezza e posizionamento sull'asse reale, dal valore incognito di θ, la funzione di verosimiglianza vale

$$L(\theta) \propto \frac{1}{\theta^n}\, \mathbf{1}_{(x_1/2,x_1)}(\theta) \times \mathbf{1}_{(x_2/2,x_2)}(\theta) \times \cdots \times \mathbf{1}_{(x_n/2,x_n)}(\theta).$$

Affinch $L(\theta)$ risulti diversa da zero, tutte le funzioni indicatrici devono valere 1: è facile convincersi che questo accade solo per quei valori di θ per i quali risulti, contemporaneamente,

$$\theta < \min_{j=1,\cdots,n} x_j = x_{(1)} \quad \text{e} \quad \theta > \frac{1}{2} \max_{j=1,\cdots,n} x_j = x_{(n)};$$

Ne segue che la statistica sufficiente per il parametro reale θ è in questo caso bidimensionale, ovvero

$$t_1(x_1, \cdots, x_n) = x_{(1)}, \quad t_2(x_1, \cdots, x_n) = x_{(n)}$$

♢

Altre volte non è possibile determinare alcuna statistica sufficiente che abbia dimensione inferiore ad n, il numero delle osservazioni.

Esempio 2.12 [*Distribuzione normale asimmetrica.*]

[3] ha introdotto la famiglia di distribuzioni normali asimmetriche: nella sua formulazione più semplice, si dice che $X \sim SN(\theta)$ se la funzione di densità vale

$$p(x;\theta) = 2\,\varphi(x)\Phi(\theta x), \tag{2.11}$$

dove $\varphi(\cdot)$ e $\Phi(\cdot)$ rappresentano, rispettivamente, la funzione di densità e quella di ripartizione di una v.a. normale standard. Supponiamo di osservare un campione di n osservazioni i.i.d. con distribuzione $SN(\theta)$; la funzione di verosimiglianza associata all'esperimento è

$$L(\theta;\mathbf{x}) \propto \prod_{i=1}^{n} \Phi(\theta x_i);$$

e non esiste alcuna funzione delle x_i, di dimensione inferiore ad n, che soddisfi la Definizione 2.9. ♢

2.5 Informazione di Fisher

In questa sezione assumeremo che il modello statistico in uso goda di alcune proprietà di regolarità. In particolare si assume che [31]

1. il supporto delle osservazioni non vari al variare del valore del parametro θ in Ω;
2. il vero velore θ_0 del parametro sia un punto interno (non di frontiera) dello spazio $\boldsymbol{\Omega}$ che ha dimensione finita;
3. non ci siano problemi di identificabilità, nel senso della Definizione 2.2;
4. esista un intorno $\mathcal{I}$ di θ_0 all'interno del quale la funzione di log-verosimiglianza è tre volte differenziabile rispetto a θ;
5. sia sempre possibile scambiare il segno di derivazione con quello di integrale nei calcoli relativi alla funzone di log-verosimiglianza.

Le condizioni di regolarità garantiscono, di fatto, una semplificazione nella trattazione matematica del modello. In pratica esse sono verificate molto spesso: fa eccezione, tuttavia, il modello uniforme

considerato nell'Esempio 2.11, nel quale il supporto della v.a. X è, per θ fissato, l'intervallo chiuso $[0, \theta]$; non risulta quindi verificata la prima delle assunzioni sopra elencate

Sia $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ un campione di n replicazioni di una v.a. con densità $f(\cdot \mid \theta)$, $\theta \in \Omega$. Per il momento assumiamo che il parametro θ sia uno scalare e, per semplicità di notazione, consideriamo il caso in cui le X_i siano dotate di densità assolutamente continua rispetto alla misura di Lebesgue: il caso discreto non presenta alcuna differenza di sostanza. Indichiamo con

$$\ell(\theta; \mathbf{x}) = \log L(\theta, \mathbf{x})$$

la funzione di log-verosimiglianza associata all'esperimento osservato $\mathbf{X} = \mathbf{x}$.

Definizione 2.10 *Si chiama funzione* score *e si indica con $U(\mathbf{x}, \theta)$ la derivata prima rispetto a θ di $\ell(\theta; \mathbf{x})$*

$$U(\mathbf{x}, \theta) = U(x_1, \cdots, x_n, \theta) = \frac{\partial}{\partial \theta} \ell(\theta; \mathbf{x}) = \frac{\frac{\partial}{\partial \theta} f(x_1, \cdots, x_n; \theta)}{f(x_1, \cdots, x_n; \theta)} = \frac{f'_\theta(x_1, \cdots, x_n; \theta)}{f(x_1, \cdots, x_n; \theta)};$$

Sotto le condizioni di regolarità sopra descritte, la funzione score è dotata di alcune proprietà notevoli.

Proposizione 2.1 *Sotto le condizioni di regolarità riportate all'inizio di questa sezione,*

A) $\mathbf{E}_\theta \left(U(\mathbf{X}; \theta) \right) = 0$
B) $\mathrm{Var} \left(U(\mathbf{X}; \theta) \right) = - \mathbf{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \ell(\theta; \mathbf{X}) \right)$

Dimostrazione 2.1 *Per quanto concerne il primo risultato si ha*

$$\mathbf{E}_\theta \left(U(\mathbf{X}; \theta) = \int_{\mathcal{X}^{(n)}} U(x_1, \cdots, x_n; \theta) f(x_1, \cdots, x_n; \theta) d\mathbf{x} \right.$$

$$= \int_{\mathcal{X}^{(n)}} \frac{\partial}{\partial \theta} f(x_1, \cdots, x_n; \theta) d\mathbf{x} = \frac{\partial}{\partial \theta} \int_{\mathcal{X}^{(n)}} f(x_1, \cdots, x_n; \theta) d\mathbf{x} = \frac{\partial}{\partial \theta} 1 = 0.$$

Per il calcolo della varianza, si ha

$$\mathrm{Var} \left(U(\mathbf{X}; \theta) \right) = \mathbf{E}_\theta \left(U^2(\mathbf{X}; \theta) \right) = \mathbf{E}_\theta \left(\left(\frac{\partial}{\partial \theta} \ell(\theta; \mathbf{X}) \right)^2 \right). \tag{2.12}$$

Inoltre

$$\frac{\partial^2}{\partial \theta^2} \ell(\theta; \mathbf{X}) = \frac{f''(\mathbf{x}; \theta) f(\mathbf{x}; \theta) - f'(\mathbf{x}; \theta) f'(\mathbf{x}; \theta)}{f^2(\mathbf{x}; \theta)};$$

perciò

$$\left(\frac{f'(\mathbf{x}; \theta)}{f(\mathbf{x}; \theta)} \right)^2 = \frac{f''(\mathbf{x}; \theta)}{f(\mathbf{x}; \theta)} - \frac{\partial^2}{\partial \theta^2} \ell(\theta; \mathbf{X}).$$

Dunque

$$\mathrm{Var} \left(U(\mathbf{X}; \theta) \right) = \int_{\mathcal{X}^{(n)}} \left(\frac{f'(\mathbf{x}; \theta)}{f(\mathbf{x}; \theta)} \right)^2 f(\mathbf{x}; \theta) d\mathbf{x} = \int_{\mathcal{X}^{(n)}} f''(\mathbf{x}; \theta) d\mathbf{x} - \mathbf{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \ell(\theta; \mathbf{X}) \right),$$

che fornisce la tesi non appena si noti che, scambiando il segno di integrale con quello di derivata, il primo addendo del membro di destra dell'ultima relazione è nullo.

La derivata seconda rispetto al parametro θ della funzione di log-verosimiglianza, con segno cambiato, si chiama *informazione osservata* nel punto $\mathbf{x} \in \mathcal{X}$, relativa al parametro θ,

$$j(\theta) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta; \mathbf{x}) \tag{2.13}$$

La $\text{Var}\,(U(\mathbf{X}; \theta))$ prende il nome di *informazione attesa di Fisher* e si indica col simbolo $I(\theta)$. Come risulta dal punto B della Proposizione 2.1, $I(\theta)$ può anche essere definita come il valore atteso dell'informazione osservata $j(\theta)$: da qui il nome. Questo suggerisce anche il seguente risultato, utile per il calcolo approssimato di molte grandezze.

Proposizione 2.2 *Sotto le condizioni di regolarità riportate all'inizio di questa sezione, si ha che*

$$\frac{J(\theta)}{n} \xrightarrow{p} I_1(\theta), \tag{2.14}$$

dove $I_1(\theta)$ rappresenta l'informazione attesa di Fisher associata ad un campione di dimensione uno.

Dimostrazione 2.2 *Si tratta di una semplice applicazione della legge dei grandi numeri.*

La proposizione precedente svolge naturalmente un ruolo importante anche nel caso di campioni finiti ma ragionevolmente numerosi; in tal caso, infatti, potremo utilizzare l'approssimazione

$$J(\theta) \approx nI_1(\theta). \tag{2.15}$$

La quantità $I(\theta)$ svolge un ruolo importante in tutte le impostazioni inferenziali: in ambito bayesiano essa risulterà centrale nell teoria delle distribuzioni a priori non informative di cui ci occuperemo nella §5.2.

Il caso multidimensionale. Nel caso generale in cui il modello statistico prevede un vettore di parametri $\boldsymbol{\theta} \in \mathbb{R}^k$, si definisce vettore **score**, o punteggio, il vettore delle derivate parziali prime della funzione di log-verosimiglianza rispetto alle componenti di θ

$$\mathbf{U}(\mathbf{X}, \boldsymbol{\theta}) = \left(\frac{\partial\,\ell(\theta; \mathbf{X})}{\partial\theta_1}, \cdots, \frac{\partial\,\ell(\theta; \mathbf{X})}{\partial\theta_k} \right).$$

La Proposizione 2.1 può essere generalizzata in modo immediato al caso multivariato, tenendo presente che, in questo caso, $\text{Var}\,(\mathbf{U}(\mathbf{X}, \boldsymbol{\theta}))$ è una matrice $k \times k$, che prende il nome di matrice di informazione attesa di Fisher. Avremo cioè

$$I(\boldsymbol{\theta}) = \text{Var}\,(\mathbf{U}(\mathbf{X}, \boldsymbol{\theta}))\,,$$

dove l'elemento generico della matrice $I(\boldsymbol{\theta})$ è dato da

$$I_{hj}(\boldsymbol{\theta}) = \mathbb{E}\left(-\frac{\partial^2}{\partial\theta_h\theta_j}\ell(\theta; \mathbf{X}) \right).$$

Il seguente teorema giustifica, in qualche modo, l'appellativo di informazione associato alla quantità $I(\cdot)$.

Teorema 2.1 *Siano $\mathbf{X} = (X_1, \cdots, X_n)$ un vettore di n replicazioni indipendenti della stessa v.a. con distribuzione $p(x, \boldsymbol{\theta})$, $\theta \in \mathbb{R}^k, k \geq 1$. Allora l'informazione attesa relativa al campione di dimensione n è pari a n volte l'informazione attesa relativa ad una sola osservazione.*

Dimostrazione 2.1 *Semplice: lasciata per esercizio.*

In altri termini il Teorema 2.1 suggerisce che ogni osservazione fornisce, in media, una prefissata quantità di informazione sul parametro $\boldsymbol{\theta}$, $I(\boldsymbol{\theta})$ appunto, e l'informazione si cumula al crescere della dimensione campionaria. Il fatto che l'informazione di Fisher dipenda, in generale, dal valore del parametro θ suggerisce altresì che l'informazione associata a ciascuna osservazione agisce in modo non uniforme sul nostro livello informativo relativo a $\boldsymbol{\theta}$: questa osservazione risulterà importante, in un'ottica bayesiana, per il calcolo di distribuzioni a priori non informative, di cui si discuterà nella §5.2.

Esempio 2.13 [*Informazione di Fisher per un campione bernoulliano.*]
Siano $X_1, X_2, \cdots, X_n$ v.a. indipendenti e somiglianti con distribuzione di tipo Be(θ). La funzione di log-verosimiglianza relativa alla prima osservazione è

$$\log p(x;\theta) = x \log\theta + (1-x)\log(1-\theta), \qquad x = 0, 1.$$

La funzione score vale allora

$$U(X_1, \theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta},$$

mentre l'informazione osservata è pari a

$$j(X_1, \theta) = \frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}.$$

Ricordando che nel caso bernoulliano $\mathbf{E}(X) = \theta$, si ha che l'informazione di Fisher associata ad una singola osservazione è pari a

$$I(\theta) = \mathbf{E}\left(\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}\right) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)},$$

e l'informazione di Fisher associata ad un campione di dimensione n è pari, per il Teorema 2.1, a $n/(\theta(1-\theta))$. ♢

Esempio 2.14 [*Informazione di Fisher per un campione gaussiano.*]
Siano $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, con entrambi i parametri incogniti. In questo caso, allora, $\boldsymbol{\theta} = (\mu, \sigma^2)$. In virtù del teorema 2.1, in presenza di osservazioni indipendenti e somiglianti, l'informazione di Fisher è pari ad n volte l'informazione di Fisher relativa ad una sola osservazione. La funzione di log-verosimiglianza relativa ad una generica osservazione è

$$\log p(x;\boldsymbol{\theta}) = -\frac{1}{2}\left(\log\sigma^2 + \frac{(x-\mu)^2}{\sigma^2}\right).$$

Il vettore score vale allora

$$U(X_1, \boldsymbol{\theta}) = \left(\frac{\partial}{\partial\mu}\log p(x;\boldsymbol{\theta}); \frac{\partial}{\partial\sigma^2}\log p(x;\boldsymbol{\theta})\right) = \left(\frac{(x-\mu)}{\sigma^2}; -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}\right).$$

La matrice d'informazione osservata vale allora

$$J(X_1, \boldsymbol{\theta}) = \begin{pmatrix} \dfrac{1}{\sigma^2} & \dfrac{(x-\mu)}{\sigma^4} \\ \dfrac{(x-\mu)}{\sigma^4} & -\dfrac{1}{2\sigma^4} + \dfrac{(x-\mu)^2}{\sigma^6} \end{pmatrix}$$

Ricordando poi che $\mathbb{E}(X-\mu)=0$ e $\mathbb{E}(X-\mu)^2=\sigma^2$, si ha che la matrice di Fisher è pari a

$$I(\boldsymbol{\theta})=\begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}.$$

♢

Esempio 2.15 [*Informazione di Fisher per campioni da leggi di Poisson.*]
Siano $X_1, X_2, \cdots, X_n$ v.a. indipendenti e somiglianti con distribuzione di tipo Poi(θ). In questo caso, la funzione di log-verosimiglianza relativa alla prima osservazione è

$$\log p(x;\boldsymbol{\theta}) = -\theta + x\log\theta - \log(x!);$$

la funzione score vale

$$U(X,\boldsymbol{\theta}) = -1 + \frac{X}{\theta},$$

e l'informazione osservata è $j(X_1,\theta) = x/\theta^2$. Ricordando che la media di una v.a. di con legge Poi(θ) vale proprio θ, l'informazione attesa di Fisher vale quindi

$$I(\theta) = \frac{1}{\theta}.$$

♢

Non sempre conviene utilizzare la parte B della Proposizione 2.1, come dimostra l'esempio seguente.
Esempio 2.16 [*Informazione di Fisher per il modello SN(θ).*]
La funzione di densità di una v.a. $X \sim SN(\theta)$ è data nella (2.11). La funzione di log-verosimiglianza associata ad una singola osservazione è dunque

$$\ell(\theta) \propto \log\Phi(\theta x),$$

Allora $U(X,\theta) = X\varphi(\theta X)/\Phi(\theta X)$, e

$$I_1(\theta) = \mathbb{E}\left(U^2(X,\theta)\right) = \int_{\mathbb{R}} 2x^2\varphi(x)\frac{\varphi^2(\theta x)}{\Phi(\theta x)}dx$$

Trattando separatamente l'integrale sul semiasse negativo e quello sul semiasse positivo e tenendo conto del fatto che $\Phi(-z) = 1-\Phi(z)$, per ogni $z>0$, che la densità di una normale standard è simmetrica rispetto allo zero, si ha, con un semplice cambio di segno nel primo integrale,

$$I_1(\theta) = \int_0^\infty 2x^2\varphi(x)\varphi^2(\theta x)\frac{1}{[\Phi(\theta x)\Phi(-\theta x)]}dx$$

L'ultimo integrale non è risolubile analiticamente ma è possibile dimostrare che la quantità $\varphi(\theta x)/\sqrt{\Phi(\theta x)\Phi(-\theta x)}$ può essere efficaciemente approssimata da $2\varphi(2\theta x/\pi)$. Perciò

$$I_1(\theta) \approx \int_0^\infty 2x^2\varphi(x)\varphi^2(2\theta x/\pi)dx;$$

calcoli standard conducono all'espressione approssimata di $I_1(\theta)$:

$$I_1(\theta) \approx \frac{2}{\pi}\left(1+\frac{8\theta^2}{\pi^2}\right)^{-3/2}. \tag{2.16}$$

♢

2.6 La divergenza di Kullback-Leibler

Definizione 2.11 *Dato il modello statistico $\mathcal{E} = \{\mathcal{X}, \mathcal{P}, \boldsymbol{\Omega}\}$, si definisce divergenza di Kullback-Leibler di una determinata distribuzione di probabilità $p(x) \in \mathcal{P}$ - che supponiamo assolutamente continua con densità $f(x)$ - rispetto ad una misura di riferimento $p_0(x) \in \mathcal{P}$, anche questa supposta assolutamente continua con densità $f_0(x)$, la quantità*

$$D_{KL}(p; p_0) = \int_{\mathcal{X}} \log \frac{f_0(x)}{f(x)} f_0(x) dx. \tag{2.17}$$

Tale quantità non è simmetrica rispetto ai due argomenti p e p_0 e va interpretata come la distanza della legge p rispetto a p_0: per questo motivo D non può essere considerata una vera e propria distanza. Elenchiamo di seguito le principali proprietà di D. Per ulteriori approfondimenti si rimanda, ad esempio, a [19].

Teorema 2.2 *La divergenza $D_{KL}(p; p_0)$ soddisfa le seguenti proprietà:*

a) $D_{KL}(p; p_0) \geq 0$
b) $D_{KL}(p; p_0) = 0$ *se e solo se* $f(x) = f_0(x)$ *quasi ovunque.*

Dimostrazione 2.2 *Basta dimostrare che*

$$\int f_0(x) \log f_0(x) dx \geq \int f_0(x) \log f(x) dx,$$

e che il segno di uguaglianza si può verificare solo quando $f(x) = f_0(x)$ quasi ovunque. Nel corso della dimostrazione si assume che i due integrali sopra scritti valgano 0 nell'insieme in cui $f_0(x) = 0$, qualunque sia $f(x)$. In maniera equivalente occorre dimostrare che

$$\int f_0(x) \log \frac{f(x)}{f_0(x)} dx \leq 0.$$

Poich, per ogni $x \geq -1$, si ha $x \geq \log(1+x)$, con il segno di uguaglianza valido solo quando $x = 0$, risulta

$$\log \frac{f(x)}{f_0(x)} = \log \left(1 + \frac{f(x)}{f_0(x)} - 1\right) \leq \frac{f(x)}{f_0(x)} - 1.$$

e il segno di uguaglianza valido solo se $f(x) = f_0(x)$. Di conseguenza,

$$\int f_0(x) \log \frac{f(x)}{f_0(x)} dx \leq \int f_0(x) \left(\frac{f(x)}{f_0(x)} - 1\right) dx = \int f(x) dx - \int f_0(x) dx = 0,$$

dove l'ultimo passaggio deriva dal fatto che sia f che f_0 sono delle densità di probabilità. Se $f_0(x) \neq f(x)$ in un insieme non trascurabile la serie di disuguaglianze sopra elencate vale in senso stretto.

Il secondo enunciato della dimostrazione afferma che se due densità di probabilità differiscono su un insieme di misura non nulla, la divergenza di Kullback -Leibler di una dall'altra è diversa da zero. In altri termini, dato un modello statistico $\{\mathcal{X}, \mathcal{P}, \boldsymbol{\Omega}\}$, esso risulta identificabile e se e solo se la divergenza di Kullback e Leibler fra due qualunque delle sue componenti risulta maggiore di zero. Quando la divergenza di Kullback-Leibler è utilizzata per calcolare la distanza tra due elementi della stessa classe parametrica, è importante sottolineare che D_{KL} è invariante rispetto a riparametrizzazioni biunivoche: se $\{p(\cdot; \theta), \theta \in \boldsymbol{\Omega}\}$ è una famiglia parametrica di distribuzioni e $\lambda = \lambda(\theta)$ è una trasformazione monotona, allora risulta, per ogni coppia di valori θ e θ_0,

$$D_{KL}(\theta;\theta_0) = D_{KL}(\lambda;\lambda_0). \tag{2.18}$$

Esempio 2.17 [*Distanza tra due distribuzioni esponenziali.*]
Consideriamo due distribuzioni di tipo Exp(θ), ovvero

$$p(x;\theta) = \theta e^{-\theta x}; \qquad p(x;\theta_0) = \theta_0 e^{-\theta_0 x};$$

Si calcola facilmente che

$$D_{KL}(\theta;\theta_0) = \mathbb{E}_{\theta_0}\left(\log\frac{\theta_0 e^{-\theta_0 X}}{\theta e^{-\theta X}}\right) = \log\frac{\theta_0}{\theta} + (\theta-\theta_0)\,\mathbb{E}_{\theta_0}(X) = \log\frac{\theta_0}{\theta} + \frac{\theta-\theta_0}{\theta_0} \tag{2.19}$$

D'altro canto utilizzando la parametrizzazione alternativa $\lambda = 1/\theta$, le densità diventano

$$p(x;\lambda) = \frac{1}{\lambda}e^{-x/\lambda}; \qquad p(x;\lambda_0) = \frac{1}{\lambda_0}e^{-x/\lambda_0};$$

e, di conseguenza,

$$D_{KL}(\lambda;\lambda_0) = \mathbb{E}_{\theta_0}\left(\log\frac{\lambda e^{-X/\lambda_0}}{\lambda_0 e^{-X/\lambda}}\right) = \log\frac{\lambda}{\lambda_0} + \frac{(\lambda_0-\lambda)}{\lambda\lambda_0}\,\mathbb{E}_{\theta_0}(X) = \log\frac{\lambda}{\lambda_0} + \frac{\lambda_0-\lambda}{\lambda},$$

che, una volta riespressa in termini di θ, coincide con la (2.19).

♢

È possibile inoltre ottenere versioni simmetrizzate della divergenza (2.17).

Definizione 2.12 *Si definisce distanza "simmetrica" di Kullback-Liebler tra due elementi di $\mathcal{P}$, p e p_0, individuate dai valori del parametro θ e θ_0, la quantità*

$$J_{KL}(p,p_0) = D_{KL}(p;p_0) + D_{KL}(p_0;p) = \int_{\mathcal{X}} \log\frac{f_0(x)}{f(x)}\,(f_0(x)-f(x))\,dx. \tag{2.20}$$

Alternativamente si può simmetrizzare D_{KL} considerando il minimo delle due divergenze [16]:

Definizione 2.13 *Si definisce discrepanza "intrinseca" di Kullback-Liebler tra due elementi di $\mathcal{P}$, p e p_0, individuate dai valori del parametro θ e θ_0, la quantità*

$$M_{KL}(p,p_0) = \min\{D_{KL}(p;p_0); D_{KL}(p_0;p)\}. \tag{2.21}$$

Le due definizioni precedenti individuano effettivamente delle distanze.

2.7 Un'approssimazione della funzione di verosimiglianza

In situazioni regolari, e per grandi valori della dimensione campionaria, è possibile approssimare l'andamento della funzione di verosimiglianza relativa con quello di una densità di probabilità gaussiana. Vediamo in dettaglio cosa avviene nel caso in cui θ sia un parametro scalare. Utilizzando infatti uno sviluppo in serie di Taylor fino al secondo ordine della funzione di log-verosimiglianza si ottiene

$$\log L(\theta) \approx \log L(\hat\theta) + U(\hat\theta;\mathbf{x})(\theta-\hat\theta) - \frac{1}{2}J(\hat\theta)(\theta-\hat\theta)^2;$$

considerando che la funzione score calcolata in $\hat\theta$ vale zero, ed utilizzando la (2.15), si ottiene

$$\ell(\theta) = \log \frac{L(\theta)}{L(\hat{\theta})} \approx -\frac{n}{2} I_1(\hat{\theta})(\theta - \hat{\theta})^2 \tag{2.22}$$

e dunque la funzione di verosimiglianza relativa può essere approssimata con una densità gaussiana con media pari a $\hat{\theta}$ e varianza $(nI_1(\hat{\theta}))^{-1}$, ovvero

$$L_R(\theta) \approx L_{appr}(\theta) \propto \exp\left\{-\frac{n}{2} I_1(\hat{\theta})(\theta - \hat{\theta})^2\right\} \tag{2.23}$$

Esempio 2.18 [*Osservazioni i.i.d. con legge di Poisson*]
Consideriamo un campione di n osservazioni indipendenti con distribuzione Poi(θ); la funzione di verosimiglianza associata al campione osservato vale allora

$$L(\theta) \propto \prod_{i=1}^{n} \left[e^{-\theta}\theta^{x_i}\right] = e^{-n\theta}\theta^{n\bar{x}},$$

e la sua versione relativa è

$$L_R(\theta) = \frac{e^{-n\theta}\theta^{n\bar{x}}}{e^{-n\hat{\theta}}\hat{\theta}^{n\bar{x}}} \tag{2.24}$$

dove $\hat{\theta} = \bar{x}$. Per calcolare l'approssimazione gaussiana occorre determinare l'informazione di Fisher in $\hat{\theta}$. Dai risultati ottenuti nell'Esempio 2.15 si ha $I(\theta) = n/\theta$, da cui

$$L_{appr}(\theta) \propto \exp\left\{-\frac{n}{2\bar{x}}(\theta - \bar{x})^2\right\} \tag{2.25}$$

Come esemplificazione abbiamo generato con **R** un campione di 20 osservazioni da una legge di Poisson con parametro 3 ed abbiamo ottenuto

```
4 5 4 1 3 5 4 3 2 2 4 4 5 1 4 2 5 2 2 2
```

La figura 2.6 rappresenta le funzioni (2.24) e (2.25) nell'esempio specifico.

♢

2.8 Proprietà frequentiste delle procedure basate su $L(\theta)$

2.8.1 Lo stimatore di massima verosimiglianza

La scelta di utilizzare, come stima puntuale, il valore che massimizza la funzione di verosimiglianza è stata giustificata, nella §2.3, sul piano intuitivo. Esistono però risultati di carattere asintotico ed altri di natura frequentista, che giustificano tale scelta. In questa sezione verranno brevemente riassunti i più importanti tra questi. In particolare vedremo come lo stimatore di massima verosimiglianza risulti, sotto condizioni molto generali, consistente e dotato, asintoticamente e opportunamente normalizzato, di una distribuzione campionaria di tipo gaussiano.

Teorema 2.3 *Sotto le condizioni di regolarità riportate all'inizio della §2.5, sia* $\mathbf{X}_n = (X_1, \cdots, X_n)$ *un campione i.i.d. estratto da una popolazione con densità del tipo* $p(x; \theta_0)$*, con* $\theta_0 \in \Omega$*, e assumiamo che il modello statistico* $\mathcal{P}$ *sia identificabile, nel senso della (2.2). Allora, per* $n \to \infty$*, deve esistere una successione consistente di stimatori* $\hat{\theta}_n$*, ovvero*

$$\lim_{n\to\infty} \hat{\theta}_n = \theta_0,$$

Dimostrazione 2.3 *La dimostrazione che segue sorvola su alcuni aspetti tecnici che il lettore più esigente può ritrovare in [86], pag.126 oppure [31], pag.123. Determinare il massimo della funzione di verosimiglianza* $L(\theta) = f(\mathbf{x}_n;\theta)$ *è equivalente a determinare il massimo della sua versione logaritmica* $\ell(\theta) = \log L(\theta)$*; il problema non viene modificato se a tale funzione si sottrae la quantità costante* $\ell(\theta_0)$ *e si divide il tutto per la costante* n*. Dunque consideriamo il problema di massimizzazione di*

$$G_n(\theta) = \frac{1}{n}\left(\ell(\theta) - \ell(\theta_0)\right) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p(x_i;\theta)}{p(x_i;\theta_0)} = \frac{1}{n}\sum_{i=1}^{n} Z_i;$$

Le v.a. $Z_i = \log p(x_i;\theta) - \log p(x_i;\theta_0)$*,* $i = 1,\cdots,n$*, sono indipendenti e somiglianti, con media pari a*

$$\int \log\frac{p(x;\theta)}{p(x;\theta_0)} p(x;\theta_0)dx = -D_{KL}(\theta,\theta_0). \tag{2.26}$$

Per la legge dei grandi numeri, allora,

$$G_n(\theta) \xrightarrow{p} -D_{KL}(\theta,\theta_0)$$

Ricordando, dalla §2.6, che $D_{KL}(\theta,\theta_0)$ *è sempre positiva e vale zero se e solo se* $\theta = \theta_0$*, si può concludere che massimizzare* $G_n(\theta)$ *equivale, asintoticamente, a minimizzare* $D_{KL}(\theta,\theta_0)$ *e tale minimo si ottiene proprio per* $\hat{\theta}_n = \theta_0$*; questo, insieme alla identificabilità del modello, garantisce la consistenza della successione di stime di massima verosimiglianza* $\hat{\theta}_n$*.*

Inoltre è possibile ricavare un risultato generale circa la distribuzione asintotica di $\hat{\theta}_n$.

Teorema 2.4 *Sotto le condizioni elencate prima della Proposizione 2.1, se* $\mathbf{x}_n = (X_1,\cdots,X_n)$ *è un campione i.i.d. estratto da una popolazione con densità del tipo* $p(x;\theta_0)$*, con* $\theta_0 \in \Omega$*, e se* $\hat{\theta}_n$ *è uno stimatore consistente di* θ_0*, allora*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \frac{1}{I_1(\theta_0)}\right) \tag{2.27}$$

dove $I_1(\theta_0)$ *rappresenta l'informazione di Fisher associata ad una singola osservazione.*

Dimostrazione 2.4 *Vedi [67], pag.240 oppure, per una dimostrazione più rigorosa, [56].*

In pratica, noi non conosciamo il vero valore θ_0 del parametro, cosicch il denominatore della varianza della legge gaussiana nella (2.27) viene stimata con $I_1(\hat{\theta}_n)$.

Esempio 2.19
esempio di media difficoltà, non calcolabile esattamente ?????????????????????????????? ♢

2.8.2 Intervalli di confidenza

Molto spesso, o quasi sempre, la stima puntuale di un parametro non fornisce sufficienti garanzie di precisione, ed è più ragionevole fornire un insieme di valori che, alla luce del campione osservato, possano essere considerati ragionevoli stime del parametro.

Definizione 2.14 *Sia* $(X_1, X_2, \cdots, X_n)$ *un campione casuale estratto da una popolazione distribuita secondo un elemento del modello* $\{\mathcal{X}, \mathcal{P}, \boldsymbol{\Omega}\}$*, ovvero con densità* $\{p(\cdot;\theta), \theta \in \boldsymbol{\Omega}\}$*, con* $\boldsymbol{\Omega}$ *intervallo sottoinsieme della retta reale.*
Siano $T^+ = t^+(X_1, X_2, \cdots, X_n)$ *e* $T^- = t^-(X_1, X_2, \cdots, X_n)$ *due statistiche tali che*

a) $t^-(x_1, x_2, \cdots, x_n) \leq t^+(x_1, x_2, \cdots, x_n)$ *per ogni possibile n-pla in* $\mathcal{X}$*;*
b) $\Pr(T^- \leq \theta \leq T^+) \geq 1 - \alpha$, *per ogni* $\theta \in \boldsymbol{\Omega}$.

Allora l'intervallo aleatorio (T^-, T^+) *viene detto intervallo di confidenza per il parametro* θ *con un livello di confidenza pari a* $1 - \alpha$.

Una volta osservato il risultato campionario le due funzioni T^- e T^+ producono due valori reali che determinano gli estremi dell'intervallo di confidenza. Una tale procedura fornisce garanzie di tipo frequentista, nel senso che, ripetutamente utilizzata, $100(1-\alpha)$ volte su 100 produce un intervallo che contiene il vero valore del parametro. In pratica, utilizzando il metodo una sola volta, si può solo affermare di avere una confidenza (da cui il nome della procedura) pari a $1-\alpha$ che l'intervallo numerico contenga effettivamente il parametro.

Esistono diverse tecniche di costruzione di intervalli di confidenza. Una trattazione esauriente è in [22]. Tuttavia, per dimensioni campionarie sufficientemente grandi, e nelle situazioni regolari in cui vale il Teorema 2.4, quel risultato viene utilizzato affermando che, approssimativamente, $\hat{\theta}_n$ ha distribuzione normale con media pari a θ_0 e varianza pari al reciproco di $n\, I_1(\hat{\theta}_n)$. In tal modo è semplice costruire un intervallo di confidenza approssimato per il parametro θ. Utilizzando la (2.27), infatti, risulta che

$$\Pr\left(-z_{1-\alpha/2} < \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{I_1(\hat{\theta}_n)} < z_{1-\alpha/2}\right) = 1 - \alpha, \tag{2.28}$$

dove z_β rappresenta il percentile di ordine β di una distribuzione normale standardizzata. Riscrivendo l'evento nella (2.28) in altro modo si ottiene dunque

$$\Pr\left(\hat{\theta}_n - \frac{1}{\sqrt{nI_1(\hat{\theta}_n)}} z_{1-\alpha/2} < \theta_0 < \hat{\theta}_n + \frac{1}{\sqrt{nI_1(\hat{\theta}_n)}} z_{1-\alpha/2}\right) = 1 - \alpha,$$

Denotando con $\text{s.e.}(\hat{\theta}_n) = 1/\sqrt{nI_1(\hat{\theta}_n)}$ l'errore standard associato a $\hat{\theta}_n$, si ha dunque che

$$\left(\hat{\theta}_n - z_{1-\alpha/2}\ \text{se}(\hat{\theta}_n);\ \hat{\theta}_n + z_{1-\alpha/2}\ \text{se}(\hat{\theta}_n)\right), \tag{2.29}$$

rappresenta un intervallo di confidenza di livello $1 - \alpha$.

esempio ??????????????????????????????????

2.8.3 Verifica di ipotesi

Supponiamo, per semplicità che il parametro θ sia reale e si vogliano confrontare le due ipotesi

$$H_0: \quad \theta = \theta_0 \qquad \text{vs.} \qquad H_1: \quad \theta \neq \theta_0$$

La quantità intuitivamente importante per misurare la compatibilità dei dati con le due ipotesi è la verosimiglianza relativa di θ_0, ovvero il rapporto di verosimiglianza

$$\frac{L(\theta_0; \mathbf{x})}{L(\hat{\theta}, \mathbf{x})}, \tag{2.30}$$

una misura dell'evidenza fornita dall'esperimento a favore dell'ipotesi nulla θ_0 relativamente al valore più "verosimile" $\hat{\theta}$. Ovviamente si può usare una qualunque trasformazione monotona di (2.30): per diversi motivi, teorici e di calcolo, si preferisce utilizzare la quantità

$$W(\mathbf{x}) = -2 \log \frac{L(\theta_0; \mathbf{x})}{L(\hat{\theta}, \mathbf{x})} = -2 \left(\ell(\theta_0) - \ell(\hat{\theta}) \right).$$

$W(\mathbf{x})$ va letta come un'indicatore dell'evidenza sperimentale contro H_0; valori grandi di $W(\mathbf{x})$ favoriscono l'ipotesi alternativa. Tuttavia, come nel caso della stima per intervalli, non è semplice definire esattamente cosa si intende per piccolo, e non è quindi semplice, sulla base del risultato $W(\mathbf{x}_0)$ osservato, operare una scelta tra le due ipotesi. Questa impasse è risolvibile solo attraverso un utilizzo delle proprietà frequentiste della quantità $W(\mathbf{x})$ o, come vedremo nei capitoli successivi, attraverso una impostazione bayesiana. Ci limitiamo qui ad illustrare brevemente come utilizzare la statistica $W(\mathbf{x})$ in ambito frequentista.

Per quanto visto nella §2.7,

$$\log \frac{L(\theta_0; \mathbf{x})}{L(\hat{\theta}, \mathbf{x})} \approx -\frac{1}{2} j(\hat{\theta})(\theta_0 - \hat{\theta})^2,$$

da cui $W(\mathbf{x}) \approx j(\hat{\theta})(\theta_0 - \hat{\theta})^2$. In problemi regolari, e per grandi numerosità campionarie, come visto nella (2.14) e nella §2.8.1, si ha che $j(\hat{\theta}) \approx nI_1(\hat{\theta})$, e $\hat{\theta} \approx \theta^*$, dove θ^* indica il vero valore di θ. Ne segue che, sotto l'ipotesi nulla H_0, la statistica $W(\mathbf{x})$ può essere scritta, con lo stesso ordine di approssimazione [2], come

$$W_e(\mathbf{x}) = nI(\hat{\theta})(\theta_0 - \hat{\theta})^2,$$

dove W_e viene chiamata solitamente la statistica di Wald. Sapendo inoltre che lo stimatore di massima verosimiglianza ha distribuzione asintotica normale $N(\theta^*, j(\theta^*))$, ed utilizzando le relazioni precedenti, è possibile dimostrare facilmente che, sotto l'ipotesi H_0, sia la statistica $W(\mathbf{x})$ che la statistica $W_e(\mathbf{x})$ hanno distribuzione asintotica di tipo χ^2_1. In ambito frequentista questo risultato viene utilizzato per "calibrare" il risultato osservato x_0 rispetto al resto dello spazio campionario $\mathcal{X}$. Si definisce infatti valore p (o p-value) la probabilità, quando è vera l'ipotesi nulla H_0, di osservare risultati dell'esperimento altrettanto o più sfavorevoli (in termini della statistica W oppure W_e) di x_0 nei confronti di H_0. In formula, si dice valore p la quantità

$$\Pr\left(W(\mathbf{X}) \geq W(x_0); H_0\right). \tag{2.31}$$

Esempio 2.20 [*Distribuzione esponenziale*]
Si osserva un campione $X_1, \cdots, X_n$ di v.a. condizionatamente a θ indipendenti e somiglianti con distribuzione esponenziale $Exp(\theta)$. La funzione di verosimiglianza è

$$L(\theta) \propto \theta^n \exp\left\{-n\bar{x}\theta\right\},$$

dove $\bar{x}$ è la media campionaria. Si vuole verificare l'ipotesi nulla $H_0 : \theta = \theta_0 = 2$ contro l'alternativa $H_1 : \theta \neq 2$.

Poich $S(\theta) = n(\theta^{-1} - \bar{x})$ e $\hat{\theta} = \bar{x}^{-1}$, si ottiene facilmente che $J(\hat{\theta}) = n\bar{x}^2$ e $I_n(\theta) = n/\theta^2$. Le due statistiche test danno allora lo stesso valore

$$W(\mathbf{x}_0) = W_e(\mathbf{x}_0) = n\bar{x}^2(\theta - \bar{x}^{-1})^2 = n(2\bar{x} - 1)^2;$$

La distribuzione campionaria di W e W_e, sotto l'ipotesi nulla, è di tipo χ^2_1; il p-value associato all'osservazione $\bar{x}$ è pertanto

$$p(\bar{x}) = \Pr\left(W \geq n(2\bar{x} - 1)^2\right),$$

rappresentato graficamente in Figura 2.7[3]

♢

I metodi di verifica di ipotesi basati sulla statistica $W(\mathbf{x})$ (o sue approssimazioni) possono essere estesi a situazioni più generali di quelle descritte fin qui; in particolare, si riesce a determinare la distribuzione asintotica di $W(\mathbf{x})$ sotto l'ipotesi nulla, ogniqualvolta l'ipotesi nulla H_0 può essere espressa come un sottospazio proprio di Ω [2].

altro esempio

2.9 Il principio di verosimiglianza

L'idea di basare le procedure inferenziali esclusivamente sull'informazione derivante dalla funzione di verosimiglianza, ovvero condizionatamente al risultato osservato, viene formalizzata nel cosiddetto *principio di verosimiglianza*, del quale esistono due versioni, denominate, rispettivamente, debole e forte.

Definizione 2.15 [**Principio di verosimiglianza debole**]*: Dato un esperimento del tipo (2.1), l'informazione relativa al parametro θ, fornita da un campione $\mathbf{x} = (x_1, \cdots, x_n)$ è interamente contenuta nella funzione di verosimiglianza. Se $\mathbf{x}_1$ e $\mathbf{x}_2$ sono due campioni diversi che producono due funzioni di verosimiglianza differenti solo per una costante moltiplicativa, ovvero*

$$L(\theta; \mathbf{x}_1) = cost \times L(\theta; \mathbf{x}_2) \qquad per\ ogni\ \theta, \tag{2.32}$$

allora i due campioni forniscono identica informazione sul parametro.

La versione forte del principio ammette che, anche qualora le due verosimiglianze provengano da modelli diversi, aventi però, lo stesso spazio parametrico, le conclusioni inferenziali debbano essere le stesse.

Definizione 2.16 [**Principio di verosimiglianza forte**]*: Dati due esperimenti statistici $\mathcal{E}_1 = (\mathcal{X}_1, \mathcal{P}_1, \Omega)$ e $\mathcal{E}_2 = (\mathcal{X}_2, \mathcal{P}_2, \Omega)$, se si osserva un campione $\mathbf{x}_1$ da $\mathcal{E}_1$ e un campione $\mathbf{x}_2$ da $\mathcal{E}_2$, e le funzioni di verosimiglianza associate sono tali che*

$$L_1(\theta; \mathbf{x}_1) = cost \times L_2(\theta; \mathbf{x}_2) \qquad per\ ogni\ \theta, \tag{2.33}$$

allora i due campioni forniscono identica informazione sul parametro.

Esempio 2.21 [*Campionamento diretto e inverso*] La popolazione di un certo comune in Sardegna presenta una specifica caratteristica genetica G con frequenza θ che si vuole stimare. Si decide allora di sottoporre ad analisi specifica due campioni i.i.d. di individui estratti dalla popolazione.

[3] La figura ha un andamento qualitativamente simile ad una funzione di verosimiglianza espressa in termini del parametro θ^{-1}. Questa interpretazione è però non corretta in quanto sull'asse delle ascisse c'è il risultato osservato e non il parametro; inoltre il p-value non rappresenta la verosimiglianza del risultato effettivamente osservato bensì, come è chiaro dalla (2.31), la probabilità associata ad un evento più generale ($W(X) \geq W(x_0)$). Riprenderemo questi argomenti a proposito della critica bayesiana all'uso del p-value

Nel laboratorio A si prestabilisce la dimensione n del campione e si osserva che k delle unità statistiche analizzate presentano la caratteristica G. Si tratta quindi di un semplice esperimento bernoulliano, e la funzione di verosimiglianza del parametro θ associabile al primo esperimento è

$$L_1(\theta) = \theta^k (1-\theta)^{n-k}.$$

Nel laboratorio B, invece, si decide di continuare ad analizzare pazienti fino a raggiungere il numero di k individui con la caratteristica G. In questo caso l'aspetto aleatorio dell'esperimento è fornito dal numero X di individui senza la caratteristica G che sarà necessario analizzare prima di osservare il k -esimo "successo". È ben noto allora che la distribuzione di X e del tipo BiNeg(k, θ). La funzione di verosimiglianza associata al parametro θ in questo secondo caso risulta allora

$$L_2(\theta) = \Pr(X = j; k, \theta) = \binom{k+j-1}{j-1} \theta^k (1-\theta)^j$$

Se, per caso, si verifica che $X = n - k$, ovvero anche nel secondo laboratorio il numero complessivo di individui analizzati è pari ad n, avremo che

$$L_2(\theta) = \binom{n-1}{n-k-1} \theta^k (1-\theta)^{n-k} \propto L_1(\theta).$$

Dal punto di vista del principio di verosimiglianza, i due esperimenti sono assolutamente equivalenti come livello d'informazione che producono su θ. Dal punto di vista del campionamento ripetuto, alla base della filosofia statistica classica, le cose sono differenti: ad esempio, la ricerca dello stimatore di minima varianza tra quelli non distorti condurrebbe, nel primo caso, all'uso della frequenza campionaria $\hat{\theta}_1 = K/n$, nel secondo caso, invece, lo stimatore da utilizzare sarebbe $\hat{\theta}_2 = (K-1)/(n-1)$. ♢

L'esempio precedente mostra chiaramente come l'utilizzo di stimatori basati sulle proprietà frequentiste di non distrosione o di minima varianza sia non sempre compatibile con il principio di verosimigilanza. [9], riprendendo sostanzialmente alcuni risultati di [18], hanno dimostrato che il principio di verosimiglianza è una conseguenza diretta dell'adozione di altri due principi ben più radicati e unanimemente riconosciuti nella letteratura statistica: il *principio di sufficienza* e il *principio di condizionamento debole*. Il principio di sufficienza, in parole semplici, sostiene che l'informazione relativa ad un parametro fornita da un campione è tutta contenuta nelle statistiche sufficienti relativamente a quel parametro. Il principio di condizionamento, invece, sostiene che lo spazio campionario di riferimento dovrebbe essere il più omogeneo possibile. Illustriamo il concetto con un famoso esempio dovuto a Cox. Supponiamo di voler stimare il valore medio θ di una grandezza X e si hanno a disposizione due laboratori, uno in cui la deviazione standard delle n misurazioni è pari a $\sigma_1 = 1$, e un altro in cui la deviazione standard è pari a $\sigma_2 = 5$. Si lancia una moneta regolare per scegliere il laboratorio e si decide per il primo laboratorio. Un'applicazione immediata del principio di condizionamento, in questo caso, ci fa affermare che l'errore standard della media campionaria sarà pari a $1/n$ ovvero verrà calcolata sulla base delle caratteristiche del solo laboratorio coinvolto effettivamente nell'esperimento.

Il risultato di Birnbaum appare paradossale in quanto fa discendere, da due principi generalmente accettati, un altro principio, quello di verosimiglianza, che risulta invece incompatibile con un'impostazione classica dell'inferenza, basata principalmente sul principio del *campionamento ripetuto*, secondo il quale una procedura statistica deve essere valutata secondo la sua efficacia

globale, ovvero non solo sul campione osservato bensì sull'insieme di tutti i possibili campioni osservabili.

La questione è di natura fondazionale e non ci dilungheremo ulteriormente. Anticipiamo solamente che l'impostazione bayesiana, adottata in questo testo è del tutto compatibile con il principio di verosimiglianza, se si eccettuano alcuni criteri di scelta delle distribuzioni a priori, che descriveremo nella §5.2.

2.10 Eliminazione dei parametri di disturbo

Un modello statistico contiene in genere molti parametri incogniti. Non sempre tuttavia essi rappresentano quantità di interesse diretto per il ricercatore. Molti spesso, parametri addizionali sono introdotti con lo scopo di costruire un modello più flessibile e in grado di adattarsi alle caratteristiche osservate nei dati. Nell'Esempio 2.7, relativo ad un problema di cattura e ricattura, il parametro di interesse era soltanto il numero N di individui che formavano la popolazione, ma evidentemente un modello statistico in grado di formalizzare al meglio la componente aleatoria delle occasioni di cattura deve tener conto del parametro p, probabilità di cattura dei singoli individui nelle singole occasioni, per quanto questo non fosse un obiettivo primario dello studio. In casi come questi sarebbe, almeno a livello teorico, ideale poter disporre di una funzione di verosimiglianza che dipenda esclusivamente dai parametri di interesse. Questo problema è noto come il problema della "eliminazione dei parametri di disturbo" ed ha suscitato un notevole interesse di ricerca soprattutto negli ultimi quindici anni del secolo scorso. Formalmente, considerato un modello statistico del tipo (2.1), sia $\theta \in \Omega$ il generico valore del parametro incognito che assumiamo, in questa sezione avere dimensione $p \geq 2$. Partizioniamo allora θ in due componenti, ovvero $\theta = (\psi, \lambda)$, dove ψ, di dimensione $q < p$ rappresenta il vettore di componenti di θ di diretto interesse, mentre λ di dimensione $p - q$ è il cosiddetto parametro di disturbo. L'obiettivo è dunque quello di costruire, a partire dalla funzione di verosimiglianza completa $L(\psi, \lambda)$, una nuova funzione di verosimiglianza che esprima l'evidenza sperimentale a favore dei possibili valori di ψ e che dipenda, il meno possibile, dai valori incogniti di λ, che viene detto *di disturbo*, proprio perch, se fosse noto, sapremmo esattamente come procedere in modo canonico. Esistono molte possibili soluzioni a questo problema, nessuna pienamente soddisfacente, se non in casi molto particolari. Per approfondimenti si possono consultare i testi di [66], [81] e [31]. Qui ci limitiamo ad illustrare la soluzione più naturale e più generale, ovvero la costruzione di una funzione di *verosimiglianza profilo*. Per ψ fissato, sia $\hat{\lambda}_\psi$ il valore che massimizza la verosimiglianza intesa come sola funzione di λ ovvero

$$\hat{\lambda}_\psi = \operatorname{argmax}_\lambda L(\psi, \lambda); \tag{2.34}$$

Assumiamo qui per semplicità espositiva che il valore $\hat{\lambda}_\psi$ esista finito per ogni ψ. Possiamo allora definire verosimiglianza profilo per il parametro d'interesse ψ in ψ_0 la funzione di verosimiglianza originale calcolata nel punto $(\psi_0, \hat{\lambda}_{\psi_0})$. Più in generale la funzione di verosimiglianza profilo sarà definita da

$$\hat{L}(\psi) = L\left(\psi, \hat{\lambda}_\psi\right). \tag{2.35}$$

La funzione di verosimiglianza profilo traccia una curva sulla superficie di verosimiglianza passando, per ogni ψ_0 fissato, per il valore di λ che massimizza la striscia di verosimiglianza definita da

$L(\psi_0, \lambda)$. Una immediata interpretazione della (2.35) si può avere in negativo: se la funzione di verosimiglianza profilo (adeguatamente normalizzata come una comune funzione di verosimiglianza) assume un valore basso per un certo valore di $\psi = \psi_0$, allora si può dire che tutte le coppie (ψ_0, λ) sono poco verosimili.

Esempio 2.22 [*Distribuzione normale con* (μ, σ^2) *incogniti.*] Sia $(X_1, \cdots, X_n) \overset{\texttt{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, e supponiamo di essere interessati al solo parametro μ. Quindi in questo caso

$$\psi = \mu; \qquad \lambda = \sigma^2;$$

La funzione di verosimiglianza completa è allora

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\} \\ &\propto \frac{1}{\sigma^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\}, \\ &= \frac{1}{\sigma^n} \exp\left\{-\frac{n}{2\sigma^2}\left[s^2 + (\bar{x} - \mu)^2\right]\right\}, \end{aligned}$$

dove s^2 è la varianza campionaria $\sum(x_i - \bar{x})^2/n$. Per μ fissato, si vede facilmente che

$$\hat{\sigma}^2 = s^2 + (\bar{x} - \mu)^2$$

La funzione di verosimiglianza profilo è dunque

$$\hat{L}(\mu) = \frac{n^{n/2} e^{-n/2}}{(s^2 + (\bar{x} - \mu)^2)^{n/2}} \propto \left[1 + \left(\frac{\bar{x} - \mu}{s}\right)^2\right]^{-\frac{n}{2}},$$

che coincide con il nucleo di una densità di tipo t di Student con $n-1$ gradi di libertà, in assonanza con la nota procedura classica in cui è lo stimatore $\bar{X}$ ad avere distribuzione t con lo stesso numero gradi di libertà. ♢

Pur di generale applicabilità, la funzione di verosimiglianza profilo non sempre fornisce risultati convincenti come mostra l'esempio seguente.

Esempio 2.23 [*Il problema di Neyman e Scott.*] Si abbia un campione composto da $2n$ osservazioni, tutte indipendenti tra loro e somiglianti a due a due, ovvero per $i = 1, \cdots, n$, siano

$$X_{i1}, X_{i2} \overset{\texttt{iid}}{\sim} \mathrm{N}(\mu_i, \sigma^2).$$

In pratica le osservazioni provengono, a coppie, da n popolazioni normali aventi tutte la stessa varianza incognita ma medie incognite e diverse. Consideriamo il caso in cui il parametro di interesse è σ^2 mentre il vettore delle n medie $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)$ è il parametro di disturbo. La funzione di verosimiglianza completa, ponendo $\bar{x}_i = (x_{i1} + x_{i2})/2$, è

$$\begin{aligned} L(\boldsymbol{\mu}, \sigma^2) &\propto \frac{1}{\sigma^{2n}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{2}(x_{ij} - \mu_i)^2\right) \\ &= \frac{1}{\sigma^{2n}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_{i1} - \mu_i)^2 + (x_{i2} - \mu_i)^2\right]\right) \\ &= \frac{1}{\sigma^{2n}} \exp\left(-\frac{1}{\sigma^2}\sum_{i=1}^{n}\left[(\bar{x}_i - \mu_i)^2 + \frac{1}{4}(x_{i1} - x_{i2})^2\right]\right) \end{aligned}$$

Si vede facilmente che per ogni valore fissato di σ^2, risulta

$$\hat{\mu}_{i,\sigma^2} = \hat{\mu}_i = \bar{x}_i,$$

per cui la funzione di verosimiglianza profilo è

$$\hat{L}(\sigma^2) \propto \frac{1}{\sigma^{2n}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\frac{x_{i1}-x_{i2}}{\sqrt{2}}\right]^2\right) \tag{2.36}$$

È semplice verificare che il valore che massimizza la funzione di verosimiglianza profilo di σ^2 è

$$\hat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n}\left[\frac{x_{i1}-x_{i2}}{\sqrt{2}}\right]^2$$

Lo stimatore appena determinato non è nemmeno consistente; infatti è un semplice esercizio verificare che la v.a

$$W_i = \frac{X_{i1}-X_{i2}}{\sigma\sqrt{2}} \sim N(0,1), \qquad i = 1, \cdots n.$$

Perciò $W_i^2 \sim \chi_1^2$ e $\mathbb{E}(W_i) = 1$. Per la legge forte dei grandi numeri allora, al crescere di n risulta

$$\hat{\sigma}^2 \to \frac{\sigma^2}{2}.$$

Da notare che questo esempio si differenzia in modo notevole da quelli studiati finora, almeno per quanto concerne le questioni asintotiche: infatti, al crescere di n, cresce anche il numero dei parametri di disturbo. ♢

2.11 La famiglia esponenziale

Un modello statistico $\mathcal{E} = (\mathcal{X}, \mathcal{P}, \Omega)$, in cui l'elemento generico di $\mathcal{P}$ può essere scritto come

$$p(x;\theta) = h(x)\,\exp\left\{-G(\boldsymbol{\theta}) + \sum_{j=1}^{k}\psi_j(\boldsymbol{\theta})t_j(x)\right\} \tag{2.37}$$

in cui il supporto della distribuzione non dipende da θ e consiste nella chiusura dell'insieme di tutti i valori $x \in \mathbb{R}^p$, tali che $h(x) > 0$, si chiama *famiglia esponenziale di ordine k*. Nella formula (2.37), θ è un vettore di $\mathbb{R}^d$, e $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k)$ è il vettore delle funzioni parametriche che compaiono nell'espressione della densità. Spesso, ma non necessariamente, la dimensione d del parametro θ coincide con l'ordine k della famiglia. In tal caso il vettore $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$ può essere considerato una riparametrizzazione (cosiddetta *canonica*), del vettore $\boldsymbol{\theta}$. In questo caso la (2.37) diventa

$$p(x;\psi) = h(x)\,\exp\left\{-\tilde{G}(\psi) + \sum_{j=1}^{k}\psi_j t_j(x)\right\} \tag{2.38}$$

Esempio 2.24 [*Distribuzione binomiale*]
La famiglia binomiale $\mathrm{Bin}(n,\theta)$, con n noto è un esempio di famiglia esponenziale di ordine 1, in quanto la distribuzione può scriversi, per $x = 0, 1, \cdots, n$, come

$$p(x;\theta) = \binom{n}{x} \exp\{x \log\theta + (n-x)\log(1-\theta)\}$$
$$= \binom{n}{x} \exp\left\{x \log\frac{\theta}{1-\theta} + n\log(1-\theta)\right\};$$

basterà dunque porre

$$G(\theta) = -n\log(1-\theta); \quad h(x) = \binom{n}{x} \quad t_1(x) = x, \quad \psi(\theta) = \log\frac{\theta}{1-\theta}.$$

Nella parametrizzazione canonica, dunque, avremo $\tilde{G}(\psi) = n\log(1+e^{\psi})$, e l'espressione della distribuzione diventa

$$p(x;\psi) = \binom{n}{x} \exp\left\{\tilde{G}(\psi) + x\psi\right\}. \tag{2.39}$$

L'espressione (2.39) non appare ora particolarmente conveniente; il suo significato e soprattutto quello del parametro canonico ψ, saranno chiari nel Capitolo 10, quando verranno trattati i modelli lineari generalizzati in un'ottica bayesiana ♢

Esempio 2.25 [*Distribuzione normale*]
La famiglia gaussiana $\mathrm{N}(\mu, \sigma^2)$ è un esempio di famiglia esponenziale di ordine 2. Basta infatti scrivere la densità come

$$p(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left(x^2 - 2\mu x + \mu^2\right)\right\}$$
$$= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right\};$$

Anche qui l'espressione (2.37) si ottiene facilmente ponendo

$$G(\mu,\sigma^2) = \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log\sigma^2\right) \quad h(x) = \frac{1}{\sqrt{2\pi}},$$

e

$$t_1(x) = x^2, \psi_1(\mu,\sigma^2) = -\frac{1}{2\sigma^2}, \quad t_2(x) = x, \quad \psi_2(\mu,\sigma^2) = \frac{\mu}{\sigma^2}$$

♢

Per una distribuzione esponenziale di ordine 1 nella sua forma canonica l'espressione della funzione generatrice dei momenti è molto semplice: se X è una v.a con distribuzione (2.38), allora

$$M_X(u) = \exp\left\{\tilde{G}(\psi+u) - \tilde{G}(\psi)\right\};$$

Esempio 2.24 (continua). In questo caso, la funzione generatrice dei momenti vale

$$M_X(u) = \frac{\exp\{n\log(1+e^{\psi+u})\}}{\exp\{n\log(1+e^{\psi})\}} = (1-\theta+\theta\,e^u)^n.$$

Più in generale, nel caso in cui $k=1$ e la famiglia è regolare, il calcolo dei momenti delle statistiche $t_j(\cdot)$ che compaiono nella (2.37) è molto semplice [2]. Ad esempio, tenendo conto che la (2.37) integra ad 1 e invertendo l'ordine degli operatori di integrazione e di derivazione si ottiene,

$$\int_{\mathcal{X}} \frac{\partial}{\partial\theta} p(y;\theta)dx = 0;$$

$$\int_{\mathcal{X}} p(y;\theta)\left(-G'(\theta) + \psi_1'(\theta)t_1(\theta)\right)dx = 0,$$

ovvero

$$\mathbf{E}\left(\mathrm{t}_1(\mathrm{X})\right) = \frac{\mathrm{G}'(\theta)}{\psi_1'(\theta)}.$$

Dato un campione di osservazioni i.i.d. con distribuzione esponenziale, la distribuzione congiunta si può scrivere come

$$p(\mathbf{x};\boldsymbol{\theta};) = \prod_{i=1}^{n}\left[h(x_i)\right]\exp\left\{-nG(\boldsymbol{\theta}) + \sum_{j=1}^{k}\psi_j(\boldsymbol{\theta})\sum_{i=1}^{n}t_j(x_i)\right\},$$

che appartiene ancora ad una famiglia esponenziale dello stesso ordine. Il vettore delle statistiche sufficienti, in questo caso, è dato da

$$\left(\sum_{i=1}^{n}t_1(x_i),\ldots,\sum_{i=1}^{n}t_k(x_i)\right).$$

2.12 Anomalie della funzione di verosimiglianza

Abbiamo più volte sottolineato come la funzione di verosimiglianza rappresenti il veicolo attraverso cui l'esperimento fornisce evidenza pro o contro le possibili spiegazioni della realtà, stilizzate nei diversi valori che il parametro può assumere. Questa posizione è condivisa da molti statistici, di orientamento bayesiano e non, primi fra tutti i sostenitori di un'impostazione neo-fisheriana che trova nel testo di [29] il suo manifesto, ma si vedano anche [2] e [67]. Tuttavia, in questa sezione si sono concentrati alcuni aspetti meno scontati, presentati spesso come controesempi all'intera costruzione inferenziale basata sulla funzione di verosimiglianza. Secondo lo scrivente, essi rappresentano interessanti spunti per un ulteriore approfondimento sul ruolo del modello statistico e della funzione di verosimiglianza nel ragionamento induttivo.

Esempio 2.11 (continua). Abbiamo già sottolineato come questa situazione risulti anomala in quanto la statistica sufficiente ha dimensione 2 mentre il parametro è scalare. Costruiamo ora la funzione di verosimiglianza. Per quanto già visto in precedenza, essa vale

$$L(\theta;\mathbf{x}) \propto \begin{cases}\theta^{-n} & x_{(n)}/2 < \theta < x_{(1)} \\ 0 & \text{altrimenti}\end{cases}; \tag{2.40}$$

Dunque, qualunque siano i valori osservati del campione, la funzione di verosimiglianza risulta decrescente in un insieme limitato; in altri termini la metà del massimo campionario osservati risulta più *verosimile* del minimo campionario. Questo risultato, per quanto banale dal punto di vista matematico e pienamente giustificabile in un'impostazione puramente frequentista (vedi oltre), non è facilmente interpretabile da un punto di vista condizionato: le n osservazioni sono variabili aleatorie con legge uniforme in un intervallo di ampiezza e posizione aleatorie e, apparentemente, nulla suggerisce che il massimo delle osservazioni sia in un qualche senso, più vicino a 2θ di quanto il minimo non sia vicino a θ. Va inoltre sottolineato che la funzione di verosimiglianza risulta decrescente nell'intervallo indicato dalla (2.40): questo significa che $x_{(1)}$ rappresenta il valore di *minima* verosimiglianza tra quelli non esclusi dal risultato campionario.

Risultati simili sono ottenibili seguendo una logica frequentista. Siano $T_1 = X_{(1)}$ e $T_2 = X_{(n)}/2$, questa volta considerati come stimatori. Risultati standard relativi alle statistiche d'ordine (vedi Appendice C.1) ci permettono di stabilire che

$$\mathbf{E}_\theta(T_1) = \frac{n+2}{n+1}\,\theta, \qquad \mathbf{E}_\theta(T_2) = \frac{2n+1}{2n+2}\,\theta$$

e

$$\mathrm{Var}_\theta(T_1) = \frac{n}{(n+1)^2(n+2)}\,\theta^2, \qquad \mathrm{Var}_\theta(T_2) = \frac{n}{4(n+1)^2(n+2)}\,\theta^2$$

e

$$\mathrm{Cov}(T_1, T_2) = \frac{1}{2(n+1)^2(n+2)}\,\theta^2.$$

Ne segue che entrambi gli stimatori sono asintoticamente non distorti, con lo stesso tasso di convergenza al vero valore di θ; inoltre, in termini di errore quadratico medio (ovvero la somma della varianza e del quadrato della distorsione) lo stimatore T_1 è uniformemente migliore di T_2. Più in generale, se si considerano tutte le combinazioni convesse di T_1 e T_2, nella forma

$$T_c = c\,T_1 + (1-c)\,T_2, \qquad 0 < c < 1,$$

si può dimostrare che il valore di c che minimizza l'errore quadratico medio è, qualunque sia il valore di n, $c_0 = 8/28$; anche questa conclusione non appare immediatamente interpretabile dal punto di vista inferenziale.

2.13 Esercizi

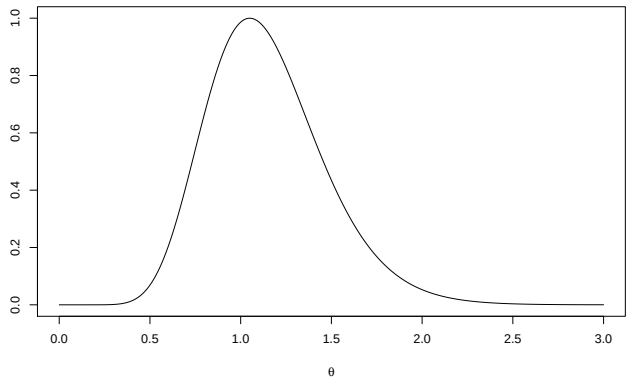

Figura 2.5. Funzione di verosimiglianza per l'Esempio 2.9:

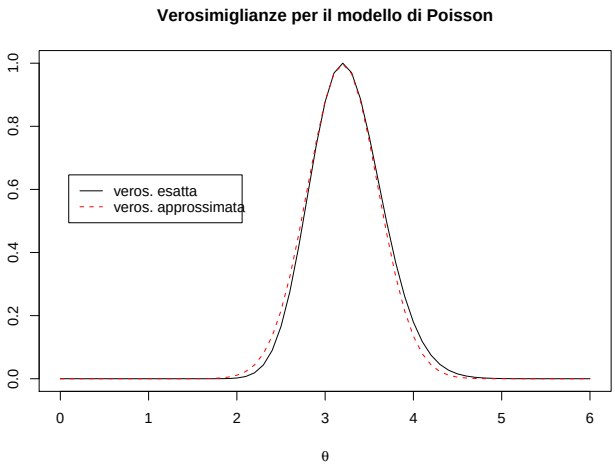

Figura 2.6. Funzioni di verosimiglianza esatta e approssimata per il modello di Poisson

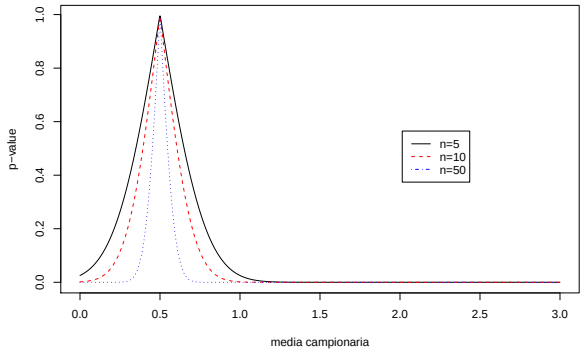

Figura 2.7. Livello del p-value per il test $H_0 : \theta = 2$ vs. $H_1 : \theta \neq 2$ al variare della media campionaria e per diversi valori di n. Si noti come la massima compatibilità del campione osservato con l'ipotesi nulla si ha per $\bar{x} = 0.5$. La cosa non è sorprendente visto che $1/\bar{X}$ è lo stimatore di massima verosimiglianza per il modello esponenziale.

.

3

Inferenza statistica da un punto di vista bayesiano

3.1 Il teorema di Bayes e il processo induttivo

Scopo primario dell'inferenza statistica, almeno nella tradizione classica, è quello di acquisire ulteriore conoscenza su quantità incerte, sulle quali spesso si hanno informazioni parziali, non sufficienti a eliminare del tutto l'incertezza: questo processo induttivo è spesso la necessaria premessa a un processo decisionale, dove le informazioni in precedenza acquisite, adeguatamente filtrate, vengono utilizzate per scegliere "quale" decisione prendere, fra diverse possibili. Gli esempi che seguono possono contribuire a chiarire questi concetti.

Esempio 3.1 [*Sondaggio pre-elettorale.*]
Per un determinato seggio elettorale, si vuole stimare la percentuale di elettori che, nelle prossime elezioni politiche, voterà per le due coalizioni di centro-destra e di centro-sinistra. ♢

Esempio 3.2 [*Previsioni finanziarie.*]
La società *Truffa.net* è regolarmente quotata in borsa e, il giorno 30 giugno 2006, ogni sua azione vale 4 euro. Avendo a disposizione i valori delle azioni della società per tutti i giorni feriali del periodo maggio '05 - maggio '06, vogliamo "prevedere" il valore delle azioni della *Truffa.net* al giorno 15 luglio 2006. ♢

Esempio 3.3 [*Numerosità di una popolazione.*]
Si vuole stimare il numero di extra-comunitari senza regolare permesso di soggiorno presenti in Italia ad una certa data. ♢

Gli esempi precedenti presentano diversi gradi di difficoltà e vanno affrontati con diversi strumenti: la caratteristica comune è però quella di voler stimare una quantità: negli esempi (3.1) e (3.2) si tratta di una grandezza futura, su cui possiamo, al più azzardare delle previsioni; nell'esempio (3.3) la quantità incognita è invece già determinata e l'incertezza è dovuta al fatto che non siamo in grado di misurarla con precisione.

Esempio 3.1 (continua). Considereremo ora in dettaglio l'esempio 3.1 in una versione ancor più semplificata: Immaginiamo cioè di essere interessati soltanto alla percentuale di voti che otterrà lo schieramento di centro-sinistra. Assumeremo che l'impostazione classica del problema sia nota

al lettore. Dalla popolazione di votanti in quel seggio elettorale si estrae un campione casuale[1] di di n elettori (il valore n è considerato qui fissato in anticipo) e a ciascuno di loro si chiede per chi voterà alle prossime elezioni; a ciascun componente del campione si associa una variabile aleatoria Y_j che vale 1 se l'elettore vota per il centro-sinistra e vale 0 altrimenti (se vota altri partiti, oppure scheda bianca o nulla). Le distribuzioni di probabilità possibili da associare alle variabili aleatorie $Y_1, \cdots, Y_n$ costituiscono, nel loro insieme, il modello statistico. In questo caso il modello si costruisce partendo dalla considerazione che il valore incognito da stimare è un valore $\theta \in [0, 1]$. Condizionatamente al valore assunto da θ, le n osservazioni $Y_1, \cdots, Y_n$ vengono supposte indipendenti e tutte con la stessa distribuzione di probabilità, ovvero, per $j = 1, \cdots, n$, si ha

$$P(Y_j = 1|\theta) = 1 - P(Y_j = 0|\theta) = \theta.$$

Per quanto visto nel Capitolo 2, siamo in presenza di un modello statistico binomiale, in cui

$$Y_1, \cdots, Y_n \sim Be(\theta), \qquad j = 1, \cdots, n;$$

La funzione di verosimiglianza associata a tale esperimento è

$$L(\theta; \mathbf{y}) = \prod_{i=1}^{n} P(Y_i = y_i|\theta),$$

dove abbiamo usato la convenzione, piuttosto comune, di indicare con lettere maiuscole, le variabili aleatorie e con lettere minuscole i valori assunti dalle stesse nell'effettiva realizzazione dell'esperimento. Senza perdere in generalità supponiamo che l'esperimento si concluda con k successi (valori di Y_i uguali a 1) e $n - k$ insuccessi (valori di Y_i uguali a 0). Si ha allora

$$L(\theta) = \theta^k (1 - \theta)^{n-k}. \tag{3.1}$$

Per fissare ancor più le idee, supponiamo, come nel Capitolo 2, che $n = 10$, e $k = 7$; ne segue che

$$L(\theta) = \theta^7 (1 - \theta)^3.$$

La funzione di verosimiglianza "pesa" i diversi valori possibili che θ può assumere, sulla base dei dati rilevati. Nell'esempio specifico, il valore $\hat{\theta} = 0.7$ appare come il più "verosimile". Prima di proseguire nella trattazione matematico-statistica dell'esempio (3.1), è bene considerare altri esempi, molto diversi tra loro, che però presentano, una volta formalizzati, una struttura identica a quella dell'Esempio 3.1. [58] considera le tre seguenti situazioni:

S1) Tizio sostiene di essere in grado di riconoscere se un brano musicale è stato scritto da Mozart oppure da Beethoven dopo appena quattro note. Gli sottoponiamo allora gli incipit di dieci brani scelti a caso dal repertorio dei due autori e verifichiamo le sue capacità.

S2) La signora Bianchi sostiene che bevendo una tazza di tè al latte, è in grado di stabilire se è stato versato prima il latte oppure il t: anche in questo caso sottoponiamo la signora a un test di 10 prove.

S3) Il signor Rossi sostiene di possedere capacità soprannaturali e di essere in grado di prevedere il risultato di un lancio di una moneta regolare; lo stesso, effettuiamo 10 prove sperimentali.

[1] nelle applicazioni concrete di tale schema, ovviamente, occorrerà ricorrere ad una preliminare stratificazione del campione secondo variabili rilevanti: trascuriamo qui questi argomenti per concentrarci esclusivamente sugli aspetti induttivi.

Dal punto di vista formale le tre situazioni sperimentali suddette non differiscono tra loro, n differiscono dall'Esempio 3.1. In tutti i casi si hanno n variabili aleatorie bernoulliane (di tipo 0-1) che assumono il valore 1 con probabilità incognita θ, che danno luogo a un vettore di dati osservati (stringa di valori 0 e 1). Assumiamo che in tutti gli esperimenti si osservino $k = 7$ successi su $n = 10$ prove. Ne segue che le funzioni di verosimiglianza associate ai tre esperimenti saranno del tutto identiche, così come le stime puntuali del parametro incognito θ: valuteremo pari a 0.7 sia la probabilità dell'esperto di musica di riconoscere un brano sia la capacità del presunto sensitivo di prevedere il futuro. Allo stesso modo, l'incertezza relativa a tale stima, espressa dalla $\text{Var}(\hat{\theta}) = \hat{\theta}(1-\hat{\theta})/n = 0.21/10 = 0.021$, sarà uguale nei tre casi. C'è qualcosa di insoddisfacente in questa coincidenza di conclusioni: per quanto l'esperimento statistico abbia fornito lo stesso risultato, è abbastanza ragionevole avere un diverso grado di fiducia sulle affermazioni dei tre personaggi in questione. Ad esempio, un appassionato di musica reputa possibile e ragionevole che un vero esperto possa riconoscere un brano dopo quattro note mentre potrebbe non aver nessuna fiducia su chi sostiene di avere doti di preveggenza, e non saranno certo 10 prove (troppo poche...) a fargli cambiare idea.

Ciò che non è stato considerato finora sono le informazioni a priori che avevamo su θ nei diversi esperimenti. Nell'Esempio 3.1 θ rappresentava la percentuale di votanti per la lista di centro-sinistra in un dato collegio elettorale, e certamente, sulla base delle informazioni relative alle elezioni precedenti, abbiamo idea dell'ordine di grandezza della percentuale di voti che ci possiamo attendere per quella lista. Allo stesso modo molte persone, certo non tutte, darebbero più fiducia all'esperto musicale che non al preveggente. Come inserire tali informazioni nell'analisi statistica? Lo si può fare attraverso il teorema di Bayes: quello che serve è una formalizzazione di tutto ciò che sappiamo su θ in termini di una distribuzione di probabilità *iniziale*. Matematicamente si tratta di determinare una distribuzione π per θ, ovvero una legge di probabilità sui sottoinsiemi di $\boldsymbol{\Omega}$, che d'ora in poi indicheremo genericamente con

$$\Pi(A), \qquad A \subset \boldsymbol{\Omega},$$

dove $\boldsymbol{\Omega}$ rappresenta l'insieme dei valori che possono essere assunti dal parametro θ (negli esempi $\boldsymbol{\Omega} = [0,1]$[2]). Se $\boldsymbol{\Omega}$ ha cardinalità più che numerabile, è prassi utilizzare una legge di probabilità assolutamente continua, che indicheremo con la lettera minuscola π. ??????

Da notare il cambiamento di status di θ che, mentre nell'impostazione classica è considerato una quantità fissata ma incognita, essa diventa, in un'ottica bayesiana, una variabile aleatoria la cui distribuzione iniziale dipende dalle informazioni in nostro possesso, in quel dato contesto geografico, temporale e culturale. Tale distribuzione è forzatamente soggettiva ovvero varia da individuo a individuo perch rappresenta la sintesi delle informazioni che il singolo individuo possiede sul problema specifico. La disponibilità di $\pi(\theta)$ permette di scrivere una formula analoga alla (1.8) valida per variabili aleatorie. La cosiddetta distribuzione *finale* o *a posteriori*, dopo aver osservato il risultato sperimentale $\mathbf{y} = (y_1, \cdots, y_n)$ è

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta)L(\theta;\mathbf{y})}{\int_{\boldsymbol{\Omega}} \pi(\theta)L(\theta;\mathbf{y})d\theta} \tag{3.2}$$

[2] Per motivi matematici si preferisce imporre che la legge di probabilità Π sia definita non proprio su "tutti" i sottoinsiemi di $\boldsymbol{\Omega}$, bensì su una classe di questi, nota come σ-algebra di Borel di $\boldsymbol{\Omega}$, e che gode di alcune proprietà specifiche, come la chiusura rispetto alle operazioni di unione numerabile, negazione e intersezione

Come nel caso della (1.8), il denominatore della (3.2) è una costante di normalizzazione che non dipende da θ e spesso viene indicata come

$$m(\mathbf{y}) = \int_{\boldsymbol{\Omega}} \pi(\theta) L(\theta; \mathbf{y}) d\theta, \tag{3.3}$$

per sottolineare che si tratta della distribuzione marginale del vettore di variabili aleatorie $\mathbf{y}$.

Esempio (3.1) (continua). Supponiamo che le nostre informazioni sul collegio in questione siano molto scarse e assumiamo per θ una distribuzione iniziale uniforme nell'intervallo $[0, 1]$ ovvero

$$\pi_1(\theta) = 1, \;\; 0 < \theta < 1$$

Combinando la distribuzione a priori con la funzione di verosimiglianza (3.1), si ottiene la distribuzione finale

$$\pi_1(\theta|\mathbf{y}) = \frac{\theta^7(1-\theta)^3}{\int_0^1 \theta^7(1-\theta)^3 d\theta} = \frac{\theta^7(1-\theta)^3}{B(8,4)} = 1320\,\theta^7(1-\theta)^3,$$

dove $B(a, b)$ è la funzione Beta di Eulero definita nell'Appendice E. La distribuzione finale in Figura 2 non è altro che la versione normalizzata della funzione di verosimiglianza (3.1): questo avviene perch la distribuzione iniziale uniforme non ha aggiunto informazioni ulteriori rispetto a quanto già fornito dalla funzione di verosimiglianza[3].

Invece di utilizzare la distribuzione uniforme π_1, in presenza di specifiche informazioni su θ (ad esempio, se il collegio elettorale è in Emilia-Romagna abbiamo ragione di credere che la percentuale di voti della lista del centro-sinistra si attesterà intorno a valori superiori al 40-50 per cento), potremmo decidere di utilizzare un'altra distribuzione iniziale, ad esempio unimodale intorno a $\theta = 0.5$. È comune (per motivi che chiariremo in seguito) ma non obbligatorio scegliere la distribuzione all'interno della famiglia di distribuzioni Beta, definita in Appendice E.2). Nel nostro caso, potremmo utilizzare, a titolo di esempio, la distribuzione Beta con parametri ($\alpha = \beta = 5$). Questo significa assegnare a θ, a priori, una media pari a 0.5 e una varianza pari a 0.023 Con calcoli del tutto simili ai precedenti si arriva alla nuova distribuzione a posteriori

$$\pi_2(\theta|\mathbf{y}) = \frac{\theta^{7+5-1}(1-\theta)^{3+5-1}}{\int_0^1 \theta^{7+5-1}(1-\theta)^{3+5-1} d\theta}$$
$$= \frac{\theta^{12-1}(1-\theta)^{8-1}}{B(12,8)} = 604656 \quad \theta^{11}(1-\theta)^7,$$

che rappresenta ancora una distribuzione di tipo Beta con parametri modificati in $\alpha^* = \alpha + 7 = 12$ e $\beta^* = \beta + 3 = 8$. Nella §4.1 verranno riprese, in maggior dettaglio, alcune analisi bayesiane elementari in presenza di dati dicotomici.

3.2 La soggettività delle conclusioni

La prima conclusione che si trae dall'esempio precedente è che le informazioni a priori hanno un ruolo importante nelle inferenza e che tali informazioni introducono nell'analisi una componente soggettiva. Un altro ricercatore, con un diverso bagaglio di conoscenze, potrebbe ben arrivare a diverse conclusioni inferenziali, soprattutto quando l'esperimento, tramite la funzione di

[3] ma la questione della non informatività associabile alla distribuzione uniforme è più sottile e verrà discussa nella§5.2.

verosimiglianza, è poco informativo (ad esempio quando il numero delle osservazioni è piccolo). Questo aspetto è stato ed è tuttora al centro di un acceso dibattito scientifico. I sostenitori dell'impostazione bayesiana affermano che il fare uso di informazioni particolari, *contingenti*, è il modo in cui ogni essere razionale opera in situazioni d'incertezza: del resto persone diverse possono benissimo prendere decisioni differenti anche sulla base di informazioni condivise uguali. Chi invece critica l'approccio bayesiano sostiene che la statistica, per conservare dignità scientifica e per poter essere proficuamente utilizzata nella pratica, deve garantire una *oggettività* delle conclusioni che si possono trarre da un esperimento e perciò queste non possono dipendere dalle informazioni di chi conduce l'esperimento.

Per favorire un compromesso *operativo* tra le due posizioni (più difficile, se non logicamente impossibile, è sperare in un compromesso filosofico), alcuni studiosi hanno proposto dei criteri per la determinazione di distribuzioni iniziali "oggettive" che non debbono rappresentare le informazioni a priori possedute dal ricercatore ma consentono comunque l'utilizzo del teorema di Bayes con quanto ne consegue: queste distribuzioni vengono spesso indicati coi termini *oggettive, convenzionali, di default* oppure *non informative*. Torneremo su questo argomento nel Cap. 5.

3.2.1 La distribuzione a posteriori è il riassunto dell'inferenza.

La formula di Bayes produce la distribuzione finale $\pi(\theta|\mathbf{y})$, che rappresenta la distribuzione di probabilità del parametro θ oggetto di interesse, condizionata al risultato dell'esperimento. In essa è racchiusa tutta l'informazione su θ e su di essa ci si basa per produrre indicatori sintetici, esattamente come si suole fare in ambito descrittivo. Ad esempio, un indicatore sintetico per θ (che, ricordiamo, è in questo contesto, una variabile aleatoria) può essere la media aritmetica, oppure la mediana della distribuzione finale; come indice di variabilità può invece essere utilizzata la varianza di tale distribuzione.

Esempio 3.1 (continua). Abbaimo visto come l'utilizzo di una distribuzione iniziale di tipo Beta$(5, 5)$ per θ produca, una volta osservati 7 successi su 10 prove, una distribuzione finale ancora di tipo Beta ma con parametri, rispettivamente pari a 12 e 8. Da qui si può immediatamente arguire che la media finale per θ è pari a $\mathbf{E}(\theta \mid \mathbf{Y}) = 0.6$ e una misura dell'incertezza intorno a questa stima è fornito dalla varianza finale, pari a $\mathrm{Var}(\theta \mid \mathbf{Y}) = 0.011$. Alternativamente una stima puntuale dei θ può essere rappresentata dalla mediana finale. Con **R**

```
qbeta(.5,shape1=12, shape2=8)
```

che fornisce 0.603, non dissimile dalla media finale.

3.3 La logica dell'induzione: evidenza, inferenza, decisioni

Per apprezzare il ruolo che l'impostazione bayesiana gioca nel ragionamento induttivo è opportuno distinguere i diversi problemi che ci poniamo di fronte ad un certo risultato sperimentale. Questa sezione raccoglie considerazioni già svolte in [78] e [69]. Nel contesto di un esperimento statistico, sia dato un certo risultato osservato x. Il modo di elaborare il risultato x dipende dagli scopi dello studio, ma si possono distinguere le seguenti classi di obiettivi potenziali

- Valutazione dell'evidenza sperimentale a favore di certe ipotesi
- Formulazione di inferenze rispetto a determinate ipotesi (inferenza strutturale) o a risultati futuri (inferenza predittiva)
- Scelta di una decisione entro un insieme predefinito, sapendo che l'effetto corrispondente dipende dalla validità o meno di determinate ipotesi.

Contrariamente alla statistica classica, l'approccio bayesiano distingue in modo naturale fra le succitate categorie. Illustreremo la questione con un esempio.

Esempio 3.4 [*Test diagnostico, [78]*]

Tizio si sottopone ad un test per una certa malattia M. Il test ha le seguenti caratteristiche illustrate nella tabella 3.4 Supponiamo che il test, applicato su Tizio, risulti positivo (T+). Le

	$T+$	$T-$
M	0.95	0.05
NM	0.02	0.98

Tabella 3.1.

conclusioni che si possono trarre da questo risultato possono essere le seguenti

(A)C'è evidenza sperimentale a favore dell'ipotesi che Tizio sia realmente affetto dalla malattia.
(B)Tizio probabilmente non è affetto dalla malattia M.
(C)Tizio dovrebbe essere curato per la malattia M.

Le tre affermazioni, a prima vista, possono sembrare contraddittorie. Ad una più attenta disamina, però, esse risultano perfettamente compatibili. Nel caso (A) occorre confrontare le verosimiglianze delle due ipotesi (M, NM). Poich $P(T^+|M) = 0.95$ e $P(T^+|NM) = 0.02$, si ha

$$\frac{P(T^+|M)}{P(T^+|NM)} = \frac{0.95}{0.02} = 47.5; \tag{3.4}$$

siamo dunque di fronte ad una notevole evidenza sperimentale a favore dell'ipotesi che Tizio sia affetto da M. La quantità (3.4), in ambito bayesiano viene in genere indicata col simbolo B e prende il nome di fattore di Bayes dell'ipotesi al numeratore (in questo caso M, ovvero aver la malattia) contro l'ipotesi al denominatore. Il fattore di Bayes verrà formalmente introdotto e ampiamente discusso nella sezione 6.3 e, più ampiamente, nel Capitolo 8. Nel caso (B) occorre calcolare la probabilità a posteriori che Tizio sia affetto da M ovvero $P(M|T^+)$. Posti $\pi = P(M)$ e $\pi^* = P(M|T^+)$, si ha

$$\pi^* = \frac{P(M)P(T^+|M)}{P(T^+)} = \frac{P(M)P(T^+|M)}{P(M)P(T^+|M) + P(NM)P(T^+|NM)}$$

$$= \frac{0.95\,\pi}{0.95\,\pi + 0.02\,(1-\pi)} = \frac{95\,\pi}{93\,\pi + 2}.$$

Dunque la risposta al quesito B dipende dalla probabilità a priori π, che in questo caso può essere interpretata come la frequenza della malattia in questione in quel dato contesto geografico. La Figura 3.4 mostra il valore di π^* in funzione del valore iniziale π.

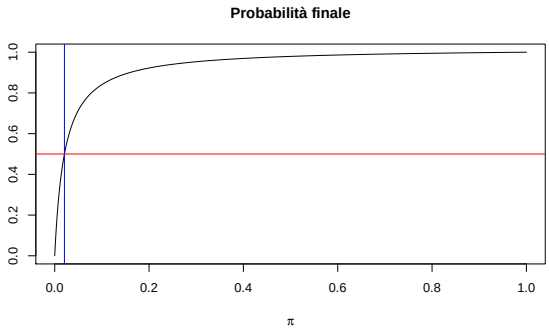

Figura 3.1. Probabilità a posteriori di aver la malattia M dopo un test positivo, al variare della probabilità a priori. Il segmento verticale individua il valore iniziale di π in corrispondenza del quale il valore di π^* è pari a 0.5

Può essere a volte più efficace ragionare in termini di disuguaglianze: infatti, molto spesso è importante conoscere per quali valori di π si ottiene, a posteriori un valore di π^* superiore ad una certa soglia, importante nel problema specifico. A titolo di esempio, affinch risulti $\pi^* > 0.5$, nel nostro caso dovrà risultare

$$\frac{95\,\pi}{93\,\pi + 2} > \frac{1}{2}$$

ovvero $\pi > 0.0206$. Questo significa che le conclusioni probabilistiche favoriscono l'ipotesi M soltanto se, a priori, siamo disposti a concedere, a tale ipotesi, una probabilità superiore al 2 per mille: e questo, ovviamente, può e deve essere stabilito caso per caso con la collaborazione fondamentale di un medico o di un epidemiologo!
Nel caso (C), per decidere se effettuare la cura o meno non basta nemmeno conoscere π. Occorre inserire, nello schema formale, anche una valutazione delle conseguenze che si possono avere a seguito delle decisioni prese. Concludendo, mentre A è supportato decisamente dai dati le affermazioni B e C possono essere compatibili con il dato sperimentale, ma occorre introdurre altre informazioni per stabilirlo. Riepilogando, i 3 obiettivi (a) valutare l'evidenza sperimentale, (b) effettuare un'inferenza, (c) prendere una decisione, sono distinti ma tra loro collegati. Qualunque impostazione che non introduca informazioni extra-sperimentali incontra serie difficoltà nel dover affrontare le questioni (b) e (c). ♢

3.4 Alcune note tecniche

3.4.1 La costante di marginalizzazione

La formula (1.8) presenta un denominatore che non dipende da θ. Per questo, in molti testi, è consuetudine presentarla nella forma seguente

$$\pi(\theta|\mathbf{y}) \propto \pi(\theta)L(\theta;\mathbf{y}) \tag{3.5}$$

In questo modo viene enfatizzato il fatto che la (1.8) dipende dal prodotto di due diverse fonti d'informazione. Il denominatore svolge il ruolo di costante di normalizzazione, per far sì che la

(1.8) sia effettivamente una distribuzione di probabilità.

Esempio 3.1 (continua). Riprendendo i dati usati in precedenza, l'esperimento forniva 7 successi su 10, ovvero la funzione di verosimiglianza risultava proporzionale a $\theta^7(1-\theta)^3$. Se la distribuzione a priori è, come prima, una Beta(5,5) essa risulta proporzionale a $\theta^4(1-\theta)^4$. Dalla (3.5) si deduce che

$$\pi(\theta|\mathbf{y}) \propto \theta^{12-1}(1-\theta)^{8-1},$$

facilmente riconoscibile come il nucleo di una distribuzione Beta(12,8). Nel seguito faremo quasi sempre riferimento alla (3.5) piuttosto che alla (1.8). L'effettivo calcolo del denominatore si presenterà invece come necessario soltanto in problemi dove il calcolo della legge marginale del dato osservato svolge un ruolo importante, come in alcuni problemi di verifica di ipotesi e di scelta del modello, che verranno discussi, rispettivamente, nella sezione 6.3 e nel Capitolo 8

3.4.2 Alcuni aspetti matematici

3.5 Esercizi

4

Analisi di semplici modelli statistici

In questo capitolo viene descritta l'analisi bayesiana dei più elementari modelli statistici: l'obiettivo è quello di fornire al lettore una sufficiente familiarit con il tipo di elaborazioni matematiche e computazionali che un'analisi bayesiana richiede. Le procedure inferenziali verranno introdotte in maniera informale, per poi essere riprese in modo più articolato nel Capitolo 6.

4.1 Dati dicotomici

In molte situazioni pratiche l'esperimento statistico consiste nell'osservare alcune replicazioni, diciamo n, di uno stesso fenomeno aleatorio che, in ogni singola prova o replicazione, può dar luogo a due possibili risultati, che codifichiamo in genere con 0 e 1; sono esempi ovvi di questa situazione i lanci ripetuti di una moneta, o la somministrazione di uno specifico trattamento ad un numero di pazienti, per poi verificare su ciascuno se è stato raggiunto un certo risultato (miglioramento, cioè 1) oppure no (nessun miglioramento, cioè 0). Assumiamo che in ogni singola prova, la probabilit di "successo" (cioè 1) sia costante e pari a θ. Volendo utilizzare una notazione più rigorosa, va ricordato che, in un'ottica bayesiana, θ è una variabile aleatoria che occorre indicare con simbolo maiuscolo Θ. Condizionatamente all'evento $\Theta = \theta$, possiamo poi considerare le prove ripetute come indipendenti tra loro. Marginalmente, a Θ verr assegnata una legge di probabilit che, nel caso di dati dicotomici, è spesso una legge di tipo Beta(α, β). In questi casi il modello statistico è così formulato. Condizionatamente all'evento $\Theta = \theta$

$$X_1, X_2, \cdots, X_n \overset{\texttt{iid}}{\sim} \text{Be}(\theta), \qquad \theta \sim Beta(\alpha, \beta), \tag{4.1}$$

dove l'espressione $X \sim \text{B}e(\theta)$ significa che X è una v.a. con distribuzione di Bernoulli con parametro θ, ovvero

$$\Pr\left(X = x \mid \Theta = \theta\right) = \theta^x (1-\theta)^{1-x}, \qquad x = 0, 1. \tag{4.2}$$

D'ora in poi, tuttavia, come è ormai consuetudine in letteratura, l'evento condizionante nel membro di sinistra della (4.2), verrà indicato semplicemente con θ piuttosto che con l'evento $\Theta = \theta$. Abbiamo già visto nel §2.1 che, nel caso di dati dicotomici, se $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ rappresenta il risultato osservato, la funzione di verosimiglianza associata è

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} \left[\theta^{x_i} (1-\theta)^{1-x_i}\right] = \theta^k (1-\theta)^{n-k},$$

dove k rappresenta il numero di successi osservati. Per quanto già visto in §2.1 la distribuzione a posteriori per θ è

$$\pi(\theta \mid \mathbf{x}) \propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1},$$

ovvero $\theta \mid \mathbf{x} \sim Beta(\alpha+k, \beta+n-k)$. Una stima puntuale del parametro θ può essere considerata la sua media a posteriori ovvero

$$\mathbb{E}(\theta \mid \mathbf{x}) = \frac{\alpha+k}{\alpha+\beta+n};$$

i dettagli sul calcolo dei momenti di una v.a. con distribuzione Beta sono riportati nell'Appendice E.2. Si noti che $\mathbb{E}(\theta \mid \mathbf{x})$ può essere scritta come

$$\mathbb{E}(\theta \mid \mathbf{x}) = \mathbb{E}(\theta)\,\frac{\alpha+\beta}{\alpha+\beta+n} + \hat{\theta}\frac{n}{\alpha+\beta+n},$$

mettendo in tal modo in luce come la media a posteriori possa essere interpretata come una media ponderata della media di θ a priori, $\alpha/(\alpha+\beta)$, e della stima di massima verosimiglianza ($\hat{\theta} = k/n$). La formula precedente chiarisce anche il ruolo della dimensione campionaria: al crescere di n il secondo termine della media ponderata prende il sopravvento e le valutazioni a priori perdono importanza; un indicatore di precisione di tale stima è la varianza a posteriori

$$\mathrm{Var}(\theta \mid \mathbf{x}) = \frac{(\alpha+k)(\beta+n-k)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}.$$

Esempio 3.1 (continua). Consideriamo di nuovo il risultato sperimentale precedente ($k = 7$ successi su $n = 10$ prove) e notiamo come esso modifiche in modo più o meno significativo diverse distribuzioni iniziali. La Figura 4.1 considera, per ogni riquadro, la legge a priori (tratteggiata) e quella finale (continua) nei casi in cui la legge iniziale è di tipo Beta con parametri (α, β) pari, rispettivamente, a $a)(1, 10)$, $b)(10, 1)$, $c)(0.5, 0.5)$, $d)(5, 5)$. Le quattro leggi a priori indicano posizioni iniziali molto diverse nei confronti della v.a. θ. L'esperimento, sia pure relativo a poche prove, modifica le opinioni iniziali e rende le leggi finali più simili tra loro. Questa idea, qui descritta in modo qualitativo può essere resa in modo più formale attraverso l'uso della divergenza di Kullback e Leibler, illustrata nella §2.6. In modo più elementare, possiamo notare, dalla tabella 4.1 come le medie a posteriori siano ben più simili che non le medie a priori.

valori di $(\alpha, beta)$ a priori	media iniziale	media finale
$(1, 10)$	0.09	0.381
$(10, 1)$	0.91	0.809
$(0.5, 0.5)$	0.5	0.682
$(5, 5)$	0.5	0.6

Tabella 4.1. Media iniziali e finali per per diverse distribuzioni iniziali.

4.2 Dati uniformi

Siano $X_1, X_2, \cdots, X_n$ n v.a. indipendenti e somiglianti (condizionatamente al valore di θ) con distribuzione uniforme nell'intervallo $(0, \theta)$; in simboli, per $j = 1, \cdots, n$ si ha $X_j \sim U(0, \theta)$. L'obiettivo

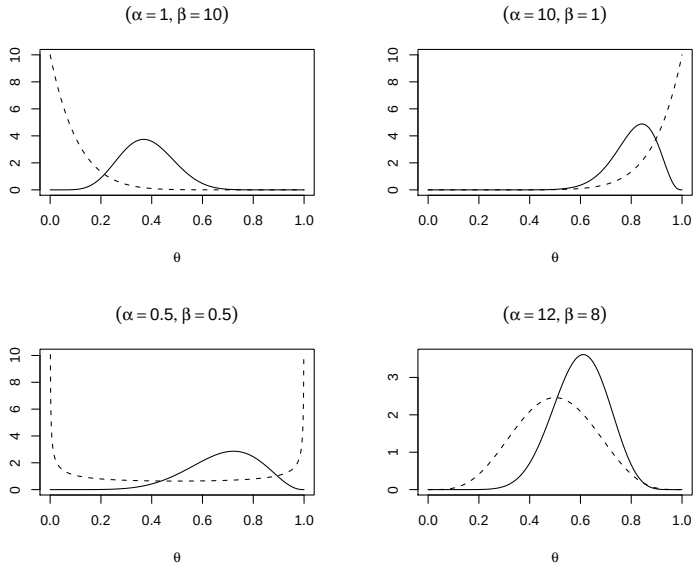

Figura 4.1. Distribuzioni iniziali (tratteggiate) e finali (continue) quando la legge iniziale è di tipo Beta con parametri (α, β) pari, rispettivamente, a $a)(1, 10)$, $b)(10, 1)$, $c)(0.5, 0.5)$, $d)(5, 5)$.

è fare inferenza sul parametro incognito θ, estremo superiore del supporto delle osservazioni. La funzione di verosimiglianza associata alle n osservazioni è dunque

$$L(\theta; \mathbf{x}) \propto \frac{1}{\theta^n} I_{(x_{(n)}, +\infty)}(\theta),$$

dove $x_{(n)} = \max\{x_1, x_2, \cdots, x_n\}$. Si noti che $L(\theta)$ ha un andamento monotono decrescente ed è maggiore di zero solo per valori del parametro superiori al massimo valore osservato nel campione, in quanto θ rappresenta l'estremo superiore del supporto delle variabili osservabili. Un'analisi inferenziale basata su $L(\theta)$ conduce allora alla stima di massima verosimiglianza $\hat{\theta} = x_{(n)}$ la cui distribuzione, in virtù di risultati generali relativi alle statistiche d'ordine (vedere Appendice C.1), è data da

$$f_{\hat{\theta}}(t\ \theta) = \frac{n t^{(n-1)}}{\theta^n} I_{(0,\theta)}(t),$$

ovvero $Y = \hat{\theta}/\theta \sim Beta(n, 1)$. Dalle proprietà della distribuzione Beta si ricava allora che

$$\mathbb{E}\left(\hat{\theta}; \theta\right) = \frac{n\,\theta}{n+1},$$

ovvero $\hat{\theta}$ è uno stimatore asintoticamente corretto. Un'ovvia correzione porta alla costruzione dello stimatore non distorto $T_{ND}(\mathbf{x}) = \hat{\theta}(n+1)/n$ Volendo poi determinare lo stimatore migliore secondo l'errore quadratico medio, tra quelli di tipo lineare, $T_c(\mathbf{x}) = c\,\hat{\theta}$, con $c \in \mathbb{R}$, occorre minimizzare rispetto a c la quantità

$$\mathbb{E}\left((cX_{(n)} - \theta)^2; \theta\right)$$

Semplici calcoli, lasciati al lettore, conducono a

$$c^* = \frac{(n+2)}{(n+1)}.$$

Un'analisi bayesiana convenzionale del problema parte dalla scelta di una famiglia di distribuzioni a priori. Per semplicità computazionale scegliamo la famiglia delle distribuzioni del tipo

$$\pi(\theta) \propto \theta^{-\alpha}, \qquad \alpha > 0. \tag{4.3}$$

Per ogni $\alpha > 0$, la distribuzione iniziale (4.3) non è una vera e propria distribuzione di probabilità, in quanto il suo integrale, sul supporto $\boldsymbol{\Omega} = (0, +\infty)$, non è finito. Si parla in questo caso di distribuzione iniziale impropria: il suo uso è tuttavia consentito, fintanto che la distribuzione finale che si ottiene a partire da questa, risulti propria, qualunque sia la dimensione campionaria e qualunque sia il risultato osservato. Riprenderemo questi aspetti nel Capitolo 5. In questo caso particolare la distribuzione finale è allora

$$\pi(\theta \mid \mathbf{x}) \propto \theta^{-(n+\alpha)} I_{(x_n, +\infty)}(\theta),$$

che risulta integrabile, e quindi ben definita, quando $\alpha + n > 1$, e questo giustifica l'aver posto il vincolo di positività su α nella scelta della famiglia di distribuzioni a priori. La costante di normalizzazione vale

$$\int_{x_{(n)}}^{\infty} \theta^{-(\alpha+n)} d\theta = \frac{1}{x_{(n)}^{\alpha+n-1}(n+\alpha-1)}.$$

Come stima puntuale bayesiana per θ,possiamo ade esempio considerare la media a posteriori per un valore di α fissato, che vale

$$\mathbf{E}(\theta \mid \mathbf{x}) = x_{(n)}^{n+\alpha-1}(n+\alpha-1) \int_{x_{(n)}}^{\infty} \theta^{-(n+\alpha-1)} d\theta = \frac{n+\alpha-1}{n+\alpha-2} x_{(n)}.$$

È facile verificare che lo stimatore di massima verosimiglianza si ottiene per $\alpha \to \infty$ mentre i casi $\alpha = 2$ e $\alpha = 3$ corrispondono, rispettivamente, allo stimatore non distorto T_{ND} e allo stimatore che minimizza l'errore quadratico medio T_{c^*}.

Da notare come, nella pratica statistica si preferisca in genere utilizzare la distribuzione a priori con $\alpha = 1$, che equivale all'uso dello stimatore T_c con $c = n/(n-1)$. La scelta di $\alpha = 1$ trova giustificazione nel tentativo di ottenere, per via bayesiana, intervalli di credibilità con buone proprietà in senso frequentista. Il lettore interessato a questi approfondimenti può consultare, ad esempio, [73]. Un'analisi bayesiana "propria"' con dati di tipo uniforme può essere condotta adottando, come legge iniziale, una distribuzione di Pareto $\mathrm{Pa}(\gamma, \beta)$, con $\beta > 0$ e $\gamma > 2$. In questo caso la legge finale risulterebbe

$$\pi(\theta \mid \mathbf{x}) \propto \frac{1}{\theta^{n+\gamma+1}} I_{(\max(x_n, \beta), +\infty)}(\theta),$$

che è ancora di tipo Pareto con parametri aggiornati

$$\beta^* = \max(x_n, \beta), \qquad \gamma^* = \gamma + n.$$

Di conseguenza,

$$\mathbf{E}(\theta \mid \mathbf{x}) = \frac{\gamma^* \beta^*}{\gamma^* + 1}, \text{ e } \mathrm{Var}(\theta \mid \mathbf{x}) = \beta^* \left(\frac{\gamma^*}{\gamma^* - 2} - \left(\frac{\gamma^*}{\gamma^* - 1} \right)^2 \right).$$

Da notare come la legge iniziale di Pareto sia in pratica un "troncamento" a β della legge iniziale utilizzata in precedenza non appena si ponga $\alpha = \gamma + 1$, e come sia proprio il troncamento a renderla una legge di probabilità propria.

4.3 La distribuzione gaussiana

In questo paragrafo assumiamo che le osservazioni campionarie possano essere considerate come replicazioni indipendenti e somiglianti (condizionatamente al valore dei parametri) di una variabile aleatoria $X \sim N(\mu, \sigma^2)$. Questa situazione, pur rappresentando un caso particolare di un contesto più generale, riveste notevole importanza, per il ruolo che la distribuzione normale ricopre nella teoria della probabilità e dell'inferenza statistica; l'assunzione di normalità delle osservazioni è infatti spesso giustificata per valori sufficientemente grandi della dimensione campionaria n, come approssimazione della vera distribuzione delle X_i. Nei paragrafi successivi considereremo il problema del calcolo della distribuzione a posteriori dei parametri d'interesse nei due casi seguenti

- μ incognita e σ^2 nota
- μ e σ^2 entrambi incogniti

4.3.1 Varianza nota

Siano $X_1, X_2, \cdots, X_n$ n variabili aleatorie che, condizionatamente al valore del parametro μ, sono indipendenti e con la stessa distribuzione $N(\theta, \sigma^2)$, con σ^2 costante nota.
La funzione di verosimiglianza è allora

$$L(\mu) = \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \right) \propto \exp\left\{ -\frac{n(\bar{x} - \mu)^2}{2\sigma^2} \right\}, \tag{4.4}$$

dove l'ultima relazione si ottiene eliminando tutti i fattori che non dipendono dal parametro di interesse μ (vedi formula (2.5)).

Occorre ora scegliere una distribuzione a priori per μ: tale scelta, nel rispetto del paradigma bayesiano, dovrebbe essere strettamente soggettiva e suggerita del tipo di informazioni che il ricercatore ha a disposizione. La difficoltà pratica di elicitare, ogni volta, la propria distribuzione a priori conduce spesso lo statistico ad operare una scelta di comodo, prediligendo alcune distribuzioni a priori per la relativa facilità dei calcoli che esse inducono. In questa trattazione elementare considereremo soltanto due casi:

- distribuzione a priori normale;
- distribuzione a priori non informativa;

la scelta di queste distribuzioni verrà illustrata più in dettaglio nelle §5.1 e §5.2. Nel primo caso assumiamo che la distribuzione a priori $\pi(\mu)$ sia di tipo $N(\alpha, \tau^2)$, con α e τ^2 costanti note, determinate in base alle conoscenze specifiche. Un modo semplice per calibrare i valori di α e τ^2 è il seguente, fondamentalmente identico a quanto proposto in [25]: si pone α pari al valore di μ ritenuto a priori più ragionevole; in pratica, una sorta di stima iniziale basata su conoscenze presperimentali; si stabilisce poi qual è, a priori, l'intervallo di semi-ampiezza q, centrato in α in cui, con pratica certezza, cade il valore di μ. Poich le realizzazioni di una distribuzione $N(\alpha, \tau^2)$ sono contenute, con una probabilità superiore a 0.995, nell'intervallo $(\alpha - 3\tau; \alpha + 3\tau)$, basta dunque porre $\tau = q/3$. Avremo così che

$$\pi(\mu) \propto \exp\left\{ -\frac{(\mu - \alpha)^2}{2\tau^2} \right\},$$

e, per il teorema di Bayes, la distribuzione a posteriori è

$$\begin{aligned}
&\pi(\mu|x_1,\cdots,x_n) \propto \pi(\mu)L(\mu)\\
&= \exp\left\{-\frac{1}{2}\left(\frac{(\mu-\alpha)^2}{\tau^2}+\frac{n(\bar{x}-\mu)^2}{\sigma^2}\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{(\mu-\alpha^*)^2}{\tau^{*2}}\right)\right\},
\end{aligned} \tag{4.5}$$

dove

$$\begin{aligned}
\alpha^* &= \frac{\alpha/\tau^2 + n\bar{x}/\sigma^2}{1/\tau^2 + n/\sigma^2} = \frac{\alpha\sigma^2 + \bar{x}n\tau^2}{\sigma^2 + n\tau^2},\\
\tau^{*2} &= \frac{1}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}} = \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}.
\end{aligned} \tag{4.6}$$

Il calcolo esplicito della (4.6) è basato sulla seguente formula, facilmente verificabile, valida per (a, b, z) reali e (A, B) positivi:

$$A(z-a)^2 + B(z-b)^2 = (A+B)\left(z - \frac{aA+bB}{A+B}\right)^2 + \frac{AB}{A+B}(a-b)^2; \tag{4.7}$$

si veda l'Appendice C.4.1 per una generalizzazione al caso multivariato della (4.7). La formula (4.5) è riconoscibile come il nucleo di una distribuzione normale di media α^* e deviazione standard τ^*, per cui

$$\mu \mid \mathbf{x} \sim N\left(\alpha^*, \tau^{2*}\right).$$

Le formule (4.6) meritano un'analisi accurata; la media a posteriori, α^*, si ottiene come media ponderata della media a priori, α, e della stima di massima verosimiglianza, $\bar{x}$, e i pesi delle due componenti sono le rispettive precisioni definite come il reciproco delle varianze. Se l'informazioni a priori è sostanziale, la distribuzione a priori risulta molto concentrata intorno a α, e il valore elicitato di τ^2 è relativamente piccolo: di conseguenza la componente a priori risulterà importante nel computo di α^*. Conclusioni opposte valgono naturalmente nel caso di τ^2 grande. Il ruolo della dimensione campionaria n è simile; per n grande la componente $\bar{x}$ prende il sopravvento e la media a posteriori si avvicina al valore campionario $\bar{x}$. Per quanto riguarda τ^* va notato come anch'essa risulti direttamente proporzionale a τ^2 e inversamente proporzionale alla dimensione campionaria. È possibile rendere più precise queste affermazioni, osservando che se la varianza di $\pi(\mu)$ tende a $+\infty$, la distribuzione finale di μ converge[1] ad una distribuzione normale con media $\bar{x}$ e varianza σ^2/n. Si noti come sia stato riottenuto, in un contesto diverso il classico risultato sulla distribuzione campionaria della media di un campione estratto da una popolazione normale. Questa analogia ha suggerito che, più in generale, l'uso di distribuzioni a priori uniformi, potesse in qualche modo rappresentare una lettura bayesiana delle procedure frequentiste, ma in realtà le cose non sono sempre così semplici. Ritorneremo su questi aspetti nella §5.2. Da un punto di vista matematico, una varianza a priori infinita corrisponde ad utilizzare una distribuzione a priori uniforme sulla retta reale

$$\pi(\mu) \propto 1. \tag{4.8}$$

Come già accadeva nel caso di campionamento di dati uniformi, la (4.8) non è una legge di probabilità, nel senso che il suo integrale su $\mathbb{R}$ non è 1. Anche in questo caso, tuttavia la distribuzione finale risultante è propria per qualunque dimensione campionaria e per qualunque valore osservato di $\bar{x}$.

[1] la nozione di convergenza qui considerata è quella debole, corrispondente alla convergenza in distribuzione delle v.a.: si veda, ad esempio [30].

4.3.2 Media e varianza incognite

Consideriamo ora un caso in cui il parametro da stimare risulti bidimensionale. Siano allora $X_1, X_2, \cdots, X_n$ n osservazioni indipendenti con distribuzione $N(\mu, \sigma^2)$ ed entrambi i parametri risultano non noti. La funzione di verosimiglianza associata all'effettivo campione osservato $\mathbf{x} = (x_1, \cdots, x_n)$ è così calcolabile:

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{j=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x_j - \mu)^2 \right\} \right] \\ &\propto \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \mu)^2 \right\} \\ &\propto \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \bar{x} + \bar{x} - \mu)^2 \right\} \\ &\propto \frac{1}{\sigma^{2n/2}} \exp\left\{ -\frac{n}{2\sigma^2} \left[s^2 + (\bar{x} - \mu)^2 \right] \right\}, \end{aligned} \tag{4.9}$$

dove $s^2 = n^{-1} \sum_{j=1}^{n} (x_i - \bar{x})^2$ rappresenta la varianza campionaria (non corretta). Per motivi meramente computazionali è conveniente modificare la parametrizzazione consueta e porre $\sigma^2 = 1/\psi$. Il parametro ψ viene di solito chiamato *precisione*. In tal modo la funzione di verosimiglianza diventa

$$L(\mu, \psi) \propto \psi^{n/2} \exp\left\{ -\frac{n}{2}\psi \left[s^2 + (\bar{x} - \mu)^2 \right] \right\}. \tag{4.10}$$

Una distribuzione bivariata che viene spesso utilizzata in questo contesto è la cosiddetta *Normale-Gamma* che denoteremo con il simbolo NoGa$(\alpha, g, \delta, \lambda)$ definita nell'Appendice E. L'uso di tale distribuzione implica che, a priori,

- per ψ fissato, $\mu \mid \psi \sim N(\alpha, (g\psi)^{-1})$
- $\psi \sim \text{Gamma}(\delta, \lambda)$;

L'uso di una distribuzione iniziale Gamma Inversa per ψ equivale ad utilizzare una legge Gamma su σ^2 (vedi Appendice E.2). Moltiplicando la funzione di verosimiglianza per la distribuzione a priori si ottiene facilmente che

$$\begin{aligned} \pi(\mu, \psi \mid \mathbf{x}) &\propto L(\mu, \psi)\pi(\mu \mid \psi)\pi(\psi) \\ &= \psi^{\frac{n}{2}+\delta+\frac{1}{2}-1} \exp\left\{ -\psi \left(\frac{ns^2}{2} + \lambda \right) \right\} \exp\left\{ -\frac{1}{2}\psi \left[n(\bar{x} - \mu)^2 + g(\mu - \alpha)^2 \right] \right\} \\ &= \psi^{\frac{n}{2}+\delta-1} \exp\left\{ -\psi \left(\frac{ns^2}{2} + \lambda + \frac{ng}{2(n+g)}(\bar{x} - \alpha)^2 \right) \right\} \times \\ &\quad \sqrt{\psi} \exp\left\{ -\frac{1}{2}\psi(n+g) \left(\mu - \frac{n\bar{x} + g\alpha}{n+g} \right)^2 \right\}, \end{aligned}$$

dove l'ultimo passaggio è stato effettuato utilizzando ancora la (4.7); si riconosce che il nucleo della distribuzione finale è ancora quello di una Normale-Gamma; più precisamente, la distribuzione finale di (μ, ψ) è $NG(\alpha^*, g^*, \lambda^*, \delta^*)$ dove

$$\alpha^* = \frac{n\bar{x} + g\alpha}{n+g}, \quad g^* = g + n,$$

$$\lambda^* = \lambda + \frac{ns^2}{2} + \frac{ng}{2(n+g)}(\bar{x}-\alpha)^2, \quad \delta^* = \delta + \frac{n}{2}.$$

La distribuzione finale marginale di $\sqrt{\frac{\delta^* g^*}{\lambda^*}}(\mu - \alpha^*)$ è del tipo t di Student con $2\delta^*$ gradi di libertà. Questo risultato è ottenibile mediante la semplice marginalizzazione rispetto a ψ del nucleo della distribuzione a posteriori congiunta e viene lasciata per esercizio, ma può essere altresì ottenuta attraverso un semplice adattamento del Teorema E.1.

L'analisi appena descritta può essere resa "non informativa" ponendo gli iperparametri della legge Normale-Gamma, α, g, λ e δ, tutti uguali a zero. Si può facilmente dimostrare che questo equivale ad utilizzare, come legge iniziale, la distribuzione impropria

$$\pi(\mu, \psi) \propto \frac{1}{\sqrt{\psi}}, \tag{4.11}$$

che corrisponde, nella parametrizzazione in termini di σ^2, all'utilizzo della legge impropria $\pi(\mu, \sigma^2) \propto 1/\sigma^3$ In questo caso, la distribuzione finale sarà allora tale che

- per ψ fissato $\mu \mid \psi, \mathbf{x} \sim N(\bar{x}, (n\psi)^{-1})$
- $\psi \mid \mathbf{x} \sim \text{Gamma}(n/2, ns^2/2)$

e la legge marginale a posteriori di μ è, sempre per il teorema E.1, una t di Student con n gradi di libertà e parametri di posizione e scala pari, rispettivamente, a $\bar{x}$ e s^2/n. Nel capitolo 5, tuttavia, vedremo che la distribuzione iniziale che fornisce soluzioni numericamente coincidenti con un'analisi classica, ed anche per questo detta *non informativa*, sarà

$$\pi(\mu, \psi) \propto \frac{1}{\psi}, \tag{4.12}$$

che corrisponde, nella parametrizzazione in termini di σ^2, all'utilizzo della legge impropria $\pi(\mu, \sigma^2) \propto 1/\sigma^2$. In questo caso, con calcoli del tutto simili si otterrà, a posteriori, che

- per ψ fissato, $\mu \mid (\psi, \mathbf{x}) \sim N(\bar{x}, (n\psi)^{-1})$
- $\psi \mid \mathbf{x} \sim \text{Gamma}\big((n-1)/2, ns^2/2\big)$.

Avremo allora che

$$\frac{\sqrt{n-1}(\mu - \bar{x})}{s} \sim St_1(n-1, 0, 1); \tag{4.13}$$

il precedente risultato, espresso in termini della varianza campionaria corretta, $\tilde{s}^2$ equivale a

$$\frac{\sqrt{n}(\mu - \bar{x})}{\tilde{s}} \sim St_1(n-1, 0, 1), \tag{4.14}$$

che rappresenta la versione bayesiana del ben noto risultato classico[2].

Esempio 4.1 [*Dimensione del cranio dell'"Homo erectus"*]. [17].

Un antropologo, specializzato nell'evoluzione umana, scopre sette scheletri di Homo erectus in un'area africana dove non erano mai stati trovati, in precedenza, altri scheletri. Egli effettua una misurazione della capacità cranica, che in centimetri cubici, fornisce i seguenti risultati

$$925, 892, 900, 875, 910, 906, 899; \quad .$$

Le statistiche campionarie valgono allora

$$\bar{x} = 901, \qquad s^2 = 206.28.$$

[2] Attenzione, però: qui la variabile aleatoria è μ e non la coppia di statistiche campionarie $(\bar{x}, \tilde{s})$!

Consideriamo il caso in cui l'antropologo, proprio perch non esistono reperti simili trovati nella zona non abbia informazioni a priori molto precise e decida dunque di utilizzare la distribuzione non informativa (4.11). In questo caso la legge finale sarà di tipo Normale-Gamma con parametri

$$\alpha^* = \bar{x} = 901, g^* = n = 7, \delta^* = \frac{n}{2} = 3.5, \lambda^* = \frac{ns^2}{2} = 722.$$

Marginalmente allora il parametro μ, capacità media del cranio degli esemplari di Homo erectus, ha distribuzione di tipo $St(7, \bar{x}, s^2/(n-1))$. ♢

4.4 Modello di Poisson

Siano $X_1, X_2, \cdots, X_n$ n v.a. indipendenti (condizionatamente al valore di θ) e somiglianti, con distribuzione di Poisson di parametro θ; in simboli,

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Po}(\theta).$$

La funzione di verosimiglianza associata all'esperimento è

$$L(\theta; \mathbf{x}) \propto e^{-n\theta}\theta^{n\bar{x}},$$

dove $\bar{x}$ rappresenta, al solito, la media campionaria. Una distribuzione iniziale computazionalmente conveniente per questo modello è fornita dala legge Gamma(δ, γ); in tal caso, la distribuzione finale risulta pari a

$$\pi(\theta; \mathbf{x}) \propto \exp\{-(n+\gamma)\theta\}\, \theta^{n\bar{x}+\delta-1},$$

riconoscibile ancora come il nucleo di una distribuzione Gamma$(n\bar{x}+\delta, n+\gamma)$. Utilizzando le formule in appendice per il calcolo dei momenti di una distribuzione Gamma, si avrà che la media e la varianza a posteriori per θ valgono

$$\mathbb{E}(\theta \mid \mathbf{x}) = \frac{n\,\bar{x}+\delta}{n+\gamma}; \qquad \text{Var}(\theta|\mathbf{x}) \propto \frac{n\,\bar{x}+\delta}{(n+\gamma)^2}.$$

Anche in questo caso la media finale risulta essere una combinazione lineare della stima di massima verosimiglianza e della media iniziale.
La scelta non informativa, in questo contesto, equivale a porre i parametri della legge iniziale, γ e δ, entrambi pari a zero. Questo equivale ad utilizzare la legge impropria

$$\pi(\theta) \propto \frac{1}{\theta}. \tag{4.15}$$

In questo caso la media finale coincide con la media campionaria $\bar{x}$ e la varianza finale è pari a $\bar{x}/n$.
Esempio 4.2 [*Visitatori della Galleria Borghese*]
L'intendenza alle Belle Arti vuole stabilire il tasso medio di visitatori mattutini della Galleria Borghese di Roma con l'obiettivo di adottare tariffe particolari. Per 10 giorni feriali consecutivi si contano quanti visitatori acquistano il biglietto durante la prima ora di apertura, e si osservano i seguenti risultati:

```
61 46 62 52 51 46 36 58 44 61
```

che forniscono una media pari a $\bar{x} = 51.7$ Si vuole utilizzare a priori una legge di tipo Gamma e, da informazioni ricavate da un'indagine simile effettuata l'anno precedente si può elicitare una media a priori pari a 48: per non dar troppo peso a questa informazione si adotta una varianza a priori elevata, diciamo 200. Dalle formule sui momenti di una distribuzione gamma, possiamo allora ricavare i parametri della legge iniziale in funzione della media e della varianza elicitate: sappiamo allora che la nostra legge iniziale può essere rappresentata da una distribuzione Gamma$(11.52, 0.24)$. Ne segue che, a posteriori,

$$\theta \,|\, \mathbf{x} \sim \text{ Gamma } (528.52, 10.24).$$

La media a posteriori vale allora 51.64, a conferma della "debolezza" delle informazioni a priori. Possiamo inoltre calcolare, ad esempio, la probabilità che il tasso di visite medio risulti maggiore o uguale a 55, come

$$\Pr\left(\theta \geq 55|\mathbf{x}\right) = \int_{55}^{\infty} \frac{(10.24)^{528.52}}{\Gamma(528.52)} e^{-10.24\,\theta} \theta^{527.52} d\theta.$$

Una semplice integrazione numerica o il ricorso a **R** (funzione `pgamma`) fornisce la risposta, pari a 0.0679.

♢

4.5 Altri esempi notevoli

4.5.1 Confronto fra due proporzioni

Sia X_1 il numero di successi su n_1 prove bernoulliane indipendenti con probabilità di successo pari a θ_1 e X_2 il numero di successi su n_2 prove bernoulliane indipendenti (e indipendenti dalle precedenti n_1) con probabilità di successo pari a θ_2. Si vuole fare inferenza su una misura di distanza tra θ_1 e θ_2. In termini concreti possiamo pensare ad X_1 e X_2 come al numero aleatorio di successi osservati rispettivamente su n_1 ed n_2 pazienti a cui sono stati somministrati due diversi farmaci. In epidemiologia è invalsa la tradizione di considerare, quale parametro di interesse, il logaritmo del rapporto delle odds dei due eventi "guarigione con l'i-esimo farmaco", $(i = 1, 2)$, ovvero

$$\psi = \log\left(\frac{\theta_1}{1-\theta_1}\frac{1-\theta_2}{\theta_2}\right)$$

Il parametro ψ è facilmente interpretabile in quanto $\psi < 0$ (> 0) corrisponde al caso $\theta_1 < \theta_2$ $(\theta_1 > \theta_2)$. Poich $X_1 \sim \text{ Bin}(n_1, \theta_1)$ e $X_2 \sim \text{ Bin}(n_2, \theta_2)$, una semplice analisi bayesiana coniugata conduce ad assumere θ_1 e θ_2 a priori indipendenti ed entrambi con legge di probabilità di tipo Beta, rispettivamente con parametri (α_1, β_1) e (α_2, β_2) ovvero

$$\pi(\theta_1, \theta_2) \propto \theta_1^{\alpha_1 - 1}(1-\theta_1)^{\beta_1 - 1}\ \theta_2^{\alpha_2 - 1}(1-\theta_2)^{\beta_2 - 1};$$

Supponiamo che il numero dei successi nei due campioni sia pari, rispettivamente, a k_1 e k_2. La funzione di verosimiglianza associata all'esperimento è allora

$$L(\theta_1, \theta_2) \propto \theta_1^{k_1 - 1}(1-\theta_1)^{n_1 - k_1 - 1}\ \theta_2^{k_2 - 1}(1-\theta_2)^{n_2 - k_2 - 1};$$

Di conseguenza , ricordando che la funzione di verosimiglianza binomiale è coniugata con la legge Beta, la distribuzione a posteriori congiunta per θ_1 e θ_2 è il prodotto di due Beta, ovvero, con una licenza di notazione,

$$\theta_1, \theta_2, \mid k_1, k_2 \sim \text{ Beta}(k_1 + \alpha_1, n_1 - k_1 + \beta_1) \times \text{ Beta}(k_2 + \alpha_2, n_2 - k_1 + \beta_2)$$

Il calcolo della distribuzione finale di ψ si configura ora come un problema di calcolo di probabilità che non ha però una soluzione in forma esplicita. È possibile tuttavia ottenere una soluzione approssimata attraverso un approccio basato sulla simulazione. Tratteremo con maggior dettaglio di questi argomenti nel Capitolo 7, ma è opportuno qui anticiparne le potenzialità. Poich la distribuzione finale di (θ_1, θ_2) è disponibile in forma esplicita ed è semplice generare, ad esempio con **R** , valori pseudo-casuali con legge Beta, sarà sufficiente allora procedere secondo il seguente algoritmo

Simulazione a posteriori per la stima del log-odds ratio.

1. genera M valori $\pi_1^{(1)}, \pi_1^{(2)}, \cdots \pi_1^{(i)}, \cdots, \pi_1^{(M)}$
   dalla distribuzione finale di θ_1

2. genera M valori $\pi_2^{(1)}, \pi_2^{(2)}, \cdots \pi_2^{(i)}, \cdots, \pi_2^{(M)}$
   dalla distribuzione finale di θ_2

3. Per $i = 1, \cdots M$, poni
$$\psi^{(i)} = \log\left(\pi_1^{(i)}(1 - \pi_2^{(i)})\right) - \log\left(\pi_2^{(i)}(1 - \pi_1^{(i)})\right),$$

Ne consegue che i valori

$$\psi^{(1)}, \psi^{(2)}, \cdots \psi^{(i)}, \cdots, \psi^{(M)}$$

rappresentano un campione, di dimensione M grande a piacere dalla distribuzione finale di ψ ! Tale campione può essere utilizzato per stimare una qualunque caratteristica della legge a posteriori di ψ. Ad esempio, una stima puntuale per ψ può essere prodotta calcolando la media degli M valori $\psi^{(i)}$. Qui di seguito riportiamo il codice in **R** per il calcolo della distribuzione a posteriori di ψ ed alcune sintesi. Riprenderemo in esame questo esempio nella § 7.3.

Codice **R** per la distribuzione finale del log odds ratio

```
odd.ratio<-function(n1,k1,n2,k2,a1=.5,b1=.5,a2=.5,b2=.5, n=10000) {
alp1<-a1+k1;  alp2<-a2+k2;  bet1<-n1-k1+b1;  bet2<-n2-k2+b2
x1<-rbeta(n,alp1,bet1)
x2<-rbeta(n,alp2,bet2)
psi<-sort(log(x1/(1-x1))-log(x2/(1-x2)))
inf<-psi[0.025*n];  sup<-psi[0.975*n]
media<-mean(psi); varianza<-var(psi)
aa<-list(''la media finale \'e'', media, ''la varianza e''', varianza,
''l'intervallo al 95 e''', c(inf,sup))
den.psi <- density(psi, width=2)
hist(psi,prob=T,xlab=expression(psi),nclass=100,
main=''Distribuzione finale del parametro di odds ratio'')
lines(den.psi)
aa}
```

Come esemplificazione numerica consideriamo il caso in cui si osservino due campioni di dimensione $n_1 = n_2 = 20$ con numero di successi pari $k_1 = 15$ e $k_2 = 10$, e che le leggi iniziali su θ_1 e θ_2 siano uniformi. Allora il comando

```
odd.ratio(n1=20,k1=15,n2=20,k2=10,a1=1,b1=1,a2=1,b2=1, n=10000)
```

fornisce una media a posteriori finale per ψ pari a 1.033 con intervallo di credibilità[3] stimato, a livello del 95%, pari a

$$(-0.2407;\ 2.357)\,.$$

4.5.2 Confronto fra due medie

Uno dei problemi standard in un corso istituzionale di inferenza è quello del confronto fra le medie di due popolazioni gaussiane. La trattazione classica di questo problema presenta diversi gradi di difficoltà a seconda delle assunzioni che si fanno sulle due varianze. Consideriamo qui il caso intermedio in cui le due varianze siano considerate uguali ma aleatorie. Il caso più generale, con varianze entrambe aleatorie e non necessariamente uguali, ovvero il famoso problema di Behrens e Fisher, verrà affrontato nel capitolo 7, attraverso metodi di simulazione. Siano allora

$$X_{11}, X_{12}, \cdots, X_{1n_1} \overset{\text{iid}}{\sim} N(\theta_1, \sigma^2)$$

e

$$X_{21}, X_{22}, \cdots, X_{2n_2} \overset{\text{iid}}{\sim} N(\theta_2, \sigma^2)$$

due campioni di numerosità rispettive n_1 ed n_2 estratti da due popolazioni normali con parametri indicati sopra. Condizionatamente ai valori di $(\theta_1, \theta_2, \sigma^2)$, le osservazioni sono tutte indipendenti tra loro. Vogliamo qui determinare la distribuzione a posteriori della funzione parametrica

$$\xi = \theta_1 - \theta_2,$$

espressione della differenza tra le due medie. Questa scelta non è l'unica ragionevole. Altri autori preferiscono ad esempio considerare la differenza standardizzata ξ/σ. Qui si è scelto di utilizzare ξ per rendere più evidenti le differenze e le analogie con l'analisi frequentista. Affronteremo il problema come uno di stima, rimandando alla §8.3 la trattazione dello stesso come problema di confronto tra ipotesi alternative. La funzione di verosimiglianza associata all'esperimento sopra descritto è allora, per una banale generalizzazione della 4.9, pari a

$$L(\theta_1, \theta_2, \sigma^2) \propto \frac{1}{\sigma^{n_1+n_2}} \exp\left(-\frac{1}{2\sigma^2}\left[n_1 s_1^2 + n_2 s_2^2 + n_1(\bar{x}_1 - \theta_1)^2 + n_2(\bar{x}_2 - \theta_2)^2\right]\right), \tag{4.16}$$

dove $\bar{x}_i, s_i^2$, $i = 1, 2$, rappresentano media e varianza campionaria relative alla popolazione i-esima. Come distribuzione iniziale adotteremo ancora una forma coniugata considerando le due medie θ_1 e θ_2 a priori indipendenti e somiglianti condizionatamente al valore di σ^2, ovvero

$$\theta_1, \theta_2 \mid \sigma^2 \overset{\text{iid}}{\sim} N(\alpha, \frac{\sigma^2}{g}). \tag{4.17}$$

[3] Gli intervalli di credibilità possono essere considerati la versione bayesiana dei classici intervalli di confidenza e verranno discussi nella § 6.2.

Al parametro σ^2 viene invece assegnata una legge di tipo Gamma Inversa,

$$\sigma^2 \sim \text{GI}(\delta, \lambda). \tag{4.18}$$

Anche qui, come nella §4.3.2, si potrebbe riparametrizzare la varianza σ^2 in termini della precisione $\psi^{-1} = \sigma^2$. I calcoli sarebbero di natura e difficoltà analoghe. Moltiplicando allora la densità dei parametri come specificate nelle (4.17) e (4.18) e la funzione di verosimiglianza (4.16), e combinando tra loro le due forme quadratiche relative a θ_1 e quelle relative a θ_2, si ottiene che la distribuzione congiunta a posteriori è proporzionale a

$$\begin{aligned}
&\pi(\theta_1, \theta_2, \sigma^2 \mid \mathbf{x}_1, \mathbf{x}_2) \propto \frac{1}{\sigma^{n_1+n_2+2\delta+4}} \\
&\exp\left(-\frac{1}{2\sigma^2}\left[2\lambda + n_1 s_1^2 + n_2 s_2^2 + (n_1+g)(\theta_1 - \alpha_1^*)^2 + \frac{n_1 g}{n_1+g}(\bar{x}_1 - \alpha)^2\right.\right. \\
&\left.\left. +(n_2+g)(\theta_2 - \alpha_2^*)^2 + \frac{n_2 g}{n_2+g}(\bar{x}_2 - \alpha)^2\right]\right),
\end{aligned} \tag{4.19}$$

dove, per $i = 1, 2$

$$\alpha_i^* = \frac{n\bar{x}_i + g\alpha}{n+g}. \tag{4.20}$$

La distribuzione finale (4.19) è dunque della stessa forma funzionale della legge a priori ed avremo

$$\theta_1 \mid \sigma^2, \mathbf{x}_1, \mathbf{x}_2 \sim \text{N}(\alpha_1^*, \frac{\sigma^2}{n_1+c}) \tag{4.21}$$

$$\theta_2 \mid \sigma^2, \mathbf{x}_1, \mathbf{x}_2 \sim \text{N}(\alpha_2^*, \frac{\sigma^2}{n_2+c}) \tag{4.22}$$

$$\sigma^2 \mid \mathbf{x}_1, \mathbf{x}_2 \sim \text{GI}(\delta^*, \lambda^*), \tag{4.23}$$

con $\delta^* = \delta + (n_1+n_2)/2$ e

$$\begin{aligned}
\lambda^* &= \lambda + \frac{n_1 s_1^2 + n_2 s_2^2}{2} + \frac{n_1 g}{2(n_1+g)}(\bar{x}_1 - \alpha)^2 + \frac{n_2 g}{2(n_2+g)}(\bar{x}_2 - \alpha)^2 \\
&= \lambda + \frac{n_1 s_1^2 + n_2 s_2^2}{2} + g\frac{n_1(n_2+g) + n_2(n_1+g)}{(n_1+g)(n_2+g)}(\alpha - \bar{x}^*)^2 \\
&+ \frac{g n_1 n_2}{n_1(n_2+g) + n_2(n_1+g)}(\bar{x}_1 - \bar{x}_2)^2,
\end{aligned}$$

con

$$\bar{x}^* = \frac{n_1(n_2+g)\bar{x}_1 + n_2(n_1+g)\bar{x}_2}{n_1(n_2+g) + n_2(n_1+g)}.$$

Da notare che il condizionamento rispetto a $\mathbf{x}_2$ nella (4.21) e quello rispetto a $\mathbf{x}_1$ nella (4.22) di fatto non operano; inoltre, condizionatamente a σ^2, le due medie restano indipendenti anche dopo l'osservazione campionaria. Questa indipendenza però cessa di sussistere se si considera la distribuzione marginale congiunta di (θ_1, θ_2) (Esercizio **??**). Inoltre, la presenza di tutte e quattro le statistiche sufficienti $(\bar{x}_1, \bar{x}_2, s_1^2, s_2^2)$ nella espressione di λ^* suggerisce come le distribuzioni marginali finali delle due medie θ_1 e θ_2 dipenderanno anche da osservazioni provenienti dall'"altro" campione. Questo è solo un esempio di un fenomeno molto più generale e tipico dell'impostazione bayesiana, riportato spesso con il nome di *borrowing strength*. Il fatto che σ^2 sia considerato aleatorio e che il suo valore influenzi sia la distribuzione di θ_1 e θ_2, che la distribuzione campionaria della statistiche sufficienti, rende le varie grandezze in gioco interdipendenti. Graficamente, ??

Ritorneremo su questi aspetti a proposito dell'analisi della varianza, una tecnica che generalizza il metodo qui illustrato al caso di k diverse popolazioni.

Analogamente a quanto visto nella §4.3.2, anche in questo caso è possibile rendere meno informativa (a priori) l'analisi ponendo pari a zero gli iperparametri della legge a priori, ovvero $\alpha = g = \lambda = \delta = 0$. È facile vedere in tal caso che le formule (4.21), (4.22) e (4.23) si semplificano nelle

$$\theta_1 \mid \sigma^2, \mathbf{x}_1, \mathbf{x}_2 \sim \mathrm{N}(\bar{x}_1, \frac{\sigma^2}{n_1}) \tag{4.24}$$

$$\theta_2 \mid \sigma^2, \mathbf{x}_1, \mathbf{x}_2 \sim \mathrm{N}(\bar{x}_2, \frac{\sigma^2}{n_2}) \tag{4.25}$$

$$\sigma^2 \mid \mathbf{x}_1, \mathbf{x}_2 \sim \mathrm{GI}((n_1 + n_2)/2, (n_1 s_1^2 + n_2 s_2^2)/2) \tag{4.26}$$

Dalle formule precedenti, utilizzando noti risultati sulle combinazioni lineari delle variabili aleatorie gaussiane (vedi Appendice) si ottiene che

$$\xi \mid (\sigma^2, \mathbf{x}_1, \mathbf{x}_2) \sim N\left(\bar{x}_1 - \bar{x}_2, \frac{\sigma^2}{n^*}\right),$$

dove n^* e la cosiddetta "numerosità campionaria effettiva" ed è pari a $n_1 n_2/(n_1 + n_2)$. Allora, per il teorema E.1, si avrà, marginalmente che

$$\xi \mid (\mathbf{x}_1, \mathbf{x}_2) \sim \mathrm{St}\left(n_1 + n_2, \bar{x}_1 - \bar{x}_2, \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 n_2}\right) \tag{4.27}$$

Anche in questo caso, adottando una legge a priori impropria differente, ovvero $\pi(\theta_1, \theta_2, \sigma^2) \propto \sigma^{-2}$, è facile verificare, attraverso calcoli del tutto simili a quelli appena presentati,che

$$\xi \mid (\mathbf{x}_1, \mathbf{x}_2) \sim \mathrm{St}\left(n_1 + n_2 - 2, \bar{x}_1 - \bar{x}_2, \frac{(n_1 s_1^2 + n_2 s_2^2)(n_1 + n_2)}{(n_1 + n_2 - 2) n_1 n_2}\right), \tag{4.28}$$

ritrovando così, in forma bayesiana, un risultato noto della statistica classica.

4.6 La normale multivariata

In questa sezione consideriamo un problema più complesso, ovvero l'inferenza sui parametri di una distribuzione normale multivariata. Sarà necessario utilizzare strumenti di algebra lineare, brevemente richiamati in Appendice.

Siano allora $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{iid}}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, n vettori aleatori p-dimensionali indipendenti e somiglianti condizionatamente al vettore $\boldsymbol{\mu}$ e alla matrice di covarianza $\boldsymbol{\Sigma}$, che assumiamo definita positiva.

La funzione di verosimiglianza associata è allora

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}.$$

La forma quadratica

$$\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

può essere trasformata nel modo seguente

$$
\begin{aligned}
&\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\\
&=\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\boldsymbol{\mu})\\
&=\sum_{i=1}^{n}[(\mathbf{x}_i-\bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})]+n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\\
&=\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'\right)+n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\\
&=\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{Q}\right)+n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}),
\end{aligned}
$$

dove $\mathbf{Q}=\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'$ è la matrice di devianza campionaria, tr () rappresenta l'operatore *traccia* di una matrice quadrata e nella quarta e quinta linea della formula precedente abbiamo utilizzato la proprietà della traccia stessa (vedi Appendice A). Perciò la funzione di verosimiglianza può essere scritta come

$$
L(\boldsymbol{\mu},\boldsymbol{\Sigma})\propto\frac{1}{|\boldsymbol{\Sigma}|^{n/2}}\exp\left\{-\frac{1}{2}\left(n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})+\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{Q}\right)\right)\right\} \tag{4.29}
$$

Una distribuzione a priori coniugata per la funzione di verosimiglianza (4.29), e che rappresenta la versione multidimensionale della legge Normale Gamma Inversa, è fornita dalla cosiddetta Normale Wishart inversa, qui presentata in una forma non del tutto generale ma che consente una certa semplificazione nei calcoli,

$$
\boldsymbol{\mu}|\boldsymbol{\Sigma}\sim N_p(\boldsymbol{\xi},c^{-1}\boldsymbol{\Sigma}),\quad \boldsymbol{\Sigma}\sim IW_p(m,\boldsymbol{\Omega}), \tag{4.30}
$$

ovvero

$$
\pi(\boldsymbol{\mu}|\boldsymbol{\Sigma})\propto\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{c}{2}(\boldsymbol{\xi}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu})\right\}
$$

e

$$
\pi(\boldsymbol{\Sigma})\propto\frac{1}{|\boldsymbol{\Sigma}|^{\frac{m+p+1}{2}}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}^{-1}\right)\right\}
$$

L'utilizzo della a priori (4.30) prevede l'elicitazione dei seguenti parametri.

- $\boldsymbol{\xi}$, ovvero la media a priori del vettore $\boldsymbol{\mu}$, il valore considerato più ragionevole prima dell'esperimento;
- c, che misura il grado di fiducia nell'elicitazione a priori di $\boldsymbol{\mu}$; piccoli valori di c rendono la legge a priori poco informativa;
- $\boldsymbol{\Omega}$ e m rappresentano i parametri della legge sulla matrice di previsione $\boldsymbol{\Sigma}^{-1}$ e possono essere calibrati tenendo conto che

$$
\mathbf{E}\left(\boldsymbol{\Sigma}\right)=\frac{\boldsymbol{\Omega}^{-1}}{p-m-1}
$$

Scrivendo la distribuzione finale di $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ come proporzionale al prodotto della a priori e della funzione di verosimiglianza, si può applicare il solito Lemma in Appendice C.4 concernente la combinazione di forme quadratiche, e ottenere la distribuzione finale di $\boldsymbol{\mu}$ condizionata a $\boldsymbol{\Sigma}$,

$$
\boldsymbol{\mu}|\boldsymbol{\Sigma},\mathbf{x}\sim N_p(\boldsymbol{\xi}^*,(c+n)^{-1}\boldsymbol{\Sigma}), \tag{4.31}
$$

dove

$$\boldsymbol{\xi}^* = \frac{c\boldsymbol{\xi} + n\bar{\mathbf{x}}}{c+n};$$

la legge marginale di $\boldsymbol{\Sigma}$ è ancora di tipo Wishart inversa, ovvero

$$\boldsymbol{\Sigma}|\mathbf{x} \sim IW_p(m+n, \boldsymbol{\Omega}^*), \tag{4.32}$$

dove

$$\boldsymbol{\Omega}^* = \left(\mathbf{Q} + \boldsymbol{\Omega}^{-1} + \frac{nc}{n+c}(\bar{\mathbf{x}} - \boldsymbol{\xi})'(\bar{\mathbf{x}} - \boldsymbol{\xi})\right)^{-1},$$

In particolare, ponendo $c = m = 0$ e $\boldsymbol{\Omega}^{-1} = \mathbf{0}$, si ottiene una distribuzione a priori impropria, ovvero qualcosa che non è una distribuzione di probabilità. Tuttavia tale scelta dà luogo ad una ben definita legge a posteriori con la quale si ottengono risultati che, seppure formalmente diversi, coincidono praticamente coi risultati ottenibili attraverso un'analisi frequentista del problema. Infatti, in tal caso le (4.31) e (4.32) diventano

$$\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathbf{x} \sim N_p(\bar{x}, n^{-1}\boldsymbol{\Sigma}),$$

e

$$\boldsymbol{\Sigma}|\mathbf{x} \sim IW_p(n, \mathbf{Q}^{-1}).$$

QUESTO VA SPOSTATO DOPO

Caso non informativo: $c \to 0, \quad \boldsymbol{\Omega}^{-1} = \mathbf{0}, \quad m = \mathbf{0}$ Equivale all'utilizzo della legge di Jeffreys

$$\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \frac{1}{|\boldsymbol{\Sigma}|^{\frac{p+1}{2}}}$$

Significato della legge di Jeffreys

La matrice simmetrica definita positiva $\boldsymbol{\Sigma}$ può esprimersi come

$$\boldsymbol{\Sigma} = \mathbf{H}'\mathbf{D}\mathbf{H}$$

dove $\mathbf{H}$ è una matrice ortogonale e $\mathbf{D}$ è la matrice diagonale degli autovettori, ovvero

$$\mathbf{H}'\mathbf{H} = \mathbf{I}_p \qquad \mathbf{D} = \text{ diag } (\lambda_1, \ldots, \lambda_p).$$

Si può allora scrivere, assumendo che gli autovalori siano diversi tra loro,

$$\pi(\boldsymbol{\Sigma})d\boldsymbol{\Sigma} = \pi(\mathbf{H}, \mathbf{D})I_{[\lambda_1 > \lambda_2 > \cdots > \lambda_n]}d\mathbf{H}d\mathbf{D}$$

e si può dimostrare che

$$\pi(\mathbf{H}, \mathbf{D}) = \pi(\mathbf{H}'\mathbf{D}\mathbf{H})\prod_{i<j}(\lambda_i - \lambda_j)$$

Perciò

$$\pi^J(\mu, \boldsymbol{\Sigma}) = \pi^J(\mu, \mathbf{H}, \mathbf{D}) \propto \frac{1}{|\mathbf{D}|^{\frac{p+1}{2}}} \prod_{i<j} (\lambda_i - \lambda_j)$$

che, senza motivo apparente, introduce un fattore che tende a tenere "separati gli autovalori della matrice $\boldsymbol{\Sigma}$. Questo è controintuitivo poiché in genere si tende a considerare "simili gli autovalori.

Questa conclusione è in accordo con le conclusioni classiche, in cui lo stimatore $\mathbf{S}$ viene in genere modificato poiché tende ad avere autovalori troppo sparsi...

4.7 Consistenza del metodo bayesiano

Teorema 4.1 *Sia lo spazio parametrico formato da un numero finito di punti, ovvero*

$$\Omega = \{\theta_0, \theta_1, \cdots, \theta_m\},$$

e consideriamo una distribuzione a priori $\pi(\theta)$ che dia peso positivo ad ogni θ_j in Ω, ovvero

$$\pi(\theta_j) = p_j > 0, \qquad j = 0, 1, \cdots, m,$$

con $p_0 + p_1 + \cdots p_m = 1$. Sia $X_1, X_2, \cdots$ una successione di v.a. indipendenti e somiglianti con distribuzione $p(\cdot \mid \theta_0)$. Allora

$$\lim_{n\to\infty} \pi(\theta_j \mid X_1, X_2, \cdots, X_n) = \begin{cases} 1 & j = 0 \\ 0 & j > 0 \end{cases}. \tag{4.33}$$

In altri termini il teorema afferma che, al crescere dell'informazione campionaria, la distribuzione finale andrà a concentrarsi, con probabilità 1, sul vero valore del parametro, indipendentemente dalla legge iniziale sul parametro stesso, a patto che, inizialmente, nessuno tra i possibili valori di θ sia escluso.

Dimostrazione 4.1 *Sarà sufficiente dimostrare la (4.33) per il caso $j = 0$. La quantità $\pi(\theta_0 \mid \mathbf{x})$ può essere riespressa nella forma*

$$\pi(\theta_0 \mid \mathbf{x}) = \frac{p_0 \prod_{i=1}^n p(x_i \mid \theta_0)}{\sum_{j=0}^m p_j \prod_{i=1}^n p(x_i \mid \theta_j)} = \frac{p_0}{\sum_{j=0}^m p_j \prod_{i=1}^n \frac{p(x_i|\theta_j)}{p(x_i|\theta_0)}} =$$

$$p_0 \left[p_0 + \sum_{j=1}^m p_j \prod_{i=1}^n \frac{p(x_i \mid \theta_j)}{p(x_i \mid \theta_0)} \right]^{-1}.$$

Occorre allora dimostrare che

$$\sum_{j=1}^m p_j \prod_{i=1}^n \frac{p(x_i \mid \theta_j)}{p(x_i \mid \theta_0)} \to 0$$

Per questo è sufficiente mostrare che

$$\prod_{i=1}^n \frac{p(x_i \mid \theta_j)}{p(x_i \mid \theta_0)} = \exp\left\{ \sum_{i=1}^n \log \frac{p(x_i \mid \theta_j)}{p(x_i \mid \theta_0)} \right\} \to 0, \qquad \forall j = 1, \cdots, m,$$

Sia

$$Z_j = \log \frac{p(x_i \mid \theta_j)}{p(x_i \mid \theta_0)} \qquad j = 1, \cdots n;$$

Allora

$$\exp\left\{\sum_{i=1}^{n}\log\frac{p(x_i \mid \theta_j)}{p(x_i \mid \theta_0)}\right\} = \exp\left\{n\frac{1}{n}\sum_{i=1}^{n} Z_j\right\}.$$

Per la legge forte dei grandi numeri

$$\frac{1}{n}\sum_{i=1}^{n} Z_j \overset{q.c.}{\to} E(Z_1).$$

Per la disuguaglianza di Jensen, inoltre,

$$E(Z_1) = E\left(\log\frac{p(x_i \mid \theta_j)}{p(x_i \mid \theta_0)}\right) < \log E\left(\frac{p(x_i \mid \theta_j)}{p(x_i \mid \theta_0)}\right) = \log 1 = 0$$

Ne segue che $\frac{1}{n}\sum_{i=1}^{n} Z_j$ *converge quasi certamente. ad un valore negativo e, di conseguenza,*

$$\exp\left\{\sum_{i=1}^{n}\log\frac{p(x_i \mid \theta_j)}{p(x_i \mid \theta_0)}\right\} \to 0, \qquad q.c.$$

Il precedente teorema non è sorprendente. Il metodo bayesiano, utilizzando tutta l'informazione inerente al parametro, fornita dalla funzione di verosimiglianza, eredita le buone proprietà asintotiche di quest'ultima. L'informazione extra-sperimentale, inserita attraverso la legge a priori, può modificare le inferenze "nel finito", per piccole dimensioni campionarie, ma il risultato asintotico viene determinato dalla funzione di verosimiglianza, a patto che la legge a priori non escluda alcuna possibilità. Lo stesso risultato vale in situazioni più generali, ad esempio quando lo spazio parametrico Ω è un sottoinsieme di uno spazio euclideo $\mathbb{R}^k$ per qualche k intero. Le questioni diventano più complesse in ambito non parametrico, ovvero quando il parametro di interesse ha dimensione infinita: in questi casi è più complesso elicitare una distribuzione a priori su uno spazio infinito dimensionale che garantisca la consistenza del metodo bayesiano. Esistono a tal proposito, alcuni controesempi notevoli. La natura introduttiva di questo testo impedisce di proseguire lungo questa strada: il lettore interessato può approfondire leggendo il testo di [44].

4.8 Esercizi

5

Scelta della distribuzione iniziale

La scelta della distribuzione iniziale per i parametri presenti nel modello statistico è stata considerata a lungo, a torto o a ragione, l'aspetto cruciale dell'impostazione bayesiana. Di fatto, la distribuzione iniziale è il vettore attraverso cui le informazioni extra-sperimentali vengono inserite nel procedimento induttivo, e l'adozione di una distribuzione iniziale rende l'analisi statistica, almeno sul piano formale, inequivocabilmente soggettiva. Secondo molti studiosi, tale soggettività rende problematica, se non impossibile, la comunicazione scientifica: lo stesso risultato sperimentale potrebbe infatti condurre a conclusioni inferenziali sostanzialmente diverse, qualora le informazioni a priori introdotte nel modello fossero anche parzialmente differenti.

Esempio 5.1 [*Parziale specifica della distribuzione iniziale*].

[10] considera il seguente esempio stilizzato. Si hanno n osservazioni da una popolazione normale con media θ incognita e varianza nota e pari a 1. La statistica sufficiente per θ è la media campionaria $\bar{x}$ e la funzione di verosimiglianza, come visto nella §4.3.1, vale

$$L(\theta) \propto \exp\left\{-\frac{1}{2}(\bar{x}-\theta)^2\right\}.$$

Le informazioni a priori si limitano alla elicitazione dei soli tre quartili che assumiamo essere

$$Q_1 = -1, \qquad Q_2 = Me = 0, \qquad Q_3 = 1. \tag{5.1}$$

Esistono ovviamente infinite distribuzioni di probabilità compatibili con la nostra parziale elicitazione dell'informazione a priori. A scopo illustrativo consideriamo le seguenti:

a)
$$\pi_1(\theta) = \frac{1}{\tau\sqrt{2\pi}} \exp\left\{-\frac{1}{2\tau^2}\theta^2\right\},$$
con $\tau^2 = 2.19$; ovvero $\theta \sim N(0, 2.19)$.

b)
$$\pi_3(\theta) = \frac{\lambda}{2} \exp\left\{-\lambda \mid \theta \mid\right\},$$
con $\lambda = 1.384$; ovvero θ ha distribuzione di Laplace La$(0, 1.384)$.

c)
$$\pi_2(\theta) = \frac{1}{\pi(1+\theta^2)},$$
ovvero $\theta \sim Ca(0,1)$, cioè θ ha distribuzione di tipo Cauchy standardizzata.

Supponiamo di voler sintetizzare l'informazione a posteriori attraverso il calcolo delle media. Questo implica il calcolo di un integrale che ha una soluzione analitica soltanto quando utilizziamo π_1. Negli altri due casi occorre ricorrere a procedure numeriche. La tabella 5.1 riporta il calcolo della media a posteriori di θ utilizzando le tre diverse distribuzioni a priori e per diversi valori di $\bar{x}$ (la dimensione campionaria può essere posta uguale a uno, senza perdere in generalità):

$\pi(\theta\|\bar{x})$	0.0	1.0	2.0	3.0	4.5	6.0	10.0
$N(0, 2.19)$	0	0.69	1.37	2.06	3.09	4.12	6.87
$La(1.384)$	0	0.39	0.92	1.68	3.12	4.61	8.61
$Ca(0, 1)$	0	0.55	1.28	2.28	4.09	5.66	9.80

Tabella 5.1. Media a posteriori per diversi valori di $\bar{x}$ e diverse distribuzioni a priori.

♢

Si può notare come la differenza tra le conclusioni aumenti all'aumentare del valore di x osservato, ovvero all'aumentare della discordanza tra le informazioni a priori (in tutti e tre i casi, centrate intorno al valore $\theta = 0$) e il risultato sperimentale. Questo esempio ci insegna che esistono casi in cui, soprattutto se le informazioni a priori contrastano con l'evidenza sperimentale, la distribuzione finale risulta fortemente instabile e molto sensibile a variazioni della legge iniziale. Va però precisato che non si tratta di un difetto intrinseco della procedura bayesiana: essa si limita a riportare, correttamente, il momento di contrasto tra le due fonti informative: si veda al proposito [63]. Le argomentazioni a favore dell'impostazione bayesiana, di contro, si basano sulle seguenti considerazioni:

1. nulla è veramente oggettivo: l'evidenza sperimentale serve solo a modificare, attraverso il meccanismo fornito dal teorema di Bayes, le valutazioni pre-sperimentali di ciascun individuo. Il teorema di Bayes rappresenta lo strumento matematico attraverso il quale un individuo razionale aggiorna le proprie conoscenze sulla base del risultato sperimentale
2. il ruolo della distribuzione iniziale spesso non è così determinante dal punto di vista pratico: soprattutto a fronte di una dimensione campionaria elevata, le distribuzioni a posteriori relative a due diverse distribuzioni iniziali risulteranno più "vicine" delle corrispondenti iniziali.

Questo dibattito ha suscitato, e suscita tuttora, un'attenzione particolare verso lo studio di distribuzioni iniziali di tipo "convenzionale" o, come spesso vengono indicate con un termine fuorviante, "non informative". Tali distribuzioni dovrebbero formalizzare, in un qualche senso, l'assenza di qualsiasi informazione sui parametri del modello: la loro determinazione è stata interpretata per qualche tempo come un ponte in grado di unire la logica bayesiana con la pretesa oggettività dei metodi classici. Oggi, più modestamente, le distribuzioni convenzionali vengono percepite come uno strumento che consenta comunque di adottare un approccio bayesiano, pure in assenza di precise informazioni a priori. Di queste tematiche discuteremo nella §5.2.

L'uso di distribuzioni iniziali comporta inoltre problemi di tipo computazionale. A meno di non utilizzare forme analitiche particolari, le cosiddette distribuzioni "coniugate" al modello statistico utilizzato, risulta spesso difficile, se non impossibile, ottenere espressioni analitiche in forma esplicita

per le distribuzioni finali, come già osservato nell'esempio 5.1. La difficoltà di gestire distribuzioni finali non in forma esplicita ha costituito un ostacolo alla diffusione delle tecniche bayesiane nella pratica statistica, almeno fino agli inizi degli anni '90 del secolo scorso: attualmente, la possibilità di ottenere campioni "simulati" dalla distribuzione finale rende questo problema meno cogente: rinviamo la discussione e la descrizione dei metodi computazionali al capitolo 7.

Possiamo quindi affermare che, attualmente, convivano almeno due diverse filosofie all'interno dell'impostazione bayesiana, che definiremo brevemente con i termini, riduttori ma espliciti, di "soggettiva" e "oggettiva". La filosofia oggettiva fa riferimento ai lavori classici di [32], [58], e afferma come un problema di inferenza statistica vada sempre inquadrato in un contesto probabilistico; la distribuzione iniziale è soggettiva; ognuno parte dalla formalizzazione delle proprie conoscenze che vengono modificate attraverso il dato sperimentale in modo coerente. L'impostazione soggettiva presenta, allo stesso tempo, solide basi fondazionali ed una non trascurabile difficolt in ambito applicativo: è infatti molto difficile, a volte impossibile esprimere in modo preciso le proprie opinioni iniziali attraverso una singola distribuzione di probabilità: si pensi ad esempio ai moderni problemi di *Machine Learning*, dove i parametri da stimare sono spesso nell'ordine di migliaia.

La seconda impostazione, quella cosiddetta "oggettiva", pur riconoscendo la validit teorica dell'approccio soggettivo[1] parte dal riconoscimento della debolezza applicativa del metodo soggettivo e propone la determinazione e l'adozione di distribuzioni a priori di tipo convenzionale, derivabili sulla base delle proprietà matematiche del modello statistico utilizzato. In tal modo si tenta di conseguire due obiettivi: da una parte conservare la pretesa "oggettività" delle procedure inferenziali che dipendono in questo modo esclusivamente dal modello statisico prescelto e dal risultato campionario; dall'altra, l'uso di una legge iniziale consente ancora di fare uso del teorema di Bayes per produrre conclusioni inferenziali nel linguaggio probabilistico, proprio dell'approccio bayesiano. Ma il perseguimento simultaneo dei due obiettivi comporta, come vedremo in questo capitolo alcuni problemi sia di forma che di sostanza, primo fra tutti il fatto che le leggi iniziali convenzionali saranno molto raramente delle vere e proprie leggi di probabilità: più spesso si tratterà di leggi cosiddette improprie, ovvero funzioni positive con integrale infinito. Torneremo più volte su queti aspetti nel corso di questo capitolo.

Il capitolo è organizzato come segue: nella §5.1 si illustra brevemente, attraverso alcuni esempi, l'utilizzo delle distribuzioni coniugate. Nella §5.2 viene illustrata, in modo più articolato, la problematica delle distribuzioni convenzionali: in particolare vengono descritte le tecniche più comunemente utilizzate in letteratura, ovvero la distribuzione iniziale invariante di Jeffreys e il metodo delle *reference priors*.

Un modo alternativo di analizzare l'influenza della distribuzione a priori sulle inferenze finali consiste nel calcolare direttamente qual è il range della stima a posteriori della quantità d'interesse quando la distribuzione a priori può variare in una classe predeterminata.

Esempio 5.1(continua). Potremmo considerare, in questo caso, la classe Γ di tutte le distribuzioni di probabilità a priori compatibili con i vincoli imposti dalla (5.1). Occorre allora calcolare l'estremo inferiore e l'estremo superiore del valore atteso a posteriori di θ (o di una qualunque altra funzione parametrica di interesse), ovvero

[1] che va comunque perseguito quando possibile, ovvero quando si è in grado di produrre una distribuzione iniziale per i parametri del modello

$$\left[\inf_{\pi\in\Gamma} \mathbb{E}\left(\theta \mid \mathbf{x}\right); \sup_{\pi\in\Gamma} \mathbb{E}\left(\theta \mid \mathbf{x}\right)\right]. \tag{5.2}$$

Questo approccio prende il nome di *robustezza globale* ed è stato analizzato a fondo negli anni 90 del secolo scorso. Un discussione non approfondita di queste tematiche si trova nella §5.3.

5.1 Distribuzioni coniugate

Siano $X_1, X_2, \cdots, X_n$, n variabili aleatorie somiglianti e indipendenti condizionatamente a un vettore di parametri $\theta \in \Omega$. Assumiamo che le v.a. siano dotate di densità (nel caso di v.a. assolutamente continue) o distribuzione di probabilità (nel caso discreto) indicata con $p(x \mid \theta)$. La funzione di verosimiglianza per θ associata a un vettore di osservazioni $(X_1 = x_1, \cdots, X_n = x_n)$ è allora, come già noto,

$$L(\theta) \propto \prod_{j=1}^{n} p(x_j \mid \theta).$$

Una distribuzione di probabilità iniziale $\pi(\theta)$ si dice coniugata al modello utilizzato o, equivalentemente, alla funzione di verosimiglianza $L(\theta)$, se la forma funzionale della distribuzione iniziale e della distribuzione finale sono uguali.

Esempio 5.2 [*Dati dicotomici*].

Abbiamo già visto nel capitolo precedente che, dato un modello bernoulliano, l'uso di una distribuzione iniziale di tipo Beta(α, β) implica che la distribuzione finale sia ancora di tipo Beta con parametri modificati dalle osservazioni campionarie, cosicch $\alpha^* = \alpha + k$ e $\beta^* = \beta + n - k$, dove k è il numero di successi nelle n prove effettuate. ♢

Esempio 5.3 [*Esponenziale-Gamma*]

Siano $X_1, \cdots, X_n$ n osservazioni indipendenti con distribuzione esponenziale di parametro λ, ovvero, per $j = 1, \cdots, n$

$$p(x_j \mid \lambda) = \lambda \exp\{-\lambda x_j\}.$$

La funzione di verosimiglianza è allora

$$L(\lambda) \propto \lambda^n \exp\left\{-\lambda \sum_{j=1}^{n} x_j\right\},$$

e una espressione per la distribuzione iniziale di λ che sia coniugata con la funzione di verosimiglianza è data dalla distribuzione di tipo Gamma(γ, δ). Si vede allora facilmente che la distribuzione a posteriori risultante è proporzionale a

$$\pi(\lambda \mid \mathbf{x}) \propto \lambda^{n+\gamma-1} \exp\left\{-\lambda(\delta + n\bar{x})\right\},$$

ovvero si tratta ancora di una distribuzione di tipo Gamma(δ^*, γ^*), dove i parametri sono stati aggiornati dall'informazione sperimentale in

$$\gamma^* = \gamma + n \text{ e } \delta^* = \delta + n\bar{x} \quad .$$

Anche qui la media a posteriori, che vale $(\gamma + n)/(\delta + n\,\bar{x})$, può essere espressa come una combinazione convessa della media a priori e della stima di massima verosimiglianza con pesi pari a $p = \delta/(\delta + n\,\bar{x})$ e $1 - p$, rispettivamente ♢

Esempio 5.4 [*Normale-normale*].
Nella §4.3.2 abbiamo considerato il caso di una popolazione normale nei due casi in cui solo la varianza oppure media e varianza siano considerate incognite. Le distribuzioni iniziali utilizzate in quegli esempi erano ancora di tipo coniugato al modello pertinente. ♢

Esempio 5.5 [*Famiglia Esponenziale*].
Gli esempi discussi in questa sezione sono in realtà tutti casi particolari della più generale famiglia esponenziale, discussa in §2.11 ♢

Le distribuzioni coniugate sono state utilizzate molto fino agli anni '80 soprattutto per la semplicità di calcolo collegata al loro uso. Tuttavia i vincoli che impongono, soprattutto nella calibrazione delle code della distribuzione iniziale, sono difficilmente giustificabili in termini di informazioni a priori. L'esplosione delle potenzialità computazionali negli ultimi anni, di cui daremo conto nel Capitolo 7, ha ridotto notevolmente l'importanza delle distribuzioni coniugate. In ogni caso, quello che segue è uno specchietto riassuntivo delle più note coniugazioni tra modelli statistici e distribuzioni iniziali, tratto da [68].

Modello	**Distrib. iniziale**	**Distr. finale**	**Notazione**
$\text{Be}(\theta)$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + k, \beta + n - k)$	$k =$ numero di successi
$\text{N}(\mu, \sigma_0^2)$	$\text{N}(\mu_0, \tau^2)$	$\text{N}(\frac{\mu_0\sigma^2 + \bar{x}n\tau^2}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2})$	σ_0^2 noto
$\text{Po}(\theta)$	$\text{Gamma}(\lambda, \delta)$	$\text{Gamma}(\lambda + n\bar{x}, \delta + n)$	
$\text{Esp}(\theta)$	$\text{Gamma}(\lambda, \delta)$	$\text{Gamma}(\lambda + n, \delta + n\bar{x})$	
$\text{U}(0, \theta)$	$\text{Pa}(\alpha, \xi)$	$\text{Pa}(\alpha + n, w)$	$w = \max(x_{(n)}, \xi)$

Tabella 5.2. Principali distribuzioni coniugate

5.2 Distribuzioni non informative

Nell'ultimo capoverso della §4.3.1 si era accennato al fatto che nel modello normale-normale, quando si fa tendere la varianza τ^2 della distribuzione iniziale a $+\infty$, si sta in pratica utilizzando una legge di probabilità impropria, che conduce, tuttavia, ad una distribuzione finale ben definita, in grado di fornire inferenze operativamente identiche a quelle frequentiste. È possibile estendere questa coincidenza di risultati a modelli statistici più generali? E' possibile, in altri termini, arrivare per vie bayesiane ad una completa rilettura della statistica frequentista? Se si potesse rispondere in modo affermativo alla precedente domanda, avremmo ridotto l'impostazione classica dell'inferenza statistica ad un caso particolare, limite, dell'impostazione bayesiana, da utilizzare quando non esista alcuna informazione extra-sperimentale e tuttavia si voglia conservare la ricchezza e la praticità dell'output di un'analisi bayesiana.

È comprensibile così l'importanza teorica e pratica dello studio e della derivazione di distribuzioni a priori convenzionali, non riferite cioè ad uno specifico livello informativo del ricercatore. Se, per ogni modello statistico, esistesse una distribuzione iniziale di questo tipo[2], l'impostazione bayesiana, per così dire, "non soggettiva, potrebbe essere considerata un quadro di riferimento

[2] Sono molti gli aggettivi con cui tali distribuzioni vengono definite: di riferimento, di default, convenzionali, non informative, oggettive.

generale per l'impostazione del problema inferenziale, nella sua accezione più generale. Molti ricercatori, ancora oggi, sono impegnati in questa direzione e, sebbene il quadro generale sia lungi dall'essere completato, è innegabile che attualmente la quasi totalità delle applicazioni pratiche dei modelli bayesiani si avvale dell'uso di distribuzioni iniziali convenzionali.

5.2.1 Notazione e motivazioni

Non è possibile dare una definizione soddisfacente del concetto di distribuzione non informativa, proprio perch è altrettanto sfuggente il concetto di *informazione* in senso probabilistico. Dal punto di vista operativo, una distribuzione non informativa dovrebbe concretizzarsi in una distribuzione iniziale basata non già sulle informazioni a disposizione del ricercatore, bensì costruita in modo convenzionale, tenendo conto solamente della struttura statistica dell'esperimento.

Il motivo primario del loro esteso utilizzo è che consentono di effettuare un'analisi statistica di carattere sì bayesiano, ma comunque non basata su informazioni o valutazioni specifiche dell'utente e per questo percepita più facilmente come condivisibile tra diversi utilizzatori. Il termine "non informative", con cui sono maggiormente note in letteratura, è un po' fuorviante poich fa pensare alla possibilità che alcune leggi di probabilità possano non contenere informazione: in realtà si tratta di distribuzioni di probabilità[3] che producono inferenze di tipo bayesiano, ovvero basate sulla distribuzione finale del parametro oggetto di interesse, che sono allo stesso tempo il più possibile "vicine" (in un senso che va precisato matematicamente) alle soluzioni frequentiste.

L'uso delle distribuzioni non informative prefigura così un approccio al processo induttivo per certi versi differente da quello bayesiano ortodosso, il quale presuppone la quantificazione matematica delle opinioni specifiche del ricercatore sulle quantità non osservabili in gioco. La tabella che segue tenta una schematizzazione delle differenze teoriche ed operative tra il punto di vista bayesiano ortodosso e quello, per dirla con un ossimoro, "oggettivo", ovvero basato sull'uso di distribuzioni a priori convenzionali.

	Imp. oggettiva	Imp. soggettiva
Motivazioni	Pragmatismo	Coerenza
Obiettivi	Inferenza	Max. Utilità Attesa
Probabilità	Non solo soggettiva	Soggettiva
Elicitazione	Automatica	Soggettiva
Uso pratico	Enorme	Limitato
Punto di forza	Facilità d'uso	Sistema coerente
Punti deboli	Definizione vaga - Possibili Incoerenze	inferenze diverse da quelle classiche

Nel corso di questa sezione denoteremo con il simbolo

$$\mathcal{E}_k = (\mathcal{X}_k, \boldsymbol{\Omega}, \mathcal{P}) \tag{5.3}$$

il consueto esperimento statistico che consiste nella osservazione di k replicazioni indipendenti e somiglianti (condizionatamente al vettore dei parametri $\omega \in \boldsymbol{\Omega} \subseteq \mathbb{R}^d$), $\mathbf{x} = (X_1, \cdots, X_k)$ provenienti da una legge di probabilità che assumiamo appartenere alla famiglia

$$\mathcal{P} = \{p(\cdot \mid \omega), \quad \omega \in \boldsymbol{\Omega}\},$$

[3] Ad essere rigorosi esse spesso non rispettano tutte le caratteristiche di una legge di probabilità: ad esempio, sono spesso improprie.

dove $\boldsymbol{\Omega}$ è lo spazio dei parametri. Con il simbolo $\pi^N(\omega)$ indicheremo una generica distribuzione non informativa su $\boldsymbol{\Omega}$. Nel corso di questa trattazione, per forza di cose introduttiva considereremo soltanto il caso in cui il problema inferenziale sia un problema di stima. La determinazione di distribuzioni convenzionali per problemi di verifica di ipotesi e scelta del modello statistico è un problema decisamente più complesso di cui ci occuperemo, sia pure in modo superficiale nella §6.3.3.

5.2.2 La distribuzione uniforme

Consideriamo inizialmente la situazione sperimentale più semplice in cui lo spazio parametrico sia composto da un numero finito di possibili valori, cioè

$$\boldsymbol{\Omega} = \{\omega_1, \omega_2, \cdots, \omega_H\}$$

In tal caso è piuttosto intuitivo ritenere che una distribuzione iniziale non informativa per questo problema debba pesare allo stesso modo le H possibili alternative cosicché avremo

$$\pi^N(\omega_j) = H^{-1}, \quad j = 1, \cdots, H.$$

L'intuizione è stata in questo caso guidata dal cosiddetto *Principio di Ragione Insufficiente* (PRI) [54]. Nonostante la semplicità del contesto, anche questo problema nasconde delle insidie: se, ad esempio, ω_1 viene ulteriormente suddiviso in due diversi valori, diciamo ω_{11} e ω_{12}, la cardinalità di $\boldsymbol{\Omega}$ diviene $H + 1$ e, di conseguenza, il PRI indurrebbe ad utilizzare una distribuzione a priori uniforme sugli $H + 1$ valori, modificando totalmente la legge iniziale. La situazione è ancora più complessa quando $\boldsymbol{\Omega}$ ha una cardinalità numerabile o addirittura superiore.

Esempio 5.6 [*Modello di Bernoulli*] Si osservano k replicazioni $(X_1, \cdots, X_k)$ di una v.a.

$$X \sim Be(\omega), \quad \omega \in [0, 1],$$

In questo esempio lo spazio parametrico, pur avendo potenza del continuo, è compatto e una naturale estensione del PRI porterebbe a supporre che la legge iniziale di default per ω sia uniforme su $\boldsymbol{\Omega}$, ovvero

$$\pi^N(\omega) = 1, \qquad 0 \leq \omega \leq 1.$$

Sebbene tale scelta sia operativamente ragionevole, vedremo che in diverse impostazioni, si preferiranno distribuzioni differenti. ♢

Il problema essenziale legato all'uso della legge uniforme (costante) su un sottoinsieme della retta reale è la mancanza di invarianza rispetto a trasformazioni biunivoche (uno ad uno) del vettore dei parametri ω.

Esempio 5.6 (continua). In alcune applicazioni, il parametro di interesse potrebbe non essere ω bensì una sua trasformazione, ad esempio, $\lambda = -\log \omega$: la legge uniforme su ω induce, per via dello Jacobiano della trasformazione, la seguente legge sul nuovo parametro λ:

$$\pi^N(\lambda) = e^{-\lambda} \ \mathbf{1}_{[0,\infty)}(\lambda),$$

riconoscibile come una distribuzione esponenziale di parametro 1. Di contro una legge uniforme su λ non ricondurrebbe alla legge uniforme su ω[4]. ♢

Quali rimedi adottare allora nel caso in cui $\boldsymbol{\Omega}$ ha una cardinalità infinita più che numerabile? Affronteremo la situazione in dettaglio nelle prossime sezioni; vale intanto la pena sottolineare che

- I difensori e gli utilizzatori del PRI sostengono che, prima di poter utilizzare tale principio, occorre scegliere con cura la cardinalità e la parametrizzazione rispetto alle quali sarà poi ragionevole adottare una distribuzione uniforme.
- Il fatto che la legge di probabilità sia impropria non è, in genere, un vero problema. Leggi improprie sono teoricamente giustificabili in termini di additività finita [71]. L'importante è verificare che la corrispondente distribuzione a posteriori ottenuta mediante applicazione del teorema di Bayes risulti propria, qualunque sia il risultato campionario osservato!!

5.2.3 Il metodo di Jeffreys

La mancanza di invarianza della legge uniforme rispetto a trasformazioni dei parametri indusse [50] a formulare un nuovo criterio di costruzione di distribuzioni a priori non informative. La definizione è piuttosto semplice: dato il modello statistico (5.3), la distribuzione non informativa di Jeffreys è

$$\pi^J(\omega) \propto \sqrt{\det(I(\omega))}, \tag{5.4}$$

dove $I(\omega)$ rappresenta la matrice d'informazione attesa di Fisher, definita in §2.5 relativa ad una singola osservazione

$$I(\omega) = \left\{ I_{a,b} = -\,\mathbb{E}_\omega \left(\frac{d^2}{d\omega_a d\omega_b} \log p(x|\omega) \right) \right\},$$

$a, b = 1, \cdots, d$. L'uso di $\pi^J(\omega)$ è ovviamente subordinato all'esistenza di tale matrice e al suo essere definita positiva, condizione peraltro verificata nella quasi totalità dei modelli statistici di uso comune.

Esempio 5.6 (continua). La densità relativa ad una singola osservazione è

$$p(x \mid \omega) = \omega^x (1-\omega)^{1-x} \mathbf{1}_{(0,1)}(x);$$

ne segue che

$$\begin{aligned} \ell(\omega) = \log p(x \mid \omega) &= x \log \omega + (1-x) \log(1-\omega); \\ \frac{d}{d\omega}\ell(\omega) &= \frac{x}{\omega} - \frac{1-x}{1-\omega}; \\ -\frac{d^2}{d\omega^2}\ell(\omega) &= \frac{x}{\omega^2} + \frac{1-x}{(1-\omega)^2}; \\ -\,\mathbb{E}_\omega \left(\frac{d^2}{d\omega^2}\ell(\omega) \right) &= \frac{1}{\omega} + \frac{1}{1-\omega} = \frac{1}{\omega(1-\omega)}; \end{aligned}$$

la legge di Jeffreys è dunque

$$\pi^J(\omega) = \frac{1}{\pi} \omega^{-1/2} (1-\omega)^{-1/2}. \tag{5.5}$$

Si tratta di una distribuzione di tipo Beta($\frac{1}{2}, \frac{1}{2}$), simmetrica e dalla caratteristica forma ad U, come si nota dalla Figura 5.1. La distribuzione (5.5) è invariante per riparametrizzazioni biunivoche di

[4] in realtà non si tratterebbe nemmeno di una distribuzione di probabilità, poich λ è definito su un supporto illimitato.

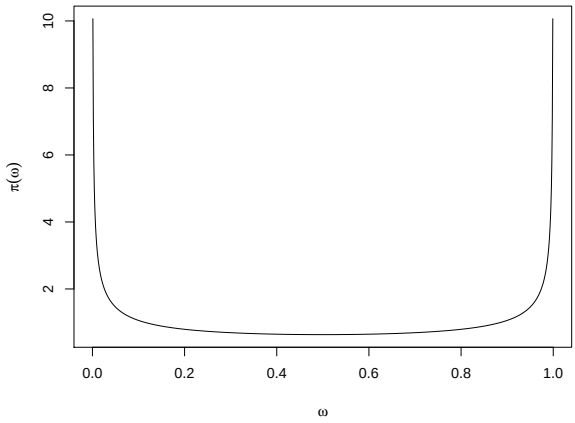

Figura 5.1. Distribuzione iniziale di Jeffreys per il modello binomiale.

ω. Questo implica che se $\lambda = \lambda(\omega)$ è una trasformazione biunivoca di ω tale che $I(\lambda)$ esiste ed è definita positiva, allora è possibile dimostrare che

$$\pi_\lambda^J(\lambda) = \pi_\omega^J(\omega(\lambda)) \mid \det(\frac{\partial \omega}{\partial \lambda}) \mid, \tag{5.6}$$

ovvero la distribuzione di Jeffreys può ottenersi sia ricalcolando l'informazione attesa di Fisher nella nuova parametrizzazione oppure attraverso la formula usuale per il calcolo della densità di una trasformazione biunivoca di variabile aleatoria. Questo risultato è garantito dall'invarianza della legge di Jeffreys.

Esempio 5.6 (continua). Consideriamo la trasformazione biunivoca $\lambda = -\log\theta$; il nuovo parametro λ ha ora come supporto il semiasse positivo e la nuova legge di Jeffreys è data da

$$\pi^J(\lambda) = \frac{1}{\pi}\pi^J(\theta(\lambda))\left|\frac{\partial\theta}{\partial\lambda}\right| = e^{\frac{\lambda}{2}}\left(1-e^{-\lambda}\right)^{-\frac{1}{2}}e^{-\lambda} = \left(e^{\lambda}-1\right)^{-\frac{1}{2}} \tag{5.7}$$

Lo stesso risultato è ottenibile se l'intero modello bernoulliano è espresso mediante il nuovo parametro λ. In questo caso, la log-verosimiglianza relativa ad una singola osservazione risulta essere

$$\ell(\lambda) = -\lambda x + (1-x)\log(1-e^{-\lambda})$$

da cui si ottiene facilmente che

$$\frac{\partial}{\partial\lambda}\ell(\lambda) = -x + (1-x)\frac{e^{-\lambda}}{(1-e^{-\lambda})}$$

e

$$-\frac{\partial^2}{\partial\lambda^2}\ell(\lambda) = (1-x)\frac{e^{-\lambda}}{(1-e^{-\lambda})^2};$$

ne segue che $I(\lambda) = e^{-\lambda}/(1-e^{-\lambda}) = (e^{\lambda}-1)^{-1}$ e la distribuzione di Jeffreys così ottenuta coincide con la (5.7). ♢

L'idea di Jeffreys fu quella di generalizzare il concetto di invarianza già popolare nel caso di modelli con parametri di posizione (o di scala, o di posizione e scala) come il modello gaussiano con μ e/o σ parametri incogniti.

Esempio 5.7 [*Parametro di posizione*]. Siano $X_1, \cdots, X_n \overset{\text{iid}}{\sim} f(x-\mu)$. La famiglia $\{f_\mu, \mu \in \mathbb{R}\}$ è invariante per traslazione, nel senso che $Y = X + a$ ha una distribuzione ancora del tipo f, qualunque sia a. μ è allora un parametro di posizione e come tale si richiede una "naturale" condizione d'invarianza

$$\pi(\mu) = \pi(\mu - a), \quad \forall a \in \mathbb{R}. \tag{5.8}$$

L'equazione (5.8) ha come unica soluzione la funzione costante $\pi(\mu) = k$, per k qualunque, che per altro, non è una distribuzione di probabilità per nessun valore di k. ♢

È facile ora vedere che l'utilizzo del metodo di Jeffreys conduce allo stesso risultato: qualunque sia la famiglia di posizione utilizzata, si ottiene sempre una $I(\mu)$ costante. Infatti, sia

$$\mathcal{F} = \{f(x-\mu), \quad \mu \in \mathbb{R}\}$$

il modello utilizzato. Allora $f'_\mu(x;\mu) = -f'_\mu(x-\mu)$ e

$$\frac{f'_\mu(x;\mu)}{f(x;\mu)} = -\frac{f'_\mu(x-\mu)}{f(x-\mu)};$$

Ricordando la definizione di informazione attesa di Fisher come varianza della funzione score, avremo

$$\begin{aligned} I(\mu) &= \mathbb{E}\left(\left(\frac{f'_\mu(x;\mu)}{f'_\mu(x;\mu)}\right)^2\right) \\ &= \int \frac{(f'_\mu(x-\mu))^2}{f(x-\mu)} dx \\ &= [\text{ con il cambio di variabile } (x-\mu) = z] \\ &= \int \frac{(f'_\mu(z))^2}{f(z)} dz, \end{aligned}$$

che chiaramente non dipende da μ. Di conseguenza, $\pi^J(\mu) \propto k$, con k costante che non dipende da μ.

Per il lettore interessato ai dettagli matematici della costruzione della distribuzione iniziale invariante di Jeffreys, si rimanda alla §C.2.1.

Esempio 5.8 [*Parametro di scala*]. Siano $X_1, \cdots, X_n \overset{\text{iid}}{\sim} f(x;\sigma)$, dove

$$f(x;\sigma) = \frac{1}{\sigma} g\left(\frac{x}{\sigma}\right).$$

La famiglia $\{f(\cdot;\sigma), \sigma \in \mathbb{R}^+\}$ è invariante per cambiamenti di scala, nel senso che la v.a. $Y = aX$ ha una distribuzione ancora di tipo f, qualunque sia a. Si dice allora che σ è un parametro di scala. Dimostriamo qui che, qualunque sia la famiglia di scala, l'informazione di Fisher è pari a k/σ^2, dove k è una costante che non dipende da σ. In questo caso

$$f'_\sigma(x) = -\frac{1}{\sigma^2}\left(g(x/\sigma) + x g'(x/\sigma)\right)$$

e

$$\frac{f'_\sigma(x;\sigma)}{f(x;\sigma)} = -\frac{1}{\sigma}\left(1 + x\frac{g'(x/\sigma)}{g(x/\sigma)}\right).$$

Perciò

$$\begin{aligned} I(\mu) &= \mathbb{E}\left(\left(\frac{f'_\mu(x;\mu)}{f'_\mu(x;\mu)}\right)^2\right) \\ &= \mathbb{E}\left(\frac{1}{\sigma^2}\left(1+\frac{xg'(x/\sigma)}{\sigma g(x/\sigma)}\right)^2\right) \\ &= \frac{1}{\sigma^2}\int_0^\infty \left(1+\frac{xg'(x/\sigma)}{\sigma g(x/\sigma)}\right)^2 \frac{1}{\sigma} g(x/\sigma)dx \\ &= [\text{ con il cambio di variabile } (x/\sigma)=z] \\ &= \frac{1}{\sigma^2}\int_0^\infty \left(1+z\frac{g'(z)}{g(z)}\right)^2 g(z)dz \end{aligned}$$

Poiché l'ultimo integrale scritto non dipende da σ si ha che $I(\sigma) = k/\sigma^2$, e la distribuzione a priori di Jeffreys per qualsiasi modello con parametro di scala pari a

$$\pi^J(\sigma) \propto \frac{1}{\sigma}.$$

♢

Esempio 5.9 [*Parametri di posizione e scala*]. Consideriamo una v.a X con densità del tipo

$$f(x;\mu,\sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right).$$

In questo caso, la trasformazione lineare $Y = aX+b$ ha densit dello stesso tipo di f. Dato un modello statistico con parametri di posizione e scala; allora il determinante della matrice di informazione di Fisher è pari a k/σ^2; i calcoli sono simili a quelli relativi ai due esempi precedenti e vengono omessi. Si conclude allora che la legge non informativa di Jeffreys per un modello di posizione e scala è ancora

$$\pi^J(\mu,\sigma) \propto \frac{1}{\sigma}.$$

♢

Esempio 5.10 [*Dati di sopravvivenza con censura*].

♢

Il metodo di Jeffreys è ancora oggi quello più comunemente utilizzato quando $d = 1$. Tuttavia lo stesso Jeffreys suggerì alcune modifiche alla sua regola generale nel caso di parametro multi-dimensionale. Soprattutto, egli considerava in modo separato eventuali parametri di posizione e scala presenti nel modello.

Esempio 5.11 Consideriamo qui un esempio in cui il numero di parametri da stimare è pari al numero di osservazioni; in casi come questi, l'uso della distribuzione iniziale di Jeffreys può indurre addirittura problemi di consistenza. Siano $X_i \sim \mathrm{N}(\mu_i, 1), \quad i = 1, \cdots, p$ indipendenti tra loro. Si vuole stimare la funzione reale dei parametri

$$\mu = \frac{1}{p}\sum_{i=1}^p \mu_i^2$$

È facile vedere, tenendo conto dell'esempio precedente, che il metodo (5.4) condurrebbe a

$$\pi^J(\mu_1, \cdots, \mu_p) \propto 1$$

Poich la verosimiglianza associata al nostro campione è

$$L(\mu_1, \cdots, \mu_p) \propto \prod_{j=1}^{n} \exp\left(-\frac{1}{2}(x_i - \mu_i)^2\right),$$

la distribuzione finale del vettore $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)$ risulta proporzionale alla verosimiglianza stessa, ovvero

$$\boldsymbol{\mu} \mid \mathbf{x} =\sim N_p(\mathbf{x}, \mathbf{1})$$

e, con semplici calcoli, oppure utilizzando le proprietà delle forme quadratiche di una distribuzione normale (vedi §**??**),

$$p\,\theta \sim \chi^2_p(\mathbf{x}'\mathbf{x}), \tag{5.9}$$

ovvero $p\,\theta$ ha distribuzione χ^2 con p gradi di libertà e parametro di non centralità pari a $\mathbf{x}'\mathbf{x}$. È facile verificare però che la (5.9) non è una *buona* distribuzione a posteriori per θ. Ad esempio,

$$\mathbf{E}\,(\theta|\mathbf{x}) = 1 + \frac{\sum x_i^2}{p},$$

e, utilizzando tale quantità come stimatore in senso classico di θ, si può notare che

$$\lim_{p\to\infty} \left[\mathbf{E}^{\pi}\,(\theta|\mathbf{x}) - \theta\right] = 2,$$

ovvero tale stimatore risulta inconsistente: lo stesso, spiacevole fenomeno, si verifica anche qualora considerassimo la moda o la mediana della legge a posteriori come stima puntuale. ♢

Questo esempio mette in risalto uno dei problemi più frequenti nella ricerca di "buone" distribuzioni non informative. Il metodo di Jeffreys cerca la distribuzione non informativa per l'intero vettore ω. Se in una specifica applicazione il parametro d'interesse è soltanto θ, di dimensione inferiore a ω, come nell'esempio 5.11, questo introduce una "distorsione" nella procedura. Questa osservazione è alla base del metodo delle reference priors proposto da Berger e Bernardo.

5.2.4 Il metodo delle *reference priors*

Una possibile misura del contenuto informativo di una legge di probabilità π è dato dall'*entropia* $\mathcal{E}$, definita, nel caso di misure di probabilità assolutamente continue, come

$$En = -\int_{\boldsymbol{\Omega}} \pi(\omega) \log \pi(\omega) d\omega.$$

Abbiamo inoltre già visto nella §2.6 che una possibile misura della "distanza" di una misura di probabilità p rispetto ad un'altra misura di probabilità q presa come riferimento, è rappresentata dalla divergenza o *numero* di Kullback-Leibler,

$$D_{KL}(p; q) = \int_{\boldsymbol{\Omega}} q(x) \log \frac{q(x)}{p(x)} dx,$$

che, tra le altre proprietà, assume il valore zero se e solo se le due densità praticamente coincidono, ovvero $q(x) = p(x)$ quasi certamente (rispetto a q)[5]. I due concetti di entropia e divergenza di Kullback-Leibler sono alla base della definizione di informazione contenuta in un esperimento, dovuto a [57].

[5] Questo si traduce nel fatto che l'insieme delle x in cui $q(x) \neq p(x)$ ha misura q pari a zero.

Definizione 5.1 [*Informazione di Shannon-Lindley.*] *Dato un esperimento*

$$\mathcal{E}_k = (\mathcal{X}_k, \boldsymbol{\Omega}, \mathcal{P})$$

si definisce Informazione contenuta in $\mathcal{E}_k$*, relativamente ad una distribuzione a priori* π*, la quantità*

$$I_{\mathcal{E}_k}(\pi) = \int_{\mathcal{X}_k} \int_{\boldsymbol{\Omega}} m(\mathbf{x}_k)\pi(\omega|\mathbf{x}_k) \log \frac{\pi(\omega|\mathbf{x}_k)}{\pi(\omega)} d\omega d\mathbf{x}_k = \mathbb{E}^{\theta} \left(\mathrm{D}_{\mathrm{KL}}(\pi(\omega); \pi(\omega|\mathbf{X}_{\mathrm{k}})) \right). \tag{5.10}$$

In altri termini, $I_{\mathcal{E}_k}(\pi)$ rappresenta il valore medio rispetto alla legge marginale m del numero di Kullback-Leibler della distribuzione iniziale rispetto alla distribuzione finale indotta.

È allora ragionevole misurare il contributo informativo di una determinata $\pi(\omega)$ in termini della (5.10) [6]. [14] ha introdotto il metodo delle reference priors. Le due novità essenziali proposte nella determinazione di una distribuzione iniziale di riferimento, che indicheremo con π^r, sono state

(a) determinazione di π^r come argomento massimizzante $\mathcal{I}_{E_k}(\pi)$;
(b) nel caso multiparametrico, distinzione esplicita tra parametro d'interesse e parametri di disturbo.

La formalizzazione rigorosa dell'algoritmo delle reference priors comporta una serie di problemi tecnici non sempre risolubili. Per la natura introduttiva di questo testo ci limiteremo a considerare situazioni di tipo *regolare* in cui, specificatamente, il modello statistico soddisfa i seguenti requisiti

- esiste una statistica sufficiente della stessa dimensione del parametro;
- lo stimatore di massima verosimiglianza del vettore dei parametri ha distribuzione asintotica di tipo gaussiano;
- la distribuzione a posteriori è asintoticamente gaussiana.

Per ulteriori approfondimenti, suggeriamo la lettura di [16] o [11]. Consideriamo per primo il caso univariato, ovvero dove il parametro ω ha supporto $\Omega \in \mathbb{R}$. Abbiamo già visto come la quantità $I_{\mathcal{E}_k}(\pi)$ rappresenti l'incremento medio di informazione che l'esperimento fornisce quando la legge a priori è $\pi(\omega)$. Per $k \to \infty$, $I_{\mathcal{E}_k}(\pi)$ assume allora il significato di *ammontare complessivo d'informazione mancante sul parametro* ω, e la legge a priori che massimizza $I_{\mathcal{E}_\infty}(\pi)$ può a ben diritto essere definita come la "meno informativa".

Il problema si traduce allora in uno di massimizzazione della quantità $I_{\mathcal{E}_\infty}(\pi)$ quando $\pi(\omega)$ varia in una predeterminata classe di distribuzioni a priori Γ[7]. Purtroppo, come illustrato in [7], questo problema spesso non ha una soluzione utile poiché, per molti modelli statistici, risulta $I_{\mathcal{E}_\infty}(\pi) = \infty$ per diverse distribuzioni $\pi(\omega)$ che, per di più, risultano tutt'altro che non informative, almeno dal punto di vista intuitivo. [14] suggerisce allora la seguente variante:

1. Si massimizza $I_{\mathcal{E}_k}(\pi)$ per k fissato, ottenendo la cosiddetta k-reference prior $\pi^r_k(\omega)$
2. Si definisce la reference prior come limite puntuale delle $\pi^r_k(\omega)$,

$$\pi^r(\omega) \lim_{k\to\infty} \frac{\pi^r_k(\omega)}{\pi^r_k(\Omega_0)} \tag{5.11}$$

dove $\Omega_0 \subset \Omega$ è uno specifico compatto.

[6] Quanto affermato è ragionevole ma non obbligatorio: ad esempio, si può notare che nelle espressioni di $\mathcal{I}_{E_k}(\pi)$ ed En la funzione integranda dipende da ω solo attraverso le misura $\pi(\omega)$. Questo fa sì che distribuzioni ottenute per *permutazioni* del supporto della legge $\pi(\omega)$ mantengano lo stesso livello di entropia.

[7] In genere si considera come classe Γ la famiglia di *tutte* le distribuzioni di probabilità

È bene però subito aggiungere alcune precisazioni relative a tale approccio: innanzitutto non è detto che il limite (5.11) esista; inoltre tale limite è puntuale e non assicura una convergenza nella metrica indotta dalla misura d'informazione di Kullback-Leibler: in altri termini, pur esistendo la distribuzione limite, non è assicurato che essa risulti "vicina", nel senso della (5.10), alle $k-$reference prior; infine, ma è il male minore, il limite (5.11) risulta spesso essere una distribuzione impropria.

Nelle condizioni di regolarità sopra illustrate, è tuttavia possibile dare una dimostrazione di tipo euristico del modo in cui l'algoritmo delle reference priors ricerca la π^r.

Per definizione si ha

$$\begin{aligned}\mathcal{I}_{E_k}(\pi) &= \int_{\boldsymbol{\Omega}} \pi(\omega) \left[\int_{\mathcal{X}_k} p(\mathbf{x}_k|\omega) \left(\log \pi(\omega|\mathbf{x}_k) - \log \pi(\omega)\right) \right] d\omega d\mathbf{x}_k \\ &= \int_{\boldsymbol{\Omega}} \pi(\omega) \log \frac{\exp\{\int_{\mathcal{X}_k} p(\mathbf{x}_k|\omega) \log \pi(\omega|\mathbf{x}_k) d\mathbf{x}_k\}}{\pi(\omega)} d\omega,\end{aligned}$$

e tale quantità deve essere massimizzata rispetto a $\pi(\omega)$. Si tratta allora di un problema di calcolo delle variazioni, del tipo

$$\sup_f \int f(x) \log \frac{g(x)}{f(x)} dx$$

la cui trattazione esula dagli obiettivi di questo testo. È però facile vedere che la soluzione (si veda ad esempio **?**]) è del tipo $f(x) \propto g(x)$. Ne segue che

$$\pi^r(\omega) \propto \exp\{\int_{\mathcal{X}_k} p(\mathbf{x}_k|\omega) \log \pi(\omega|\mathbf{x}_k) d\mathbf{x}_k\}; \tag{5.12}$$

la (5.12) fornisce la soluzione solo in maniera implicita; in casi regolari, tuttavia, lo stimatore di massima verosimiglianza, di dimensione scalare (per assunzione) $\hat{\omega}_k$, è tale che

$$\exp\{\int_{\mathcal{X}_k} p(\mathbf{x}_k|\omega) \log \pi(\omega|\mathbf{x}_k) d\mathbf{x}_k\} = \exp\{\int_{\mathbb{R}} p(\hat{\omega}_k|\omega) \log \pi(\omega|\hat{\omega}_k) d\hat{\omega}_k\}. \tag{5.13}$$

Inoltre, per ipotesi, la distribuzione finale di ω è asintoticamente normale, ovvero

$$\pi(\omega|\hat{\omega}_k) \sim N(\omega; \hat{\omega}_k, [kI(\hat{\omega}_k)]^{-1})$$

e la (5.13) risulta approssimatamente uguale a

$$\exp\left\{\int_{\mathbb{R}} p(\hat{\omega}_k|\omega)[\log I(\hat{\omega}_k)^{1/2} - \frac{k}{2} I(\hat{\omega}_k)(\omega - \hat{\omega}_k)^2] d\hat{\omega}_k\right\}$$

Assumendo infine che $\hat{\omega}_k$ risulti uno stimatore consistente di ω si avrà allora

$$\pi^r(\omega) \propto I(\omega)^{1/2},$$

ovvero la distribuzione non informativa proposta da Jeffreys. Va sottolineato come, questa coincidenza sia garantita soltanto nel caso univariato e sotto condizioni di regolarità. Una formalizzazione più rigorosa di questa tecnica condurrebbe a risultati poco rassicuranti. Si può infatti dimostrare che, in certi casi, la distribuzione a priori che massimizza la $I_{\mathcal{E}_k}$ risulta concentrata su un numero finito di punti [7].

Analizziamo ora il caso in cui $\omega = (\theta, \lambda)$ e il solo θ rappresenta il parametro di interesse. Per semplicità di notazione si assume che θ e λ siano parametri unidimensionali: con semplici adattamenti alle seguenti formule è possibile trattare il caso generale.

Sia ora $\omega = (\theta, \lambda)$ e sia θ il solo parametro di interesse. In questo caso l'informazione di Fisher è una matrice quadrata I di dimensione 2: indichiamo con il simbolo I_{ij}, $i, j = 1, 2$, l'elemento (i, j)-esimo di I. In contesti multiparametrici la distribuzione iniziale di Jeffreys assume la forma seguente,

$$\pi^J(\theta, \lambda) \propto \det\left[I(\theta, \lambda)\right]^{1/2}.$$

Il metodo delle reference prior provvede alla determinazione di quella distribuzione $\pi(\theta, \lambda)$ che massimizzi la divergenza d'informazione tra $\pi(\theta|\mathbf{x}_k)$ e $\pi(\theta)$, ovvero

$$I_{\mathcal{E}_k}(\pi(\theta, \lambda)) = \int_{\boldsymbol{\Theta}} \pi(\theta) \left[\int_{\mathcal{X}_k} p(\mathbf{x}_k|\theta) \left(\log \pi(\theta|\mathbf{x}_k) \; - \; \log \pi(\theta)\right)\right] d\theta d\mathbf{x}_k. \tag{5.14}$$

L'equazione, ovviamente, non dipende direttamente da λ, che è stato già eliminato per integrazione e, analogamente a quanto visto prima,

$$\pi_k^r(\theta) \propto \exp\{\int_{\mathcal{X}_k} p(\mathbf{x}_k|\theta) \log \pi(\theta|\mathbf{x}_k) d\mathbf{x}_k\}$$

Tale risultato vale qualunque sia la scelta per $\pi(\lambda|\theta)$. L'algoritmo delle reference priors nella versione proposta da Berger e Bernardo, suggerisce la seguente strategia:

A) Scegli

$$\pi^r(\lambda|\theta) \propto I_{22}(\theta, \lambda)^{1/2},$$

ovvero la legge iniziale di Jeffreys qualora θ fosse noto;

B) Massimizza la (5.14) utilizzando $\pi^r(\lambda|\theta)$.

In questo modo, analogamente al caso univariato, risulta

$$\pi_k^r(\theta) \propto \exp\left\{ \int_{\hat{\theta}} \int_{\hat{\lambda}} p(\hat{\theta}, \hat{\lambda}|\theta) \right\} \times \log N(\theta; \hat{\theta}, S_{11}(\hat{\theta}, \hat{\lambda}) d\hat{\theta} d\hat{\lambda}$$

dove abbiamo posto $S = I^{-1}$, cosicch $S_{11} = I_{22}/\det(I)$. Inoltre, per le assunte condizioni di regolarità, $S(\hat{\theta}, \hat{\lambda}) \to S(\theta, \lambda)$ in probabilità e

$$\begin{aligned}
\pi_k^r(\theta) &\propto \exp\left\{ \int_{\hat{\theta}} \int_{\hat{\lambda}} \int_{\Lambda} p(\hat{\theta}, \hat{\lambda}|\theta, \lambda) \pi^r(\lambda|\theta) \times \log N(\theta; \hat{\theta}, S_{11}(\hat{\theta}, \hat{\lambda}) d\lambda d\hat{\theta} d\hat{\lambda} \right\} \\
&= \exp\left\{ \int_{\hat{\theta}} \int_{\hat{\lambda}} p(\hat{\theta}, \hat{\lambda}|\theta, \lambda) \int_{\Lambda} \pi^r(\lambda|\theta) \times \log N(\theta; \hat{\theta}, S_{11}(\hat{\theta}, \hat{\lambda}) d\lambda d\hat{\theta} d\hat{\lambda} \right\} \\
&\cong \exp\left\{ \frac{1}{2} \int_{\Lambda} \pi^r(\lambda|\theta) \log S_{11}^{-1}(\theta, \lambda) d\lambda \right\}.
\end{aligned}$$

Perciò

$$\pi^r(\theta, \lambda) = \pi(\lambda|\theta) \exp\{\frac{1}{2} \int_{\Lambda} \pi(\lambda|\theta) \log \frac{\det(I)}{I_{22}} d\lambda\}.$$

Fin qui abbiamo trascurato il problema della non integrabilità delle eventuali distribuzioni a priori. All'interno dell'algoritmo delle reference priors tale problema si aggira considerando una successione di insiemi compatti e annidati che "invadono" lo spazio parametrico e sui quali definiamo una successione di reference priors. Diamo qui di seguito la versione più generale dell'algoritmo, quando il parametro è suddiviso in due blocchi: l'estensione al caso con più blocchi è immediata, Si veda ad esempio [15]

Algoritmo delle reference prior per il caso bidimensionale

1 Si parte dal nucleo della distribuzione non informativa per λ, per θ fissato, ovvero

$$\pi^*(\lambda|\theta) \propto \sqrt{I_{22}(\theta,\lambda)}$$

2 Normalizzazione di $\pi^*(\lambda|\theta)$:

- se $\pi^*(\lambda|\theta)$ è integrabile (ovvero la sua versione normalizzata è propria), poni

  $$\pi(\lambda|\theta) = \pi^*(\lambda|\theta)k(\theta)$$

  con

  $$k(\theta)^{-1} = \int_{\Lambda} \pi^*(\lambda|\theta)d\lambda;$$

- se $\pi^*(\lambda|\theta)$ non è integrabile (cioè la distribuzione è impropria), si determina una successione non decrescente di sezioni di sottoinsiemi di Λ convergenti all'intero Λ

  $$\Lambda_1(\theta), \Lambda_2(\theta), \cdots, \Lambda_m(\theta), \cdots \to \Lambda,$$

  la cui forma può dipendere da θ, e sui quali è possibile definire, per ogni $m \in \mathbb{N}$,

  $$\pi_m(\lambda|\theta) = \pi^*(\lambda|\theta)\, k_m(\theta)$$

  con $k_m(\theta)^{-1} = \int_{\Lambda_m(\theta)} \pi^*(\lambda|\theta)d\lambda$.

3 Si determina la distribuzione marginale di θ, condizionatamente a Λ_m,

$$\pi_m(\theta) \propto \exp\left\{\frac{1}{2}\int_{\Lambda_m(\theta)} \pi_m(\lambda|\theta)\log\frac{\det I(\theta,\lambda)}{I_{22}(\theta,\lambda)}\right\}d\lambda.$$

4 Si pone

$$\pi^r(\theta,\lambda) = \lim_{m\to\infty}\frac{k_m(\theta)\pi_m(\theta)}{k_m(\theta_0)\pi_m(\theta_0)}\frac{\pi_m(\lambda|\theta)}{\pi(\lambda|\theta_0)}$$

Esempio 5.6 (continua). Abbiamo già ottenuto, in questo contesto, la legge iniziale di Jeffreys. Essendo ω un parametro scalare, la reference prior qui coincide con la 5.5. Va notato come la distribuzione a priori suggerita per questo problema non sia la distribuzione uniforme, quella intuitivamente più adattata a rappresentare il nostro presunto stato di ignoranza relativamente al parametro θ; inoltre, la distribuzione (5.5) è una vera e propria distribuzione di probabilità nel senso che il suo integrale sul supporto $[0,1]$ vale uno. Questa situazione è atipica; in genere le distribuzioni iniziali ottenute mediante il metodo delle refernce priors risultano improprie.

Esempio 5.12 [*Modello Trinomiale*]. Riconsideriamo ora l'esempio precedente ma suddividiamo i risultati possibili non più in due categorie bensì in tre, ovvero, ad esempio, X_i assume i valori $-1, 0$ e 1 con rispettive probabilità pari a ω_1, ω_2 e $1-\omega_1-\omega_2$. Supponiamo di essere interessati, come nel caso bernoulliano, al parametro $\theta = \omega_1$; ora tuttavia, nel modello è presente anche il parametro di disturbo $\lambda = \omega_2$. Per calcolare la matrice di informazione di Fisher, scriviamo la verosimiglianza relativa ad una singola osservazione come

$$L(\theta,\lambda) = p(x;\theta) = \theta\delta_{(1)}(x) + \lambda\delta_{(0)}(x) + (1-\theta-\lambda)\delta_{(-1)}(x);$$

dove $\delta_{(u)}(v)$ rappresenta la funzione Delta di Dirac che vale 1 solo quando $u = v$ e altrimenti vale 0, ovvero

$$\delta_{(u)}(v) = \begin{cases} 1 & u = v \\ 0 & u \neq v \end{cases}.$$

Allora, la derivata della log-verosimiglianza è

$$\frac{\partial}{\partial\theta}\ell(\theta,\lambda)=\frac{\delta_{(1)}(x)-\delta_{(-1)}(x)}{\theta\delta_{(1)}(x)+\lambda\delta_{(0)}(x)+(1-\theta-\lambda)\delta_{(-1)}(x)}$$

e

$$\frac{\partial^2}{\partial\theta^2}\ell(\theta,\lambda)=\frac{\big(\delta_{(1)}(x)-\delta_{(-1)}(x)\big)^2}{\big(\theta\delta_{(1)}(x)+\lambda\delta_{(0)}(x)+(1-\theta-\lambda)\delta_{(-1)}(x)\big)^2},$$

da cui si ottiene, calcolando il valor medio rispetto alla $p(x;\theta,\lambda)$,

$$I_{\theta\theta}=\frac{1-\lambda}{\theta(1-\theta-\lambda)}.$$

Gli altri elementi della matrice si ottengono in modo analogo:

$$I(\theta,\lambda)=\frac{1}{1-\theta-\lambda}\times\begin{pmatrix}\dfrac{1-\lambda}{\theta} & 1\\ 1 & \dfrac{1-\theta}{\lambda}\end{pmatrix}$$

La distribuzione a priori di Jeffreys è dunque

$$\pi^J(\theta,\lambda)\propto\frac{1}{\sqrt{\theta\lambda(1-\lambda-\theta)}}$$

Per quanto concerne il calcolo della reference prior, applicando l'algoritmo precedente si ottiene

1. $\pi^*(\lambda|\theta)\propto\sqrt{I_{22}(\theta,\lambda)}=\frac{1}{\sqrt{\lambda(1-\lambda-\theta)}}$
2. $\pi^(\lambda|\theta)=k(\theta)\frac{1}{\sqrt{\lambda(1-\lambda-\theta)}}\ I_{[0,1-\theta]}(\lambda)$
3. $\pi(\theta)=\exp\left\{\frac{1}{2}\int_{\Lambda(\theta)}k(\theta)\frac{1}{\sqrt{\lambda(1-\lambda-\theta)}}\log\frac{1}{\theta(1-\theta)}d\lambda\right\}=\frac{1}{\sqrt{\theta(1-\theta)}}$
4. $\pi^r(\theta,\lambda)\propto\frac{1}{\sqrt{\theta\lambda(1-\theta)(1-\theta-\lambda)}}$

Analizziamo ora comparativamente le due distribuzioni a priori $\pi^r(\theta,\lambda)$ e $\pi^J(\theta,\lambda)$ ottenute coi due metodi. La differenza più significativa si apprezza considerando le due distribuzioni marginali a priori per θ.

- $\pi^r(\theta)=\frac{1}{\pi}\frac{1}{\sqrt{\theta(1-\theta)}}$
  ovvero la reference prior è una distribuzione Beta$(1/2,1/2)$, la cui media è $E^r(\theta)=1/2$
- $\pi^J(\theta)=\frac{1}{2}\theta^{-1/2}$
  ovvero la distribuzione marginale di θ, nel senso di Jeffreys, è una distribuzione Beta$(1/2,1)$, con densità decrescente, la cui media è $E^J(\theta)=1/3$.

Questa differenza risulta ancor più accentuata nel caso generale in cui la variabile X è politomica con h possibili valori. Si può infatti dimostrare che, in questo caso, denotando i parametri della distribuzione di X con $\theta_{(1)},\theta_{(2)},\cdots,\theta_{(h)}$, risulta

$$E^r(\theta_{(i)})=\frac{1}{2^i}\quad i=1,\cdots,h;\qquad E^r(\theta_{(i)})=\frac{1}{h}\ i=1,\cdots,h.$$

Dunque, la distribuzione di Jeffreys, essendo non informativa per l'intero vettore ω, mantiene la sua natura "scambiabile" e tutte le componenti del vettore hanno la stessa distribuzione marginale. Al contrario la distribuzione di riferimento "ordina" le componenti di ω secondo la loro rilevanza inferenziale nel problema specifico, inducendo così un ordinamento anche tra le distribuzioni

marginali; ad esempio, la distribuzione marginale di riferimento per θ nel caso tridimensionale coincide con la distribuzione iniziale di Jeffreys per lo stesso parametro nel caso binomiale (Esempio 5.6. Questo può essere più o meno accettabile a seconda dei casi: quel che è certo è che, in casi multiparametrici, la ricerca di una distribuzione non informativa per un parametro di interesse o per l'intero vettore dei parametri, conduce a soluzioni differenti. Il messaggio dell'esempio è che non è possibile, allo stesso tempo, pretendere che le componenti del vettore dei parametri siano a priori tutte valutate sullo stesso piano[8] e che la distribuzione a priori risulti coerente per marginalizzazione. Per certi versi, questo risultato riconduce a quanto detto a proposito del Principio di Ragione Insufficiente. ♢

Esempio 5.13 *rapporto di normali*

♢

Esempio 5.14 *caso non iid* AR(1)?

♢

5.3 La sensibilità delle conclusioni rispetto alla distribuzione a priori

5.3.1 Cenni al problema della robustezza

5.3.2 Il ruolo della dimensione campionaria

5.4 Esercizi

[8] esiste un modo formale di spiegare questo concetto, la scambiabilità di cui ci occuperemo nella §6.4.

6

Procedure inferenziali bayesiane

Nella teoria classica dell'inferenza le procedure vengono generalmente suddivise in quattro categorie:

- problemi di stima puntuale
- problemi di stima per intervallo
- problemi di verifica delle ipotesi
- problemi di previsione.

Se questa classificazione trova piena giustificazione in ambito classico, dove gli strumenti con cui si affrontano i suddetti problemi sono spesso molto diversi, in un contesto bayesiano l'esigenza di distinguere tra diverse problematiche inferenziali è avvertita in maniera inferiore. Come già osservato nel Capitolo 3, il risultato conclusivo di un'analisi bayesiana è la determinazione della distribuzione a posteriori dei parametri di interesse; le successive elaborazioni di questa riguardano perlopiù la sfera della statistica descrittiva, oppure, introducendo strumenti logici e matematici ulteriori che esulano dagli scopi di questo scritto, della teoria delle decisioni statistiche, alle quali accenneremo brevemente nel §6.6: i lettori interessati possono consultare [68] e, per chi non teme l'inglese, [10] o [76].

Nel corso di questo capitolo si mostrerà come le procedure bayesiane per la determinazione di stime puntuali e per intervallo difficilmente producono risultati operativamente distanti dai risultati classici; in particolare, in molti esempi di elevata importanza applicativa, l'uso di distribuzioni a priori "oggettive" condurrà esattamente alle stesse conclusioni di un' analisi classica. Diversa è la situazione per quanto concerne i problemi di verifica di ipotesi. Qui le diverse impostazioni possono fornire risultati discrepanti, in alcuni casi, anche molto elementari, addirittura contrastanti. I problemi di previsione la situazione può definirsi intermedia. Da un lato essi rappresentano il contesto dove maggiormente può essere apprezzata la necessità dell'impostazione bayesiana. Esistono situazioni dove una distribuzione predittiva è ottenibile solo attraverso l'uso del teorema di Bayes. Tuttavia, in situazioni regolari, dove i metodi classici forniscono una risposta approssimata, essa non è in genere molto distante da quella bayesiana.

6.1 Stima puntuale

Un esperimento statistico consiste nell'osservare un punto (il campione) $\mathbf{x}$ nello spazio di tutti i possibili campioni $\mathcal{X}$. Le leggi di probabilità che possono aver generato tale campione sono raccolte

in una famiglia, già introdotta nella § 2.2,

$$\mathcal{P} = \{p(\mathbf{x} \mid \theta), \quad \theta \in \boldsymbol{\Omega}\}.$$

Noi assumiamo che una sola di queste leggi abbia agito ma non sappiamo quale: scopo di una procedura di stima puntuale è quello di selezionare un valore $\tilde{\theta} \in \boldsymbol{\Omega}$ quale indicatore della "vera" legge di probabilità

In ambito bayesiano, una volta determinata la distribuzione a posteriori per θ, la sintesi più naturale è fornita da un qualche indicatore sintetico di posizione. Per semplicità di calcolo, e per il suo ruolo anche al di fuori della statistica, il valore atteso o media, che indicheremo con il simbolo $\mathbf{E}(\theta \mid \mathbf{x})$, si è imposto come il più popolare.

Esempio 6.1 [*Distribuzione normale.*]
Nella § 4.3.1 abbiamo già visto come, in presenza di un vettore di osservazioni $\mathbf{x}$ proveniente da una distribuzione $N(\mu, \sigma^2)$ con varianza nota, e quando la distribuzione iniziale per μ è anch'essa di tipo gaussiano $N(\mu_0, \tau^2)$, la distribuzione finale di μ è a sua volta gaussiana $N(\mu^*, \tau^{2*})$, dove i valori μ^* e τ^{2*} sono forniti dalle (4.6). In questo caso la stima puntuale per il parametro μ è senz'altro il valore μ^* che rappresenta, in questo specifico contesto, moda, media e mediana della distribuzione finale per μ. La coincidenza delle tre sintesi è però dovuta alle proprietà di simmetria e unimodalità della distribuzione normale; non vi è motivo di credere che essa si estenda anche ad altre distribuzioni.

♢

Esempio 6.2 [*Distribuzione esponenziale.*]
Si vuole acquisire informazioni sulla durata media di servizi allo sportello in una banca. Possiamo considerare i vari tempi di servizio indipendenti tra loro e con distribuzione di tipo esponenziale di parametro θ. Si osserva un campione di n durate e sia X_i, $i = 1, \cdots, n$, la variabile aleatoria che indica la durata del generico servizio. La funzione di verosimiglianza associata all'osservazione dei dati $\mathbf{x} = (x_1, \cdots, x_n)$ è

$$L(\theta; \mathbf{x}) \propto \prod_{i=1}^{n} \left(\theta \exp^{-\theta x_i}\right) = \theta^n e^{-\theta \sum_i x_i};$$

Se le informazioni a priori su θ possono essere espresse attraverso una distribuzione Gamma(δ, γ), allora in virtù del coniugio tra funzione di verosimiglianza esponenziale e distribuzione iniziale gamma si ottiene una distribuzione finale ancora di tipo Gamma$(\delta + n, \gamma + \sum_i x_i)$. Per risultati noti intorno alla distribuzione Gamma, segue che la media a posteriori di θ è pari a

$$E(\theta \mid \mathbf{x}) = \frac{\delta + n}{\gamma + \sum_i x_i}.$$

Nel caso di un'analisi non informativa si utilizzerebbe come legge a priori per il parametro θ, la distribuzione iniziale di Jeffreys che, nel caso esponenziale - caso particolare di un parametro di scala (Esempio 5.8 è pari a

$$\pi^J(\theta) \propto \frac{1}{\theta},$$

corrispondente ad una distribuzione iniziale coniugata con $\delta = \gamma = 0$. In tal caso la distribuzione finale del parametro θ è dunque di tipo Gamma$(n, \sum_i x_i)$ e la media a posteriori vale $1/\bar{x}$, coincidente con la stima di massima verosimiglianza.

♢

Ma l'uso del valore atteso a posteriori non è sempre la soluzione più semplice o più ragionevole.

A volte il calcolo di $\mathbb{E}(\theta \mid \mathbf{x})$ è complicato; altre volte esso, addirittura, non esiste. In questi casi è auspicabile ricorrere a stime puntuali alternative come, ad esempio, la mediana o la moda della distribuzione a posteriori. La scelta dell'indicatore sintetico da utilizzare è un problema che può essere affrontato in modo formale attraverso la teoria delle decisioni statistiche; tale disciplina fornisce uno schema generale di riferimento al cui interno è possibile rileggere gran parte dell'inferenza bayesiana; in questo testo non percorreremo questa strada e soltanto alcuni accenni all'impostazione decisionale verranno forniti nella §6.6.

Occorre tuttavia sottolineare una differenza nell'uso bayesiano di indicatori sintetici come la media e la mediana dall'uso che se ne fa nella statistica classica, nell'ambito della teoria degli stimatori. Per chiarire le idee consideriamo il semplice esempio seguente: supponiamo che i dati siano replicazioni indipendenti e somiglianti di una variabile aleatoria con densità del tipo $f(x-\theta)$, ovvero θ è un parametro di posizione. La teoria classica dell'inferenza, in questo caso, distingue l'uso dello stimatore $T_1(\mathbf{x}) = \bar{X}$, cioè la media aritmetica campionaria, da, ad esempio, lo stimatore $T_2(\mathbf{x}) = Me(\mathbf{x})$, la mediana campionaria, in termini di robustezza dei due stimatori rispetto a dati anomali. In ambito bayesiano, l'inferenza è basata sulla distribuzione finale di θ, e la scelta di una sintesi di tale distribuzione risulta una decisione meno critica. Al limite, se la distribuzione finale è approssimativamente normale, media, moda e mediana forniscono lo stesso risultato.

Queste considerazioni non devono comunque far pensare che l'analisi bayesiana sia esente dal problema dell'influenza di dati anomali sulle conclusioni inferenziali o che le conclusioni bayesiane siano più robuste: l'esempio precedente deve servire semplicemente a riflettere sul diverso significato che le sintesi dell'informazione hanno in ambito classico e bayesiano; nel primo caso la tecnica di stima viene concepita in modo da limitare i rischi di un'eventuale assenza di robustezza; in ambito bayesiano tali precauzioni vanno inserite all'interno del processo di "elicitazione" della legge iniziale: una volta ottenuta la legge a posteriori essa viene "riassunta", mediante tecniche di statistica descrittiva.

Consideriamo il seguente semplice esempio, per certi versi paradigmatico.

Esempio 6.3 [*Osservazioni influenti*]

Abbiamo già visto come in presenza di osservazioni indipendenti e identicamente distribuite $X_1, \cdots, X_n$ con legge $N(\theta, \sigma^2)$ con σ^2 noto, e qualora si adotti per θ una distribuzione iniziale coniugata di tipo $N(\mu, \tau^2)$, la distribuzione finale di θ dipende dalle osservazioni solo attraverso la media campionaria $\bar{x}$, cosicché[1]

$$[\theta \mid \mathbf{x}] \stackrel{d}{=} [\theta \mid \bar{x}] \sim N\left(\frac{n\bar{x}\tau^2 + \mu\sigma^2}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right).$$

Scriviamo la media campionaria come combinazione lineare della media delle prime $n-1$ osservazioni, denotata $\bar{x}_{(n-1)}$, e dell'ultima osservazione x_n, ovvero

$$\bar{x} = \frac{n-1}{n}\bar{x}_{(n-1)} + \frac{1}{n}x_n.$$

È possibile allora studiare il comportamento della distribuzione finale di θ in funzione di x_n. Poiché la distribuzione finale, almeno nel caso in esame in cui σ^2 è noto, dipende da x_n solo attraverso la propria media, è sufficiente limitarci ad analizzare il comportamento di tale media come funzione di x_n. Si ha allora

[1] il simbolo $\stackrel{d}{=}$ indica uguaglianza in distribuzione.

$$E(\theta \mid \mathbf{x}) = E\left(\theta \mid \bar{x}_{(n-1)}, x_n\right) = \frac{x_n}{\sigma^2 + n\tau^2} + \frac{(n-1)\bar{x}_{(n-1)} + n\tau^2 + \mu\sigma^2}{\sigma^2 + n\tau^2}.$$

Ceteris paribus, all'aumentare di x_n la media a posteriori, così come la mediana e la moda, cresce indefinitamente, a testimonianza della forte sensibilità della media a posteriori rispetto a dati anomali. In un'ottica bayesiana questa mancanza di robustezza va letta come un limite del modello gaussiano-gaussiano, e non come un problema della media come indicatore statistico. Per produrre procedure più robuste all'interno del paradigma bayesiano, in questo contesto, è necessario utilizzare distribuzioni a priori più diffuse, con code più consistenti, ad esempio una densità t di Student con pochi gradi di libertà: questo impedirebbe tuttavia la possibilitàdi ottenere una risposta in forma esplicita e occorrerìricorrere a metodi numerici, che rimandiamo al prossimo capitolo. ♢

Esempio 6.4 [*Distribuzione normale asimmetrica.*]

Abbiamo già introdotto la famiglia delle densità normali asimmetriche nell'Esempio 2.12; ricordiamo qui che, nella sua formulazione più semplice, si dice che $X \sim SN(\theta)$ se la funzione di densità vale

$$p(x;\theta) = 2\,\varphi(x)\Phi(\theta x),$$

dove $\varphi(\cdot)$ e $\Phi(\cdot)$ rappresentano, rispettivamente, la funzione di densità e quella di ripartizione di una v.a. normale standard. Supponiamo di osservare un campione di n osservazioni i.i.d. con distribuzione $SN(\theta)$; la funzione di verosimiglianza associata all'esperimento è

$$L(\theta;\mathbf{x}) \propto \prod_{i=1}^{n} \Phi(\theta x_i);$$

essa è dunque proporzionale al prodotto di n funzioni monotone di θ, crescenti o decrescenti a seconda del segno delle x_i (possiamo trascurare, perché avviene con probabilità nulla, il caso in cui una o più delle x_i valga esattamente zero!). Se ne deduce che, qualora il campione contenga tutte osservazioni con lo stesso segno, la funzione di verosimiglianza risulta monotona e la stima di massima verosimiglianza vale $+\infty$ o $-\infty$, a seconda del segno delle x_i. Un' analisi bayesiana non informativa di questo problema prevede il calcolo della distribuzione iniziale invariante di Jeffreys, che in questo contesto, non ha una forma analitica esplicita e vale [59]

$$\pi^J(\theta) \propto I^{\frac{1}{2}}(\theta), \tag{6.1}$$

dove

$$I(\theta) = \mathbf{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(z;\theta)\right)^2 = \int 2z^2\varphi(z)\frac{\varphi^2(\theta z)}{\Phi(\theta z)}dz.$$

Si può verificare che questa legge iniziale è integrabile e quindi propria; inoltre la (6.1) ha code che tendono a zero alla stessa velocità di $(1+\theta)^{-3/2}$ per cui i suoi momenti non esistono: ne segue allora che, nel caso in cui tutte le osservazioni abbiano lo stesso segno e la funzione di verosimiglianza risulti di conseguenza monotona, una delle code della distribuzione finale $\pi^J(\theta|\mathbf{x})$ tende a zero con la stessa velocità della distribuzione iniziale; pertanto il valore atteso a posteriori, in questo caso, non esisterebbe. ♢

In casi come questi è certamente consigliabile utilizzare la mediana della legge finale, la cui esistenza, anche se non l'unicità, è sempre garantita. **Esempio 6.5** [*Binomiale*]

♢

Esempio 6.6 [*senza distribuzione standard*] ♢

6.2 Stima per intervallo

Poiché l'operazione di stima di un parametro incognito mediante un singolo valore $\tilde{\theta}$ è destinata a fornire un risultato sbagliato con pratica certezza, appare molto più ragionevole sintetizzare il risultato dell'inferenza attraverso la produzione di un insieme di valori C, sottoinsieme di $\boldsymbol{\Omega}$ che, con una certa probabilità a posteriori, contiene il vero valore di θ.

Partendo dalla distribuzione a posteriori $\pi(\theta \mid \mathbf{x})$, il modo più ragionevole di costruire un insieme di valori di θ di livello $1-\alpha$, ovvero che contenga il valore vero di θ con probabilità non inferiore a $1-\alpha$, è quello di determinare un insieme *HPD*, acronimo inglese che sta per *Highest Posterior Density*; si tratta in pratica di inserire nell'insieme C tutti i valori di θ la cui densità a posteriori risulta più elevata, fino a raggiungere una probabilità complessiva non inferiore al livello prescelto $1-\alpha$. Illustriamo la procedura con due esempi canonici per poi ritornare alla procedura generale.

Il caso normale-normale.

Nel modello Normale-Normale, già analizzato in (6.1), quando il parametro da stimare è la media μ, si ottiene una distribuzione finale di tipo $N(\mu^*, \tau^{2*})$, e quindi unimodale e simmetrica. Dunque, qualunque sia il livello di credibilità $1-\alpha$, l'insieme HPD risulterà un intervallo simmetrico intorno alla media finale μ^*: si ottiene così un intervallo del tipo $(\mu^* - h, \mu^* + h)$, dove $\mu^* + h$ rappresenta il quantile d'ordine $1-\alpha/2$ della distribuzione finale. Per la determinazione esplicita di h basta notare che, a posteriori, $(\mu - \mu^a st)/\tau^a st \sim \mathrm{N}(0,1)$. Poiché $\Pr(\mu \leq \mu^a st + h) = 1 - \alpha/2$,

$$\Pr\left(\frac{\mu - \mu^*}{\tau^*} \leq \frac{h}{\tau^*}\right) = 1 - \frac{\alpha}{2},$$

risulterà $h = \tau^* z^2_{1-\alpha/2}$, dove $z^2_{1-\alpha/2}$ rappresenta il quantile d'ordine $1-\alpha/2$ della distribuzione normale standardizzata. Ne consegue che l'insieme di credibilità bayesiano per la media di una distribuzione normale è dato dall'intervallo

$$(\mu^* - \tau^* z_{1-\frac{\alpha}{2}}, \mu^* + \tau^* z_{1-\frac{\alpha}{2}}),$$

dove μ^* e τ^{2*} rappresentano la media e la varianza della distribuzione finale. Giova ricordare che, nel caso si utilizzi una distribuzione iniziale non informativa (nella fattispecie $\pi(\mu) \propto 1$), risulterà $\mu^* = \bar{x}$ e $\tau^{2*} = \sigma_0^2/n$ e l'insieme bayesiano coincide numericamente con l'intervallo di confidenza ottenuto attraverso l'approccio frequentista.

Esempio 6.7 [*Punteggio ad un test*]

Un'agenzia per l'assunzione sottopone dei candidati ad un test attitudinale. La traduzione del test in un punteggio numerico è soggetta ad errori di misurazione cosicché possiamo supporre che il risultato di un certo candidato sia una v.a. $X \sim N(\mu, 100)$, dove μ rappresenta l'effettivo valore del candidato. Da indagini passate si può considerare che, nella popolazione dei potenziali candidati si possa assumere che $\mu \sim N(100, 225)$. Consideriamo un candidato che ottiene un punteggio $x = 115$. Si vuole determinare un insieme di credibilità per il valore μ. Calcoli già visti ci consentono di affermare che la distribuzione finale per μ è ancora $N(\tilde{\mu}, \tilde{\tau}^2)$, dove

$$\tilde{\mu} = \frac{100100 + 115225}{100 + 225} = 110.38, \quad \tilde{\tau}^2 = \frac{100225}{100 + 225} = 69.23.$$

L'intervallo di credibilità di livello 0.95 è allora

$$(\tilde{\mu} \pm \tilde{\tau} z_{0.975}) = (110.38 \pm \sqrt{69.23}\, 1.96) = (94.07, 126.68)$$

Come già osservato, l'intervallo di confidenza frequentista in questo esempio si ottiene ponendo $\tilde{\mu} = \bar{x} = x$ (notare che $n = 1$) e $\tilde{\tau}^2 = \sigma_0^2 = 100$. Il risultato numerico per il nostro esempio sarebbe dunque

$$(95.40, 134.60).$$

♢

La determinazione di un insieme HPD non è sempre possibile per via analitica. Può accadere ad esempio che la distribuzione finale risulti multimodale cosicché l'insieme HPD potrebbe essere addirittura formato da intervalli disgiunti di valori. Se la distribuzione finale è monotona decrescente, l'insieme HPD di livello $1 - \alpha$ si ottiene determinando il quantile di ordine $1 - \alpha/2$ della distribuzione finale. Nel caso, piuttosto generale, in cui la distribuzione finale $\pi(\theta \mid \mathbf{x})$ è unimodale, occorre ricorrere a metodi numerici. In grandi linee la procedura consiste in tre passi

1. Per un valore k fissato [k deve essere minore del massimo valore di $\pi(\theta \mid \mathbf{x})$)], determinare le due radici $\theta_1(k)$ e $\theta_2(k)$ dell'equazione $\pi(\theta \mid \mathbf{x}) = k$;
2. Calcolare $C(k) = \Pr(\theta \in (\theta_1(k), \theta_2(k)))$
3. Determinare il valore $\tilde{k}$ che risolve l'equazione $C(k) = 1 - \alpha$;

l'intervallo bayesiano HPD sarà allora

$$\left(\theta_1(\tilde{k}), \theta_2(\tilde{k})\right).$$

Esempio 6.8 [*Parametro di una legge esponenziale*]
Per un lotto di componenti elettroniche occorre determinare il tempo medio di vita. Si può assumere che ciascuna componente abbia una durata di vita esponenziale di parametro θ (e dunque media $1/\theta$). Da informazioni passate possiamo affermare che $\theta \sim \text{Gamma}(2, 20)$. Vengono osservate cinque lampadine e il vettore delle osservazioni è $\mathbf{x} = (16, 12, 14, 10, 12)$, per una media campionaria $\bar{x} = 12.8$. Dalla Tabella (5.2) si deduce che la distribuzione finale per θ è ancora di tipo Gamma$(\tilde{\delta}, \tilde{\lambda})$, dove $\tilde{\delta} = 2 + 5 = 7$ e $\tilde{\lambda} = 20 + 5 \times 12.8 = 84$.

Il seguente codice in R fornisce la soluzione al nostro problema; basta inserire, come input il valore α e i parametri della distribuzione finale. Il codice è facilmente adattabile ad altre situazioni.

Codice **R** per la determinazione di un'intervallo di credibilità
INTRODURRE CORREZIONE PER IL CASO DELTA ¡ 1

```
cred.gamma<-function(alpha=0.05, lambda, delta){
soglia<-1-alpha
if(delta>1){
argma<-(delta-1)/lambda
mode<-dgamma(argma, delta, rate=lambda)
pp<-function(t,k){dgamma(t, delta, rate=lambda)-k}
kk<-seq(0,mode,length=500)
cont<-0
```

```
risp<- -1
while(risp<0){
cont<-cont+1
t1<-uniroot(pp, c(0,argma), tol=.0001, k=kk[cont])$root
t2<-uniroot(pp, c(argma, 100), tol=.0001, k=kk[cont])$root
valore<-pgamma(t2, delta, rate=lambda)- pgamma(t1, delta, rate=lambda)
if (valore < soglia) {risp<-1}
}}
list( "estremo inferiore" =t1, "estremo superiore" =t2, "prob."  =valore)
}
```

L'intervallo di credibilità di livello 0.95 per θ così calcolato vale

$$(0.028, 0.146).$$

Va notato che, essendo la distribuzione finale asimmetrica, l'intervallo non risulta centrato né sulla media a posteriori (pari in questo caso a $\tilde{\delta}/\tilde{\lambda} = 0.083$) né sulla moda.

Per questo esempio specifico è possibile una soluzione alternativa che non conduce ad un intervallo HPD ma ad una sua buona approssimazione: è noto che, se una v.a. $Z \sim Gamma(\delta, \gamma)$, allora la sua trasformazione $Y = 2\delta X$ ha distribuzione $\chi^2_{2\gamma}$: nel nostro caso questo implica che $Y = 168\,\theta \sim \chi^2_{14}$. È sufficiente allora determinare i quantili di ordine 0.025 e 0.975 (rispettivamente 5.62 e 26.11) di tale distribuzione per ottenere un intervallo di credibilità di livello 0.95 costruito eliminando le code della distribuzione finale. Ne segue che $\Pr(5.62 \leq 168\,\theta \leq 26.11) = 0.95$, cosicché

$$(0.033, 0.155)$$

è un intervallo di credibilità di livello 0.95 per θ; non essendo un intervallo HPD l'intervallo risulta più ampio di quello ottenuto in precedenza. ♢

Un'approssimazione basata sulla simulazione a posteriori

Può accadere che la distribuzione finale sia di una forma nota o comunque di un tipo da cui siamo in grado di generare valori pseudo-casuali; di queste metodologie si discuterà ampiamente nel Capitolo 7 ma conviene qui darne una prima esemplificazione. In questo caso esiste un modo alternativo per costruire un'*approssimazione* dell'insieme HPD, valida qualora la distribuzione finale sia di tipo unimodale. La tecnica consiste nel generare un numero G molto grande di valori dalla distribuzione finale, diciamo $(y_1, \ldots, y_G)$ e approssimare l'istogramma relativo a tale campione mediante una stima non parametrica di densità (in **R** questo è possibile mediante la funzione `density`). A questo punto ogni punto generato y_g, $g = 1, \ldots, G$, ha associato un valore approssimato della sua densità, ed è sufficiente allora ordinare tali punti in base a questa densità e inserire nell'HPD di livello $1-\alpha$, i punti di massima densità a posteriori, in modo che la densità complessiva dei punti selezionati raggiunga un livello almeno pari a $G\,(1-\alpha)$.
Una versione utilizzabile con **R** può essere la seguente funzione[2]:

Codice **R** per la costruzione approssimata di intervalli di credibilità

[2] il codice riportato non è dell'autore, ma è stato "scaricato" dalla rete, da un sito di cui non è rimasta traccia...

```
hpd<-function(x,p){
#genera un insieme hpd con probabilit\‘a finale p, basato
#su un campione a posteriori x dalla finale
dx<-density(x)
md<-dx$x[dx$y==max(dx$y)]
px<-dx$y/sum(dx$y)
pxs<--sort(-px)
ct<-min(pxs[cumsum(pxs)< p])
list(hpdr=range(dx$x[px>=ct]),mode=md) }
```

Esempio 6.8(continua). Utilizziamo il precedente codice per ottenere una un intervallo HPD per θ. Si può generare un campione di dimensione $G = 10000$ dalla legge a posteriori attraverso il comando

```
xx<-rgamma(10000,shape=7, rate=84);
```

Usiamo poi il comando

```
hpd(xx,p=0.95).
```

L'intervallo così ottenuto vale $(0.0267; 0.1460)$, praticamente identico a quello ottenuto in modo numerico.[3]

Le metodologie qui riportate si riferiscono a problemi unidimensionali. Nel caso generale in cui lo spazio parametrico $\boldsymbol{\Omega}$ è un sottoinsieme di $\mathbb{R}^k$, non emergono nuovi problemi teorici ma l'implementazione delle tecniche sopra descritte può diventare complessa. In questi casi, è ragionevole costruire l'intervallo di credibilità in modo approssimato ricorrendo a metodi numerici (vedi Capitolo 7). In questo caso sarà ovviamente più semplice rinunciare all'intervallo HPD per limitarci alla determinazione di un intervallo *equal tailed*. Alternativamente si può ricorrere ad un'approssimazione basata sulla distribuzione normale multivariata.

INVARIANZA RISPETTO A RIPARAMETRIZZAZIONI. (BERGER 1985 BOOK)

6.3 Verifica di ipotesi

In un problema di verifica delle ipotesi, si confrontano due ipotesi alternative:

$$H_0 : \quad \theta \in \boldsymbol{\Omega}_0, \text{detta } \textit{ipotesi nulla}$$

e

$$H_1 : \quad \theta \in \boldsymbol{\Omega}_1, \text{detta } \textit{ipotesi alternativa};$$

Si ha inoltre che

$$\boldsymbol{\Omega}_0 \cup \boldsymbol{\Omega}_1 = \boldsymbol{\Omega}$$

[3] Ovviamente il risultato qui riportato è frutto della generazione di valori pseudo casuali e il lettore che riprodurrà i comandi suddetti otterrà, quasi certamente, un risultato sempre diverso, ma sempre molto simile a quello qui riportato.

e, ovviamente,

$$\boldsymbol{\Omega}_0 \cap \boldsymbol{\Omega}_1 = \emptyset.$$

In pratica, occorre stabilire se il vero valore del parametro θ, è un elemento del sottoinsieme $\boldsymbol{\Omega}_0$ oppure si trova in $\boldsymbol{\Omega}_1$. In ambito classico questo schema formale può essere trattato in almeno due stili differenti che, almeno nelle intenzioni di chi propose tali approcci (Fisher da un lato, Neyman e Pearson dall'altro), dovevano corrispondere a situazioni differenti. Schematizzando un poco, possiamo classificare i due approcci nel modo seguente:

1. test di significatività: si veda ad esempio, [29];
2. impostazione decisionale: si veda, ad esempio, [68];

Nel primo caso si considera in realtà una sola ipotesi (H_0), considerando l'alternativa come contenente qualsiasi altra spiegazione del fenomeno in esame non prevista da H_0. In quest'ottica, i dati osservati vengono utilizzati per verificare la loro *conformità* [70] o *compatibilità* con l'ipotesi nulla attraverso il calcolo del valore p, ovvero la probabilità, essendo vera l'ipotesi nulla, di osservare un campione che fornisca un risultato ancora "più lontano" dall'ipotesi nulla rispetto a quello osservato.

Nella seconda impostazione, invece, si fa uso esplicito dell'ipotesi alternativa tanto che la bontà di una procedura di test viene misurata in termini di probabilità di commettere due tipi di errore, i ben noti $\alpha = \Pr(\text{Rifiutare } H_0 \mid H_0)$, detta anche probabilità di errore del I tipo, e $\beta = \Pr(\text{Accettare } H_0 \mid H_1)$, la probabilità di errore del II tipo. In questo senso le due ipotesi vengono considerate di pari dignità e soltanto la procedura usuale di ricerca del test ottimo, che in genere fissa la probabilità di errore di primo tipo e cerca di minimizzare quella di secondo tipo, introduce una distorsione verso l'ipotesi nulla.

Esempio 6.9 [*Confronto fra due ipotesi semplici*].
Consideriamo il seguente esempio, tratto da [6]. Si ha ha disposizione un campione $(X_1, \cdots, X_n)$ estratto da una popolazione normale con varianza nota e pari ad 1. Si vogliono confrontare le due ipotesi per la media

$$H_0 : \theta = -1 \qquad \text{vs.} \qquad H_1 : \theta = 1;$$

supponiamo inoltre che la media campionaria osservata sia pari a $\bar{x} = 0$. Da un punto di vista intuitivo il risultato osservato appare equidistante dalle due ipotesi in competizione; la Figura (6.9) riporta le distribuzioni campionarie della statistica $\bar{X}$ media campionaria condizionatamente alle due ipotesi alternative nei casi $n = 4$ ed $n = 25$. FIGURA ORRENDA
Il risultato osservato non appare in alcun modo discriminare, a livello di evidenza sperimentale, le due ipotesi. Torneremo più avanti su questo esempio. ♢

Abbiamo già avuto modo di osservare che, in ambito bayesiano, tutte le grandezze presenti nel modello statistico sono dotate di legge di probabilità; questo implica che un problema di verifica di ipotesi in ambito bayesiano non può prescindere dalla esplicita definizione di una ipotesi alternativa e da una sua completa probabilizzazione. Dal punto di vista operativo, il modo più naturale di quantificare, a posteriori, il peso delle due ipotesi, la nulla H_0 e l'alternativa H_1, è quello di calcolarne la probabilità $P(H_i \mid \mathbf{x})$, $i = 0, 1$, definita da

$$P(H_i \mid \mathbf{x}) = P(\boldsymbol{\Omega}_i \mid \mathbf{x}) = \int_{\boldsymbol{\Omega}_i} \pi(\theta \mid \mathbf{x}) d\theta, \qquad i = 0, 1 \tag{6.2}$$

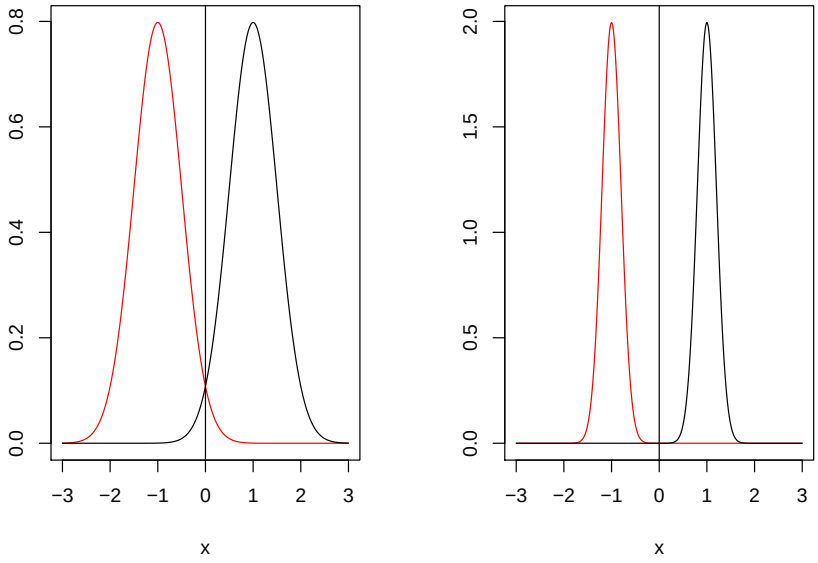

Figura 6.1. Distribuzioni campionarie alternative: nel primo caso il risultato osservato appare compatibile con entrambe le ipotesi a confronto, mentre nel secondo esempio entrambe le ipotesi dovrebbero essere riconsiderate. In entrambi i casi, tuttavia, il risultato sperimentale non fornisce alcuna discriminazione *tra le due ipotesi.*

La formula (6.2) assume in modo implicito che il parametro aleatorio θ sia dotato di densità rispetto alla misura di Lebesgue; ovvie modifiche si applicano nel caso in cui θ è una v.a. discreta, oppure una mistura di due componenti, una discreta e l'altra assolutamente continua. In questa sezione tratteremo solo alcuni esempi canonici; la trattazione generale del problema del confronto fra due ipotesi viene rinviata al Capitolo 8, in quanto, secondo l'impostazione bayesiana, non esistono differenze logiche né procedurali nel confrontare due modelli statistici oppure due ipotesi relative allo stesso modello statistico.

6.3.1 Il caso di due ipotesi semplici

I ruoli svolti dalle informazioni a priori e della funzione di verosimiglianza nella verifica di ipotesi bayesiana sono più chiaramente espressi nel caso artificialmente semplice di due ipotesi puntuali $H_0 : \theta = \theta_0$ e $H_1 : \theta = \theta_1$. In questo caso, calcoli elementari mettono in luce la natura della (6.2); siano infatti

$$\pi_0 = P(H_0) \qquad \text{e} \qquad \pi_1 = P(H_1) = 1 - \pi_0;$$

Il peso relativo a posteriori delle due ipotesi è dato dal rapporto

$$\frac{P(H_1 \mid \mathbf{x})}{P(H_0 \mid \mathbf{x})} = \frac{\pi(\theta_1 \mid \mathbf{x})}{\pi(\theta_0 \mid \mathbf{x})} = \frac{1-\pi_0}{\pi_0} \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})},$$

Tale rapporto è il prodotto di due quantità: la prima, $(1-\pi_0)/\pi_0$, rappresenta il peso relativo delle due ipotesi prima di osservare i dati; la seconda viene in genere denotata con

$$B_{10} = \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})}$$

e si chiama fattore di Bayes. Esso rappresenta il fattore moltiplicativo che trasforma il rapporto di probabilità a priori in quello a posteriori: in questo senso, B_{10} è una misura dell'evidenza sperimentale a favore dell'ipotesi H_1 o contro l'ipotesi nulla H_0. Se $B_{10} < 1$ l'esperimento fornisce maggior evidenza ad H_0 rispetto ad H_1 e il rapporto di probabilità a posteriori è minore di quello a priori. Conclusioni opposte valgono quando $B_{10} > 1$. È interessante notare che, nel caso di due ipotesi semplici, le probabilità frequentiste di errore di primo e secondo tipo hanno una diretta interpretazione in termini delle probabilità a priori assegnate alle due ipotesi π_0 e $1 - \pi_0$.

Esempio 6.10 [*Kass e Wasserman (1997)*]

Nel 1919, durante un'eclissi solare, l'astronomo Eddington effettuò il seguente esperimento: da due posizioni diverse egli misurò il grado di piegatura della luce emessa in funzione della posizione intorno al sole: egli effettuò $n_A = 5$ misurazioni dal posto A e $n_B = 7$ misurazioni dal posto B. La teoria di Newton, denotata qui con H_0, prevede una deflezione di luce di circa 0.875 secondi di arco. Al contrario, la teoria della relatività generale di Einstein, diciamo H_1, conduce qui ad una previsione di 1.75 secondi di arco. Le osservazioni riportarono un valore medio pari a $\bar{x}_A = 1.98$ secondi, con un errore standard pari a $s_A = \sqrt{\frac{\sum_i (x_{iA} - \bar{x}_A)^2}{n_A(n_A - 1)}} = 0.16$ e $\bar{x}_B = 1.61$ con errore standard pari a $s_B = 0.40$. Possiamo supporre che in entrambi i siti, i dati possano essere considerati avere distribuzione normale con media μ e deviazione standard σ incognite. Sia inoltre, a priori, $\pi_0 = \pi_1 = 0.5$, ovvero si dà la stessa probabilità iniziale alle due teorie.

Consideriamo il primo esperimento, in cui $n_A = 5$; per semplicità assumiamo che, essendo gli esperimenti in fisica in genere molto accurati, s_A possa essere considerato una buona stima puntuale di $\sigma/\sqrt{n_A}$. Allora, la media campionaria $\bar{X}_A \sim N(\mu, 0.16^2)$ con $\mu = 0.875$ secondo l'ipotesi H_0 e $\mu = 1.75$ secondo l'ipotesi H_1. Il fattore di Bayes vale

$$B_{10}^A = \frac{\frac{1}{0.16\sqrt{2\pi}} \exp\left\{-(1.98 - 1.75)^2/(2 * 0.16^2)\right\}}{\frac{1}{0.16\sqrt{2\pi}} \exp\left\{-(1.98 - 0.875)^2/(2 * 0.16^2)\right\}} = 0.81 \times 10^{10}.$$

In pratica, sono sufficienti 5 osservazioni per fornire un'evidenza inconfutabile (odds pari circa ad otto miliardi ad uno) in favore di H_1; è però il caso di chiarire bene qual è la conclusione che B suggerisce: il fattore di Bayes non dice che H_1 è vera, ma solo che l'evidenza sperimentale in suo favore, comparativamente con H_0, è un miliardo di volte superiore... Consideriamo adesso i dati relativi al secondo punto di osservazione. Calcoli analoghi conducono a

$$B_{10}^B = \frac{\frac{1}{0.40\sqrt{2\pi}} \exp\left\{-(1.61 - 1.75)^2/(2 * 0.40^2)\right\}}{\frac{1}{0.40\sqrt{2\pi}} \exp\left\{-(1.61 - 0.875)^2/(2 * 0.40^2)\right\}} = 5.076$$

In questo caso le conclusioni sono molto più incerte e il rapporto di odds è solo di 5 a 1 in favore di H_1. I risultati sono in linea con l'intuizione: in questo problema si ha $\mu_0 < \mu_1$: quando l'esperimento fornisce, come nel primo caso, una media campionaria superiore a μ_1, l'evidenza sperimentale a favore di H_1 appare decisiva. Nel secondo caso, invece, la media campionaria è tale che $\mu_0 < \bar{x}_B < \mu_1$; in questi casi le conclusioni sono ovviamente più deboli. Va sottolineato come il fattore di Bayes pesi l'evidenza a favore delle due ipotesi sulla sola base delle loro verosimiglianze. Più avanti vedremo che, in situazioni più complesse, anche il fattore di Bayes conterrà informazioni extra-sperimentali. ♢

L'esigenza di produrre inferenze in qualche modo "oggettive", o che comunque dipendano in modo significativo dai dati osservati, rende il fattore di Bayes importante *per sé*. Quando, come nell'esempio precedente, $\pi_0 = 0.5$, esso è equivalente al rapporto di odds a posteriori che rappresenta la "vera" risposta bayesiana al problema di verifica di ipotesi. Una definizione più generale di fattore di Bayes è la seguente:

Definizione 6.1 *Si chiama fattore di Bayes, e si indica con B_{10}, il rapporto tra le odds a posteriori e quelle a priori delle ipotesi a confronto H_1 e H_0*

$$B_{10} = \frac{\Pr(H_1 \mid \mathbf{x})}{\Pr(H_0 \mid \mathbf{x})} \Big/ \frac{\Pr(H_1)}{\Pr(H_0)}.$$

Se le due ipotesi sono entrambe semplici, B_{10} coincide con il rapporto di verosimiglianza e non contiene alcuna componente soggettiva, se non la scelta del modello statistico adottato, scelta peraltro condivisa da tutte le impostazioni inferenziali.

Esempio 6.9 (continua).

DIRE CHE IN QUESTO CASO IL TEST CLASSICO FORNIREBBE RISPOSTE DIVERSE PER DEI VALORI DI ALPHA RAGIONEBOLI

PARALLELISMO TRA ERRORI ALPHA E BETA E PO E PI A PRIORI

DISCORSO DI PICCINATO SUL CONDIZIONAMENTO

6.3.2 Il caso dell'ipotesi alternativa composta

Consideriamo adesso un caso più generale in cui l'ipotesi nulla puntuale $H_0 : \theta = \theta_0$ viene posta a confronto con un'alternativa composta $H_1 : \theta \neq \theta_0$; questa situazione corrisponde, da un punto di vista metodologico, al classico test di significatività. Anche in questo caso assumiamo che i dati possano essere considerati realizzazioni di v.a. $X_1, X_2, \cdots, X_n$ con distribuzione normale $N(\theta, \sigma^2)$ con σ noto, per semplicità espositiva. L'assegnazione delle probabilità a priori richiede qui un minimo di attenzione; poiché H_0 è una ipotesi semplice, la probabilità π_0 risulterà concentrata sul punto θ_0. Di contro, la probabilità di H_1 deve essere distribuita su tutti i valori di θ diversi da θ_0. Dal punto di vista della notazione, un modo conveniente per descrivere la distribuzione a priori è il seguente

$$\pi(\theta) = \pi_0 \delta_{\theta_0}(\theta) + (1 - \pi_0) g(\theta) \mathbf{1}_{\{\theta \neq \theta_0\}}(\theta); \tag{6.3}$$

la funzione δ di Dirac è stata già introdotta nella § 5.2.4; infine $g(\cdot)$ rappresenta la legge di probabilità a priori quando è vera H_1, ovvero la legge di θ condizionata all'informazione che $\theta \in \boldsymbol{\Omega}_1$. Risulta allora

$$\frac{P(H_1 \mid \mathbf{x})}{P(H_0 \mid \mathbf{x})} = \frac{1 - \pi_0}{\pi_0} \frac{\int_{\theta \neq \theta_0} L(\theta; \mathbf{x}) g(\theta) d\theta}{L(\theta_0; \mathbf{x})}$$

mentre il fattore di Bayes è pari a

$$B_{10} = \frac{P(H_1 \mid \mathbf{x})}{P(H_0 \mid \mathbf{x})} \Big/ \frac{(1 - \pi_0)}{\pi_0} = \frac{\int_{\theta \neq \theta_0} L(\theta; \mathbf{x}) g(\theta) d\theta}{L(\theta_0; \mathbf{x})} = \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})},$$

dove l'espressione $m_j(\mathbf{x})$ rappresenta la distribuzione marginale del vettore $\mathbf{x}$ condizionatamente all'ipotesi H_j, $j = 0, 1$. Per motivi computazionali, nel caso del modello normale, spesso si utilizza

una distribuzione a priori $g(\cdot)$ di tipo coniugato, e quindi anch'essa normale, ovvero $\theta \sim N(\xi, \tau^2)$. In genere, è ragionevole, ma non obbligatorio, supporre che $\xi = \theta_0$, in quanto si presume che l'incertezza intorno a θ_0 sia tale da considerare, inizialmente più probabili piccole distanze, in un senso o nell'altro, da θ_0 rispetto a distanze maggiori. La trattazione che segue è comunque facilmente generalizzabile al caso in cui $\xi \neq \theta_0$. Per il calcolo di B_{10} è forse opportuno ricordare che, poiché la funzione di verosimiglianza dipende dal campione osservato solo attraverso le statistiche sufficienti (in questo caso la media campionaria, il numeratore e il denominatore di B_{10} dipendono dai dati solo attraverso $\bar{x}$, la cui densità, condizionatamente ad un θ generico, risulta ancora normale con media θ e varianza σ^2/n. Avremo così

$$B_{10} = \frac{\int_{\theta \neq \theta_0} \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \exp\left\{-n(\bar{x}-\theta)^2/(2\sigma^2)\right\} \frac{1}{\tau\sqrt{2\pi}} \exp\left\{-(\theta-\theta_0)^2/(2\tau^2)\right\} d\theta}{\frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \exp\left\{-n(\bar{x}-\theta_0)^2/(2\sigma^2)\right\}}$$

L'integrale al numeratore può essere risolto in modo analitico utilizzando la (4.7); è più agevole però ottenerne il valore utilizzando il seguente risultato.

Lemma 6.1. *Se $X \mid \theta \sim N(\theta, \sigma^2)$, e inoltre $\theta \sim N(\mu, \tau^2)$. Allora, marginalmente,*

$$X \sim N(\mu, \sigma^2 + \tau^2).$$

Dimostrazione. Sia $Y = X - \theta$. La legge di $Y \mid \theta$ è ovviamente $N(0, \sigma^2)$, qualunque sia θ, perciò Y è indipendente da θ. La quantità $X = Y + \theta$ è allora una combinazione lineare di normali indipendenti ed è quindi normale con media pari a $\mathbf{E}(Y) + \mathbf{E}(\theta) = 0 + \mu = \mu$ e varianza pari a

$$\mathrm{Var}(Y) + \mathrm{Var}(\theta) = \sigma^2 + \tau^2.$$

da cui la tesi. $\diamond$

Ritornando al calcolo di B_{10}, allora, l'utilizzo del Lemma 6.1 permette di esprimerne il numeratore come

$$\int_{\boldsymbol{\Omega}} \varphi(\bar{x}, \theta, \frac{\sigma^2}{n}) \varphi(\theta, \theta_0, \tau^2) d\theta = \varphi(\bar{x}, \theta_0, \frac{\sigma^2}{n} + \tau^2),$$

dove si è indicata con il simbolo $\varphi(a, b, s^2)$ la densità gaussiana nel punto a, di media b e varianza s^2. Risulta allora,

$$B_{10} = \frac{\sqrt{n} \exp\left\{-n(\bar{x}-\theta_0)^2/(2(\sigma^2+n\tau^2))\right\} / \sqrt{\sigma^2+n\tau^2}\sqrt{2\pi}}{\sqrt{n} \exp\left\{-n(\bar{x}-\theta_0)^2/(2\sigma^2)\right\} / \sigma\sqrt{2\pi}}.$$

Semplici elaborazioni conducono a

$$B_{10} = \frac{\sigma}{\sqrt{\sigma^2+n\tau^2}} \exp\left\{+\frac{n}{2\sigma^2}\left[(\bar{x}-\theta_0)^2\left(1 - \frac{\sigma^2}{\sigma^2+n\tau^2}\right)\right]\right\}$$

Denotando con

$$u = \frac{\sqrt{n}(\bar{x}-\theta_0)}{\sigma}$$

il valore della usuale statistica test, e ponendo $\rho^2 = \tau^2/\sigma^2$, si ottiene

$$B_{10} = \frac{1}{\sqrt{1+n\rho^2}} \exp\left\{\frac{u^2}{2}\frac{n\rho^2}{1+n\rho^2}\right\} \tag{6.4}$$

La formula (6.4) merita qualche approfondimento, per la sua importanza come strumento bayesiano per la scelta tra ipotesi ed anche perché, come vedremo è stata oggetto di un interessante ed

acceso dibattito teorico. Dunque, la (6.4) mostra come, per u fissato, il fattore di Bayes risulti essere una funzione decrescente sia di n che di τ^2. Questo implica che, per grandi valori di n, un valore pure alto della statistica u - che intuitivamente dovrebbe fornire evidenza contro H_0 - viene invece considerato dal fattore di Bayes B_{10} come favorevole all'ipotesi nulla. Questo fenomeno è noto in letteratura come il paradosso di Jeffreys-Lindley ed ha suscitato un notevole dibattito fondazionale. Analizziamo però con più attenzione che cosa realmente avviene, per $n \to \infty$. Per saggiare la consistenza di una procedura di test occorre verificarne il comportamento asintotico sia nel caso in cui sia vera l'ipotesi nulla H_0 sia nel caso in cui sia vera l'ipotesi alternativa. Quando è vera H_0, è facile rendersi conto che la statistica test U ha legge normale standard; perciò il secondo fattore della (6.4) è quasi certamente limitato e dunque B_{10} converge a 0, correttamente, al crescere della dimensione campionaria. Al contrario, se i dati sono realizzazioni di una legge $N(\theta_1, \sigma^2)$ con $\theta_1 \neq \theta_0$, allora

$$U = \frac{\sqrt{n}}{\sigma}\left(\bar{X} - \theta_1 + \theta_1 - \theta_0\right) \sim N\left(\frac{\sqrt{n}}{\sigma}(\theta_1 - \theta_0), 1\right).$$

Perciò U^2 ha distribuzione di tipo χ^2_1 con parametro di non centralità pari a $n(\theta_1-\theta_0)^2/\sigma^2$; ne segue che, per grandi valori di n, la v.a. U^2 diverge quasi certamente a $+\infty$; di conseguenza, $B_{10} \to +\infty$, in accordo con l'intuizione.

Va inoltre ricordato come la convergenza a zero di B_{10} per grandi valori di n può bene essere imputata al fatto che, per dimensioni campionarie elevate, l'adozione di una ipotesi puntuale cessa di rappresentare un'adeguata approssimazione ad una più ragionevole ipotesi intervallare del tipo $H_0 : \theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$; il lettore interessato a questi approfondimenti può consultare [12].

Il comportamento di B_{10}, per grandi valori di τ^2 ci dice invece che nei problemi di verifica di ipotesi non è possibile, se non in circostanze molto speciali, utilizzare distribuzioni a priori improprie; infatti, al crescere di τ^2, la distribuzione a priori di θ "tende" ad una distribuzione uniforme sull'intera retta reale, ovvero ad una distribuzione impropria. La spiegazione di questo fenomeno è semplice. Nei problemi di stima l'analisi bayesiana è basata sull'intera distribuzione finale e nessun punto nello spazio $\boldsymbol{\Omega}$ gode di una natura speciale come quella di θ_0. Per questo, pur utilizzando una legge iniziale impropria, del tipo

$$\pi(\theta) \propto k \times h(\theta),$$

dove $h(\theta)$ è una funzione positiva non integrabile su $\boldsymbol{\Omega}$ e - di conseguenza - la costante k non è esattamente determinabile, si giunge ad una legge a posteriori

$$\pi(\theta \mid \mathbf{x}) = \frac{k \times h(\theta) L(\theta; \mathbf{x})}{k \times \int_{\boldsymbol{\Omega}} h(\theta) L(\theta; \mathbf{x})},$$

che, di fatto, non dipende dalla costante. Al contrario, nel caso di verifica di ipotesi nulla semplice, la costante non scompare; ci limitiamo qui a considerare il nostro esempio guida ma il problema è del tutto generale. Il fattore di Bayes in questo caso sarebbe pari a

$$B_{10} = \frac{k \int_{\theta \neq \theta_0} f(\mathbf{x}|\theta) h(\theta) d\theta}{f(\mathbf{x}|\theta_0)}$$

e dipenderebbe dalla costante non determinabile k. Inoltre la probabilità finale associata all'ipotesi nulla H_0 dipende dal fattore di Bayes; infatti,

$$\Pr(H_0|\mathbf{x}) = \frac{\pi_0 f(\mathbf{x}|\theta_0)}{\pi_0 f(\mathbf{x}|\theta_0) + (1-\pi_0) k \int_{\theta \neq \theta_0} f(\mathbf{x}|\theta) h(\theta) d\theta};$$

dividendo per il numeratore si ottiene allora

$$\Pr(H_0|\mathbf{x}) = \left(1 + \frac{1-\pi_0}{\pi_0} B_{10}\right)^{-1}, \tag{6.5}$$

e neanche $\Pr(H_0|\mathbf{x})$ è calcolabile in presenza di distribuzioni iniziali improprie. Questo problema impedisce dunque l'uso di distribuzioni improprie nelle procedure bayesiane di verifica di ipotesi, in particolare quando le due ipotesi hanno diversa dimensione. Esistono tuttavia recenti sviluppi che risolvono questo problema mediante l'uso dei cosiddetti campioni di prova o "training samples", il cui uso discuteremo nel Capitolo 8. Il lettore interessato può consultare ii lavori originali di [8] e [65] oppure il testo di [76]. Per approfondimenti sul paradosso di Lindley si vedano, tra gli altri, [12] e [38].

Confronti con l'impostazione classica. Lo stesso problema, affrontato da un punto di vista classico, avrebbe comportato il calcolo della statistica u e del suo relativo valore p a due code. La tabella mostra il calcolo di B_{10} e del corrispondente valore p per specifici valori della statistica u e per diverse numerosità campionarie. I confronti vengono effettuati ponendo $\tau^2 = \sigma^2$: intuitivamente questo significa considerare la distribuzione a priori alla stregua di un'ulteriore osservazione campionaria.

invertire la tabella

Valore di u	valore-p	$n = 1$	$n = 10$	$n = 50$	$n = 100$	$n = 1000$
1.64	0.10	1.38	1.02	0.52	0.38	0.12
1.96	0.05	1.84	1.73	0.92	0.66	0.22
2.56	0.01	3.63	5.92	3.48	2.55	0.83
3.29	0.001	10.58	41.31	28.21	21.13	7.04

Tabella 6.1. Valori di B_{10} per diversi valori di n, nel caso $\rho^2 = 1$, in corrispondenza a valori notevoli del p-value.

Esempio 6.11 [*Capacità paranormali*]

L'esempio che segue è ispirato da Berger (2000) ma i numeri sono leggermente differenti. Esso mostra con estrema chiarezza come l'analisi bayesiana di un problema di verifica di ipotesi possa condurre a conclusioni in completo disaccordo con l'analisi classica. I dati si riferiscono ad un esperimento nel campo della psicocinesi, descritto in [79]. Per verificare se un certo soggetto è dotato di abilità psico-cinetiche, venne utilizzato un generatore di eventi casuali basato su principi della meccanica quantistica: l'esperimento consiste nel verificare se il soggetto è in grado di influenzare il generatore. Semplificando un po', supponiamo che il generatore "spari" delle particelle verso una porta "quantistica": qui le particelle possono proseguire lungo la luce rossa oppure la luce verde. Il soggetto tenta di spingere le particelle lungo la luce rossa; nel caso in cui egli non riesca ad influenzare le particelle il meccanismo aleatorio è tale per cui ogni particella ha probabilità 0.5 di proseguire lungo ciascuno dei due fasci di luce. Vennero effettuate un milione di prove ($n = 10^6$), ognuna delle quali poteva essere considerata una prova bernoulliana in cui si osserva 1 (se la particella sceglie la luce rossa) oppure 0 (luce verde). Sia $\theta = P(X_i = 1)$ e sia $Y = \sum X_i$ il numero di particelle che prende la strada *rossa*. Per quanto visto nella §2.1 si ha che $Y \sim Bin(n, \theta)$. Si vuole verificare dunque il sistema di ipotesi

$$\begin{cases} H_0 : \theta = \frac{1}{2} & \text{Il soggetto non influenza il generatore} \\ H_1 : \theta \neq \frac{1}{2} & \text{Il soggetto influenza il generatore} \end{cases}$$

Si potrebbe obiettare che una ipotesi alternativa unilaterale ($\theta > \frac{1}{2}$) sarebbe in questo caso più ragionevole: le conclusioni di fondo tuttavia non cambierebbero e i calcoli numerici risultano più semplici adottando un'ipotesi alternativa bilaterale. Il valore osservato della statistica Y è pari a 501550. Data l'enorme numerosità campionaria è lecito utilizzare l'approssimazione normale per cui, sotto l'ipotesi nulla H_0,

$$Z = \frac{Y - n/2}{\sqrt{n/4}} \sim N(0, 1).$$

Il valore di Z osservato è dunque pari a $z_{oss.} = 3.1$ con conseguente valore p bidirezionale pari a $P(|Z| \geq 3.1) = 2 \times\ 0.000967\ = 0.00192$.

L'analisi classica conduce quindi ad un chiaro rifiuto dell'ipotesi nulla. Vediamo ora cosa avviene attraverso un approccio bayesiano. Sia dunque $\pi_0 = P(H_0)$ la probabilità iniziale dell'ipotesi nulla, che lasciamo per ora non specificata. Condizionatamente a H_1, ovvero $\theta \neq 1/2$, sia $g(\theta)$ la distribuzione iniziale per θ. L'analisi non informativa qui sceglierebbe $\pi_0 = 1/2$ e $g(\theta) = 1$ (distribuzione uniforme) oppure la distribuzione iniziale di Jeffreys $g(\theta) \propto 1/\sqrt{\theta(1-\theta)}$. La probabilità a posteriori di H_0 risulta così pari a

$$P\left(H_0 \mid y\right) = \frac{\pi_0 P(Y = y \mid \theta = \frac{1}{2})}{\pi_0 P(Y = y \mid \theta = \frac{1}{2}) + (1 - \pi_0) \int_{\theta \neq \frac{1}{2}} P(Y = y \mid \theta) g(\theta) d\theta}. \tag{6.6}$$

Ad esempio, utilizzando $\pi = 1/2$ e $g(\theta)$ costante si ottiene $P(H_0 \mid y) = 0.87$. Un'analisi bayesiana completa prevede il calcolo, oltre che di $P(H_0 \mid y)$, dell'insieme HPD condizionato ad H_1, ovvero quel sottoinsieme di $\boldsymbol{\Omega}_1 : \{\theta : \theta \neq \frac{1}{2}\}$ più "probabile" nell'ipotesi che H_1 sia vera. Calcoli simili a quelli illustrati nella §6.2 conducono all'intervallo di credibilità di livello 0.95

$$C = (0.50015, 0.5029).$$

L'alta probabilità finale di H_0 è chiaramente in contrasto con le conclusioni raggiunte secondo un'impostazione classica; essa potrebbe essere imputata al fatto che, inizialmente avevamo posto $\pi_0 = \Pr\left(H_0\right) = 1/2$. Per indebolire tale conclusione possiamo seguire due strade

- calcolare la probabilità finale al variare della probabilità iniziale π_0;
- utilizzare il fattore di Bayes;

Nel primo caso, utilizzando la (6.6) come funzione di π_0, e tenendo fissata la $g(\theta)$, si ha

$$P\left(H_0 \mid y\right) = \left(1 + \frac{b(1 - \pi_0)}{a\pi_0}\right)^{-1}, \tag{6.7}$$

dove, nel nostro caso, usando l'approssimazione normale,

$$a = P(Y = y_{oss} \mid \theta \cong \frac{1}{2}) = 6.53 \times 10^{-6}, \qquad b = \int_0^1 P(Y = y \mid \theta) d\theta = \frac{1}{n+1} \approx 10^{-6}.$$

Un elementare studio di funzione mostra come la (6.7) sia una funzione crescente di π_0, come è lecito attendersi: va notato che la probabilità finale di H_0 è maggiore di 0.5 non appena la probabililità iniziale π risulti maggiore di 0.132.

Nel secondo caso si indeboliscono le assunzioni a priori utilizzando semplicemente le quantità a e b appena definite. In quest'ottica, il fattore di Bayes rappresenta una generalizzazione del rapporto di verosimiglianza dove al denominatore viene considerata una sorta di media della funzione di verosimiglianza, ponderata con la distribuzione iniziale $g(\theta)$; nell'esempio specifico si ottiene, ovviamente $B_{10} = b/a = 0.153$. Un'analisi più accurata deve includere anche il controllo della sensibilità delle risposte al variare della distribuzione $g(\theta)$. La tabella che segue mostra i valori estremi che assume B_{01} quando la densità $g(\theta)$ varia in opportune classi di densità.

MANCA LA TABELLA

Questo esempio è istruttivo per diversi motivi

- Il valore p è comunque molto diverso dal fattore di Bayes: anche se i due indicatori misurano effettivamente cose differenti, va sottolineato come l'evidenza fornita contro l'ipotesi nulla dai due metodi sia fortemente contrastante, e la tabella mostra come tale contrasto non dipenda dalla distribuzione $g(\theta)$;
- in casi come questi ha senso considerare un'ipotesi nulla puntuale come $H_0 = \frac{1}{2}$, poiché tale valore ha un preciso significato nel contesto in esame e corrisponde al fatto che il soggetto in esame non sia in grado di modificare il flusso delle particelle. In altre situazioni, dove non esistono valori parametrici, per così dire, *privilegiati*, tale modellizzazione non è giustificabile e dovremmo ricorrere alla formulazione di due ipotesi composte.
- abbiamo considerato un esempio in cui l'ipotesi alternativa H_1 è bilaterale (o, come spesso si dice, a due code): questo è relativamente importante; le stesse conclusioni si raggiungono in esempi in cui l'ipotesi alternativa è unilaterale.

♢

6.3.3 Uso di distribuzioni improprie nei problemi di test

Come già sottolineato in §6.3.2, non è possibile utilizzare, in maniera diretta, distribuzioni improprie quando si confrontano due ipotesi "annidate". Questo problema ha suscitato particolare attenzione nel corso dell'ultimo decennio ed ha condotto alla costruzione di diverse proposte. Daremo conto di questi sviluppi in modo più articolato in §8.13; qui ci limitiamo ad illustrare brevemente una delle possibili soluzioni, il fattore di Bayes intrinseco, proposto da [8], in un contesto molto semplice. Siano $(X_1, X_2, \cdots, X_n) \overset{\text{iid}}{\sim} p(x_i; \theta)$. Si vuole verificare il sistema di ipotesi

$$H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta \neq \theta_0,$$

e si desidera non introdurre informazioni soggettive sul parametro θ. Poiché l'uso diretto della legge uniforme $\pi_1^N(\theta) \propto$ cost. non è possibile, [8] propongono l'utilizzo dei cosiddetti "campioni di prova", o "training samples"; in grandi linee, si tratta di utilizzare una "parte" del campione per aggiornare la legge iniziale impropria e renderla propria, e sfruttare la parte rimanente del campione per il confronto tra le due ipotesi. In questo caso specifico assumiamo, per semplicità, che sia sufficiente utilizzare una singola osservazione del campione per ottenere una legge su θ propria, pur partendo dall'impropria $\pi_1^N(\theta)$. Per fissare le idee, sia $m_j^N(\mathbf{x})$, per $j = 0, 1$, la legge marginale del vettore $\mathbf{x}$ quando è vera l'ipotesi H_j, ottenuta come

$$m_j^N(\mathbf{x}) = \int p(\mathbf{x};\theta)\pi_j^N(\theta)d\theta;$$

nel caso specifico, ovviamente,

$$m_0(\mathbf{x}) = p(\mathbf{x};\theta_0) = \prod_{i=1}^{n} m_0(x_i).$$

Supponiamo allora di utilizzare la prima osservazione, x_1, per aggiornare la legge iniziale e le altre $n-1$ per valutare l'evidenza a favore delle due ipotesi. Avremo così

$$\pi_1^N(\theta \mid x_1) = \frac{\pi_1^N(\theta)p(x_1;\theta)}{m_1^N(x_1)},$$

che può essere utilizzata come "nuova" legge iniziale, questa volta propria, per il calcolo del fattore di Bayes. Denotiamo inoltre col simbolo $\mathbf{x}_{(-i)}$ il vettore delle osservazioni privato della unità i-esima. Avremo allora che il fattore di Bayes può essere espresso come

$$B_{10}(\mathbf{x}_{(-1)}) = \int_{\boldsymbol{\Omega}} \frac{p(\mathbf{x}_{(-1)};\theta)\pi_1^N(\theta)p(x_1;\theta)d\theta}{m_1^N(x_1)m_0(\mathbf{x}_{(-1)})}. \tag{6.8}$$

Moltiplicando e dividendo per $p(x_1;\theta_0) = m_0(x_1)$, si può scrivere il fattore di Bayes così ottenuto come funzione esplicita di x_1, ovvero

$$B_{10}(\mathbf{x}_{(-1)}) = \frac{m_1^N(\mathbf{x})}{m_0^N(\mathbf{x})}\,\frac{m_0^N(x_1)}{m_1^N(x_1)} = \frac{B_{10}^N(\mathbf{x})}{B_{10}^N(\mathbf{x}_{(-1)})}, \tag{6.9}$$

dove B_{10}^N rappresenta il fattore di Bayes ottenuto mediante l'uso della legge impropria π^N. Tuttavia, questa procedura dipende, in modo ingiustificato, dal particolare valore x_1 scelto per aggiornare la legge impropria iniziale. Questa dipendenza può essere superata in diversi modi; [8] propongono di calcolare il fattore di Bayes *intrinseco* come il valore medio degli n possibili fattori di Bayes, ottenuti secondo la (6.9), ovvero

$$B_{10}^I = \frac{1}{n}\sum_{i=1}^{n} B(\mathbf{x}_{(-i)}) \tag{6.10}$$

Alternativamente, come vedremo in seguito è possibile selezionare altre sintesi dei vari $B(x_i)$, come ad esempio, quello mediano. Un approccio diverso, che mira a garantire una maggiore stabilità del fattore di Bayes intrinseco, consiste poi nel considerare il fattore di Bayes (6.12) come una variabile aleatoria, funzione del dato campionario, denotata con $B_{10}(X_1)$ e considerarne il valore atteso, nel caso in cui la distribuzione campionaria di X_1 sia calcolata sotto l'ipotesi meno restrittiva, in questo caso H_1, e assumendo che θ sia uguale alla stima di massima verosimiglianza (ottenuta utilizzando l'intero campione). Avremo così il cosiddetto fattore di Bayes intrinseco *atteso*

$$B_{10}^E = \mathbb{E}_{\hat{\theta}}\left[B_{10}(X_1)\right]. \tag{6.11}$$

Esempio 6.12 [*Ipotesi nulla puntuale nel caso gaussiano*].
Siano $(X_1, X_2, \cdots, X_n) \overset{\texttt{iid}}{\sim} N(\theta, \sigma^2)$, con σ^2 noto. Si vuole verificare il sistema di ipotesi

$$H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta \neq \theta_0,$$

Essendo

$$p(x_1;\theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp\{-(x_1-\theta)^2/(2\sigma^2)\}$$

e $\pi_1^N(\theta) =$ cost., semplici calcoli conducono facilmente a

$$m_1^N(x_1) = \int \text{cost.} p(x_1;\theta)d\theta = \text{cost.},$$

e

$$B_{10}^N(x_1) = \frac{m_1^N(x_1)}{p(x_1;\theta_0)} = \text{cost.} \times \sigma\sqrt{2\pi}\exp\left\{\frac{1}{2\sigma^2}(x_1-\theta_0)^2\right\}.$$

Analogamente, utilizzando come distribuzione campionaria di $\mathbf{x}$ la (2.5), si ottiene che

$$m_0^N(\mathbf{x}) = p(\mathbf{x};\theta_0) = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\theta_0)^2\right\}$$

e

$$m_1^N(\mathbf{x}) = \int \text{cost.} p(\mathbf{x};\theta)d\theta = \text{cost.}\ .$$

Perciò

$$\begin{aligned} B_{10}(x_1) &= \frac{\text{cost.}\frac{\sigma\sqrt{2\pi}}{\sqrt{n}}}{\exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\theta_0)^2\right\}}\frac{\exp\left\{-\frac{1}{2\sigma^2}(x_1-\theta_0)^2\right\}}{\sigma\sqrt{2\pi}\ \text{cost.}} \\ &= \frac{1}{\sqrt{n}}\exp\left\{\frac{1}{2\sigma^2}\left[n\left(\bar{x}-\theta_0\right)^2\right]\right\}\exp\left\{-\frac{1}{2\sigma^2}\left(x_1-\theta_0\right)^2\right\}. \end{aligned} \tag{6.12}$$

Dunque il fattore di Bayes intrinseco aritmetico vale

$$B_{10}^I = \frac{1}{\sqrt{n}}\exp\left\{\frac{1}{2\sigma^2}\left[n\left(\bar{x}-\theta_0\right)^2\right]\right\}\frac{1}{n}\sum_{j=1}^{n}\left(\exp\left\{-\frac{1}{2\sigma^2}\left(x_j-\theta_0\right)^2\right\}\right) \tag{6.13}$$

Per ottenere invece il fattore di Bayes intrinseco atteso, essendo $\hat{\theta} = \bar{x}$, occorre integrare, rispetto ad una legge $\mathrm{N}(\bar{x},\sigma^2)$, l'ultimo fattore della (6.12); utilizzando il Lemma 6.1 si ottiene

$$\begin{aligned} B_{10}^E &= \frac{1}{\sqrt{n}}e^{\left\{+\frac{n}{2\sigma^2}\left[(\bar{x}-\theta_0)^2\right]\right\}}\int e^{\left\{-\frac{1}{2\sigma^2}(x_1-\theta_0)^2\right\}}\frac{1}{\sigma\sqrt{2\pi}}e^{\left\{-\frac{1}{2\sigma^2}(x_1-\bar{x})^2\right\}}dx_1 \\ &= \frac{1}{\sqrt{2n}}\exp\left\{+\frac{2n-1}{2\sigma^2}\left(\bar{x}-\theta_0\right)^2\right\} \end{aligned}$$

Ricordando l'espressione (6.4), si può notare come lo stesso risultato avrebbe potuto essere ottenuto adottando, come legge iniziale per θ sotto l'ipotesi alternativa, una distribuzione normale con media θ_0 e varianza $\sigma^2(2n-1)/n$. Quando è possibile, come in questo caso, giungere ad una conclusione equivalente a quella ottenibile mediante l'utilizzo di una $\pi(\theta)$ propria, quest'ultima distribuzione prende il nome di distribuzione a priori *intrinseca*. Essa svolge, nei problemi di verifica di ipotesi, il ruolo di distribuzione non informativa. ♢

6.4 L'impostazione predittiva

L'impostazione finora adottata, che possiamo definire *ipotetica* è certamente, oggi, la più popolare. Tuttavia, lo spirito iniziale da cui emerse la critica definettiana ha condotto gli studiosi più ortodossi a sviluppi diversi: In questa sezione descriveremo brevemente le conseguenze di un'impostazione completamente *predittiva* o *previsiva* cercando comunque di sottolineare i molti punti di contatto con lo schema ipotetico, che si rivelerà pienamente interpretabile da un punto di vista predittivo non appena vengano aggiunte ulteriori condizioni sul processo di osservazione dei dati.

Nell'impostazione bayesiana ipotetica, l'idea classica del modello statistico viene ereditata e arricchita attraverso una probabilizzazione dello spazio parametrico: in altri termini il parametro θ diventa una variabile (o un vettore, nel caso multiparametrico) aleatorio. La critica di de Finetti a tale impostazione è radicale: il processo di apprendimento statistico avviene mediante la concreta osservazione di realizzazioni di variabili aleatorie: al contrario il parametro è qualcosa di virtuale, non osservabile se non in circostanze molto speciali.

Esempio 6.13 [*[32]*]
Consideriamo i due problemi seguenti:
1. si lancia una moneta n volte e si registra il risultato ottenuto. Sulla base di tale informazione si vuole prevedere il risultato del lancio $n+1$-esimo;
2. da un'urna che contiene palline bianche e nere in proporzione non nota, si estraggono, con ripetizione, n palline. Sulla base del risultato osservato si vuole prevedere il colore della $n+1$-esima pallina estratta.

All'interno dell'impostazione ipotetica, le due situazioni precedenti non presentano differenze sostanziali: una volta definito il parametro θ (che nel primo caso può rappresentare ad esempio, la probabilità che la moneta dia testa, mentre nel secondo caso può essere definito come la frequenza relativa di palline bianche nell'urna), le n osservazioni si configurano, in entrambi i casi come realizzazioni di n variabili aleatorie indipendenti e somiglianti condizionatamente al valore di θ, ed è elementare condurre un'analisi frequentista o bayesiana. de Finetti però sottolinea una differenza sostanziale tra le due situazioni: nella prima, la grandezza θ non ha un significato fisico ben definito, e dunque verificabile: in altre parole, non saremo mai in grado di stabilire il vero valore di θ. Nel caso dell'urna, al contrario, è sufficiente "aprire" l'urna, per stabilire il valore di θ. Se si accetta il punto di vista secondo cui la probabilità è qualcosa che possiamo associare soltanto ad eventi *verificabili*, ecco che cade la possibilità di impostare in chiave ipotetica il problema della moneta, e soltanto un approccio predittivo appare ragionevole. ♢

Una discussione dettagliata di questi argomenti è al di là degli scopi introduttivi di questo testo e al lettore interessato si suggerisce la lettura del testo classico di Teoria della Probabilità (1970) dello stesso de Finetti. Vale la pena però sottolineare la differenza sostanziale che de Finetti opera fra quantità osservabili e quantità non osservabili: i *fatti*, ovvero gli eventi, proposizioni per le quali si può stabilire se siano effettivamente vere o false, sono gli unici enti che è possibile valutare in senso probabilistico.

Sempre a de Finetti [32] è però dovuta una costruzione teorica che permette di reinterpretare l'approccio ipotetico in chiave predittiva, sotto alcune assunzioni speciali. Per descrivere questo approccio, occorre introdurre il concetto di scambiabilità.

Definizione 6.2 *Le variabili aleatorie* $(X_1, X_2, \cdots, X_n)$ *si dicono scambiabili se la funzione di ripartizione congiunta di una qualunque permutazione* $(X_{i_1}, X_{i_2}, \cdots X_{i_n})$ *delle n v.a. coincide con quella delle v.a.* $(X_1, X_2, \cdots, X_n)$.

Esempio 6.14 [*Normale bivariata*]
Consideriamo una v.a. (X, Y) normale a due dimensioni, con marginali standardizzate e coefficiente di correlazione pari ρ. Allora la funzione di ripartizione di (X, Y) vale

$$F_{X,Y}(u,v) = \int_{-\infty}^{u}\int_{-\infty}^{v} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(x^2+y^2-2\rho xy\right)\right\} dxdy.$$

Basta operare nell'integrale il cambio di variabile $x = y$ e $y = x$ per verificare che

$$F_{X,Y}(u,v) = F_{Y,X}(u,v).$$

che garantisce la scambiabilità di X e Y. ♢

Definizione 6.3 *La successione infinita di v.a. $X_1, X_2, \cdots$ si dice scambiabile se ogni n-pla di variabili aleatorie scelte da tale successione risulta scambiabile, qualunque sia n.*

Dalle definizioni precedenti si deduce, in particolare, che variabili aleatorie scambiabili sono necessariamente somiglianti. Inoltre è facile verificare che variabili aleatorie indipendenti sono scambiabili, mentre l'Esempio 6.14 dimostra che il viceversa non è necessariamente vero.

Illustriamo ora qui il risultato fondamentale di [32], noto come *teorema di rappresentazione per successioni scambiabili*, che stabilisce un ponte tra l'impostazione classica e quella bayesiana predittiva dell'inferenza, nell'ipotesi in cui si disponga, potenzialmente, di una successione infinita di v.a. scambiabili. Considereremo in dettaglio il caso in cui le variabili osservabili possano assumere solo i valori 0 e 1, limitandoci ad enunciare solamente il caso più generale.

Teorema 6.1 *Sia $X_1, X_2, \cdots$ una successione di variabili scambiabili che assumono solo i valori 0 e 1, con legge di probabilità P, e sia, per ogni intero positivo n, $S_n = X_1 + X_2 + \cdots + X_n$. Allora esiste una funzione di ripartizione Q tale che, per ogni n e per ogni n-pla $(x_1, x_2, \cdots, x_n)$, si ha*

$$P(X_1 = x_1, \cdots, X_n = x_n) = \int_0^1 \prod_{i=1}^{n} \left[\theta^{x_i}(1-\theta)^{1-x_i}\right] dQ(\theta) = \int_0^1 \theta^{s_n}(1-\theta)^{n-s_n} dQ(\theta),$$

dove

$$Q(\theta) = \lim_{n\to\infty} \Pr\left(\frac{S_n}{n} \le \theta\right), \qquad \theta = \lim_{n\to\infty} \frac{S_n}{n}. \tag{6.14}$$

Dimostrazione 6.1 *Si veda l'appendice (C.3.1).*

Il teorema precedente è uno dei risultati più importanti della statistica matematica e merita alcune riflessioni. Innanzitutto esso ci dice che, condizionatamente ad una variabile aleatoria θ, le osservabili $X_1, \cdots, X_n$ possono essere considerate indipendenti e somiglianti: in tal senso si recupera, almeno parzialmente, e gli si dà giustificazione, l'approccio classico alla costruzione del modello statistico; la differenza principale, però, è rappresentata dal fatto che θ non rappresenta una costante ignota da stimare, bensì un oggetto aleatorio, la cui distribuzione Q è indotta dalla legge P, attraverso la quale abbiamo imposto la condizione di scambiabilità. La relazione (6.14) afferma che Q può essere interpretata come la legge che esprime le nostre valutazioni sui valori che può assumere la frequenza relativa limite di successi, al crescere delle osservazioni: tale grandezza, a differenza del parametro (nel senso classico) ha un'interpretazione diretta: almeno a livello asintotico essa può essere *osservata*.

Si può perciò affermare che l'ipotesi di scambiabilità *recupera* le altre impostazioni, riducendole a possibili, (non necessarie: la scambiabilità non è in alcun modo un'assunzione obbligatoria)

costruzioni di tipo predittivo. Quando si opera con variabili dicotomiche, dunque, e si assume la scambiabilità, non è necessario elicitare la legge P, ma è sufficiente stabilire la legge $Q(\theta)$ che, nel linguaggio usuale, rappresenterà la distribuzione iniziale per θ; essa, in virtù del teorema di rappresentazione, induce la legge di probabilità sulla successione $\{X_1, X_2, \cdots ..\}$. In tale contesto è allora semplice produrre la distribuzione predittiva, ad esempio, del vettore $X_{n+1}, \cdots, X_{n+m}$ sulla base delle osservazioni $X_1, \cdots, X_n$. Infatti

$$\begin{aligned} &\Pr\left(X_{n+1} = x_{n+1}, \cdots, X_{n+m} = x_{n+m} | X_1 = x_1, \cdots, X_n = x_n\right) \qquad (6.15)\\ &= \frac{\Pr\left(X_1 = x_1, \cdots, X_n = x_n, X_{n+1} = x_{n+1}, \cdots, X_{n+m} = x_{n+m}\right)}{\Pr\left(X_1 = x_1, \cdots, X_n = x_n\right)} \\ &= \frac{\int_0^1 \theta^{S_{n+m}}(1-\theta)^{n+m-S_{n+m}} dQ(\theta)}{\int_0^1 \theta^{S_n}(1-\theta)^{n-S_n} dQ(\theta)}. \end{aligned}$$

Separando nel numeratore la componente relativa alle future m osservazioni, l'integrale precedente può allora essere espresso come un valore atteso rispetto alla legge di θ condizionata alle prime n osservazioni, ovvero

$$\begin{aligned} &\frac{\int_0^1 \theta^{S_{n+m}-s_n}(1-\theta)^{m-(S_{n+m}-s_n)} L(\theta, \mathbf{x}_n) dQ(\theta)}{\int_0^1 \theta^{s_n}(1-\theta)^{n-s_n} dQ(\theta)} \\ &= \int_0^1 \theta^{S_{n+m}-s_n}(1-\theta)^{m-(S_{n+m}-s_n)} dQ(\theta \mid \mathbf{x}_n), \end{aligned}$$

dove $L(\theta, \mathbf{x}_n)$, come sempre, denota la funzione di verosimiglianza associata alle prime n osservazioni, indicate in modo vettoriale col simbolo $\mathbf{x}_n$. In definitiva, il calcolo di una distribuzione predittiva è formalmente il calcolo del valore atteso di una funzione della legge a posteriori. Nel caso precedente, ad esempio, la probabilità condizionata (6.15) può leggersi come

$$\mathbf{E}^{\pi(\theta|\mathbf{x}_n)}\left[\theta^{S_{n+m}-s_n}(1-\theta)^{m-S_{n+m}+s_n}\right].$$

Inoltre, nell'assunzione di scambiabilità è in qualche modo implicita la dichiarazione di non interesse al risultato di una particolare realizzazione X_j: siamo interessati soltanto a *quanti* successi si avranno su n prove: è immediato allora ottenere, mediante il Teorema 6.1, la legge della v.a. S_n; infatti, per ogni $s = 0, 1, \cdots, n$,

$$\Pr\left(S_n = s\right) = \binom{n}{s} \Pr\left(X_1 = x_1, \cdots, X_n = x_n\right) = \int_0^1 \binom{n}{s} \theta^s (1-\theta)^{n-s} dQ(\theta),$$

dove la n-pla di risultati al secondo membro, $(x_1, x_2, \cdots, x_n)$ è tale che $\sum x_i = s$. Tutto questo ci fornisce una giustificazione ad agire, quando si esprimono pareri su Y, *come se* avessimo a disposizione una funzione di verosimiglianza di tipo bernoulliano e una legge iniziale $Q(\theta)$ sul parametro.

Quanto detto si riferisce esclusivamente al caso in cui le variabili osservabili sono dicotomiche. Qui di seguito accenniamo brevemente alle estensioni al caso in cui le osservabili abbiano come supporto l'intera retta reale. Siano allora $X_1, X_2, \cdots, X_n$ v.a. scambiabili, a valori reali con legge di probabilità congiunta P; sia inoltre $\mathcal{F}$ lo spazio di tutte le funzioni di ripartizione su $\mathbb{R}$. È allora possibile dimostrare che esiste una misura di probabilità Q definita su $\mathcal{F}$, tale che la funzione di ripartizione F_P del vettore $(X_1, \cdots, X_n)$ si può scrivere

$$F_P(x_1, \cdots, x_n) = \int_{\mathcal{F}} \prod_{i=1}^{n} F(x_i) dQ(F),$$

dove $Q(F)$ rappresenta di nuovo la legge di probabilità iniziale, interpretabile dualmente come la misura limite del processo delle funzioni di ripartizione empiriche $F_n(\cdot)$.
Pur utilizzando una formalizzazione matematica necessariamente più sofisticata, il risultato suddetto ha una interpretazione simile al Teorema 6.1: in presenza di scambiabilità si può agire *come se* le X_i fossero osservazioni indipendenti e somiglianti con legge F, la quale è però aleatoria con distribuzione Q indotta da P: tale legge può essere interpretata come descrittiva delle nostre opinioni sul comportamento asintotico di F_n. Anche in questo caso viene recuperato lo schema bayesiano ipotetico.

Va però sottolineato un aspetto: mentre, nel caso dicotomico, il Teorema 6.1 *porge* il modello binomiale come il contesto naturale per reinterpretare l'ipotesi di scambiabilità, nel caso generale il teorema stesso si limita a stabilire che le variabili osservabili possono essere considerate come indipendenti e somiglianti con funzione di ripartizione F (dipendente dalla legge soggettivamente espressa P) non meglio specificata. Tutto ciò suggerisce che, in un'impostazione predittiva dell'inferenza, se non si introducono ulteriori ipotesi (in aggiunta o in alternativa alla scambiabilità), la pratica di limitarci all'uso di modelli parametrici è ingiustificata: non a caso, i cultori di tale impostazione, sottolineano l'esigenza di un'impostazione non parametrica della statistica bayesiana, argomento peraltro troppo avanzato per essere discusso in questo testo. Il lettore interessato può consultare, per un'introduzione, [25] oppure [64]. Va però detto che è possibile *recuperare* i più importanti e popolari modelli parametrici, come la distribuzione gaussiana o quella esponenziale, semplicemente rinforzando l'ipotesi di scambiabilità con altre forme, più restrittive, di simmetria distribuzionale; per dettagli si rimanda a [15].

6.4.1 Il concetto di sufficienza nell'impostazione predittiva

All'interno dell'impostazione completamente predittiva che, come abbiamo visto, fa a meno della definizione esplicita del parametro θ, è ancora possibile parlare di sufficienza: in questo caso, però, non si tratterà più di sufficienza per il parametro bensì di *sufficienza a fini previsivi*. Come già visto, l'obiettivo di una previsione è quello di calcolare la probabilità di un evento futuro sulla base della sequenza osservata, ovvero, con la notazione delle v.a. discrete,

$$\Pr\left(X_{n+1} = x_{n+1}, \cdots, X_{n+m} = x_{n+m} \mid X_1 = x_1, \cdots, X_n = x_n\right) \tag{6.16}$$

È possibile allora che esista una opportuna funzione delle osservazioni passate, diciamo $T(X_1, \cdots, X_n)$ tale che la (6.16) risulti uguale a

$$\Pr\left(X_{n+1} = x_{n+1}, \cdots, X_{n+m} = x_{n+m} \mid T(X_1, \cdots, X_n) = t\right)$$

Questa uguaglianza suggerisce che la previsione effettuata sulla base della conoscenza di T non è meno precisa di quella ottenuta sulla base della conoscenza dell'intera sequenza di osservazioni $(x_1, \cdots, x_n)$. In tal senso si può affermare allora che T è una statistica sufficiente ai fini previsivi. Un modo alternativo per definire questo tipo di sufficienza è il seguente [25].

Definizione 6.4 *La funzione $T(X_1, \cdots, X_n)$ è una statistica sufficiente ai fini previsivi del vettore $\mathbf{X}_n = (X_1, \cdots, X_n)$ se, per ogni m, il vettore $(X_{n+1}, \cdots, X_{n+m})$ è indipendente da $\mathbf{X}_n$, condizionatamente a T.*

6.4.2 Calcoli predittivi

Dal punto di vista operativo l'impostazione predittiva si concretizza nel calcolo della distribuzione congiunta relativa ad un futuro insieme di osservazioni condizionatamente all'informazione contenuta in quelle già osservate.
Sia dunque $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_m)$ un vettore composto da m variabili aleatorie condizionatamente indipendenti e somiglianti con distribuzione $p(y;\theta)$, rappresentanti le "future" osservazioni. Supponiamo inoltre che le informazioni sul parametro θ siano sintetizzate attraverso la sua legge finale, $\pi(\theta \mid \mathbf{x})$, basata su una legge iniziale $\pi(\theta)$ e sull'osservazione di un campione $\mathbf{x} = (x_1, x_2, \cdots, x_n)$. Si può allora scrivere, assumendo per semplicità di esposizione che tutte le variabili aleatorie osservabili siano dotate di funzione di densità,

$$\pi(\mathbf{y} \mid \mathbf{x}) = \int_{\boldsymbol{\Omega}} p(\mathbf{y} \mid \theta, \mathbf{x}) \pi(\theta \mid \mathbf{x}) d\theta; \tag{6.17}$$

Se, come spesso accade le osservazioni $\mathbf{y}$ rappresentano un successivo campione dalla stessa popolazione che ha generato il vettore $\mathbf{x}$, estratto in modo indipendente a $\mathbf{x}$, condizionatamente a θ, allora la (6.17) si semplifica nella

$$\pi(\mathbf{y} \mid \mathbf{x}) = \int_{\boldsymbol{\Omega}} p(\mathbf{y} \mid \theta) \pi(\theta \mid \mathbf{x}) d\theta. \tag{6.18}$$

Esempio 6.15 [*Distribuzione esponenziale*]
Sia $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ un campione estratto da una popolazione governata da una legge esponenziale di parametro θ e supponiamo che la legge a priori su θ sia di tipo non informativo, ovvero la legge di Jeffreys,

$$\pi^J(\theta) \propto \frac{1}{\theta}.$$

Per quanto visto a proposito delle distribuzioni coniugate sappiamo che la legge finale su θ è di tipo $Gamma(n, n\bar{x})$ con media pari al reciproco della media campionaria. Vogliamo ora calcolare la distribuzione predittiva di un'ulteriore osservazione Y estratta dalla stessa popolazione esponenziale. Avremo così

$$\begin{aligned} p(y \mid \mathbf{x}) &= \int_0^\infty \theta \exp\{-\theta\, y\} \frac{(n\bar{x})^n}{\Gamma(n)} \theta^{n-1} \exp\{-\theta\, n\bar{x}\} d\theta \\ &= \frac{(n\bar{x})^n}{\Gamma(n)} \int_0^\infty \theta^n \exp\{-\theta(y + n\bar{x})\} d\theta = \frac{n^{n+1} \bar{x}^n}{(y + n\bar{x})^{n+1}} \end{aligned} \tag{6.19}$$

La distribuzione (6.19) è nota come legge di Pearson di tipo VIII [51]

♢

Esempio 6.16 [*Distribuzione normale*].
Sia $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ un campione estratto da una popolazione governata da una legge normale con parametri incogniti (μ, σ^2) e supponiamo che la legge a priori sia di tipo non informativo, ovvero la legge di Jeffreys, $\pi^J(\mu, \sigma^2) \propto \sigma^{-2}$. Riparametrizzando in termini di precisione ($\psi^{-1} = \sigma^2$) si ottiene

$$\pi^J(\mu, \psi) \propto \frac{1}{\psi}.$$

Per quanto visto nella §4.3.2, sappiamo che la legge finale su (μ, σ^2) è di tipo Normale-Gamma con parametri $(\bar{x}, 1/\sqrt{n\psi}, (n-1)/2, nS^2/2)$, dove $\bar{x}$ ed S^2 rappresentano come di consueto la media e la varianza campionarie osservate. Vogliamo allora calcolare la distribuzione predittiva relativa ad un'ulteriore osservazione Y estratta dalla stessa popolazione normale e, condizionatamente ai

valori di (μ, ψ), indipendente dal vettore delle precedenti osservazioni. Utilizzando allora la (6.18) si ottiene

$$\begin{aligned} p(y \mid (x_1, \dots x_n)) &= \int_\psi \int_\mu p(y \mid \mu, \psi) \pi(\mu, \psi \mid (x_1, \dots x_n)) d\mu d\psi \\ &\propto \int_\psi \int_\mu \sqrt{\psi} \exp\left(-\frac{\psi}{2}(y-\mu)^2\right) \\ &\times \sqrt{\psi} \exp\left(-\frac{n\psi}{2}(\bar{x}-\mu)^2\right) \psi^{(n-1)/2-1} \exp\left(-\frac{nS^2\psi}{2}\right) d\mu d\psi. \end{aligned}$$

Facendo ricorso al lemma 4.7, le due forme quadratiche in μ possono essere facilmente riespresse in modo da risolvere l'integrale rispetto al parametro μ in forma esplicita.

$$\begin{aligned} p(y \mid (x_1, \dots x_n)) &= \int_0^\infty \psi^{n/2} \exp\left\{-\frac{\psi}{2}(nS^2 + (y-\bar{x})^2)\right\} d\psi \\ &\propto \left(\frac{nS^2}{2} + \frac{n}{2(n+1)}(y-\bar{x})^2\right)^{-(\frac{n}{2}+1)} \\ &\propto \left(1 + \frac{(y-\bar{x})^2}{S^2(n+1)}\right)^{-(\frac{n}{2}+1)}, \end{aligned}$$

riconoscibile come il nucleo di una distribuzione di tipo $\mathrm{St}_1(n+1, \bar{x}, S)$. Abbiamo dunque dimostrato che, in una situazione non informativa, la legge predittiva della $(n+1)$-esima osservazione Y da una distribuzione normale con parametri incogniti è una t di Student con (n+1) gradi di libertà e parametri di posizione e scala forniti dai valori campionari. In altri termini,

$$\frac{Y - \bar{x}}{S} \mid (x_1, \dots, x_n) \sim \mathrm{St}_1(n+1, 0, 1) \tag{6.20}$$

♢

Esempio 6.17 [*Dal giornalaio.*] Gigi lavora ogni giorno in un'edicola per quattro ore. Ogni giorno, in quell'intervallo di tempo, il numero di copie del quotidiano "Repubblica" vendute può essere considerata una v.a X di Poisson di parametro θ. Le v.a. relative ai diversi giorni possono essere considerate indipendenti condizionatamente al valore di θ. Gigi registra il numero di copie vendute in dieci giorni consecutivi, che risultano essere

$$\mathbf{X} = \texttt{16 27 10 32 14 27 23 21 18 19}.$$

L'undicesimo giorno, Gigi è costretto ad aprire l'edicola con un'ora di ritardo e vuole stimare con quale probabilità riuscirà a non perdere alcun cliente.

Sia dunque X_{11} il numero di copie di "Repubblica" vendute nelle quattro ore di lavoro del giorno in questione e sia Y il numero di clienti che arrivano nella prima ora. Per θ fissato, avremo che

$$X_{11} \sim \mathrm{Po}(\theta), \qquad \text{e } Y \sim \mathrm{Po}(\theta/4);$$

Gigi è interessato alla quantità $\Pr(Y = 0 \mid \mathbf{X})$ che possiamo scrivere come

$$\Pr(Y = 0 \mid \mathbf{X}) = \int_0^\infty \Pr(Y = 0 \mid \theta)\, \pi(\theta \mid \mathbf{x}) d\theta, \tag{6.21}$$

Per calcolare l'integrale occorre conoscere dunque la distribuzione finale di θ e, dunque, la sua legge iniziale. Assumiamo allora che, inizialmente, $\theta \sim \mathrm{Ga}(\delta, \lambda)$. La verosimiglianza associata all'osservazione nei primi $n = 10$ giorni è

$$L(\theta) \propto e^{-n\theta} \theta^{\sum x_i} = e^{10\theta} \theta^{207};$$

La legge finale di θ è ancora di tipo Gamma con parametri aggiornati $\delta^* = \delta + 207$, e $\lambda^* = \lambda + 10$. La (6.21) diventa

$$\int_0^\infty e^{-\theta/4} \frac{(\lambda^*)^{\delta^*}}{\Gamma(\delta^*)} e^{-\lambda^*\theta} \theta^{\delta^*-1} d\theta,$$

facilmente riconoscibile come la funzione generatrice dei momenti associata ad una legge $\mathrm{Ga}(\delta^*, \lambda^*)$ calcolata in $1/4$. Essa vale pertanto

$$\Pr(Y = 0 \mid \mathbf{X}) = \left(\frac{4\lambda^*}{4\lambda^* + 1}\right)^{\delta^*}.$$

Utilizzando a priori una legge "oggettiva" alla Jeffreys ($\delta = \lambda = 0$), il valore stimato è pari a 0.006 $\diamondsuit$

6.5 La modellizzazione gerarchica

Il peso della scelta della distribuzione a priori in una procedura d'inferenza può essere in qualche modo "diluito" qualora si costruisca il modello statistico bayesiano in modo gerarchico. Fino ad ora abbiamo sempre considerato la situazione in cui le osservazioni $\mathbf{x} = (x_1, \ldots, x_n)$ rappresentavano un campione i.i.d. estratto da una distribuzione $p(x; \boldsymbol{\theta})$, e che il vettore di parametri $\boldsymbol{\theta}$ fosse dotato, a sua volta, di una specifica legge di probabilità $\pi(\boldsymbol{\theta})$. Un modo per "indebolire" queste assunzioni è quello di assumere che il parametro θ sia distribuito secondo una legge $\pi(\boldsymbol{\theta}|\boldsymbol{\omega})$, dipendente cioè da un iperparametro $\boldsymbol{\omega}$, al quale, a sua volta viene associata una legge di probabilità, di secondo stadio, che indicheremo con $\xi(\boldsymbol{\omega})$. In questo modo, lo stesso modello statistico può essere rappresentato in vari modi, tutti equivalenti, ma in grado di mettere in risalto aspetti diversi del modello stesso. Lo schema sopra delineato può infatti essere reinterpretato affermando che il nostro campione $\mathbf{x}$ è una realizzazione n-pla dalla legge

$$p(\mathbf{x}; \boldsymbol{\omega}) = \int_{\boldsymbol{\Omega}} p(\mathbf{x}; \boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\boldsymbol{\omega}) d\boldsymbol{\theta} \tag{6.22}$$

con distribuzione a priori per $\boldsymbol{\omega}$ fornita da $\xi(\boldsymbol{\omega})$, oppure ancora potremo dire, più semplicemente, che il modello statistico è quello di partenza, rappresentato da $p(\mathbf{x}; \boldsymbol{\theta})$, mentre la legge a priori per $\boldsymbol{\theta}$ si scrive come

$$\pi(\boldsymbol{\theta}) = \int_{\Omega} \pi(\boldsymbol{\theta} \mid \boldsymbol{\omega}) \xi(\boldsymbol{\omega}) d\boldsymbol{\omega}$$

L'una o l'altra rappresentazione sono più o meno convenienti a seconda di quale sia, nella specifica applicazione, il reale parametro di interesse, $\boldsymbol{\theta}$ oppure $\boldsymbol{\omega}$. I motivi per cui è utile costruire uno schema gerarchico sono diversi. Nel capitolo 11 illustreremo il loro uso principale, quando il contesto sperimentale è tale per cui le osservazioni non sono del tutto scambiabili, perché magari rilevate in condizioni differenti. Qui ci limitiamo a sottolineare la loro efficacia per le situazioni in cui le informazioni a priori sono deboli e il processo di elicitazione viene così diluito in due stadi.

Esempio 6.18 [*Modello gerarchico normale.*]
Sia $\mathbf{x} = (x_1, \ldots, x_n)$ un campione di osservazioni i.i.d. con distribuzione $\mathrm{N}(\theta, \sigma_0^2)$ con σ_0^2 noto. La legge a priori su θ condizionatamente ad un iperparametro τ^2, è ancora di tipo $\mathrm{N}(0, \tau^2)$; infine

$$\tau^2 \sim \mathrm{GI}(\frac{\delta}{2}, \frac{\gamma}{2}). \tag{6.23}$$

Tale costruzione equivale di fatto a stabilire che la legge iniziale marginale per τ^2 è

$$\pi(\theta) = \int \pi(\theta|\tau)\xi(\tau^2)d\tau^2,$$

che, per il teorema E.1, sappiamo corrispondere ad una distribuzione di tipo $\mathrm{St}(\delta, 0, \gamma/\delta)$. In questo senso, la costruzione gerarchica, indebolendo la precisione elicitativa al primo stadio, ha prodotto una legge a priori, più robusta, con code più pesanti.
Lo stessa situazione può essere interpretata, in alternativa, tenendo conto che essa equivale a considerare $\mathbf{x}$ come una singola osservazione n dimensionale proveniente dalla legge

$$p(\mathbf{x}|\tau) = \int_{\boldsymbol{\Omega}} p(\mathbf{x}|\theta)\pi(\theta|\tau)d\theta \tag{6.24}$$

con distribuzione iniziale $\xi(\tau^2)$. Questo equivale a dire, per il Lemma 6.1, che la funzione di verosimiglianza associata a tale modello è

$$L(\tau^2; \mathbf{x}) \propto \int_{\boldsymbol{\Omega}} \exp\left\{-\frac{1}{2}\left(\frac{n(\bar{x}-\theta)^2}{\sigma_0} - \frac{\theta^2}{\tau^2}\right)\right\} d\theta \propto \sqrt{\frac{n}{\sigma_0^2 + n\tau^2}} \exp\left\{-\frac{n\bar{x}^2}{2(\sigma_0^2 + n\tau^2)}\right\}. \tag{6.25}$$

Da notare che questa nuova formulazione, elimina l'indipendenza condizionata delle osservazioni. Qualunque formulazione si scelga, l'espressione delle leggi marginali finali di τ^2 o di θ non hanno una forma esplicita calcolabile. Ad esempio, la legge marginale di τ è proporzionale al prodotto della (6.25) e della densità relativa alla (6.23), mentre la legge finale di θ, per la (4.5), è

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \int_{\tau^2} \pi(\theta|\tau^2, \mathbf{x})\xi(\tau^2|\mathbf{x})d\tau^2 \\ &\propto \frac{\sqrt{\sigma_0^2 + n\tau^2}}{\tau^2} \exp\left\{-\frac{\sigma_0^2 + n\tau^2}{2\,\sigma_0^2\tau^2}\left(\theta - \frac{n\bar{x}\tau^2}{\sigma_0^2 + n\tau^2}\right)^2\right\} L(\tau^2; \mathbf{x})\xi(\tau^2)d\tau^2\end{aligned}$$

In ogni caso, per produrre inferenze sintetiche da questo modello, occorre utilizzare metodi numerici: vedremo nel Capitolo 7 che per questo e per modelli molto più complessi, esiste una soluzione di tipo Monte Carlo basata sulle proprietà delle catene di Markov di estrema semplicità. ♢

L'idea che una costruzione gerarchica della distribuzione a priori conduca ad un indebolimento del peso delle informazioni iniziali può essere formalizzata nel seguente risultato [56]:

Teorema 6.2 *Dato il modello statistico gerarchico definito all'inizio di questa sezione, qualunque sia il risultato osservato* $\mathbf{x}$*, si ha che*

$$D_{KL}\left(\xi(\boldsymbol{\omega}); \xi(\boldsymbol{\omega} \mid \mathbf{x})\right) < D_{KL}\left(\pi(\theta); \pi(\boldsymbol{\theta} \mid \mathbf{x})\right) \tag{6.26}$$

Dimostrazione 6.2 *Si veda [56], pag. 260.*

Il significato del teorema (6.2) è chiaro. Se misuriamo il contenuto informativo dell'esperimento come distanza di Kulback-Leibler tra la legge iniziale e quella finale di un vettore di parametri, la (6.26) ci dice che l'esperimento fornisce più informazione sul parametro θ che non sull'iperparametro $\boldsymbol{\omega}$; in altri termini l'inferenza condotta su $\boldsymbol{\omega}$ risulta meno sensibile a variazioni nella distribuzione iniziale.

6.5.1 L'approccio bayesiano empirico

Da un punto di vista fondazionale, il modo in cui si è introdotto l'utilizzo della modellizzazione gerarchica è certamente eccentrico. Laddove l'impostazione bayesiana suggerisce di utilizzare la

distribuzione iniziale per inserire nell'analisi tutte le informazioni extra-sperimentali, lì si voleva al contrario limitare l'influenza di queste informazioni, per rendere l'inferenza, in qualche modo, condivisibile. Questo strada può essere percorsa in modo ancora più deciso rinunciando di fatto ad elicitare una legge iniziale per l'iperparametro $\boldsymbol{\omega}$ e sostituendo ad esso una stima ottenuta mediante le osservazioni stesse: si tratta del cosiddetto approccio bayesiano empirico parametrico, proposto inizialmente da [75]. In pratica si stima $\boldsymbol{\omega}$ mediante quel valore $\hat{\boldsymbol{\omega}}(\mathbf{x})$ che massimizza la (6.22), intesa qui come fosse una vera e propria funzione di verosimiglianza (spesso in letteratura $\hat{\boldsymbol{\omega}}(\mathbf{x})$ viene definito come *Type II Maximum Likelihood Estimator*: si veda [10]). L'approccio bayesiano empirico ha ricevuto molta considerazione soprattutto in ambito non bayesiano come strumento per la costruzione di stimatori "robusti". Ovviamente, non è necessario che la stima $\hat{\boldsymbol{\omega}}(\mathbf{x})$ sia ottenuta mediante la massimizzazione della (6.22). È possibile ottenere tale stima anche attraverso metodi alternativi come, ad esempio, il metodo dei momenti. Tratteremo più approfonditamente di questo prima nell'esempio 6.19 e poi nel capitolo 11.

Esempio 6.18 (continua).

Volendo ottenere una stima empirica del parametro τ^2, occorre massimizzare, rispetto a τ^2, la distribuzione marginale del campione fornita dalla (6.25). Per le proprietà d'invarianza delle stime di massima verosimiglianza è conveniente determinare il massimo della (6.25) rispetto a $\sigma_0^2 + n\tau^2$. Tenendo presente che $\sigma_0^2 + n\tau^2$ è positivo per definizione, si calcola allora facilmente che

$$\widehat{\sigma_0^2 + n\tau^2} = \max\left\{\sigma_0^2; n\bar{x}^2\right\},$$

A questo punto, utilizzando le(4.6), si ottiene che lo stimatore bayesiano empirico per θ è dato da

$$\mathbf{E}\theta|\mathbf{x} = \frac{\bar{\mathrm{x}}\,\mathrm{n}\hat{\tau}^2}{\widehat{\sigma_0^2 + \mathrm{n}\tau^2}} = \left(1 - \frac{\sigma_0^2}{\max\left\{\sigma_0^2; \mathrm{n}\bar{\mathrm{x}}^2\right\}}\right)\bar{\mathrm{x}};$$

è importante notare come la stima appena calcolata sia pari esattamente a 0 ogni volta che $\sigma_0^2 > n\bar{x}$. Come ulteriore esempio della flessibilità e dell'utilità pratica di un'impostazione gerarchica consideriamo una rilettura di un esempio discusso in [76], pag.482.

Esempio 6.19 [*Modello gerarchico per variabili dicotomiche.*]

[21] STORIELLA

I dati sono dunque rappresentabili come realizzazioni di variabili aleatorie indipendenti con distribuzione binomiale, ovvero

$$X_1, \cdots, X_n \overset{\mathtt{ind}}{\sim} \mathrm{Bin}(m_i, \theta_i), \quad i = 1, \cdots, n.$$

Le componenti del vettore $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$ rappresentano le "intenzioni di acquisto" da parte delle diverse unità osservate. È ragionevole pensare allora che le diverse θ_i, ognuna associata ad una diversa unità, rappresentino la risultante della somma di una componente individuale e di una componente comune, presente perché magari gli individui condividono riferimenti culturali o sociali comuni. In termini statistici questo si traduce nell'assunzione che le varie θ_I siano realizzazioni indipendenti dalla stessa distribuzione di probabilità. Si assume cioè che

$$\theta_1, \cdots, \theta_n \overset{\mathtt{iid}}{\sim} f(\theta),$$

ovvero esse risultano indipendenti ma tutte con la stessa legge di probabilità. Un'analisi bayesiana di questo tipo di modello dipende dal livello di elicitazione che siamo in grado di produrre sulla

legge f. Consideriamo quattro diversi scenari

1. Elicitazione completa. Si assume la conoscenza perfetta della legge f. Questo può avvenire quando l'oggetto di interesse è il vettore $\boldsymbol{\theta}$ e si vuole produrne la legge finale.

2. Elicitazione gerarchica. Si assume che la legge f appartenga ad una famiglia specifica, indicizzata da un iperparametro il quale, a sua volta, è dotato di legge iniziale $\xi(\cdot)$. Ad esempio, in questo esempio, potremmo assumere che che le componenti di $\boldsymbol{\theta}$ siano distribuite secondo una legge Beta(α, β), mentre gli iperparametri (α, β) sono dotati di una legge iniziale $\xi(\alpha, \beta)$. Questo approccio è ragionevole sia quando l'oggetto d'interesse è il vettore θ sia, e soprattutto, quando l'interesse verte sulla legge di probabilità di secondo livello, e in questo caso il parametro d'interesse diventa il vettore (α, β)

3. Approccio bayesiano empirico non parametrico. Si cerca di ottenere una stima non parametrica della legge f basata sulle osservazioni a disposizione.

3. Approccio bayesiano empirico parametrico. Si considera la costruzione di cui al punto 2, ma invece di elicitare la legge iniziale degli iperparametri (α, β), essi vengono "stimati" dai dati.

In tutti gli approcci, le osservazioni $x_1, \cdots, x_n$ sono indipendenti condizionatamente ai valori di $(\theta_1, \cdots, \theta_n)$, che a loro volta sono indipendenti e somiglianti con legge $f(\theta)$. Ne segue che la legge finale del vettore $\boldsymbol{\theta}$ è proporzionale a

$$\pi(\theta_1, \cdots, \theta_n) \propto \prod_{i=1}^{n} f(\theta_i)\theta^{x_i}(1-\theta)^{m_i - x_i};$$

Le elaborazioni di questa legge dipendono dalla natura della legge $f(\cdot)$. Nel caso 2, avremo ad esempio che, condizionatamente al valore degli iperparametri (α, β), la legge finale è un prodotto di densità di tipo Beta, ovvero

$$(\theta_1, \cdots, \theta_n \mid \mathbf{x}, \alpha, \beta) \overset{\text{ind}}{\sim} \text{Beta}(\alpha + x_i, \beta + m_i - x_i), \qquad i = 1, \cdots, n;$$

Questo fornisce semplici stime condizionate per la media e la varianza a posteriori delle componenti del vettore $\boldsymbol{\theta}$. Non esistono invece "comode" forme analitiche per la legge iniziale di (α, β). In un contesto simile [43] propongono l'uso della legge impropria

$$\xi(\alpha, \beta) \propto (\alpha + \beta)^{-\frac{5}{2}}$$

e discutono alcuni approcci numerici di stima per α e β. Il terzo approccio consiste nella stima empirica della legge f. Qui in pratica si assume che le singole osservazioni siano indipendenti con legge marginale

$$p(x_i) = \int_0^1 \binom{m_i}{x_i} \theta^{x_i}(1-\theta)^{m_i - x_i} f(\theta) d\theta; \tag{6.27}$$

si tratta dunque di un esempio di modello mistura di tipo non parametrico. L'approccio 4 è invece basato su una stima dei coefficienti (α, β) Assumiamo per semplicità che tutti i valori m_i siano pari ad m. In questo caso, la (6.27) assume la forma esplicita di una distribuzione Beta-Binomiale (vedi Appendice E), e la funzione di verosimiglianza per gli iperparametri (α, β) assume la forma

$$L(\alpha, \beta) \propto \prod_{i=1}^{n} \frac{B(\alpha + x_i, \beta + m - x_i)}{B(\alpha, \beta)}$$

Per ottenere una stima degli iperparametri occorre massimizzare numericamente $L(\alpha, \beta)$; alternativamente, tenendo conto dell'espressione (vedi appendice) dei momenti di una v.a. Beta-Binomiale,

si può determinare una stima col metodo dei momenti risolvendo il seguente sistema, dove si è posto $\bar{x} = \sum x_i/m$ e $s_x^2 = \sum(x_i - \bar{x})^2/m$,

$$\begin{cases} \bar{x} = m\alpha/(\alpha+\beta) \\ s_x^2 = m\alpha\beta/\left[(\alpha+\beta)(\alpha+\beta+1)\right] \end{cases}$$

♢

6.6 Cenni alla teoria delle decisioni

Abbiamo già accennato al fatto che l'impostazione bayesiana può essere formalmente inquadrata all'interno della teoria delle decisioni. Pur non essendo questa la strada percorsa in questo testo, è opportuno accennare brevemente al modo naturale in cui la metodologia bayesiana emerge come formalizzazione matematica del comportamento di un essere razionale quando egli debba prendere decisioni in condizioni di incertezza. Il lettore interessato può consultare il testo esauriente di [68]. Un modello di teoria delle decisioni può essere visto come un gioco tra lo statistico e la "Natura"; occorre definire i seguenti ingredienti:

- Un insieme $\boldsymbol{\Omega}$ degli stati di natura, ovvero lo spazio dei possibili valori che la Natura può scegliere.
- Un insieme $\mathcal{A}$, che include tutte le possibili azioni a che lo statistico può intraprendere, sulla base delle informazioni fornite dall'esperimento.
- Una funzione di perdita $L(\theta, a)$, $a \in \mathcal{A}$, $\theta \in \boldsymbol{\Omega}$, che associa una conseguenza numerica ad ogni coppia (θ, a).

Questi ingredienti, insieme al modello statistico, permettono di formulare un problema di inferenza come un gioco statistico in cui lo statistico deve scegliere, sulla base del risultato sperimentale $\mathbf{x}$, una azione $a(\mathbf{x})$ in modo da minimizzare una funzione obiettivo basata sulla perdita $L(\theta, a)$. In un contesto bayesiano, l'incertezza relativa al valore di θ viene gestita attraverso la legge finale $\pi(\theta \mid \mathbf{x})$; la decisione bayesiana ottima, allora, è, quando esiste, quel valore $a^* \in \mathcal{A}$ che minimizza il valore atteso, a posteriori, della perdita. In pratica, sia $W^*(a)$ il valore atteso a posteriori della perdita associata all'azione $a \in \mathcal{A}$

$$W^*(a) = \mathbb{E}\left[\mathrm{L}(\theta, \mathrm{a}(\mathbf{x})) \mid \mathbf{x}\right] = \int_{\boldsymbol{\Omega}} \mathrm{L}(\theta, \mathrm{a}(\mathbf{x}))\pi(\theta \mid \mathbf{x})\mathrm{d}\theta. \tag{6.28}$$

L'azione ottima bayesiana, rispetto ad una specifica legge iniziale π allora l'azione

$$a^\star = \arg\ \min_{\mathcal{A}} W^*(a). \tag{6.29}$$

Pur non essendo di centrale interesse in un contesto bayesiano, è spesso importante, soprattutto per un confronto tra le procedure frequentiste e bayesiane, introdurre i concetti di funzione di decisione e di "rischio frequentista" associato ad una funzione di decisione. Si chiama funzione di decisione un'applicazione $d : \mathcal{X} \to \mathcal{A}$, che, ad ogni possibile risultato sperimentale $\mathbf{x}$, associa una azione $a = d(\mathbf{x}) \in \mathcal{A}$. La famiglia di tutte le possibili funzioni d, viene denotata con $\mathcal{D}$. A differenza di quanto avviene in ottica bayesiana in cui l'azione ottima viene selezionata "condizionatamente

al risultato osservato", in ambito classico si tende a determinare una strategia di ottimizzazione pre-sperimentale, che garantisca un buon comportamento nel suo eventuale uso ripetuto: per questo risulta necessario selezionare una intera funzione di decisione $d^* \in \mathcal{D}$. La scelta di d^* non è tuttavia semplice in quanto non esiste un ordinamento naturale completo all'interno dello spazio $\mathcal{D}$. Per ovviare, ma solo parzialmente, a questo problema, si usa associare ad ogni $d \in \mathcal{D}$ una sorta di rischio ferquentista associato, definito da

$$R_d(\theta) = \int_{\mathcal{X}} L(\theta, d\mathbf{x}) p(\mathbf{x}; \theta) d\mathbf{x}. \tag{6.30}$$

Il rischio R_d, tuttavia, una funzione del parametro incognito θ e questo rende possibile che, date due funzioni di decisione d_1 e d_2, una possa avere rischio inferiore all'altra per alcuni θ mentre per altri valori del parametro la situazione ribaltata. **Esempio 6.20** [ ] È noto che, per una v.a. di Poisson, sia la media che la varianza assumono lo stesso valore θ. Si pu allora pensare, in ambito classico, di stimare θ sia attraverso l'uso della media campionaria oppure della varianza campionaria (eventualmente corretta). Supponiamo allora che l'esperimento consista nell'osservare $X_1, X_2, \ldots, X_n \overset{\texttt{iid}}{\sim} \text{Po}(\theta)$: si tratta allora di confrontare le decisioni

$$d_1(x_1, x_2, \ldots, x_n) = \frac{1}{n} \sum x_i = \bar{x},$$

e

$$d_2(x_1, x_2, \ldots, x_n) = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = s^2.$$

in termini di rischio. Assumiamo, per esemplificazione, di utilizzare una perdita quadratica (vedi §6.6); avremo allora che

$$R_{d_1}(\theta) = \mathbb{E}\left((\bar{X} - te)^2\right) = \text{Var}\left(\bar{X}\right) = \frac{\theta}{n}$$

mentre

$$R_{d_2}(\theta) = \mathbb{E}\left((S^2 - te)^2\right) = \text{Var}\left(S^2\right) = ???$$

♢

Per risolvere questa enpasse, una strategia frequentemente adottata (e mutuata dalla teoria dei giochi) è quella del *minimax*: ad ogni decisione d si associa una perdita definita come la massima perdita possibile al variare di θ, ovvero

$$d_M = sup_{\theta \in \Omega} R_d(\theta);$$

si sceglierà poi la decisione d^* che minimizza la quantità d_M al variare di d in $\mathcal{D}$

La strategia del minimax viola facilmente elementari principi di razionalità. Consideriamo il seguente esempio, artificiale ma significativo, tratto da [10]. Nella Figura **??**, primo riquadro, vengono rappresentate due diverse funzioni di rischio per le decisioni d_1 e d_2. In questo caso la strategia minimax (e probabilmente qualunque altra strategia ragionevole!) condurrebbe a scegliere la decisione d_2. Supponiamo ora di venire a sapere che, quando il parametro θ si trova nell'intervallo (a, b), allora si incorre, qualunque sia la decisione adottata in un'ulteriore perdita pari a 1. Le nuove funzioni di perdita sono rappresentate nel riquadro a destra della Figura **??**. Ora, la strategia minimax, selezionerebbe la decisione d_1 che fornisce una massima perdita inferiore rispetto a quella di d_2. Tale conclusione irragionevole: la strategia minimax seleziona una diversa decisione sulla base di informazioni esogene che modificano in modo omogeneo, le perdite di **tutte** le possibili decisioni.

Questi tipo di incoerenze logiche non si verificano in ambito bayesiano, ovvero quando si utilizza la (6.29) per la selezione dell'azione ottima.

Nella teoria delle decisioni, a volte appare, oltre alle formule del rischio frequentista (6.30) e della perdita attesa a posteriori (6.28), anche una forma ibrida di rischio, che può essere interpretato come il valore atteso atteso di una certa funzione di decisione rispetto a tutti i possibili risultati sperimentali e rispetto ad una specifica legge iniziale π. Si definisce allora "rischio di Bayes" la quantit

$$r(\pi, d) = \mathbb{E}_{\boldsymbol{\Omega}}\mathbb{E}_{\mathcal{X}}\left[\mathrm{L}(\theta, \mathrm{d}(\mathbf{x}))\right] = \int_{\Omega} \mathrm{R_d}(\theta)\pi(\theta)\mathrm{d}\theta. \tag{6.31}$$

in sostanza $r(\pi, d)$ considera il comportamento di una procedura decisionale sulla base del valore atteso della perdita rispetto alla distribuzione congiunta dei dati e del parametro. Sotto condizioni di regolarità molto generali, si può facilmente mostrare che

$$\begin{aligned} r(d) &= \int_{\boldsymbol{\Omega}} \left[\int_{\mathcal{X}} L(\theta, d(\mathbf{x}))p(\mathbf{x};\theta)d\mathbf{x}\right] \pi(\theta)d\theta \\ &= \int_{\mathcal{X}} \left[\int_{\boldsymbol{\Omega}} L(\theta, d(\mathbf{x}))\pi(\theta \mid \mathbf{x})d\theta\right] m(\mathbf{x})d\mathbf{x} \\ \int_{\mathcal{X}} W(a(\mathbf{x}))m(\mathbf{x})d\mathbf{x}, \end{aligned}$$

dove $m(\mathbf{x})$ rappresenta la legge marginale di $\mathbf{x}$, in pratica la quantità che appare al denominatore nell'espressione della legge a posteriori di θ. Dall'ultima espressione appare evidente che, come $m(\mathbf{x}) > 0$, per ogni $\mathbf{x} \in \mathcal{X}$, la funzione di decisione d^* che minimizza la (6.31) si ottiene come quella che fa corrispondere, ad ogni $bx \in \mathcal{X}$, l'azione a^* definita nella (6.29). Ne segue che, per ottenere una funzione di decisione "ottima" in senso classico quello di calcolare, per ogni $\mathbf{x} \in \mathcal{X}$, la decisione ottima bayesiana: si definisce così una funzione d^*_π tale che

$$d^*_\pi(\mathbf{x}) = a^*(\mathbf{x});$$

Ovviamente, tale procedura dipende dalla legge π iniziale. Si può dimostrare che l'insieme delle funzioni di decisione così ottenute forma, al variare di π, l'insieme delle decisioni *ammissibili*, ovvero quelle il cui rischio frequentista $R_d(\theta)$ non uniformemente non inferiore al rischio di una qualsiasi altra decisione.

Stima puntuale e decisioni

La funzione di perdita $L(\theta, a)$ è uno degli ingredienti essenziali per una formalizzazione decisionale e dovrebbe essere scelta sulla base del problema specifico. Per la natura introduttiva di questo paragrafo, ci limitiamo ad assumere funzioni di perdita notevoli. La più frequente è senza dubbio quella di tipo quadratico, ovvero

$$L(\theta, a) = (\theta - a)^2; \tag{6.32}$$

In questo modo la perdita è nulla se e solo se $\theta = a$ e le conseguenze peggiorano all'aumentare della distanza tra θ e a, indipendentemente dal segno della differenza tra i due valori. È facile dimostrare che, in caso di funzione di perdita (6.32), la decisione bayesiana ottima è data dalla media della distribuzione finale di θ, ovvero

$$a^\star = \mathbb{E}\,(\theta \mid \mathbf{x})\,.$$

La dimostrazione del precedente asserto è una diretta conseguenza di una delle più note proprietà del valore atteso e cioè che la media della distribuzione di una v.a. X è quella quantità c^* che minimizza il valore atteso di $(X-c)^2$.

Un'altra perdita spesso utilizzata è quella basata sulla distanza in valore assoluto, ovvero

$$L(\theta, a) = \mid \theta - a \mid; \tag{6.33}$$

la perdita (6.33) è ancora di tipo simmetrico ma, rispetto alla perdita quadratica, dà meno peso a grandi differenze tra θ ed a. La decisione ottima bayesiana, in presenza di perdita (6.33), è data dalla mediana della legge a posteriori. La dimostrazione di questo risultato è molto semplice. Limitandoci al caso reale , occorre determinare la quantità a^* che minimizza la quantità

$$\mathbb{E}\,(\mid \theta - a \mid \mid \mathbf{x}) = \int_\Omega \mid \theta - a \mid \pi(\theta \mid \mathbf{x}) d\mathbf{x};$$

ma, denotando con M una delle mediane della distribuzione[4], risulta

$$\mid \theta - a \mid = \mid \theta - M + M - a \mid = \begin{cases} \mid \theta - M \mid + M - a & \text{se } M > a \\ \mid \theta - M \mid + a - M & \text{se } M \leq a \end{cases};$$

Dunque

$$\mathbb{E}\,(\mid \theta - a \mid \mid \mathbf{x}) = \begin{cases} \mathbb{E}\,(\mid \theta - M \mid \mid \mathbf{x}) + (M - a) & \text{se } M > a \\ \mathbb{E}\,(\mid \theta - M \mid \mid \mathbf{x}) + (a - M) & \text{se } M \leq a \end{cases};$$

in entrambi i casi il primo addendo al secondo membro non dipende da a, mentre il secondo membro è sempre non negativo e si annulla, in entrambi i casi ponendo $a = M$.

In alcuni casi, potrebbe non essere ragionevole utilizzare una perdita simmetrica, in quanto una sovra-stima di θ potrebbe avere conseguenze più o meno serie di una sotto-stima. In questi casi la perdita (6.33) si generalizza nella perdita

$$L(\theta, a) = \begin{cases} k_1(\theta - a) & \text{se } \theta > a \\ k_2(a - \theta) & \text{se } \theta \leq a \end{cases}. \tag{6.34}$$

In questo caso è facile dimostrare che la decisione bayesiana ottima risulta essere il quantile di ordine $k_1/(k_1 + k_2)$ della legge a posteriori[5]

Esempio 6.21 [*Distribuzione di Poisson*]
Si osservano $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Po}(\theta)$. La verosimiglianza associata dunque

$$L(\theta) \propto e^{-n\theta} \theta^{\sum x_i}$$

Adottando la legge iniziale di Jeffreys, che nel caso Poisson risulta essere $\pi(\theta) \propto \theta^{-1/2}$, avremo un legge a posteriori di tipo $\text{Ga}(n, \sum x_i + 0.5)$. Supponiamo inoltre che, nel contesto specifico una sopra-stima del parametro θ abbia conseguenze meno gravi rispetto ad una sua sotto-stima; tali differenze vengono quantificate nella perdita

[4] ovviamente, nella quasi totalità dei casi, la legge a posteriori di θ risulta essere assolutamente continua e la mediana risulta univocamente determinata.

[5] se $k_1 = k_2$, si torna al caso precedente e la decisione ottima torna ad essere la mediana a posteriori.

$$L(\theta, a) = \begin{cases} 3(\theta - a) & \text{se } \theta > a \\ 2(a - \theta) & \text{se } \theta \leq a \end{cases}.$$

Per quanto detto, allora la decisione ottima bayesiana corrisponde al quantile di ordine $3/(3+2) = 0.6$ della legge finale; poiché in questo caso le conseguenze di una sotto-stima erano considerate meno più gravi, il metodo bayesiano seleziona, anziché la mediana, come nel caso simmetrico, un quantile di ordine superiore, in questo caso 0.6

♢

Verifica di ipotesi e decisioni

Poiché il risultato finale di una analisi bayesiana è raccolto nella distribuzione finale del parametro θ, che supponiamo per semplicità scalare, una sintesi *puntuale* di questa distribuzione può essere rappresentata da quel valore che minimizza, in media, una particolare distanza dalla distribuzione finale. Sia allora $D(\theta, \theta^*)$ una distanza generica tra un valore generico del parametro e il valore θ^* stimato. La stima bayesiana per θ è allora quel valore θ_B che minimizza il valor medio della distanza $D(\theta, \theta^*)$, ovvero

$$\theta_B = \arg \min_{\theta} .... \tag{6.35}$$

ESEMPIO DEL TEST DI IPOTESI due stati di natura due decisioni, caso NORMALE BILATERALE caso esponenziale point null

6.7 Esercizi

7

Metodi computazionali

7.1 Introduzione

Uno dei maggiori ostacoli alla diffusione dell'impostazione bayesiana tra coloro che concretamente applicano il metodo statistico è stato, storicamente, la necessità di elaborazioni computazionali non banali: al di fuori degli accademici modelli, con distribuzioni a priori coniugate, non è quasi mai possibile ottenere, in forma esplicita, la distribuzione finale di un parametro di interesse e, tanto meno, alcune sintesi di questa distribuzione: ad esempio, risolvere in modo analitico gli integrali necessari al calcolo di medie o mediane a posteriori risulta il più delle volte impossibile.

Esempio 7.1 [*Regressione logistica*].
FORSE OCCORRE USARE DEI DATI Supponiamo di osservare $Y_1, Y_2, \cdots, Y_n$, indipendenti tra loro e tali che, per $i = 1, \cdots, n$, $Y_i \sim Be(\theta_i)$, dove la probabilità di successo[1] per l'i-esimo individuo è considerata funzione di un insieme di covariate $\mathbf{x}_i = (x_{i1}, \cdots x_{ip})$, in modo che

$$\theta_i = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}},$$

dove $\boldsymbol{\beta}$ è un vettore di coefficienti di regressione il cui valore quantifica il ruolo delle covariate nella determinazione delle probabilità di successo. La funzione di verosimiglianza vale allora

$$\begin{aligned} L(\boldsymbol{\beta}; \mathbf{y}, \mathbf{x}) &= \prod_{i=1}^{n} \theta_i^{y_i} (1-\theta_i)^{1-y_i} = \prod_{i=1}^{n} \left[\frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}\right]^{y_i} \left[\frac{1}{1 + \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}\right]^{1-y_i} \\ &= \exp\left\{\sum_{i=1}^{n} \left[y_i \mathbf{x}_i'\boldsymbol{\beta} - \log(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}})\right]\right\} \end{aligned} \tag{7.1}$$

Per semplicità consideriamo una legge a priori impropria uniforme su $\boldsymbol{\beta}$, $\pi(\boldsymbol{\beta}) \propto 1$. La distribuzione a posteriori allora è proporzionale alla (7.1) e non è possibile, con mezzi analitici, né calcolare la costante di normalizzazione né tantomeno calcolare indici sintetici. ♢

Fino agli anni '80 del secolo scorso, questo problema ha rappresentato il collo di bottiglia per lo sviluppo e la diffusione della statistica bayesiana, anche in settori applicativi che ne riconoscevano, sul piano teorico, l'estremo interesse e le notevoli potenzialità. Per molti anni, dunque, l'unica alternativa all'utilizzo di modelli troppo semplificati e poco adatti a spiegare fenomeni complessi, è stata rappresentata dall'uso di approssimazioni analitiche, basate essenzialmente sul teorema del limite centrale: di questo tratteremo nella §7.2.1. Con l'avvento del personal computer e il conseguente (e

[1] Per una trattazione più articolata di questo modello si veda il § ??

ancora in corso) sviluppo di metodi di calcolo numerico, i metodi analitici hanno lasciato sempre più spazio a tecniche di integrazione numerica, le quali, per dimensioni del parametro non eccessive (in genere, non più di dieci), ancora oggi rappresentano una valida soluzione. Ma la vera rivoluzione nella pratica statistica si è compiuta negli ultimi 15 anni, prima con l'utilizzo delle tecniche di tipo Monte Carlo, e più ancora con i metodi Monte Carlo basati sulle proprietà delle catene di Markov, che d'ora in poi indicheremo con l'acronimo MCMC. Queste nuove metodologie hanno in pratica consentito che l'approccio bayesiano non solo diventasse possibile anche con modelli particolarmente complessi, ma che, anzi, rappresentasse in molte occasioni il migliore, se non *l'unico* approccio possibile. In questo capitolo descriveremo brevemente le idee e le tecniche più importanti che si sono imposte negli ultimi anni, facendo per lo più uso di esempi per descrivere i vari metodi introdotti. Nella §7.3 verranno introdotti i metodi inferenziali basati sulla simulazione a posteriori, insistendo soprattutto sui vantaggi che essi offrono rispetto ad un approccio, quand'anche disponibile, completamente analitico. Nelle §7.5 e §7.7 verranno discussi i metodi oggi più utilizzati: la natura introduttiva del testo non consente una trattazione esaustiva dell'argomento. Per maggiori approfondimenti il lettore interessato può consultare, ad esempio, [21] o, più recente, [42].

Prima di addentrarci nelle descrizione delle tecniche, è bene comunque cercare di tracciare le linee principali lungo le quali l'uso dei metodi di simulazione si è sviluppato all'interno della metodologia bayesiana. Il problema di base, come già accennato, è stato quello di determinare distribuzioni finali in contesti complessi, o almeno, di calcolare valori attesi associati alla distribuzione finale: questi valori attesi, dal punto di vista matematico sono integrali del tipo

$$V = \int_{\boldsymbol{\Omega}} g(\theta)\pi(\theta \mid \mathbf{x})d\theta\,; \tag{7.2}$$

nelle elaborazioni bayesiane, integrali così compaiono un po' ovunque, ad esempio per il calcolo del k-esimo momento a posteriori (con $g(\theta) = \theta^k$), o per determinare la probabilità finale di un insieme B: in tal caso $g(\theta) = \mathbf{1}_B(\theta)$. Altre forme integrali appaiono nel calcolo del fattore di Bayes (cfr. §6.3). Laddove la dimensione parametrica sconsigli l'uso di tecniche di integrazione numerica, è pratica comune utilizzare metodi di tipo Monte Carlo o i cosiddetti metodi Monte Carlo Importance Sampling (cfr. §7.5), che, in grandi linee, consistono nello riscrivere l'integrale (7.2) come

$$V = \int_{\boldsymbol{\Omega}} \frac{g(\theta)}{f(\theta)}\pi(\theta \mid \mathbf{x})f(\theta)d\theta,$$

per una opportuna densità di probabilità $f(\theta)$, da scegliere in base ad alcune caratteristiche (cfr. §7.5) e poi approssimare la (7.2), generando un campione $\mathbf{w}$ di valori pseudo casuali $w_1, w_2, \cdots, w_M$ (con M in genere molto grande) dalla legge f con la quantità

$$\hat{V} = \frac{1}{M}\sum_{h=1}^{M} \frac{g(w_h)}{f(w_h)}\pi(w_h \mid \mathbf{x})$$

Non sempre questi metodi sono direttamente applicabili: ad esempio potrebbe non essere disponibile, in forma analitica la distribuzione finale $\pi(w \mid \mathbf{x})$ correttamente normalizzata oppure potrebbe non essere possibile ottenere un campione pseudo-casuale $\mathbf{w}$ da una legge f sufficientemente *affidabile*. In questi casi occorre allora generare il campione $\mathbf{w}$ secondo delle tecniche, basate sulle proprietà delle catene di Markov, che solo in termini asintotici garantiscono che il campione stesso $\mathbf{w}$ possa essere considerato un campione di valori generati dalla legge f. Questi argomenti sono ancora

oggi argomento di ricerca avanzata, che coinvolge competenze di probabilità e analisi matematica: in questo testo ci limiteremo ad una illustrazione delle tecniche più affermate, che verranno illustrate soprattutto attraverso degli esempi.

7.2 Approssimazioni analitiche

La possibilità di utilizzare il metodo bayesiano in situazioni complesse è stata, come detto, molto limitata, prima dell'avvento delle tecniche Monte Carlo: le procedure bayesiane comportano un uso notevole di calcolo integrale, ad esempio per ottenere momenti a posteriori dei parametri, o per determinare distribuzioni marginali o predittive, e quasi mai tali integrali possono essere risolti in forma analitica. Oltre alla possibilità di ricorrere a tecniche di integrazione numerica, una strada molto praticata fino agli anni '80 del secolo scorso è stata quella delle approssimazioni analitiche della distribuzione a posteriori, basate su risultati di tipo asintotico e quindi utilizzabili solo per grandi campioni. In questo paragrafo ci limiteremo ad descrivere in modo euristico i metodi di approssimazione gaussiana della distribuzione finale che, sotto condizioni piuttosto generali, conducono alla tecnica di approssimazione di integrali detta di Laplace. Concluderemo con alcuni cenni a tecniche più accurate, basate sui cosiddetti sviluppi in serie di Edgeworth e del *punto di sella*: il lettore interessato può fare riferimento a [72], [66] oppure, per un contesto più propriamente bayesiano, [85].

7.2.1 Comportamento asintotico della distribuzione finale

Consideriamo dunque la seguente situazione: siano $Y_1, \cdots, Y_n$ n replicazioni di una v.a. con densità $p(y \mid \boldsymbol{\theta})$ dove $\boldsymbol{\theta}$ è un parametro k-dimensionale la cui distribuzione iniziale indichiamo con $\pi(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^k$. Allora la distribuzione finale può scriversi come

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto \exp\left\{\log \pi(\boldsymbol{\theta}) + \log p(\mathbf{y} \mid \boldsymbol{\theta})\right\},$$

dove $\mathbf{y} = (y_1, \cdots, y_n)$. Sviluppando i due logaritmi in serie di Taylor intorno ai due rispettivi massimi m_0 e $\hat{\boldsymbol{\theta}}_n$, si ottiene

$$\log \pi(\boldsymbol{\theta}) = \log \pi(m_0) - \frac{1}{2}(\boldsymbol{\theta} - m_0)' \mathbf{H}_0^{-1} (\boldsymbol{\theta} - m_0) + R_0$$

e

$$\log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y} \mid \hat{\boldsymbol{\theta}}_n) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)' \mathbf{H}(\hat{\boldsymbol{\theta}}_n)^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) + R_n$$

dove

$$\mathbf{H}_0^{-1} = \left(-\frac{\partial^2 \log \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \right) \Big|_{\boldsymbol{\theta} = m_0}, \quad \text{e} \quad \mathbf{H}(\hat{\boldsymbol{\theta}}_n)^{-1} = \left(-\frac{\partial^2 \log p(\mathbf{y} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \right) \Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_n}$$

sono le inverse delle matrici di varianza e covarianza per le approssimazioni gaussiane della legge a priori e della funzione di verosimiglianza; inoltre R_0 e R_n rappresentano i termini residuali di secondo ordine. Assumendo condizioni di regolarità che rendono R_0 e R_n trascurabili si ottiene, per n grande e trascurando costanti che non dipendono da $\boldsymbol{\theta}$,

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - m_0)' \mathbf{H}_0^{-1} (\boldsymbol{\theta} - m_0) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)' \mathbf{H}(\hat{\boldsymbol{\theta}}_n)^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) \right\}$$

$$\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-m_n)'\mathbf{H}_n^{-1}(\boldsymbol{\theta}-m_n)\right\}$$

dove

$$H_n^{-1} = \mathbf{H}_0^{-1} + \mathbf{H}(\hat{\boldsymbol{\theta}}_n)^{-1}$$

e, in virtù del lemma (C.1) sulla combinazione di forme quadratiche (vedi §C.4.1),

$$m_n = \mathbf{H}_n\left(\mathbf{H}_0^{-1}m_0 + \mathbf{H}(\hat{\boldsymbol{\theta}}_n)^{-1}\hat{\boldsymbol{\theta}}_n\right).$$

L'espressione precedente stabilisce che, al crescere di n, la distribuzione finale assume una forma gaussiana intorno a m_n con matrice di varianze e covarianze $\mathbf{H}_n$. A loro volta,

$$\lim_{n\to\infty}\frac{m_n}{\hat{\boldsymbol{\theta}}_n} = 1$$

e, per la legge forte dei grandi numeri, e per ogni $i, j = 1, \cdots, k$,

$$\lim_{n\to\infty}\left\{\frac{1}{n}\left(-\frac{\partial^2 \log p(\mathbf{y}\mid\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}\right)\right\} = \lim_{n\to\infty}\left\{\frac{1}{n}\sum_{k=1}^{n}\left(-\frac{\partial^2 \log p(x_k\mid\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}\right)\right\}$$

$$= \int p(\mathbf{y}\mid\boldsymbol{\theta})\left(-\frac{\partial^2 \log p(\mathbf{y}\mid\boldsymbol{\theta}}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}\right)dx.$$

Ne segue che $\mathbf{H}(\hat{\boldsymbol{\theta}}_n) \to nI(\hat{\boldsymbol{\theta}}_n)$ dove $I(\boldsymbol{\theta})$ è la ben nota informazione attesa di Fisher relativa ad una osservazione (cfr. §2.5).

Possiamo allora concludere che, sotto condizioni di regolarità piuttosto generali.

$$\boldsymbol{\theta}\mid\mathbf{y} \approx N\left(\hat{\boldsymbol{\theta}}_n, \mathbf{H}(\hat{\boldsymbol{\theta}}_n)^{-1}\right)$$

oppure

$$\boldsymbol{\theta}\mid\mathbf{y} \approx N\left(\hat{\boldsymbol{\theta}}_n, \frac{1}{n}I(\hat{\boldsymbol{\theta}}_n)^{-1}\right).$$

I risultati ora illustrati si estendono facilmente al caso in cui le osservazioni risultino indipendenti ma non somiglianti.

Esempio 7.1 (continua).
Utilizzando una legge a priori uniforme, la moda a posteriori coincide con la stima di massima verosimiglianza $\hat{\beta}$. Inoltre, per $j = 1, \cdots, p$,

$$\frac{\partial}{\partial\beta_j}\log\pi(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n}\left(y_i - \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)x_{ij},$$

e, per $j, k = 1, \cdots, p$,

$$\frac{\partial^2}{\partial\beta_j\beta_k}\log\pi(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n}x_{ij}x_{ik}\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}};$$

Possiamo allora scrivere il vettore delle derivate prime come

$$\frac{\partial}{\partial\beta}\log\pi(\boldsymbol{\beta}|\mathbf{y}) = \mathbf{X}'(\mathbf{y}-\boldsymbol{\theta})$$

e la matrice delle derivate seconde come

$$H(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{V}(\boldsymbol{\beta})\mathbf{X},$$

dove V è una matrice diagonale con elemento generico

$$v_{ii} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{\left(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}\right)^2}$$

Per n grande la distribuzione finale convergerà dunque ad una normale multivariata con vettore delle medie pari a $\hat{\beta}$ e matrice di varianze e covarianze pari a $\left(\mathbf{X}'\mathbf{V}(\hat{\boldsymbol{\beta}})\mathbf{X}\right)^{-1}$.

7.2.2 Metodo di Laplace

La tecnica precedente è alla base del metodo di Laplace per le approssimazioni di integrali del tipo 7.2, o più in generale di integrali del tipo

$$\int_B f(\boldsymbol{\theta}, \mathbf{y}) d\boldsymbol{\theta}\,,$$

con f non negativa e, ovviamente, integrabile. La tecnica consiste nell'esprimere $f(\boldsymbol{\theta}, \mathbf{y})$ in forma esponenziale del tipo

$$f(\boldsymbol{\theta}, \mathbf{y}) = \exp\{nh(\boldsymbol{\theta}, \mathbf{y})\},$$

per una opportuna funzione $h(\cdot)$, dove n rappresenta in genere la dimensione campionaria; si sviluppa poi la funzione $h(\cdot)$ in serie di Taylor fino al secondo ordine intorno al punto di massimo, $\hat{\theta}_{\mathbf{y}}$, ottenendo

$$\int_B e^{nh(\boldsymbol{\theta},\mathbf{y})} d\boldsymbol{\theta} = e^{nh(\hat{\boldsymbol{\theta}}_{\mathbf{y}},y)} \int_B e^{-\frac{n}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\mathbf{y}})' H(\hat{\boldsymbol{\theta}}_{\mathbf{y}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\mathbf{y}})} d\boldsymbol{\theta}\,(1 + O(1/n))\,,$$

che rappresenta l'integrale di una funzione proporzionale alla densità di una normale multivariata. Così, ad esempio, quando la zona d'integrazione è tutto lo spazio $\mathbf{R}^k$

$$\int_{\mathbb{R}^k} e^{nh(\boldsymbol{\theta},\mathbf{y})} d\boldsymbol{\theta} = e^{nh(\hat{\boldsymbol{\theta}}_{\mathbf{y}},\mathbf{y})} \sqrt{\frac{(2\pi)^k}{n}} \frac{1}{\sqrt{\det(H(\hat{\boldsymbol{\theta}}_{\mathbf{y}}))}} (1 + O(1/n))\,.$$

L'approssimazione di Laplace fornisce dunque approssimazioni del primo ordine. Qualora però essa venga utilizzata per approssimare un rapporto tra due integrali, allora è possibile dimostrare [85] che il livello di accuratezza migliora e si ottiene un'approssimazione di secondo ordine. Questo è particolarmente utile in ambito bayesiano dove le stime finali di una grandezza si esprimono spesso attraverso rapporti di integrali. Consideriamo ad esempio il caso in cui si voglia approssimare il valore della media a posteriori di una funzione $g(\theta)$,

$$\mathbf{E}\,(g(\theta|\mathbf{y}) = \frac{\int g(\theta)\pi(\theta)L(\theta;\mathbf{y})d\theta}{\int \pi(\theta)L(\theta;\mathbf{y})d\theta}.$$

Se utilizzassimo sia al numeratore che al denominatore la tecnica di Laplace con

$$nh(\theta) = \pi(\theta)L(\theta;\mathbf{y}),$$

è facile vedere che otterremmo

$$\mathbf{E}\,[g(\theta|\mathbf{y})] = g(\hat{\theta})\,(1 + O(1/n))\,,$$

dove $\hat{\theta}$ è il punto di massimo di $nh(\theta)$. Con un semplice accorgimento è invece possibile ottenere stime più accurate. Supponiamo inizialmente che $g(\theta) > 0$: allora la funzione integranda al numeratore può essere espressa come

$$\int g(\theta)\pi(\theta)L(\theta;\mathbf{y}) = \exp\{\log g(\theta) + nh(\theta)\} = \exp\{n\tilde{h}(\theta)\}.$$

Nulla modificando nel trattamento del denominatore, è possibile allora dimostrare che

$$\mathbf{E}\left[g(\theta|\mathbf{y})\right] = \frac{\det(\tilde{H})^{\frac{1}{2}}\exp\{n\tilde{h}(\tilde{\theta})\}}{\det(\hat{H})^{\frac{1}{2}}\exp\{nh(\hat{\theta})\}}\left(1+O(1/n^2)\right),$$

dove $\tilde{\theta}$ è il punto di minimo della funzione $n\tilde{h}(\theta)$ e $\tilde{H}$ è la matrice delle derivate seconde della stessa funzione calcolate nel punto di massimo. In questo modo, in pratica, le componenti di ordine $O(1/n)$ si semplificano e l'approssimazione diviene più accurata. Quando $g(\theta)$ non è positiva per ogni θ, sarà sufficiente aggiungere a $g(\theta)$ una costante c tale che risulti $g(\theta)+c>0$ uniformemente, applicare il metodo di Laplace, e poi sottrarre la stessa costante c al risultato ottenuto.

Esempio 7.1 (continua).

Riprendiamo in esame il modello di regressione logistica e supponiamo di voler approssimare il valore a posteriori di di un singolo coefficiente di regressione, ad esempio β_1; avremo così

$$\mathbf{E}\left[g(\boldsymbol{\beta}|\mathbf{y})\right] = \mathbf{E}\left(\beta_1|\mathbf{y}\right) = \frac{\int_{\boldsymbol{\beta}}\beta_1 L(\boldsymbol{\beta};\mathbf{x},\mathbf{y})d\boldsymbol{\beta}}{\int_{\boldsymbol{\beta}} L(\boldsymbol{\beta};\mathbf{x},\mathbf{y})d\boldsymbol{\beta}};$$

??????????????????????????????????

La tecnica di Laplace può essere utilizzata anche per il calcolo approssimato della distribuzione marginale del vettore delle osservazioni (3.3). Supponiamo, in tutta generalità, che $\boldsymbol{\Omega} = \mathbb{R}^k$: ponendo

$$m(\mathbf{y}) = \int_{\boldsymbol{\Omega}}\pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\theta = \int_{\boldsymbol{\Omega}}\exp\left\{n\left(\frac{1}{n}\log\pi\boldsymbol{\theta}) + \frac{1}{n}\log p(\mathbf{y}|\boldsymbol{\theta})\right)\right\}d\theta,$$

e quindi

$$h(\boldsymbol{\theta}) = \frac{1}{n}\log\pi(\boldsymbol{\theta}) + \frac{1}{n}\log p(\mathbf{y}|\boldsymbol{\theta}),$$

l'approssimazione di Laplace fornisce

$$m(\mathbf{y}) = \sqrt{\frac{(2\pi)^k}{n}}\frac{1}{\sqrt{\det(H(\tilde{\boldsymbol{\theta}}_n))}}\pi(\tilde{\boldsymbol{\theta}}_n)L(\tilde{\boldsymbol{\theta}}_n,\mathbf{y})\left(1+O(1/n)\right), \tag{7.3}$$

dove $\tilde{\boldsymbol{\theta}}_n$ rappresenta la moda a posteriori e

$$H(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(\boldsymbol{\theta}).$$

Tenendo conto che la quantità $\det(H(\tilde{\boldsymbol{\theta}}_n))^{\frac{1}{2}}$ è trascurabile a livello di primo ordine e che, per n grande, la moda a posteriori è approssimabile dalla stima di massima verosimiglianza[2], un'approssimazione del primo ordine del logaritmo di $m(\mathbf{y})$ è data da

$$\log m(\mathbf{y}) = \frac{k}{2}\log(2\pi) + \frac{k}{2}\log n + \ell(\hat{\boldsymbol{\theta}}) + \log\pi(\hat{\boldsymbol{\theta}})\left(1+O(1/n)\right)$$

[80] propose di trascurare l'ultimo fattore, dipendente solo dalla distribuzione a priori e quindi, in qualche modo, poco influente, per grandi dimensioni campionarie. Eliminando anche il primo fattore, comune a qualunque modello, si arriva a stabilire la cosiddetta approssimazione BIC, utilizzata per produrre un fattore di Bayes approssimato, in pratica un indice pseudo bayesiano per il confronto tra modelli statistici. Più esattamente si definisce

$$BIC = -2\log m(\mathbf{y}) = k\log n - \ell(\hat{\boldsymbol{\theta}}_n,\mathbf{y}) + O(\frac{1}{n}). \tag{7.4}$$

Ritorneremo su questi aspetti nella § 8.3.1. Per ulteriori approfondimenti si veda [52].

[2] nel senso che $\tilde{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n + O(1/n)$.

7.2.3 Altri tipi di approssimazioni

Un'alternativa al metodo di Laplace è fornita dalla tecnica del *punto di sella*, che risulta particolarmente efficace quando occorre approssimare una funzione espressa attraverso un integrale, del tipo

$$g(\psi) = \int_A f(\theta, \psi) d\psi,$$

e la funzione deve essere valutata per diversi valori di ψ. L'idea di fondo è di utilizzare la tecnica di Laplace, cambiando però il punto intorno al quale si opera lo sviluppo di Taylor, che in questo caso dipenderà da ψ, e che denoteremo con $\hat{\theta}_\psi$. In tal modo si aumenta certamente la precisione dell'approssimazione, ma anche l'onere computazionale.

Esempio 7.2 [*Densità t_ν come mistura di normali*]
È noto che la densità di una v.a. t di Student con $\nu \geq 1$ gradi di libertà può esprimersi come una mistura, rispetto al parametro di scala, di distribuzioni normali con media nulla, e utilizzando come densità misturante, quella di χ^2_ν. In altri termini, si può dimostrare che

$$f_\nu(t) = \int_0^\infty \frac{\sigma}{\sqrt{2\pi\nu}} \exp\left\{-\frac{t^2\sigma^2}{2\nu}\right\} \frac{\sigma^{\nu-2} exp\{-\sigma^2/2\}}{\Gamma\left(\frac{\nu}{2}\right) 2^{\frac{\nu}{2}}} d\sigma^2.$$

♢

LIBRO SEVERINI
SADDLE POINT
EDGEWORTH

7.3 Simulazione a posteriori

L'utilizzo dei metodi basati sulla simulazione a posteriori è praticabile solo quando la distribuzione finale del parametro d'interesse è di una tipologia per la quale è possibile generare facilmente valori pseudo-casuali da quella distribuzione. L'utilità di queste tecniche è ben chiarita da un esempio.

Esempio 7.3 [*Calcolo dell'*odds ratio.]
Supponiamo di avere due campioni indipendenti di pazienti di dimensione n_1 ed n_2; ai membri del primo campione viene somministrato il farmaco **F** per la cura di un particolare tipo di allergia, mentre ai membri del secondo campione viene somministrato un placebo **P**. Il risultato della prova consiste nel contare quanti individui (rispettivamente denotati con k_1 e k_2), nei due campioni, ottengono un risultato positivo dall'uso di **F** o **P**. L'obiettivo è la stima dell'efficacia relativa del farmaco **F** rispetto al placebo **P**. Potendo assumere che ogni individuo nel primo campione rappresenti una prova bernoulliana con probabilità di successo pari a θ_1 e ogni individuo a cui si somministra il placebo rappresenti una prova bernoulliana con probabilità di successo pari a θ_2, l'obiettivo dell'inferenza si formalizza attraverso una stima di una qualche funzione della distanza tra θ_1 e θ_2, ad esempio la loro differenza, o il loro rapporto. Anche per motivi storici, legati alle approssimazioni discusse nei paragrafi precedenti, tra gli epidemiologi è comune l'utilizzo della quantità

$$\psi = \log\left(\frac{\theta_1}{1-\theta_1}\frac{1-\theta_2}{\theta_2}\right).$$

denominata *log odds ratio*. La logica sottostante alla scelta di ψ quale parametro di interesse risiede nella statistica classica; è possibile dimostrare, infatti, che l'equivalente campionario di ψ, diciamo

$$T = \log\left(\frac{\hat{\theta}_1}{1-\hat{\theta}_1}\frac{1-\hat{\theta}_2}{\hat{\theta}_2}\right), \tag{7.5}$$

ha distribuzione, sullo spazio campionario, asintoticamente normale con media pari a ψ e varianza pari a

$$\sigma_T^2 = \frac{1}{k_1} + \frac{1}{n_1 - k_1} + \frac{1}{k_2} + \frac{1}{n_2 - k_2}. \tag{7.6}$$

Questo risultato consente di ottenere stime puntuali e per intervallo del parametro di interesse ψ, almeno per grandi campioni. [47] ha ottenuto la versione bayesiana del risultato suddetto dimostrando che, qualora si scelga come distribuzione iniziale per θ_1 e θ_2 la distribuzione impropria (detta, appunto, di Haldane)

$$\pi_H(\theta_1, \theta_2) \propto \frac{1}{\theta_1(1-\theta_1)}\frac{1}{\theta_2(1-\theta_2)}, \tag{7.7}$$

allora la distribuzione finale di ψ ha distribuzione asintotica normale con media pari alla quantità (7.5) osservata sul campione e varianza come nella (7.6). È bene sottolineare che né la distribuzione campionaria di T né la distribuzione finale di ψ (qualunque sia la legge iniziale, purché diffusa) sono ottenibili in forma esplicita, per valori finiti della dimensione campionaria. Per dettagli sulla stima del *log odds ratio* si veda [55].

Vediamo ora come procedere attraverso la tecnica della simulazione a posteriori. I dati in nostro possesso possono essere formalizzati nel modo seguente:

- *numero di successi nel primo campione:* $K_1 \sim Bin(n_1, \theta_1)$
- *numero di successi nel primo campione:* $K_2 \sim Bin(n_2, \theta_2)$
- *distribuzioni iniziali su* θ_1 *e* θ_2*:* Si assumono i due parametri indipendenti a priori e inoltre

$$\theta_1 \sim Beta(\alpha_1, \beta_1), \qquad \theta_2 \sim Beta(\alpha_2, \beta_2)$$

Come già osservato in precedenza la scelta di distribuzioni iniziali di tipo Beta in problemi con dati dicotomici non è obbligatoria ma semplifica notevolmente i calcoli e garantisce una buona flessibilità: gli iperparametri possono essere fissati a piacimento, in base alle informazioni disponibili; ad esempio essi possono essere posti tutti pari a 0.5, come nel caso non informativo. Ne segue, per quanto visto nella §4.1, che le distribuzioni finali di θ_1 e θ_2 sono

$$\theta_1 \mid k_1 \sim Beta(\alpha_1 + k_1, \beta_1 + n_1 - k_1), \qquad \theta_2 \mid k_2 \sim Beta(\alpha_2 + k_2, \beta_2 + n_2 - k_2)$$

Poiché è oggi elementare, utilizzando un qualsiasi pacchetto statistico, generare valori pseudo casuali da una distribuzione di tipo Beta, possiamo dunque:

- generare un numero enorme M di valori pseudo casuali dalle leggi a posteriori di θ_1 e θ_2, che chiameremo
$$\theta_1^{(1)}, \theta_1^{(2)}, \cdots, \theta_1^{(M)}; \quad \theta_2^{(1)}, \theta_2^{(2)}, \cdots, \theta_2^{(M)};$$
- per ogni $i = 1, \cdots, M$, calcolare
$$\psi^{(i)} = \log\left(\frac{\theta_1^{(i)}}{1-\theta_1^{(i)}}\frac{1-\theta_2^{(i)}}{\theta_2^{(i)}}\right).$$

L'insieme dei valori $\psi^{(1)}, \psi^{(2)}, \cdots, \psi^{(M)}$ rappresenta dunque un campione di valori pseudo casuali generati dalla legge finale del parametro d'interesse ψ: sarà allora sufficiente scegliere M grande abbastanza per ottenere una *stima* buona quanto si voglia della legge finale di ψ. Consideriamo ora un esempio numerico: Siano $n_1 = n_2 = 10$, e si osservi $k_1 = 6$ e $k_1 = 4$. Assumiamo inoltre di utilizzare leggi a priori indipendenti e non informative [formula (5.5)][3] per i parametri θ_1 e θ_2. In questo caso, le dimensioni campionarie esigue non suggeriscono l'uso di metodi basati su approssimazioni asintotiche. Il metodo della simulazione a posteriori conduce ad una stima puntuale di ψ, attraverso la media a posteriori, pari a 0.805, mentre una stima dell'incertezza è fornita dalla varianza a posteriori pari a 0.821. È facile anche calcolare un'intervallo di credibilità finale per ψ costituito da tutti i valori compresi tra il 2.5 percentile e il 97.5 percentile del campione simulato di valori di ψ. Nel nostro caso si ottiene $(-0.956, 2.644)$. A titolo di confronto vengono riportati anche i risultati ottenuti con gli approcci classico e bayesiano basato sulla approssimazione di Haldane. In entrambi i casi si ottiene una stima puntuale per ψ pari a 0.810 e una varianza (nel primo caso relativa alla distribuzione di T, nel secondo caso relativa alla distribuzione finale di ψ) pari a 0.833. Il codice in **R** relativo a tale procedura è riportato nella § 4.5.1.
Ma il grande vantaggio dei metodi basati sulla simulazione a posteriori rispetto alle approssimazioni analitiche non risiede tanto nella tutto sommato piccola differenza tra i risultati ottenuti, quanto nel fatto che i campioni simulati dalle due distribuzioni finali possono essere riutilizzati per approssimare la distribuzione finale di un qualunque altro parametro di interesse, ad esempio la differenza $\delta = \theta_1 - \theta_2$. A livello di programma di calcolo basterà aggiungere la riga

```
delta<-x1-x2
```

e ripetere per δ i comandi già scritti per ψ. Al contrario, non c'è alcuna garanzia che sia possibile ottenere facili approssimazioni asintotiche né per la distribuzione campionaria di uno stimatore di δ, né per la distribuzione finale di δ. ♢

Va notato un altro aspetto importante: il calcolo di statistiche piuttosto comuni come i percentili risulta in questo contesto del tutto immediato: ad esempio, nel caso precedente, il valore della mediana a posteriori di ψ verrà stimato considerando la versione ordinata (in senso crescente) del campione di valori pseudo casuali $(\psi_1, \cdots, \psi_M)$ e prendendo il valore che occupa la posizione mediana.

Più in generale, dunque, quando si tratta di risolvere un integrale del tipo

$$V = \int_a^b f(x)\pi(x)dx,$$

e π è una distribuzione da cui è semplice ottenere un campione di valori pseudo-casuali, $x_{(1)}, x_{(2)}, \cdots, x_{(M)}$, sotto la sola ipotesi che V esista finito, una semplice stima di V è fornita da

$$\hat{V} = \frac{1}{M}\sum_{j=1}^{M} x_{(j)},$$

Lo stimatore $\hat{V}$ è consistente in virtù della legge dei grandi numeri.

Esempio 7.4 [*Previsioni*]

[3] si noti che la (7.7) non coincide con la (5.5), ma ne rappresenta, a meno di costanti di normalizzazione, il quadrato.

Dato un campione $X_1, \cdots, X_n$ di osservazioni i.i.d. $N(\theta, \tau^{-1})$ (τ rappresenta la precisione della variabile aleatoria), e una legge a priori non informativa (si veda § 4.3.2) $\pi(\theta, \tau) \propto \tau^{-1}$, si vuole calcolare

$$\Pr(X_{n+1} > w|\mathbf{y}),$$

ovvero la probabilità che la successiva osservazione superi una certa soglia w. È noto dalla § 4.3.2 che, a posteriori,

$$\theta|\tau, \mathbf{y} \sim N\left(\bar{x}, \frac{1}{n\tau}\right), \qquad \tau|\mathbf{y} \sim \text{ Gamma } \left(\frac{n-1}{2}, \frac{ns^2}{2}\right)$$

(qui s^2 rappresenta la varianza campionaria osservata). Questo implica che, generando M coppie di valori (θ_i, τ_i) dalle distribuzioni suddette, si ottiene un campione rappresentativo della distribuzione finale congiunta. Inoltre la quantità d'interesse può scriversi come

$$\Pr(X_{n+1} > w|\mathbf{y}) = \int_{-\infty}^{\infty}\int_{0}^{\infty} \Pr(X_{n+1} > w|\theta, \tau)\, \pi(\theta, \tau|\mathbf{y}) d\theta d\tau$$

e

$$\Pr(X_{n+1} > w|\theta, \tau) = 1 - \Phi\left(\frac{w-\theta}{\tau^{-\frac{1}{2}}}\right).$$

Ne segue allora che,

$$\Pr(X_{n+1} > w|\mathbf{y}) = \mathbb{E}_{(\theta,\tau)|\mathbf{y}}\left[1 - \Phi\left(\frac{w-\theta}{\tau^{-\frac{1}{2}}}\right)\right],$$

che potrà essere approssimata dalla quantità

$$\frac{1}{M}\sum_{j=1}^{M}\left[1 - \Phi\left(\frac{w-\theta_{(j)}}{\tau_{(j)}^{-\frac{1}{2}}}\right)\right]$$

♢

Esempio 7.5 [*Il Value at risk.*]
In finanza è prassi comune stimare un dato percentile (detto appunto *value at risk*) della distribuzione del rendimento di un dato portafoglio su un certo intervallo di tempo. Ad esempio, il quinto percentile di detta distribuzione fornirà il valore minimo del rendimento che, con una probabilità pari a 0.95, possiamo attenderci dal nostro investimento.

La modellizzazione dei rendimenti è questione attualmente molto dibattuta e la necessità di costruire modelli affidabili conduce inevitabilmente a distribuzioni dei rendimenti piuttosto complesse e non trattabili in modo analitico. Anche in questo caso, da un punto di vista bayesiano, è sufficiente poter generare un campione di valori pseudo-casuali dalla distribuzione finale del rendimento in esame, pur senza conoscerne la forma analitica, per poter ottenere stime affidabili del percentile in questione.
Sia infatti θ la variabile aleatoria che denota il rendimento del portafoglio P e sia $\pi(\theta \mid \mathbf{x})$ la distribuzione finale di θ sulla base di un dato campione di osservazioni: vogliamo determinare il *value at risk* al livello α, ovvero determinare il percentile di ordine α della distribuzione h. Supponiamo di essere in grado di generare un campione di M valori $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)}$ dalla distribuzione π. È sufficiente allora ordinare il campione ottenuto e considerarne il percentile di ordine α per ottenere una stima della nostra quantità di interesse la cui precisione dipenderà dal numero M di

valori generati. Qualora non siamo in grado di generare direttamente valori dalla distribuzione π, occorrerà utilizzare strategie più complesse, illustrate nei prossimi paragrafi.

ESEMPIO DAL LIBRO DI COLES O DA QUALCHE TESI ♢

7.4 Data Augmentation

??????????????

7.5 Metodi Monte Carlo

Vogliamo qui descrivere alcune tecniche di simulazione atte al calcolo approssimato di integrali del tipo

$$V = \int_a^b f(x)\pi(x)dx. \tag{7.8}$$

Nella statistica bayesiana (ma non solo) tali integrali appaiono ovunque. Interpretando $\pi(x)$ come la distribuzione finale, la (7.8) rappresenta ad esempio il valor medio finale della funzione f del parametro X. Tali metodi si rendono necessari quando l'integrale non è risolubile analiticamente. Per semplicità di esposizione qui consideriamo soltanto il caso univariato in cui X rappresenta una variabile reale, ma i metodi qui descritti rappresentano una valida alternativa ai metodi basati sull'integrazione numerica soprattutto nel caso di integrazione di funzioni multidimensionali.

7.5.1 Campionamento per importanza

Il metodo *Monte Carlo Importance Sampling* o di *Monte Carlo con campionamento per importanza* (MCIS, d'ora in poi) si basa sulla seguente, semplice idea: sia $\xi(x)$, una funzione di densità con supporto che contiene (o coincide con) $[a, b]$, e scriviamo l'integrale (7.8) come

$$V = \int_a^b f(x)\frac{\pi(x)}{\xi(x)}\xi(x)dx.$$

La quantità V può essere riespressa come $\mathbb{E}_\xi\left[f(x)\pi(x)/\xi(x)\right]$ ovvero il valore atteso di una specifica funzione di x rispetto alla densità di probabilità $\xi(\cdot)$. È ragionevole allora approssimare la media V con un suo corrispondente campionario, generando un campione di M valori pseudo-casuali $x^{(1)}, x^{(2)}, \cdots, x^{(M)}$ dalla legge $\xi(\cdot)$, calcolare, per ogni $i = 1, \cdots, M$, la quantità $Z_i = f(x^{(i)})\pi(x^{(i)})/\xi(x^{(i)})$ e ottenere così la stima

$$\hat{V} \approx \frac{1}{M}\sum_{i=1}^{M} Z_i$$

Prima di illustrare il metodo MCIS nel contesto bayesiano, consideriamo innanzitutto un esempio classico, trattato in [74] che qui riportiamo nella versione di [21] e che illustra chiaramente l'efficacia del metodo.

Esempio 7.6 [*Probabilità di una coda della distribuzione di Cauchy*]
Data una v.a. $X \sim Ca(0, 1)$, si vuole calcolare $V = \Pr(X > k)$, per un certo $k > 0$. Si tratta dunque di calcolare l'integrale

$$V = \int_k^\infty \frac{1}{\pi(1+x^2)} dx\,;$$

Il metodo più ovvio consiste nel generare un grande numero M di valori pseudo casuali dalla legge $Ca(0,1)$ e calcolare una prima stima di V:

$$\hat{V}_1 = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}_{(X>k)}(x_j).$$

Da notare che tale stimatore non appare efficiente dal punto di vista del numero di simulazioni in quanto tutti i valori generati che risultano minori di k vengono scartati. Senza perdere in generalità assumiamo d'ora in poi che $k = 2$. Poiché il vero valore di Vèc in questo esempio calcolabile analiticamente ed è pari a $V = 0.147$, lo stimatore $\hat{V}_1$ avrà varianza pari a $V(1-V)/M = 0.125/M$. Un primo miglioramento si può ottenere tenendo conto della simmetria della distribuzione di Cauchy e utilizzando così anche i valori minori di $-k$: avremo così lo stimatore

$$\hat{V}_2 = \frac{1}{2M} \sum_{j=1}^{M} \mathbf{1}_{(|X|>k)}(x_j).$$

In tal caso, è facile verificare che per $k = 2$ si ottiene una varianza pari a $V(1-2V)/(2M) = 0.052/M$ Un modo alternativo per calcolare V si basa sul fatto che

$$V = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)} dx$$

che può essere scritta come

$$V = \frac{1}{2} - \mathbb{E}_X\left[h(X)\right],$$

dove $h(x) = 2/(\pi(1+x^2))$ e $X \sim U(0,2)$. Si possono perciò generare valori $U_{(1)}, \cdots U_{(M)}$ da una legge uniforme in $(0,2)$ e ottenere lo stimatore

$$\hat{V}_3 = \frac{1}{2} - \frac{1}{M} \sum_{j=1}^{M} \frac{2}{\pi(1+u_{(j)}^2)}$$

A differenza di $\hat{V}_1$ e $\hat{V}_2$ lo stimatore $\hat{V}_3$ utilizza tutti i valori generati: la varianza di $\hat{V}_3$ risulta inoltre inferiore. Si può dimostrare infatti, con semplici calcoli, analitici o numerici (Esercizio 7.8.1) che $\mathrm{Var}\left(\hat{V}_3\right) = 0.028/M$. Va però notato che lo stimatore $\hat{V}_3$ non è uniformemente migliore di $\hat{V}_1$ o di $\hat{V}_2$ per qualunque valore di k: ad esempio, per $k = 3$ lo stimatore $\hat{V}_2$ è il migliore dei tre considerati. (Esercizio 7.8.2). Una quarta soluzione si basa sulla seguente proprietà della distribuzione di Cauchy: se $X \sim$ Ca $(0,1)$, allora, per ogni $k > 0$ risulta $\Pr(X > k) = \Pr(0 < X < 1/k)$ (Esercizio 7.8.3). La suddetta relazione porge un modo alternativo per scrivere V,

$$V = \int_0^{\frac{1}{2}} \frac{1}{\pi(1+x^2)} dx = \mathbb{E}_Z\left[\frac{h(Z)}{4}\right],$$

dove, stavolta, $Z \sim U(0,1/2)$; è perciò possibile considerare lo stimatore

$$\hat{V}_4 = \frac{1}{2M} \sum_{j=1}^{M} \frac{1}{\pi(1+w_{(j)}^2)},$$

dove $W_{(1)}, \cdots W_{(M)}$ sono generate da una legge uniforme in $(0,1/2)$. Anche in questo caso è possibile calcolare (Esercizio 7.8.4)) la varianza dello stimatore che vale $\mathrm{Var}\left(\hat{V}_4\right) = 0.0056/M$.

Un'ulteriore soluzione, più nello spirito MCIS, consiste nel cercare una distribuzione avente come supporto la semiretta $\{x : x > 2\}$ e con code simili alla densità di Cauchy, per utilizzarla come densità da cui generare le osservazioni. Una possibile proposta è la legge associata alla v.a. $Y = 2/U$, con $U \sim U(0,1)$. Si vede facilmente che la densità di Y (Esercizio 7.8.5) è

$$f_Y(t) = \frac{2}{t^2} \qquad t > 2\,,$$

cosicché uno stimatore alternativo è dato da

$$\hat{V}_5 = \frac{1}{M}\sum_{j=1}^{M} \frac{t_{(j)}^2}{2\pi(1+t_{(j)}^2)}\,,$$

dove $t_{(j)} = 2/u_{(j)}$, $j = 1, \cdots, M$ e le $u_{(j)}$ sono M realizzazioni pseudo casuali di una v.a. uniforme in $(0,1)$. Si può verificare (Esercizio 7.8.2) che $\mathrm{Var}\left(\hat{V}_5\right) = ????/M$. ♢

Le tante possibili soluzioni per questo semplice calcolo illustrano chiaramente la questione: se in teoria qualunque soluzione Monte Carlo risolve in qualche modo il problema, la strategia di calcolo è lungi dall'essere ininfluente e, in problemi più complessi, la scelta può essere cruciale.

Alle stime prodotte col metodo MCIS, che risultano consistenti in virtù della legge dei grandi numeri, è possibile associare, come visto nell'esempio, un errore standard. Poiché si è considerato un campionamento semplice senza ripetizione, si ha banalmente che l'errore standard di $\hat{V}$, denotato con $e.s.(\hat{V})$, è pari alla deviazione standard campionaria divisa per $\sqrt{M}$, ovvero

$$e.s.(\hat{V}) = \frac{1}{\sqrt{M}}\sqrt{\frac{\sum_{i=1}^{M}(Z_i - \hat{V})^2}{M-1}}.$$

Ovviamente, quando la funzione $\pi(x)$ è una densità di probabilità da cui è semplice generare valori pseudo-casuali, si può semplificare ulteriormente la procedura, generando i valori $x^{(1)}, x^{(2)}, \cdots, x^{(M)}$ direttamente da π e stimando V con la quantità

$$\hat{V} \approx \frac{1}{M}\sum_{i=1}^{M} f(x^{(i)}).$$

In questo caso parleremo di semplice metodo Monte Carlo (MC) piuttosto che di MCIS. Dietro questa semplice tecnica si nascondono però alcune insidie che è bene mettere in luce.

Come detto, la scelta della funzione ξ è in teoria senza alcun vincolo: in realtà non tutte le scelte forniscono buoni risultati. Consideriamo il caso, frequente, in cui l'intervallo $[a,b]$ è l'intera retta reale e supponiamo di scegliere una densità ξ le cui code decadono a zero più velocemente di quelle della densità π: ne segue che il campione di valori pseudo-casuali tenderà ad avere pochi valori di x sulle code, e quei pochi valori contribuiranno con dei pesi Z_i resi instabili dai rapporti $\pi(x^{(i)})/\xi(x^{(i)})$ con denominatori molto piccoli, in quanto la densità ξ ha code più *sottili* rispetto alla legge π. La conseguenza di tutto ciò è che le stime così ottenute possono avere una varianza altissima, addirittura infinita. Per ovviare a tale inconveniente, è buona norma scegliere una densità ξ con code più pesanti rispetto a π, ma non troppo diversa dalla π stessa, cosicché i rapporti $\pi(x^{(i)})/\xi(x^{(i)})$ non risultino troppo grandi o troppo vicini a zero.

Esempio 7.7 [*Stima robusta della media di una popolazione normale.*]
Siano $X_1, X_2, \cdots, X_n$ n osservazioni indipendenti con distribuzione $N(\theta, 1)$; si vuole stimare il parametro θ. Invece di utilizzare una distribuzione a priori coniugata di tipo gaussiano è preferibile

irrobustire la procedura considerando una densità a priori con code più pesanti; consideriamo ad esempio una densità a priori di tipo $Ca(\mu_0, 1)$, ovvero

$$\pi(\theta) = \frac{1}{\pi(1 + (\theta - \mu_0)^2)};$$

per quanto già visto nei capitoli precedenti, la media a posteriori di θ è

$$\mu^* = \mathbb{E}(\theta \mid \mathbf{y}) = \frac{\int \theta \frac{1}{(1+(\theta-\mu_0)^2} \exp\left\{-\frac{n}{2}(\bar{x} - \theta)^2\right\}}{\int \frac{1}{\pi(1+(\theta-\mu_0)^2} \exp\left\{-\frac{n}{2}(\bar{x} - \theta)^2\right\}},$$

dove $\bar{x}$ rappresenta, come sempre, la media campionaria osservata. Come già visto nell'esempio (5.1) tale integrale non è risolubile analiticamente. Una possibile soluzione è fornita dal metodo Monte CArlo: si generano M valori pseudo-casuali $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)}$ da una densità di Cauchy con parametro di posizione μ_0 e parametro di scala pari a 1 e si calcola la stima

$$\hat{\mu}_1^* = \frac{\sum_{i=1}^M \theta^{(i)} \exp\left\{-\frac{n}{2}(\bar{x} - \theta^{(i)})^2\right\}}{\sum_{i=1}^M \exp\left\{-\frac{n}{2}(\bar{x} - \theta^{(i)})^2\right\}};$$

alternativamente, sfruttando il fatto che la funzione di verosimiglianza è proporzionale ad una densità di tipo $N(\bar{x}, n^{-1})$ avremmo potuto invertire nei calcoli il ruolo della densità a priori e della funzione di verosimiglianza, generando M valori pseudo-casuali $\lambda^{(1)}, \lambda^{(2)}, \cdots, \lambda^{(M)}$ da una densità $N(\bar{x}, n^{-1})$ e ottenere la stima

$$\hat{\mu}_2^* = \frac{\sum_{i=1}^M \lambda^{(i)} / \left[1 + (\lambda^{(i)} - \mu_0)^2\right]}{\sum_{i=1}^M 1 / \left[1 + (\lambda^{(i)} - \mu_0)^2\right]};$$

Consideriamo un esempio numerico in cui $n = 1, \bar{x} = 3, \mu_0 = 0$ ed $M = 1000$. Le stime $\hat{\mu}_1^*$ e $\hat{\mu}_2^*$ possono ottenersi mediante i seguenti comandi in **R** che permettono anche di produrre la figura 7.7 in cui si può apprezzare la diversa stabilità delle stime.

```
theta<-rnorm(1000,3,1)
sum(theta/(1+theta^2))/sum(1/(1+theta^2))
plot(cumsum(theta/(1+theta^2))/cumsum(1/(1+theta^2)),
type="l",ylab="I",xlab="num. iter.")
lambda<-rt(1000,df=1)
sum(lambda * exp(-0.5*(3-lambda)^2))/sum(exp(-0.5*(3-lambda)^2))
lines(cumsum(lambda *exp(-0.5*(3-lambda)^2))/cumsum(exp(-0.5*(3-lambda)^2)), lty=3)
```

e

```
lambda<-rt(1000,df=1)
sum(lambda * exp(-0.5*(3-lambda)^2))/sum(exp(-0.5*(3-lambda)^2))
plot(cumsum(lambda *exp(-0.5*(3-lambda)^2))/cumsum(exp(-0.5*(3-lambda)^2)),
type="l",ylab="I",xlab="k")
```

Nel primo caso si ottiene una stima di μ pari a 2.311 e nel secondo si ottiene invece 2.2812. La differenza tra le due stime suggerisce di analizzare il loro comportamento all'aumentare del numero di iterazioni. La tabella seguente mostra una diversa stabilità dei due stimatori per diversi valori di M

valore di M	10^3	10^4	10^5	10^6
$\hat{\mu}_1^*$	2.310	2.292	2.281	2.288
$\hat{\mu}_2^*$	2.289	2.283	2.281	2.284

Tabella 7.1. Stime Monte Carlo al variare della dimensione del campione simulato

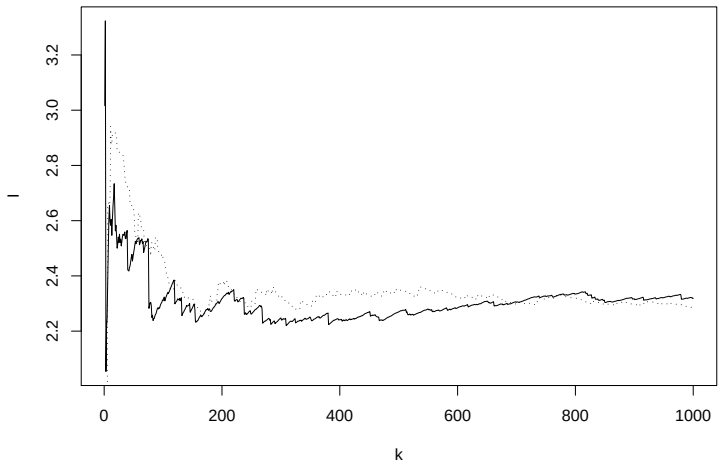

Figura 7.1. Andamento delle stime $\hat{\mu}_1^*$ e $\hat{\mu}_2^*$ al variare di M

La tabella dimostra come le due stime abbiano efficacia comparabile, sebbene $\hat{\mu}_2^*$ appaia certamente più stabile: una possibile spiegazione è il fatto che la densità da cui vengono generati i dati in questo secondo caso, la densità gaussiana, ha code più simili alla funzione integranda, ovvero il prodotto della densità normale e quella di Cauchy. Un modo alternativo per apprezzare la differenza tra i due metodi è di analizzare il range dei valori che viene coperto dal campione simulato: come mostra la figura 7.7, generare un campione dalla legge di Cauchy garantisce un'esplorazione più precisa dello spazio parametrico.

Una terza soluzione al problema verrà discussa successivamente nella §MCMC.

Purtroppo non esistono criteri generali per il calcolo dell'errore standard delle stime suddette; il più delle volte occorre elaborare tecniche matematiche specifiche, basate sulle proprietà delle densità coinvolte nel problema. Resta dunque il rischio che i pesi con cui le stime MCIS vengono calcolate possano essere mal calibrate con pochi valori molto alti che rendono lo stimatore molto variabile. Come parziale strumento di diagnostica, [43] suggeriscono di effettuare un istogramma dei pesi utilizzati: se l'istogramma presenta pochi valori anomali la stima è considerata affidabile; laddove l'istogramma presenti diversi valori isolati più grandi degli altri, c'è il rischio di ottenere una stima con alta variabilità e in tal caso l'unico correttivo può essere l'utilizzo di un M più elevato. Nella Figura **??** vengono riportati, con riferimento alla stima più erratica, $\hat{\mu}_1^*$, gli istogrammi dei valori generarti dalla densità di Cauchy nei due casi $M = 10^4$ e $M = 10^6$. ♢

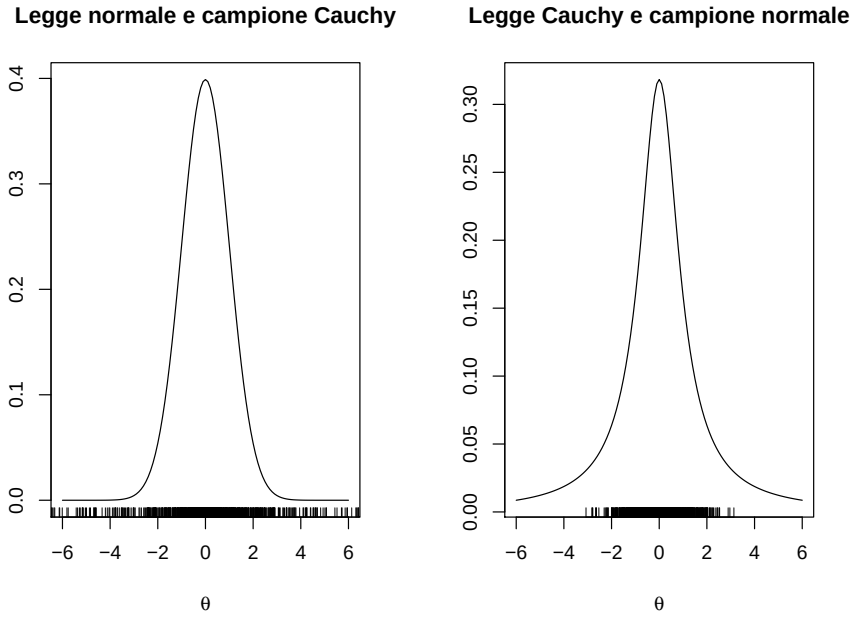

Figura 7.2. bla bla

7.5.2 Metodi accettazione-rifiuto

Una tecnica alternativa, e molto generale, per il calcolo di integrali quando non si è in grado direttamente di simulare dalla distribuzione di interesse, è quella basata sul criterio accettazione rifiuto. Supponiamo, come al solito che nostro obiettivo sia il calcolo di

$$V = \int g(\theta)\pi(\theta)d\theta$$

e che un approccio diretto di tipo Monte Carlo, basato su simulazioni da $\pi(\theta)$ non sia praticabile. Allora, sotto alcune ipotesi qui di seguito esplicitate, è possibile operare nel modo seguente: supponiamo che la distribuzione di interesse $\pi(\theta)$ sia assolutamente continua ed esprimibile come il rapporto $f(\theta)/K$, dove $f(\theta)$ rappresenta una densità non normalizzata, che sappiamo valutare[4] mentre K, in genere non nota, rappresenta la costante di normalizzazione. Supponiamo inoltre che esista un'altra distribuzione assolutamente continua $h(\theta)$ da cui è possibile simulare valori e tale che, per una costante grande a piacere c risulti $f(\theta) < ch(\theta)$, per ogni $\theta \in \boldsymbol{\Omega}$. Allora è possibile dimostrare che il seguente algoritmo genera valori dalla distribuzione *target* $\pi(\theta)$:

Algoritmo di accettazione-rifiuto (AR)

1. genera un valore *candidato* $W = w \sim h(w)$ e un valore $Y = y \sim U(0, 1)$
2. Se
$$y \leq \frac{f(W)}{ch(W)},$$
poni $\theta = w$, altrimenti rifiuta il candidato e torna al passo 1.

[4] ovvero, per ogni valore di θ siamo in grado di calcolare il valore $f(\theta)$.

Teorema 7.1

(a) I valori accettati dall'algoritmo AR si distribuiscono secondo la legge target $\pi(\theta)$.
(b) La probabilità marginale che un singolo candidato sia accettato è pari a $\frac{K}{c}$.

Dimostrazione 7.1 *(a) La funzione di ripartizione associata alla variabile aleatoria*

$$W\left|\left[Y<\frac{f(w)}{ch(w)}\right]\right.$$

può scriversi come

$$F_W(\theta)=\frac{\Pr\left(W<\theta,Y<\frac{f(w)}{ch(w)}\right)}{\Pr\left(Y<\frac{f(w)}{ch(w)}\right)}=\frac{\int_W\Pr\left(W<\theta,Y<\frac{f(w)}{ch(w)}|w\right)h(w)dw}{\int_W\Pr\left(Y<\frac{f(w)}{ch(w)}|w\right)h(w)dw}$$

$$=\frac{\int_{-\infty}^{\theta}\Pr\left(Y<\frac{f(w)}{ch(w)}|w\right)h(w)dw}{\int_{-\infty}^{\infty}\Pr\left(Y<\frac{f(w)}{ch(w)}|w\right)h(w)dw}=\frac{\int_{-\infty}^{\theta}\frac{f(w)}{c}dw}{\int_{-\infty}^{\infty}\frac{f(w)}{c}dw}$$

$$=\frac{\int_{-\infty}^{\theta}\frac{K\pi(w)}{c}dw}{\int_{-\infty}^{\infty}\frac{K\pi(w)}{c}dw}=\int_{-\infty}^{\theta}\pi(w)dw.$$

(b) La probabilità che un generico candidato $W=w$ sia accettato è pari a

$$\Pr\left(W\ \textit{accettato}\ \right)=\Pr\left(Y<\frac{f(W)}{ch(W)}\right)=\int_W\Pr\left(Y<\frac{f(W)}{ch(W)}|W=w\right)h(w)dw$$

$$=\int_W\frac{f(w)}{c}dw=\int_W\frac{K}{c}\pi(w)dw=\frac{K}{c}.$$

Il punto (b) del teorema suggerisce che, qualora K fosse nota, la costante c andrebbe scelta proprio pari a K per minimizzare il numero di candidati rifiutati. In assenza di informazioni su K occorrerà scegliere c in modo che la condizione $f(\theta)<ch(\theta)$ sia verificata per ogni θ e quindi porre

$$c=\sup_\theta\frac{f(\theta)}{h(\theta)}.$$

Esempio 7.8 [*Simulazione di una legge Beta*]
Si vuole generare un campione di valori da una distribuzione Beta$(5,2)$ e si dispone soltanto di un generatore di numeri casuali in $(0,1)$. In questo caso allora

$$\pi(\theta)=\frac{f(\theta)}{K}=\frac{\theta^8(1-\theta)}{\text{Beta}(5,2)}$$

mentre $h(\theta)=1$, per ogni θ. Poiché il valore di massimo di $f(\theta)$ è circa $c_0=0.082$ occorre scegliere un valore di $c>c_0$ ed utilizzare l'algoritmo #??, scritto in **R**.

ALGORITMO # ?? (ACCETTAZIONE - RIFIUTO)

```
ac.rej<-function(c=1, nsim=1000){
theta<-c()
w<-runif(nsim)
y<-runif(nsim)
for (j in 1:nsim) {
```

```
if(c*y[j]<(w[j]^4)*(1-w[j]))
theta<-c(theta,w[j])
}
freq<-length(theta)/nsim
hist(theta,nclass=20, prob=T,xlab=expression(theta))
curve(dbeta(x,5,2),add=T)
return(freq)
}
```

La figura 7.3 mostra il risultato che si ottiene utilizzando l'algoritmo con diversi valori della costante c e con diverso numero di candidati proposti. Nell'esempio specifico è facile vedere che il valore ottimale di c è pari a $c_0 = 8/9 = 0.043$ ed infatti le percentuali di valori accettati (e con esse, la qualità delle approssimazioni) migliora al diminuire di c verso c_0. ♢

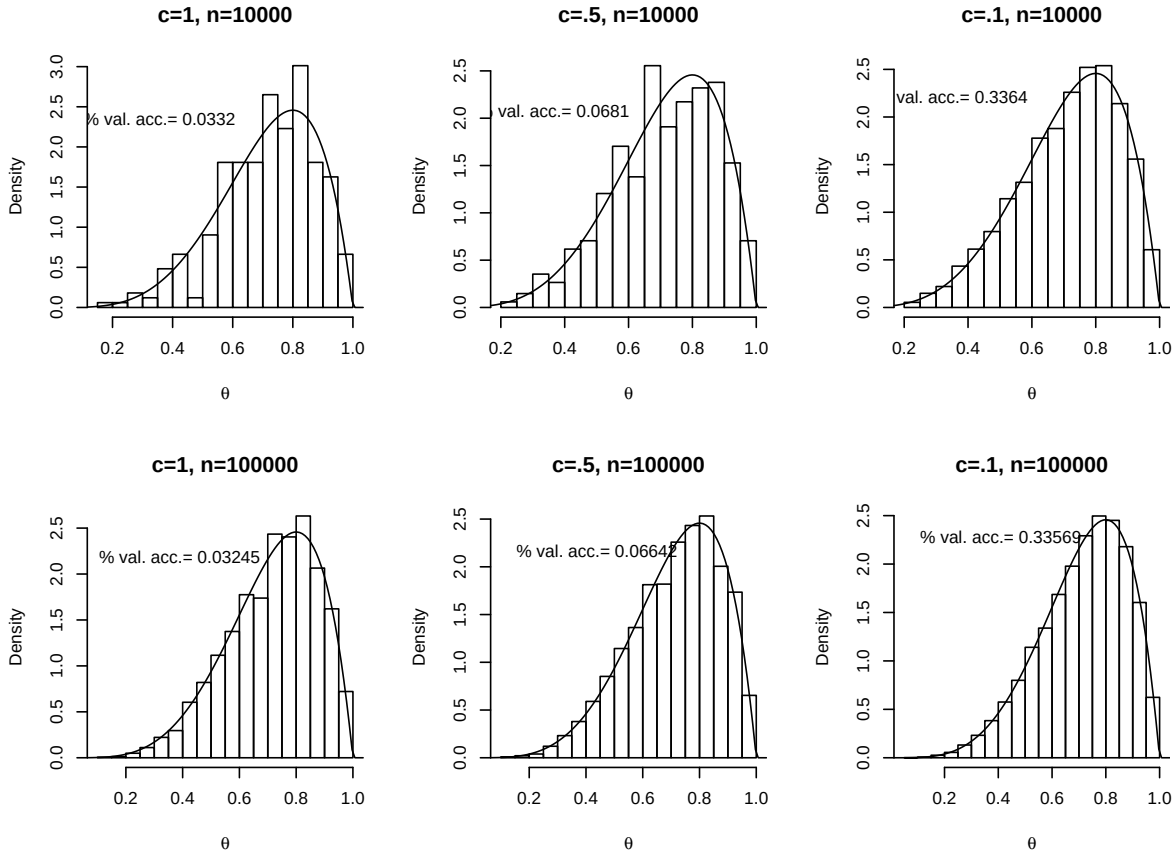

Figura 7.3. Diversa efficacia del metodo accettazione rifiuto al variare della costante c e del numero di simulazioni

7.5.3 Distribuzioni log-concave

7.6 Algoritmi adattivi

Sequential importance sampling

Population MC

7.7 Metodi MCMC

Consideriamo nuovamente il problema

$$V = \int_a^b f(x)\pi(x)dx,$$

e supponiamo ora che, oltre a non essere in grado di risolvere analiticamente l'integrale, né di generare valori pseudo-casuali da π, non si riesca nemmeno a determinare, come nei metodi MCIS e AR, una densità $\xi(x)$ da cui sia possibile generare valori e che, in qualche modo, *somigli* alla $\pi(x)$.

In situazioni del genere è possibile ricorrere ai cosiddetti metodi di tipo Monte Carlo basati sulle proprietà delle catene di Markov (MCMC). Questi metodi consentono di costruire successioni di valori pseudo-casuali che solo approssimativamente possono essere considerati come realizzazioni indipendenti dalla densità obiettivo $\pi(x)$. La descrizione e soprattutto la comprensione di tali metodi non possono fare a meno di alcune rudimentali nozioni sulle catene di Markov, che vengono concentrate nella Appendice D.

Se effettivamente si fosse in grado di simulare una realizzazione $(x^{(1)}, x^{(2)}, \cdots,)$ di una catena di Markov la cui distribuzione di equilibrio è proprio la nostra densità obiettivo $\pi(x)$, allora, per il cosiddetto teorema di ergodicità - si veda l'Appendice D, o per una trattazione più esauriente [21] - una estensione al caso di variabili aleatorie non indipendenti della legge dei grandi numeri, è possibile affermare che lo stimatore

$$\hat{V}_{MC} = \frac{1}{M}\sum_{i=1}^{M} f(x^{(i)})$$

è uno stimatore consistente di V.

Il problema generale dei metodi MCMC è dunque quello di costruire una catena di Markov $X_1, X_2, \cdots, X_n, \cdots$ la cui distribuzione limite sia proprio la legge $\pi(x)$ e, da non trascurare, che la velocità di convergenza verso tale limite sia la più alta possibile. È evidente che, nei problemi di inferenza bayesiana, la legge $\pi(x)$ rappresenta la distribuzione finale sullo spazio parametrico $\boldsymbol{\Omega}$. L'obiettivo è quello di generare una catena di Markov a tempo discreto le cui realizzazioni, dopo un periodo iniziale di assestamento, possano essere considerate, almeno in modo approssimativo, come replicazioni *quasi* indipendenti di una stessa v.a. con distribuzione $\pi(x)$. In questa trattazione, per forza di cose superficiale, ci limiteremo a illustrare, nella §7.7.1, le tecniche matematiche che consentono di costruire una simile catena e a descrivere brevemente, nei paragrafi successivi, gli algoritmi più diffusi che implementano tali tecniche.

7.7.1 Aspetti matematici

Nell'appendice D viene introdotta la proprietà di *reversibilità*, altrimenti detta dell'*equilibrio in dettaglio*. Sia π una distribuzione sullo spazio degli stati, e sia $\mathbf{P} = \{p_{ij}, i, j \in S\}$ la matrice di transizione in un passo di una catena omogenea. Se per ogni coppia di stati i e j, la probabilità di trovarsi in i e spostarsi in j è uguale alla probabilità di essere in j e spostarsi in i, si dice che la catena è reversibile. In formule, deve verificarsi che

$$\pi_i p_{ij} = \pi_j p_{ji}. \tag{7.9}$$

. Si può dimostrare che se una catena di Markov $\{X_n; n = 0, 1, \cdots\}$ è irriducibile ed esiste una legge π che soddisfa la proprietà 7.9, allora la distribuzione π risulta stazionaria per $\{X_n; n = 0, 1, \cdots\}$ (si veda ad esempio [84]).

Nelle applicazioni statistiche reali, lo spazio degli stati risulta quasi sempre continuo e la relazione (7.9) va riletta, per due generici stati w e z, come

$$\pi(w)p(w, z) = \pi(z)p(z, w), \tag{7.10}$$

dove adesso π va interpretata come una densità di probabilità sullo spazio (non numerabile) degli stati mentre $p(w, z)$ rappresenta la probabilità di transizione dallo stato w allo stato z. Si può dimostrare [21] che la proprietà di reversibilità di una distribuzione è condizione sufficiente per garantire la stazionarietà della distribuzione stessa anche nel caso continuo cosicché, per verificare se una catena converge, prima o poi, alla distribuzione stazionaria nostro obiettivo, sarà sufficiente verificare che essa soddisfi la proprietà (7.10) per ogni coppia di stati w e z: gli algoritmi MCMC vengono costruiti proprio con questo obiettivo.

Supponiamo dunque, per un momento, di essere in grado di simulare una realizzazione di una catena di Markov avente come distribuzione di equilibrio la nostra distribuzione *target* $\pi(\theta)$. Tutto quello che la teoria matematica ci garantisce è che, asintoticamente, i valori generati nella simulazione possono essere considerati realizzazioni dalla distribuzione $\pi(\theta)$. Restano però in piedi due ordini di problemi:

1. I valori generati non sono indipendenti ma legati, appunto, da una struttura markoviana.

2. Non è chiaro *fin da quando* possiamo iniziare a considerare i valori come realizzazioni da $\pi(\theta)$: In altre parole non possiamo sapere con certezza quando abbiamo toccato la terra asintotica ...

Per quanto concerne il primo problema, la sua importanza è relativa: come già detto, infatti, per il teorema ergodico [21], l'approssimazione di funzionali integrali del tipo (7.2) risulta valida anche in presenza di valori simulati da una catena di Markov. Il secondo problema ha una soluzione meno chiara: dal punto di vista delle applicazioni occorre stabilire dei criteri diagnostici, di monitoraggio della realizzazione della catena che ci consentano di stabilire, con una certa approssimazione, se e quando la catena ha raggiunto la convergenza. Tra i pacchetti statistici che implementano tale strumentazione segnaliamo CODA e BOA, disponibili anche in versione R.

7.7.2 Gli algoritmi di tipo Metropolis-Hastings

Il primo criterio generale per generare una catena di Markov con una determinata distribuzione di equilibrio è stato proposto da [60], e generalizzato poi da [49]; è basato su una successione di valori *candidati* generati da una distribuzione *di proposta* $q(\cdot)$ e accettati o meno secondo una

regola che andiamo a descrivere. Sia $p(x, y)$ una probabilità di transizione di Markov, ovvero una funzione di due argomenti, che rappresenta, per ogni (x, y) nello spazio degli stati della catena, la densità di probabilità che la catena si sposti dallo stato x allo stato y. Sia invece $q(y|x)$ la densità di probabilità che l'algoritmo *proponga* uno spostamento da x a y. Allora l'algoritmo di Metropolis e Hastings può essere descritto nel modo seguente.

Algoritmo di Metropolis-Hastings (MH)

1. Al tempo 0, la catena si trova in $X_0 = x_0$
2. Al tempo t genera un valore *candidato* $Z = z \sim q(z|x_{t-1})$
3. Calcola il rapporto $\alpha(z, x_{t-1})$, dove
$$\alpha(z, x_{t-1}) = \min\left(\frac{q(x_{t-1}|z)\pi(z)}{q(z|x_{t-1})\pi(x_{t-1})}, 1\right)$$
4. genera $\xi \sim U(0, 1)$ e poni
$$X_t = \begin{cases} z & \text{se } \xi \leq \alpha(z, x_{t-1}) \\ X_{t-1} & \text{se } \xi > \alpha(z, x_{t-1}) \end{cases}$$

Per dimostrare che l'algoritmo effettivamente ha come distribuzione di equilibrio la nostra distribuzione target $\pi(\theta)$ è sufficiente verificare che la condizione (7.10) risulti soddisfatta. Nel caso degli algoritmi di tipo MH il nucleo di transizione può scriversi come

$$p(w, z) = q(z|w)\alpha(z, w) + h(w)\delta_w(z),$$

dove

$$h(w) = \int (1 - \alpha(t, w))q(t|w)dt;$$

la formula precedente si spiega osservando che, per andare da w a z occorre prima *proporre* e poi *accettare* un valore z, oppure, nel caso in cui $w = z$, lo spostamento è possibile generando un qualsiasi valore, purché poi esso venga rifiutato. Allora il primo membro della (7.10) prende la forma

$$\pi(w)p(w, z) = \pi(w)q(z|w)\min\left(1, \frac{\pi(z)q(w|z)}{\pi(w)q(z|w)}\right) + \pi(w)h(w)\delta_w(z); \tag{7.11}$$

Senza perdere in generalità supponiamo che

$$\pi(z)q(w|z) < \pi(w)q(z|w);$$

dunque $\alpha(z, w) < 1$ e, di conseguenza, $\alpha(w, z) = 1$. Allora,

$$\begin{aligned} \pi(w)p(w, z) &= \pi(w)q(z|w)\frac{\pi(z)q(w|z)}{\pi(w)q(z|w)} + \pi(w)h(w)\delta_w(z) \\ &= \pi(z)q(w|z) + \pi(z)h(z)\delta_z(w) \\ &= \pi(z)q(w|z)\alpha(w, z) + \pi(z)h(z)\delta_z(w) = \pi(z)p(z, w). \end{aligned}$$

Per ulteriori dettagli e approfondimenti sugli algoritmi MH si possono consultare [24] e [77].

Si può evincere dalla descrizione dell'algoritmo che la sua effettiva applicabilità è subordinata alla possibilità di poter valutare la densità $\pi(x)$ per ogni valore x; più precisamente, è necessario

poter calcolare il rapporto $\pi(x)/\pi(y)$ per ogni coppia di stati (x, y). Questa è la situazione tipica in cui ci si trova in un problema di inferenza bayesiana, dove la distribuzione *target* π è la legge a posteriori che sappiamo calcolare a meno di una costante di normalizzazione: nel calcolo del rapporto $\pi(x)/\pi(y)$ la costante di normalizzazione si semplifica e l'algoritmo MH può essere implementato. Si può notare che quando la distribuzione propositiva $p(x, y)$ è simmetrica nei due argomenti, ovvero

$$p(x, y) = p(\mid x - y \mid),$$

il calcolo del rapporto α si semplifica ed avremo

$$\alpha(z, x_{t-1}) = \min\left(\frac{\pi(x_{t-1})}{\pi(z)}, 1\right).$$

Questo caso particolare è quello proposto inizialmente da [60]. La scelta della distribuzione propositiva è l'unico aspetto *soggettivo* dell'algoritmo e richiede una buona dose di esperienza; qui di seguito elenchiamo alcune tra le scelte più comuni.

Proposta *Random walk*: si passa da uno stato x ad uno y secondo una passeggiata aleatoria $y = x + Z$, con Z generato da una qualche distribuzione simmetrica rispetto allo 0, ad esempio una distribuzione $N(0, \sigma^2)$: in questo caso il valore di σ^2 governerà la tendenza dall'algoritmo a spostarsi più o meno velocemente nello spazio degli stati. Alternative alla legge normale sono la distribuzione di Cauchy $(0, \sigma^2)$ oppure una distribuzione uniforme $U(-c, c)$ con $c > 0$ da specificare. In genere i parametri della distribuzione propositiva (σ^2 oppure c) vengono determinati in modo che il tasso di valori accettati dall'algoritmo si aggiri intorno al $25 - 30\%$.

Proposta indipendente: in questo caso si pone $q(y \mid x) = g(y)$ per ogni x dove $g(\cdot)$ è una specifica distribuzione: in questo approccio, i nuovi valori proposti non dipendono dalla posizione effettiva della catena.

In generale, qualunque sia la distribuzione di proposta, è importante notare come l'algoritmo MH proponga dei nuovi valori per la distribuzione target e questi valori vengono accettati certamente se il rapporto α è pari ad 1, ovvero se il nuovo punto ha densità π eventualmente riscalata rispetto a q, superiore; se invece $\alpha < 1$ allora i punti vengono accettati proprio con probabilità pari ad α. La caratteristica di accettare anche valori candidati con densità "inferiore" a quello attuale è tipica degli algoritmi il cui obiettivo è quello di esplorare il supporto di una distribuzione e non di localizzarne il punto di massimo.

Consideriamo ora alcuni semplici esempi.

Esempio 7.9 [*Simulazione da una legge Gamma*(δ, λ)]

Si vuole costruire un algoritmo che generi valori aventi distribuzione limite del tipo

$$\pi(\theta) = \frac{\lambda^\delta}{\Gamma(\delta)} e^{-\lambda\theta}\theta^{\delta-1};$$

Utilizzeremo in questo esempio diverse distribuzioni propositive: la prima appartiene alla classe delle proposte *random walk*, con distribuzione esponenziale di media fissata pari al valore precedente della catena: risulta quindi che, al passo t, con la catena in θ_{t-1}, il rapporto $\alpha(z, \theta_{t-1})$ vale

$$\begin{aligned}\alpha(z, \theta_{t-1}) &= \frac{e^{-\lambda z} z^{\delta-1} e^{-x_{t-1}/z} x_{t-1}}{e^{-\lambda x_{t-1}} x_{t-1}^{\delta-1} e^{-z/x_{t-1}} z} \\ &= \left(\frac{z}{x_{t-1}}\right)^{\delta-2} \exp\left\{\lambda(x_{t-1} - z)\right\} \exp\left\{-(x_{t-1}/z + z/x_{t-1})\right\}.\end{aligned}$$

Il corrispondente codice in **R** è il seguente.

Metropolis Hasting con proposta di tipo *random walk*

```
met.has <- function(n=100000,x1=2, burn=.3,lambda, delta){
vet <- array(0,n)
vet[1] <- x1
for(t in 2:n){
 z <- rexp(1,rate=1/(vet[t-1]))
 accept <- exp(lambda*(vet[t-1] - z))*
 ((z/vet[t-1])^(delta-2))*exp(-vet[t-1]/z+z/vet[t-1])
 alpha <- min(1,accept)
 u <- runif(1)
 if (u <= alpha) vet[t] <- z
 else vet[t] <- vet[t-1]
 }
subs<-seq(burn*n,n,by=1)
hist(vet[subs],prob=T,nclass=100,xlab="valori di x",ylab="",
main="MH Random Walk")
curve(dgamma(x,delta,rate=lambda), add=T)
return(vet)
}
```

Una seconda possibile distribuzione propositiva è di tipo *indipendente* e genera valori esponenziali aventi *tutti* la stessa media pari alla media della distribuzione target, ovvero δ/λ. L'algoritmo vine modificato solo per i comandi di generazione dalla distribuzione propositiva e di calcolo di α.

Codice per Metropolis Hasting con proposta *indipendente* (MH_{EXP2})

```
met.has.ind<-function(n=100000,x1=2, burn=.3,lambda, delta){
vet <- array(0,n)
vet[1] <- x1
for(t in 2:n){
 z <- rexp(1,rate=lambda/delta)
 accept <- exp(lambda*(vet[t-1] - z))*((z/vet[t-1])^(delta-1))*
 exp((-lambda/delta)*(vet[t-1]-z))
 alpha <- min(1,accept)
 u <- runif(1)
 if (u <= alpha) vet[t] <- z
 else vet[t] <- vet[t-1]
 }
sub<-seq(burn*n,n,1)
hist(vet[sub],prob=T,nclass=100,xlab="valori di x",ylab="",
```

```
main="MH Independence Sampler")
curve(dgamma(x,delta,rate=lambda), add=T)
return(vet)
}
```

Nella figura **??**, si riportano i risultati dei comandi

```
met.has(n=100000,xi=2, burn=0.3, lambda=3, delta=6)
met.has.ind(n=100000,xi=2, burn=0.3, lambda=3, delta=6)
```

per ottenere, attraverso i due algoritmi, realizzazioni da una distribuzione Gamma$[6, 3]$. ♢

7.7.3 L'algoritmo di Gibbs

L'algoritmo di Gibbs, o Gibbs sampler può essere considerato un caso particolare dell'algoritmo Metropolis-Hastings, dove i valori proposti sono accettati con probabilità pari a 1. Tuttavia esso non gode della enorme applicabilità dell'altro e può essere utilizzato soltanto quando valgono certe specifiche condizioni sulla distribuzione a posteriori. L'idea di fondo è di costruire la catena di Markov utilizzando, per generare i valori di ogni singola componente del vettore dei parametri, la distribuzione a posteriori di quella componente del parametro condizionata al valore assunto, nelle precedenti iterazioni, dalle altre componenti del parametro. Si tratta quindi di generare valori casuali da distribuzioni univariate, compito spesso (ma non sempre) piuttosto semplice. Prima di descrivere in dettaglio l'algoritmo, consideriamo un semplice esempio **Esempio 7.10** [*Normale bivariata.*] Supponiamo di voler generare valori da una distribuzione normale bivariata, associata cioè ad una v.a. $\mathbf{Y} = (Y_1, Y_2)$ con media $(0, 0)$ e e matrice di correlazione

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Naturalmente, per risolvere questo problema esistono criteri più diretti ed elementari ma la semplicità del problema ci aiuterà a comprendere l'utilità del Gibbs sampling. Da proprietà elementari della distribuzione normale sappiamo che le distribuzioni di ciascuna componente di $\mathbf{Y}$ condizionata all'altra sono ancora normali; più precisamente,

$$Y_1 \mid Y_2 = u \sim \mathrm{N}(\rho u, 1 - \rho^2) \tag{7.12}$$

e

$$Y_2 \mid Y_1 = v \sim \mathrm{N}(\rho v, 1 - \rho^2) \tag{7.13}$$

È sufficiente allora scegliere un valore di partenza $\mathbf{y}^{(0)} = (y_1^{(0)}, y_2^{(0)})$ e generare una catena di Markov secondo il seguente schema

Gibbs sampling per Normale bivariata

- parti da $\mathbf{y}^{(0)} = (y_1^{(0)}, y_2^{(0)})$
- al passo t-esimo genera

$$y_1^{(t)} \sim N(\rho y_2^{(t-1)}, 1-\rho^2), \quad y_2^{(t)} \sim N(\rho y_1^{(t)}, 1-\rho^2)$$

I valori generati dal precedente algoritmo formeranno così una catena di Markov avente come distribuzione di equilibrio proprio la normale bivariata nostro obiettivo. Vedremo più avanti perché l'algoritmo funziona; intanto, c'è da notare come, ad ogni iterazione il valore di $y_2^{(t)}$ viene generato condizionatamente all'ultimo valore disponibile di y_1, ovvero $y_1^{(t)}$ e non $y_2^{(t-1)}$. Il seguente codice in **R** implementa l'algoritmo precedente:

Codice **R** per Gibbs sampling (Normale bivariata)

```
gibbs.binorm<-function(niter,rho){
warm<-0.7*niter
gibbssample<-c()
x2<-0
for(g in 1:niter){
x1<-rnorm(1,mean=rho*x2,sd=sqrt(1-rho^2))
x2<-rnorm(1,mean=rho*x1,sd=sqrt(1-rho^2))
gibbssample<-c(gibbssample,x1,x2)}
gibbssample<-matrix(gibbssample,byrow=T,ncol=2)
biv.norm<-function(y1,y2){
nucl<-0.5*(y1^2-2*rho*y1*y2+y2^2)/(1-rho^2)
dens<- 1/(2*pi)*(1/sqrt(1-rho^2))*exp(-nucl)
return(dens)}
asc<-seq(-4,4,by=.1)
z<-outer(asc,asc,biv.norm)
contour(asc,asc,z,xlab="theta1",ylab="theta2",col=4)
points(gibbssample[warm:niter,],cex=.1)}
```

I comandi

```
par(mfrow=c(1,2))
gibbs.binorm(niter=10000, rho=0.2)
gibbs.binorm(niter=10000, rho=0.8)
```

producono la Figura 7.4. ♢

Il motivo per cui l'algoritmo di Gibbs funziona risiede nel fatto che la distribuzione di transizione della catena di Markov costruita attraverso le distribuzioni condizionate effettivamente possiede, come distribuzione di equilibrio, la distribuzione congiunta di tutte le componenti. Verifichiamo, nel dettaglio cosa avviene nel caso bivariato. Per dimostrare che una catena di Markov è dotata di una distribuzione di equilibrio e che tale distribuzione sia proprio la distribuzione congiunta, deve

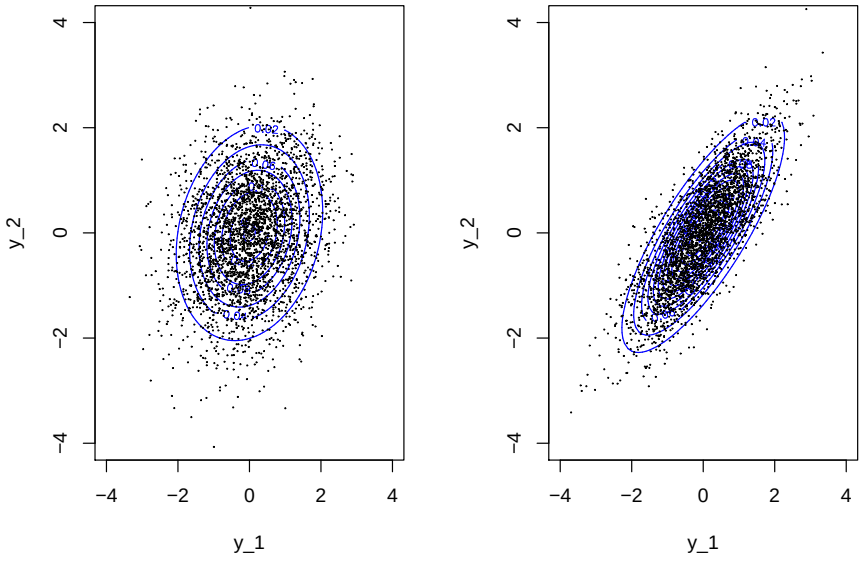

Figura 7.4. Curve di livello e valori generati mediante l'algoritmo di Gibbs per l'esempio relativo alla normale bivariata; il primo grafico si riferisce al caso $\rho = .2$, il secondo al caso $\rho = 0.8$.

verificarsi la condizione di invarianza ovvero, per ogni t,

$$\int_{\mathbb{R}^2} q(\mathbf{y}^{(t)}|\mathbf{y}^{(t-1)})\pi(\mathbf{y}^{(t-1)})d\mathbf{y}^{(t-1)} = \pi(\mathbf{y}^{(t)})$$

dove $q(\cdot \mid \cdot)$ è la distribuzione di transizione, che nel nostro caso è data dal prodotto delle due condizionate, ovvero

$$q(\mathbf{y}^{(t)}|\mathbf{y}^{(t-1)}) = \pi(y_1^{(t)} \mid y_2^{(t-1)})\pi(y_2^{(t)} \mid y_1^{(t)})$$

Infatti,

$$\begin{aligned}
&\int_{\mathbb{R}}\int_{\mathbb{R}} f(y_1^{(t)}, y_2^{(t)}|y_1^{(t-1)}, y_2^{(t-1)})\pi(y_1^{(t-1)}, y_2^{(t-1)})dy_1^{(t-1)}dy_2^{(t-1)} = \\
&\int_{\mathbb{R}}\int_{\mathbb{R}} \pi(y_1^{(t)}|y_2^{(t-1)})\pi(y_2^{(t)}|y_1^{(t)})\pi(y_1^{(t-1)}|y_2^{(t-1)})\pi(y_2^{(t-1)})dy_1^{(t-1)}dy_2^{(t-1)} = \\
&\pi(y_2^{(t)}|y_1^{(t)}) \int_{\mathbb{R}} \pi(y_1^{(t)}|y_2^{(t-1)})\pi(y_2^{(t-1)})dy_2^{(t-1)} = \\
&\pi(y_2^{(t)}|y_1^{(t)}) \int_{\mathbb{R}} \pi(y_1^{(t)}, y_2^{(t-1)})dy_2^{(t-1)} = \\
&\pi(y_2^{(t)}|y_1^{(t)})\pi(y_1^{(t)}) = \pi(y_1^{(t)}, y_2^{(t)}).
\end{aligned}$$

Nel caso generale in cui si hanno k parametri, la logica dell'algoritmo è identica. Si sceglie un punto iniziale $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$ per la catena e si aggiornano una ad una tutte le componenti.

Algoritmo Gibbs sampler

- parti da $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$

- al passo t-esimo genera

$$\begin{aligned}
\theta_1^{(t)} &\sim \pi(\theta_1|\theta_2^{(t-1)},\ldots,\theta_k^{(t-1)})\\
\theta_2^{(t)} &\sim \pi(\theta_2|\theta_1^{(t)},\ldots,\theta_k^{(t-1)})\\
\theta_3^{(t)} &\sim \pi(\theta_3||\theta_1^{(t)},\theta_2^{(t)},\ldots,\theta_k^{(t-1)})\\
&\ldots\ldots\ldots\\
\theta_k^{(t)} &\sim \pi(\theta_k|\theta_1^{(t)},\ldots,\theta_{k-1}^{(t)})
\end{aligned}$$

Esempio 7.11 [*Modelli cattura-ricattura: approccio bayesiano*]
Riconsideriamo qui l'Esempio 2.7, e mostriamo come, sebbene una soluzione bayesiana in termini analitici al problema della stima congiunta di N e p sia complessa, la soluzione basata sull'algoritmo di Gibbs risulti elementare. Riconsideriamo dunque la funzione di verosimiglianza già calcolata in (2.8),

$$L(N,p) \propto \frac{N!}{(N+m-n_1-n_2)!}p^{n_1+n_2}(1-p)^{2N-n_1-n_2},$$

e assumiamo che, a priori, p ed N risultino indipendenti con distribuzioni

$$p \sim \text{Beta}(\alpha,\beta); \qquad N \sim \text{Poi}(\lambda),$$

ovvero

$$\pi(p,N) \propto p^{\alpha-1}\,(1-p)^{\beta-1}\exp(-\lambda)\,\lambda^N/N!. \tag{7.14}$$

Combinando la funzione di verosimiglianza con la (7.14), indicando con $\mathbf{x} = (n_1, n_2, m)$ il vettore delle statistiche sufficienti, e con $r = n_1 + n_2 - m$ il numero effettivo di unità differenti osservate nelle due catture, si ottiene che

$$\pi(p,N \mid \mathbf{x}) \propto p^{n1+n2+\alpha-1}\,(1-p)^{2N-n_1-n_2+\beta-1}\frac{\lambda^N}{(N-r)!}. \tag{7.15}$$

È semplice allora verificare che

- la distribuzione a posteriori di p condizionata al valore di N è di tipo Beta$(\alpha+n_1+n_2, \beta+2N-n_1-n_2)$
- la distribuzione a posteriori di $N-r$ condizionata al valore di p è di tipo Poi(λ), dove la traslazione del supporto della v.a. N è dovuta al fatto che, dopo l'esperimento, sappiamo con certezza che $N \geq r$.

È sufficiente allora implementare un algoritmo di Gibbs per ottenere un campione di valori generato, approssimativamente, dalla distribuzione congiunta di p ed N.

FARE ESEMPIO CON STESSI NUMERI DEL CAPITOLO 2.

♢

- e' un MH con accettazione 1
- esempi N2 e Norm coniugata e Beherens Fisher CAttura ricattura??
- completamento t di student da una normale
- Rao Blackwell

```
#GIBBS CONIUGATO N(TE, SIGMA)
#
gib.con<-function(nits,y,muini=0,tauini=1,burn=0.3,alpha=0,xi=0.001,lambda=0.001, delta=0.001){
     x <- array(0,c(nits+1,3))
     x[1,1] <- muini
     x[1,2] <- tauini
     x[1,3] <- 1/tauini
     n <- length(y)
     elle<-burn*nits+1
     ybar <- mean(y)
     post<-(n*ybar+xi*alpha)/(n+xi)
     esse2 <- var(y)*(n-1)/n
     for(t in 2:(nits)){
       x[t,1]<-rnorm(1,post,sqrt(1/(x[t-1,2]*(n+xi))))
  lambdanuovo<-0.5*(n+xi)*(x[t,1]-post)^2 +
 n*esse2/2 +lambda +0.5*n*xi*(ybar-alpha)^2/(n+xi)
deltanuovo<-(n+1)/2 + delta
          x[t,2]<-rgamma(1,rate=lambdanuovo,shape=deltanuovo)
          x[t,3]<-1/x[t,2]
}
# RICORDA CHE
#The Gamma distribution with parameters shape = a and scale = s has density
#f(x)= 1/(s^a Gamma(a)) x^(a-1) e^-(x/s)
# for x > 0, a > 0 and s > 0. The mean and variance are E(X) = a*s and Var(X) = a*s^2.
# invece, rate = 1/scale
consi.te<-x[elle:nits,1]
consi.tau<-x[elle:nits,2]
consi.sig<-x[elle:nits,3]
      par(mfrow=c(2,2))
     plot(1:length(consi.te),consi.te,type='l',
            lty=1,xlab='t',ylab=expression(theta))
   hist(consi.te,nclass=30,prob=T)
#curve(dt(x,n-1,ybar),add=T)
     plot(1:length(consi.sig),consi.sig,type='l',
            lty=1,xlab='t',ylab=expression(sigma))
        hist(consi.sig,nclass=30,prob=T)
x
}
```

7.7.4 Altri algoritmi

Questo è un esempio di algoritmo ibrido dove si possono avere due diverse proposal che vengono scelte randomly ogni volta

```
hyb.met <- function(n,x0){
    x <- array(0,n)
    x[1] <- x0
    for(t in 2:n){
       u1 <- runif(1)
       if(u1 <= 0.5){
          y <- rnorm(1,0,0.25)
          alpha_((3+x[t-1]^2)/(3+y^2))^2*
                  exp(2*(x[t-1]^2 - y^2))}
       else{
          y <- rnorm(1,x[t-1],0.25)
          alpha <- ((3+x[t-1]^2)/(3+y^2))^2
       }
       accept <- min(1,alpha)
       u <- runif(1)
       if (u <= accept) x[t] <- y
       else x[t] <- x[t-1]
    }
    plot(1:length(x),x,type='b',
         lty=1,xlab=''t'',ylab=''x'')
    x
}
```

Esempio di Metropolis within Gibbs (copia il senso da Wasserman)

7.7.5 Convergenza degli algoritmi MCMC

7.8 Esercizi

7.8.1 *Fai due +due*

7.8.2 *Fai due +due*

7.8.3 *Fai due +due*

7.8.4 *Fai due +due*

7.8.5 *Fai due +due*

7.8.6 *Fai due +due*

7.8.7 *Utilizzare l'approssimazione di Laplace per riottenere la formula di Stirling (E.2).*

8

Scelta del modello statistico

8.1 Introduzione

Nei capitoli precedenti abbiamo sempre tacitamente assunto che i dati osservati durante l'esperimento statistico fossero stati prodotti da un "meccanismo generatore" dotato di una componente casuale. Allo stesso tempo, abbiamo assunto che tale meccanismo fosse uno di quelli previsti all'interno del nostro modello statistico. Quando si dice, ad esempio, che $(X_1, \cdots, X_n)$ sono realizzazioni i.i.d. di una legge $\mathrm{N}(\mu, \sigma^2)$, si sta vincolando, in modo esogeno e soggettivo, il meccanismo generatore ad appartenere alla famiglia normale; ciò che resta da fare, attraverso l'osservazione dei dati, è solo un aggiornamento della valutazione probabilistica intorno ai valori di (μ, σ^2). In un tale contesto, stiamo dunque assumendo che esista una *vera* coppia di valori (μ_0, σ_0^2), che caratterizza la legge aleatoria che regola il meccanismo generatore. Occorre qui notare il doppio livello di assunzioni che entrano in gioco. Innanzitutto, si assume l'esistenza di un vero meccanismo generatore; in secondo luogo, si assume che tale meccanismo sia contenuto nel modello statistico da noi utilizzato. La prima di queste assunzioni è certamente trasversale alle varie impostazioni statistiche e la sua messa in discussione ci porterebbe su un terreno epistemologico troppo avanzato per il livello di questo testo. Il lettore interessato può trovare approfondimenti al riguardo in [15] e [64].

La seconda assunzione è invece al centro del dibattito metodologico. Molti studiosi criticano l'approccio *parametrico* all'inferenza statistica proprio per l'impossibilit di vincolare il meccanismo generatore ad appartenere ad una famiglia così ristretta, che però garantisce spesso una maggiore facilit di calcolo e di elaborazione. Ma è fuori di dubbio che qualunque modello statistico parametrico rappresenta un costrutto matematico, teorico, il cui scopo è "approssimare", in un senso da precisare, il vero meccanismo generatore. In altre parole, tutti i modelli statistici che si utilizzano nella pratica sono invariabilmente[1] sbagliati. Tuttavia essi sono di fondamentale importanza per l'interpretazione e la modellizzazione di fenomeni reali, ed è compito primario dello statistico individuare quello o quelli più adatti nella situazione specifica. Per dirla con le parole di un grande statistico, G.E.P. Box, "*tutti i modelli sono sbagliati, ma alcuni lo sono meno di altri*".

Esempio 8.1 [*Rendimenti di titoli*]

Il dataset RETURN contiene le differenze dalla media di 80 titoli quotati in borsa al momento della chiusura di un giorno qualunque. In Figura 8.1 vengono riportati gli adattamenti relativi a quattro possibili parametrici: il modello t di Student con 10 gradi di libert, il modello gaussiano (che può naturalmente essere visto come un modello t con infiniti gradi di libert, e le loro ver-

[1] casi particolari

sioni asimmetriche[2] definite rispettivamente in [**?**] e [3]. Da una prima analisi esplorativa non è facile capire se occorre introdurre nell'analisi un parametro di asimmetria oppure no, oppure se è sufficiente utilizzare un modello gaussiano oppure il livello di extra-curtosi suggerisce un modello t.

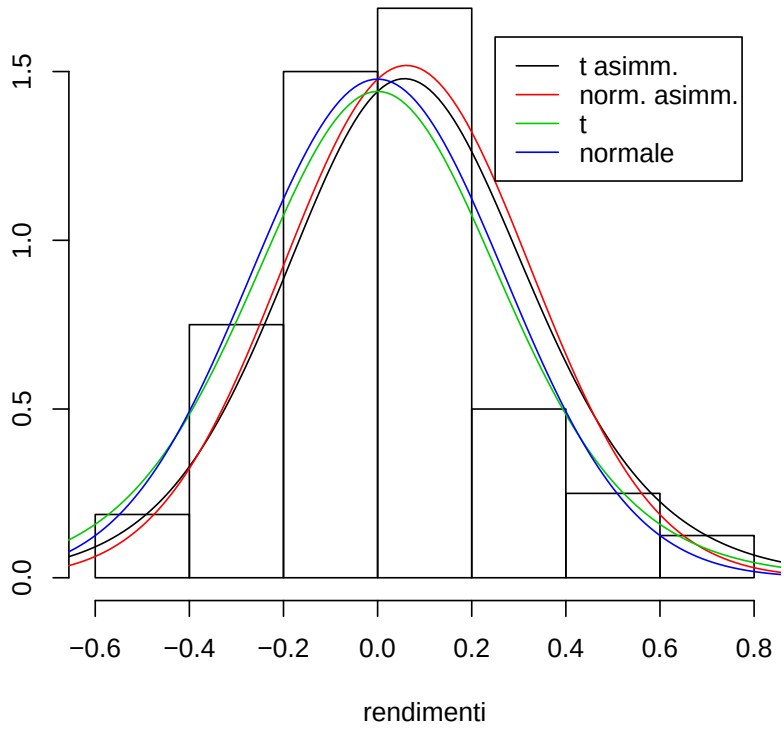

Figura 8.1. Istogramma dei dati RETURN ed adattamento di quattro possibili modelli: (a) modello t di Student asimmetrico con 10 gradi di libert , (b) modello normale asimmetrico , (c) t di Student con 10 gradi di libert; (d) modello gaussiano.

♢

Esempio 8.2 [*Confronto tra Poisson e Binomiale negativa*]
Da *JRSS*, B, n.83, pag.225 (1920): si registra il numero di incidenti sul lavoro subiti da un campione di lavoratrici di un particolare settore di produzione

Num. incidenti	0	1	2	3	4	≥ 5	totale
frequenze	447	132	42	21	3	2	647

[2] Pur senza addentrarci nel dettaglio probabilistico relativo a queste distribuzioni, basta qui dire che esse rappresentano delle generalizzazioni del modello normale e della t di Student, in quanto introducono un parametro di asimmetria, che vale 0 nel caso simmetrico.

Un semplice e naturale modello statistico per questi dati è il modello di Poisson. Si può stimare il valore di λ con la media di incidenti osservata nel campione, ovvero, in modo approssimato (si sono posti i due valori maggiori di 5 pari a 6), $\hat{\lambda} = 0.468$. La varianza campionaria vale invece 0.723; come è noto, il modello di Poisson pone il vincolo che media e varianza della popolazione siano uguali. Se il campione fornisce evidenza contraria può essere ragionevole adottare un modello in grado di catturare questo eccesso di variabilit. Un'alternativa possibile è allora il modello binomiale negativo. Si può dimostrare (Esercizio ??) che, se $X \mid \lambda \sim \text{Po}(\lambda)$, e $\lambda \sim \text{Gamma}(\alpha, 1/\beta)$, allora, marginalmente,

$$X \sim \text{BiNeg}(\alpha, 1/(\beta + 1)).$$

Dunque, adottando un modello binomiale negativo, si avrebbe $\mathbf{E}(X) = \alpha\beta$ e $\text{Var}(X) = \alpha\beta(\beta+1)$, e il modello appare così in grado di catturare una eventuale extra variabilit rispetto al valore medio. I modelli a confronto, in questo caso, sarebbero allora

$$M_1 : X \sim \text{Po}(\lambda), \quad \lambda > 0; \qquad \text{vs.} \qquad X \sim \text{BiNeg}(\alpha, 1/(\beta + 1)), \quad \alpha, \beta > 0.$$

Riprenderemo in considerazione l'esempio nel corso di questo capitolo. ♢

Esempio 8.3 [*Rendimento degli studenti in test ripetuti*]
La Figura **??** riporta i voti in trentesimi conseguiti da una classe di studenti di primo anno in una facolt di Economia in due prove parziali di statistica: alla seconda prova hanno naturalmente partecipato solo coloro che avevano superato una soglia di voto nella prima prova posta pari 16/30 (meccanismo di selezione). Se assumiamo, per iniziare, un modello gaussiano bivariato, il coefficiente di correlazione, pari a 0.28, pare indicare una mancanza di associazione tra le due prove, come se gli studenti avessero risposto in modo casuale alle domande dei due test.

Questo sistema di selezione degli studenti genera ovviamente una asimmetria nei risultati relativi al campione sottoposto alla seconda prova, formato da studenti prevedibilmente migliori di quelli eliminati dopo la prima prova. Questo fenomeno può essere catturato in modo probabilistico attraverso l'adozione di modelli più sofisticati come quelli asimmetrici, gi citati nell'esempio precedente.

CONtINUARE

♢

La scelta di un modello dipende, ovviamente, anche dall'uso che di quel modello si vuole fare. Un modello statistico può servire a descrivere un fenomeno in modo preciso, oppure può essere utile per rappresentare in modo stilizzato un fenomeno troppo complesso, cercando di coglierne gli aspetti essenziali. Oppure ancora, un modello statistico ha come obiettivo primario la previsione di risultati futuri. È altamente probabile che, nella pratica statistica, non esista un modello uniformemente migliore, ma la scelta dipender dall'uso che se ne vorr fare. L'approccio bayesiano all'inferenza statistica consente una trattazione formale di tali questioni. Nel prossimo paragrafo verranno illustrati gli strumenti necessari per una formalizzazione completa del problema della scelta del modello. Nella § 8.3 verr discusso, in maggior dettaglio, lo strumento primario di discriminazione tra modelli, ovvero il fattore di Bayes, gi utilizzato nell'ambito della verifica di ipotesi (si veda la § 6.3). Nell'ultima sezione verranno discussi alcuni approcci alternativi.

8.2 Impostazione formale del problema

Supponiamo di avere a disposizione un insieme di dati $\mathbf{x} = (x_1, \cdots, x_n)$ ed una famiglia $\mathcal{M} = \{M_1, M_2, \cdots, M_K\}$ di modelli statistici di riferimento. Ognuno di questi modelli rappresenta una famiglia di distribuzioni indicizzate da un vettore di parametri, ovvero, per $k = 1, \cdots, K$, il modello M_k rappresenta l'insieme delle distribuzioni di probabilit

$$p_k(\mathbf{x}; \boldsymbol{\theta}_k, M_k), \qquad \boldsymbol{\theta}_k \in \boldsymbol{\Omega}_k,$$

e il vettore $\boldsymbol{\theta}_k$ rappresenta il vettore dei parametri incogniti presenti nel modello M_k. Condizionatamente al modello M_k, sia $\pi_k(\boldsymbol{\theta}_k)$ la distribuzione a priori sui parametri del modello stesso[3]. Indichiamo con $p_k(\mathbf{X}; M_k)$ la distribuzione marginale o predittiva del vettore $\mathbf{X}$ condizionatamente al modello M_k, ovvero

$$p_k(\mathbf{X}; M_k) = \int_{\boldsymbol{\Omega}_k} p_k(\mathbf{X}; \boldsymbol{\theta}_k, M_k)\, \pi_k(\boldsymbol{\theta}_k)\, d\boldsymbol{\theta}_k, \qquad k = 1, \cdots, K \tag{8.1}$$

Inoltre, occorre elicitare una distribuzione di probabilit sull'insieme $\mathcal{M}$ dei modelli: essa rappresenta il nostro grado di fiducia a priori sull'efficacia esplicativa dei vari modelli a disposizione. Indichiamo tale distribuzione mediante il vettore

$$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_K), \tag{8.2}$$

dove

$$\gamma_k = \Pr(M_k), \qquad k = 1, \cdots, K.$$

Si badi bene che questa impostazione implica che stiamo assumendo l'esistenza di un solo modello "vero" e che tale modello vero è contenuto in uno degli M_k. Questa è ovviamente una forzatura: in linea di principio dovremmo interpretare la generica γ_k (ed in seguito il suo aggiornamento a posteriori) come la probabilit che il modello M_k sia quello più prossimo al vero meccanismo generatore. Inoltre, l'introduzione del vettore $\boldsymbol{\gamma}$ automaticamente implica che i vari M_k siano tra loro incompatibili, come gi osservato nell'Esempio 8.1.

Anche questa è una forzatura: molto spesso alcuni dei modelli in competizione sono annidati uno nell'altro, ovvero non sono incompatibili.

Esempio 8.4 [*Legge Normale*]
Siano $\mathbf{X} = (X_1, \cdots, X_n) \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$. Si vogliono confrontare i due modelli

$$M_1 : \mu = 0, \qquad \text{versus} \qquad M_2 : \mu \in \mathbb{R}.$$

Chiaramente, M_1 è un caso particolare di M_2 (è il caso in cui $\mu = 0$). ♢

La difficolt matematica illustrata nell'esempio precedente si supera, però, facilmente escludendo il caso particolare dal caso più generale. Così, ad esempio,nel caso precedente si definisce M_2 come la famiglia di tutte le distribuzioni normali con media $\mu \neq 0$ e varianza qualunque.

A questo punto è elementare ottenere, mediante il teorema di Bayes, le probabilit a posteriori dei singoli modelli. Infatti, per $k = 1, \cdots, K$,

[3] in realt la scrittura formalmente corretta sarebbe qui $\pi_k(\boldsymbol{\theta}_k \mid M_k)$, ma si preferir una notazione più agile, almeno nei casi in cui questo non generi confusione.

$$\Pr(M_k \mid \mathbf{X}) = \frac{\gamma_k \, p_k(\mathbf{X}; M_k)}{\sum_{j=1}^{K} \gamma_j \, p_j(\mathbf{X}; M_j)} \tag{8.3}$$

Come gi visto nella § 1.2, è spesso istruttivo considerare le odds a posteriori relative a coppie di modelli. Se si pongono a confronto i modelli M_j ed M_h, avremo allora

$$\frac{\Pr(M_j \mid \mathbf{X})}{\Pr(M_h \mid \mathbf{X})} = \frac{\gamma_j}{\gamma_h} \frac{p_j(\mathbf{X}; M_j)}{p_h(\mathbf{X}; M_h)}. \tag{8.4}$$

L'ultimo fattore della (8.4) è il fattore di Bayes che confronta i modelli M_j e M_h, che indicheremo col simbolo B_{jh}, e che quantifica come l'informazione sperimentale abbia trasformato le odds a priori (γ_j/γ_h) nelle odds a posteriori fornite dalla (8.4). Nel caso, molto comune in pratica, in cui i modelli siano considerati a priori equiprobabili, ovvero $\gamma_k = K^{-1}$, per $k = 1, \cdots, K$, la (8.3) si semplifica in

$$\Pr(M_k \mid \mathbf{X}) = \frac{p_k(\mathbf{X}; M_k)}{\sum_{j=1}^{K} p_j(\mathbf{X}; M_j)},$$

oppure, dividendo numeratore e denominatore per $p_k(\mathbf{X}; M_k)$, in

$$\Pr(M_k \mid \mathbf{X}) = \left(\sum_{j=1}^{K} B_{jk} \right)^{-1}, \tag{8.5}$$

che esprime in modo ancora più chiaro il ruolo centrale del fattore di Bayes nei problemi di scelta del modello statistico.

Il fattore di Bayes ha assunto notevole importanza pratica anche perch sembra rispettare, in modo automatico, un principio epistemologico, il cosiddetto "rasoio" di Occam, enunciato dal monaco francescano William Occam, vissuto intorno al 1300, il quale sosteneva come

> "Pluralitas non est ponenda sine necessitate".

La traduzione letterale, nel contesto in cui ci muoviamo potrebbe suonare, ad esempio, come "Non occorre usare modelli troppo sofisticati senza che sia necessario: tra due modelli che forniscono le stesse previsioni o lo stesso adattamento, scegli sempre quello più semplice, ovvero con meno parametri".

8.3 Il fattore di Bayes

Abbiamo gi introdotto, nella Definizione 6.1, il fattore di Bayes come il rapporto fra le odds a posteriori e quelle a priori. Esso rappresenta quindi il modo in cui le informazioni sperimentali contribuiscono alla discriminazione tra i diversi modelli a confronto. In situazioni regolari, i fattori di Bayes B_{jh}, con j e $h = 1, \cdots, K$, non sono altro che il rapporto fra le distribuzioni predittive (8.1) dei dati osservati condizionatamente ai due modelli M_j ed M_h. Ovviamente, tali grandezze dipendono anche dalla legge a priori sui parametri presenti nei modelli: perciò il fattore di Bayes stesso, pur essendo considerato uno strumento "debolmente" bayesiano, di fatto, può dipendere in maniera significativa dalla distribuzione a priori scelta. Inoltre, come gi precisato nella §6.3.3, questo problema di sensitivit non può essere risolto, almeno in modo diretto, dall'uso delle distribuzioni improprie. Riprenderemo il discorso su questi aspetti nella § 8.3.2.

Esempio 8.2 (continua).

Consideriamo, a priori i due modelli equiprobabili, ovvero $\gamma_1 = \gamma_2 = 0.5$. In tal caso la probabilit a posteriori associata al modello M_1 sar esprimibile, adattando la (8.5), come

$$\Pr(M_1 \mid \mathbf{x}) = \frac{1}{1 + B_{21}}.$$

Indichiamo con $\bar{x}$ la media campionaria. La funzione di verosimiglianza secondo i due modelli vale, rispettivamente, per M_1,

$$L(\lambda; \mathbf{x}) = \frac{e^{-n\lambda}\, \lambda^{n\bar{x}}}{\prod_{j=1}^{n} x_j!}, \tag{8.6}$$

e, per M_2,

$$L(\beta; \mathbf{x}) = \prod_{j=1}^{n} \binom{x_j + \alpha_0 - 1}{x_j} \frac{\beta^{n\bar{x}}}{(1+\beta)^{n(\alpha_0 + \bar{x})}}. \tag{8.7}$$

Per il calcolo di B_{21} occorre stabilire quali leggi a priori utilizzare sul parametro λ in M_1 e sui parametri (α, β) in M_2. A scopo illustrativo assumiamo che α sia noto e posto pari ad $\alpha_0 = 5$. Inoltre,

$$\lambda \mid M_1 \sim \text{Gamma}(1,1); \qquad \frac{1}{\beta + 1} \mid M_2 \sim \text{Beta}(1,1);$$

l'ultima assunzione implica tra l'altro che $\pi_2(\beta \mid M_2) = (1+\beta)^{-2}$. I valori di $p_1(\mathbf{x}; M_1)$ e $p_2(\mathbf{x}; M_1)$ possono essere ottenuti in modo analitico; infatti

$$p_1(\mathbf{x}; M_1) = \frac{1}{\prod_{j=1}^{n} x_j!} \int_0^\infty e^{-(n+1)\lambda}\, \lambda^{n\bar{x}} d\lambda = \frac{1}{\prod_{j=1}^{n} x_j!} \frac{\Gamma(n\bar{x}+1)}{(n+1)^{n\bar{x}+1}};$$

per quanto riguarda $p_2(\mathbf{x}; M_2)$,

$$p_2(\mathbf{x}; M_2) = \prod_{j=1}^{n} \binom{x_j + \alpha_0 - 1}{x_j} \int_0^\infty \frac{\beta^{n\bar{x}}}{(1+\beta)^{2+n(\alpha_0+\bar{x})}} = \prod_{j=1}^{n} \binom{x_j + \alpha_0 - 1}{x_j} \frac{\Gamma(n\bar{x})\Gamma(n\alpha_0)}{\Gamma(n\bar{x} + n\alpha_0)}.$$

Il fattore di Bayes vale dunque

$$B_{21} = \frac{\prod_j (x_j + \alpha_0 - 1)! \Gamma(n\alpha_0)(n+1)^{n\bar{x}+1}}{\prod_j (\alpha_0 - 1)!\, \Gamma(n\alpha_0 + n\bar{x})\, n\bar{x}}.$$

Utilizzando l'approssimazione di Stirling, si ottiene poi

$$B_{21} = \frac{1}{en\bar{x}} \prod_j (x_j + \alpha_0 - 1)^{x_j + \alpha_0 - 1/2} \frac{(n+1)^{n\bar{x}+1}\, (n\alpha_0 - 1)^{n\alpha_0 - 1/2}}{(\alpha_0 - 1)^{n(\alpha_0 - 1/2)}\, (n\alpha_0 + n\bar{x} - 1)^{n\alpha_0 + n\bar{x} - 1/2}}.$$

CONCLUDERE L'ESEMPIO CON I NUMERI RELATIVI ALL'ESEMPIO

Dagli esempi precedenti emergono alcune costanti che vanno sottolineate: il fattore di Bayes è di fatto il rapporto di due densit, quelle del vettore delle osservazioni sotto i due modelli a confronto. Ciò implica che non è possibile, nel calcolo di B, trascurare le costanti come si fa abitualmente nel calcolo della legge a posteriori. Nel calcolo del fattore di Bayes, le costanti che non dipendono dai parametri devono dunque essere riportate e non eliminate! Come conseguenza del punto precedente, è possibile che il fattore di Bayes possa dipendere da statistiche non sufficienti così come avviene, ad esempio, in ambito classico, quando si analizza l'adeguatezza di un modello mediante l'analisi dei residui.

L'Esempio 8.2 illustra inoltre altri aspetti che meritano attenzione: innanzitutto non è semplice condurre una trattazione bayesiana completa del problema della scelta del modello, in quanto essa comporta

- l'elicitazione di molte grandezze, prime fra tutti le leggi di probabilit a priori sui parametri relativi ai vari modelli.
- verificare come la dipendenza del risultato dalla scelta delle leggi a priori non sia, in qualche modo predominante.
- calcolare integrali che spesso non hanno una soluzione analitica

I tre punti sopra indicati sono stati al centro della ricerca in campo bayesiano negli ultimi venti anni. Una buona approssimazione che risponde, sia pure in modo parziale, alle tre questioni, è quella fornita da [80], nota con il nome di BIC e verr illustrata nella § 8.3.1. Come vedremo, però, la tecnica BIC rimuove il problema della scelta della legge a priori semplicemente trascurando, a livello asintotico il contributo di questa al calcolo delle leggi marginali. Un'altra linea di ricerca si è allora concentrata sulla possibilit di recuperare, in qualche modo, quelle leggi improprie non direttamente utilizzabili, per produrre i cosiddetti fattori di Bayes parziali. Di questo si dir nella § 8.3.2.

8.3.1 Approssimazioni del fattore di Bayes

BIC: Bayes Information Criterion

Abbiamo gi osservato nella § 7.2.2 come l'utilizzo della tecnica di approssimazione di integrali di Laplace possa condurre ad una forma approssimata del valore della distribuzione marginale dei dati condizionatamente all'uso di un certo modello M_j (vedi formula 7.3). In quest'ottica, consideriamo allora due modelli,

$$M_1 : (X_1, \cdots, X_n) \overset{\texttt{iid}}{\sim} p_1(x; \theta_1), \quad \theta_1 \in \boldsymbol{\Omega}_1 \subset \mathbf{R}^{d_1}$$

e

$$M_2 : (X_1, \cdots, X_n) \overset{\texttt{iid}}{\sim} p_2(x; \theta_2), \quad \theta_2 \in \boldsymbol{\Omega}_2 \subset \mathbf{R}^{d_2}$$

e siano $\pi_1(\theta_1)$ e $\pi_2(\theta_2)$ le distribuzioni di probabilit a priori sui rispettivi vettori dei parametri. Un'approssimazione del fattore di Bayes è fornita da

$$\tilde{B}_{12} \approx \left(\frac{2\pi}{n}\right)^{\frac{d_1-d_2}{2}} \left(\frac{\det(H_2(\tilde{\theta}_2))}{\det(H_1(\tilde{\theta}_1))}\right)^{1/2} \frac{\pi_1(\tilde{\theta}_1)L_1(\tilde{\theta}_1)}{\pi_2(\tilde{\theta}_2)L_2(\tilde{\theta}_2)}, \tag{8.8}$$

dove, rispettivamente, H_j, $L_j(\cdot)$ e $\tilde{\theta}_j$, per $j = 1, 2$, rappresentano la matrice Hessiana, la funzione di verosimiglianza, e la moda a posteriori relative al modello M_j. Quando esiste una relazione di inclusione tra i due modelli, ad esempio $M_1 \subset M_2$, l'errore che si commette tralasciando il fattore relativo ai determinanti degli Hessiani è trascurabile; [80] va oltre e, trascurando anche il fattore relativo alle distribuzioni a priori, propone in questo caso il criterio che da lei prende il nome e che in genere viene espresso in termini logaritmici,

$$S \approx -\log \tilde{B}_{12} = -\log \frac{L_1(\tilde{\theta}_1)}{L_2(\tilde{\theta}_2)} - \frac{d_2 - d_1}{2} \log n. \tag{8.9}$$

La quantit S è la somma di due addendi; il primo non è altro che il classico rapporto di verosimiglianza; il secondo va interpretato come un fattore di penalizzazione verso i modelli con molti parametri. È infatti facile dimostrare [vedi ad esempio 76], che, sotto l'ipotesi che sia vero il modello annidato, ovvero M_1, il criterio basato su S, asintoticamente fa preferire il modello corretto con probabilit

1, mentre questo non avviene utilizzando il solo rapporto di verosimiglianza. Il criterio di Schwarz fornisce dunque una risposta approssimata al problema del confronto tra modelli ed ha ricevuto negli ultimi anni notevole attenzione, soprattutto in ambito non bayesiano. Appare evidente però che la sua semplicit è dovuta semplicemente al trascurare, in modo spesso ingiustificato la componente relativa alla distribuzione a priori che, indipendente dall'argomento asintotico, produrrebbe un valore costante, che viene trascurato nel calcolo di BIC.

Il fatto che S risulti asintoticamente consistente, sotto il modello più semplice (mentre non lo è il semplice rapporto di verosimiglianza), è un modo formale per ribadire come il fattore di Bayes rispetti il principio di Occam.

Esempio 8.5 [*Modello Esponenziale.*]
Siano $(X_1, \cdots, X_n) \overset{\text{iid}}{\sim} \text{Exp}(\theta)$ e si vogliano porre a confronto i due modelli annidati

$$M_1 : \theta = \theta_0 \qquad M_2 : \theta > 0;$$

Detta $\bar{x}$ la media campionaria, la funzione di log-verosimiglianza vale

$$\ell(\theta) = n \log \theta - n\bar{x}\theta.$$

La quantit (8.9) vale allora

$$-\log \tilde{B}_{12} = \ell(\hat{\theta}) - \ell(\theta_0) - \frac{1}{2}\log n.$$

Poich $\hat{\theta} = 1/\bar{x}$, si avr

$$-\log \tilde{B}_{12} = -n\left[\log(\theta_0\bar{x}) + (1 - \theta_0\bar{x})\right] - \frac{1}{2}\log n, \tag{8.10}$$

Si può dimostrare, con pochi calcoli, che $\tilde{B}_{12}$ è consistente sia nell'ipotesi che $\theta = \theta_0$, sia nell'ipotesi alternativa (vedi Esercizio ??) ♢

DIC: Deviance Information Criterion

Un modo classico di valutare la bont di adattamento di un modello statistico è fornito dal calcolo della devianza, definita da

$$D(\boldsymbol{\theta}) = -2 \log \frac{p(\mathbf{x}; \boldsymbol{\theta})}{g(\mathbf{x})}, \tag{8.11}$$

dove $p(\mathbf{x}; \boldsymbol{\theta})$ è la generica legge di probabilit relativa al particolare modello statistico in esame e $g(\mathbf{x})$ è in genere la legge di probabilit relativa ad un modello saturato di riferimento, in cui i parametri sono tanti quante le osservazioni a disposizione. [40] suggerì di esaminare, come indicatore di adattamento, la distribuzione a posteriori di $D(\boldsymbol{\theta})$, per ottenere una interpretazione bayesiana della devianza. Sulla base di questa suggerimento, [83] hanno proposto un nuovo criterio, semplice da implementare, noto come *Deviance Information Criterion* (DIC), e definito come

$$DIC = \mathbf{E}\left(D(\boldsymbol{\theta}) \mid \mathbf{x}\right) + \nu, \tag{8.12}$$

dove ν è chiamato il numero effettivo di gradi di libert del modello ed è espresso come la differenza tra $\mathbf{E}\left(D(\boldsymbol{\theta}) \mid \mathbf{x}\right)$ e $D(\hat{\boldsymbol{\theta}})$, dove $\hat{\boldsymbol{\theta}}$ è una opportuna stima puntuale del vettore dei parametri presenti nel modello, ad esempio la media a posteriori oppure la stima di massima verosimiglianza. Una caratteristica importante del DIC è che fornisce un indicatore di adattamento relativamente al modello saturato e può dunque essere calcolato separatamente per i vari modelli in competizione.

8.3.2 Uso di distribuzioni non informative

Abbiamo gi discusso, nella §6.3.3, del fatto che non è possibile utilizzare distribuzioni improprie nei problemi di verifica di ipotesi, in quanto il fattore di Bayes risentirebbe della presenza di costanti di normalizzazione non eliminabili. Questo problema si presenta nella stessa forma nell'ambito più generale dei problemi di scelta del modello. D'altra parte, il problema di scelta del modello è, per certi versi, un problema a monte, che si vorrebbe affrontare evitando di intraprendere un cospicuo sforzo elicitativo, esteso cioè a tutti i parametri di tutti i modelli in gioco. Appare quindi fortemente auspicabile la possibilit di utilizzare distribuzioni a priori non informative in tali scenari, tanto più che spesso accade che il calcolo esplicito del fattore di Bayes sia fortemente sensibile alla scelta della legge a priori. Per ovviare a questa esigenza, si è sviluppato, soprattutto negli anni '90, un filone di ricerca molto prolifico che ha proposto diverse metodologie, perlopiù basate sul concetto di *campione di prova* o *training sample*. In poche parole, l'idea è quella di dividere l'informazione sperimentale in due parti, per poi utilizzarne una parte per aggiornare la legge di probabilit iniziale impropria e renderla propria, mentre la parte restante viene utilizzata in modo specifico per discriminare tra i modelli in competizione. L'implementazione pratica di questa semplice idea comporta una serie di scelte e di difficolt logiche e matematiche da affrontare. Attualmente esistono almeno tre diversi criteri generali, che illustreremo brevemente. Per maggiori dettagli e per una valutazione comparata delle diverse proposte si rimanda al testo di [64] o ai lavori di [37] e [13]. Assumiamo dunque di avere a disposizione K modelli $M_1, M_2, \cdots, M_K$, con

$$M_k : (X_1, \cdots, X_n) \overset{\text{iid}}{\sim} p_k(x; \boldsymbol{\theta}_k), \boldsymbol{\theta}_k \in \boldsymbol{\Omega}_k \subset \mathbb{R}^{d_k}$$

e sia la legge iniziale su $\boldsymbol{\theta}_k$, $\pi_k(\boldsymbol{\theta}_k)$ impropria in almeno qualche componente del vettore dei parametri. In quanto segue assumeremo che la distribuzione impropria utilizzata sia la *reference prior* (vedi § 5.2.4), per quanto altre scelte siano del tutto ammissibili.

Il fattore di Bayes intrinseco

Un modo naturale di suddividere l'informazione campionaria è quello di dividere il campione $(x_1, \cdots, x_n)$ in due parti, denotate con $\mathbf{x}_{(t)}$ e $\mathbf{x}_{(-t)}$; $\mathbf{x}_{(t)}$ rappresenta il cosiddetto campione di prova e viene usato per aggiornare la legge a priori impropria; $\mathbf{x}_{(-t)}$ è la parte restante del campione e viene utilizzata per discriminare tra i vari modelli. Per $k = 1, \cdots K$, si calcola la legge di probabilit su $\boldsymbol{\theta}_k$ aggiornata mediante $\mathbf{x}_{(t)}$, ovvero

$$\pi_k(\boldsymbol{\theta}_k \mid \mathbf{x}_{(t)}) = p_k(\mathbf{x}_{(t)}; \boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)/m_k(\mathbf{x}_{(t)}),$$

dove $m_k(\mathbf{x}_{(t)})$ rappresenta la densit marginale del campione di prova[4]. A questo punto è possibile utilizzare il fattore di Bayes per il confronto tra due modelli generici, basandoci sul resto del campione ovvero, per $j, h = 1, \cdots, K$,

$$B_{hj}(\mathbf{x}_{(-t)}) = \frac{\int_{\boldsymbol{\Omega}_h} p_h(\mathbf{x}_{(-t)}; \theta_h)\pi_h(\boldsymbol{\theta}_h \mid \mathbf{x}_{(t)})}{\int_{\boldsymbol{\Omega}_j} p_j(\mathbf{x}_{(-t)}; \theta_j)\pi_j(\boldsymbol{\theta}_j \mid \mathbf{x}_{(t)})};$$

Esprimendo le due leggi a posteriori che compaiono in B_{hj} attraverso la formula di Bayes si vede facilmente che vale la relazione

[4] in pratica si tratta dell'espressione (8.1) della §**??**.

$$B_{hj}(\mathbf{x}_{(-t)}) = B_{hj}^N(\mathbf{x}) \frac{m_j(\mathbf{x}_{(t)})}{m_h(\mathbf{x}_{(t)})} = B_{hj}^N(\mathbf{x}) / B_{hj}^N(\mathbf{x}_{(t)}), \tag{8.13}$$

dove $B_{hj}^N(\mathbf{x})$ rappresenta il fattore di Bayes standard, calcolato attraverso l'uso delle leggi improprie, mentre $B_{jh}^N(\mathbf{x}_{(t)})$ è la stessa quantit ma calcolata con riferimento al solo campione di prova. L'espressione (8.13), proprio perch un rapporto di due fattori di Bayes "impropri", chiarisce il motivo per cui il fattore di Bayes intrinseco non dipenda dalle costanti di normalizzazione, che si elidono nel rapporto.

Ovviamente, la dipendenza è stata sostituita da quella relativa al particolare campione di prova adottato. [13] suggeriscono di considerare campioni di prova di dimensione minimale, ovvero della numerosit minima in grado di rendere proprie tutte le leggi iniziali $\pi_1(\boldsymbol{\theta}_1), \cdots, \pi_K(\boldsymbol{\theta}_K)$. Esisteranno poi molteplici campioni di prova minimali. Nel caso classico in cui si hanno n osservazioni i.i.d. e la dimensione del campione di prova è pari a t, ad esempio, se ne avranno $\binom{n}{t}$. Tale indeterminazione si risolve calcolando un opportuno indicatore sintetico della distribuzione di tutti i possibili $B_{hj}^N(\mathbf{x}_{(t)})$, al variare di $\mathbf{x}_{(t)}$. Diversi indicatori individuano diverse versioni del fattore di Bayes intrinseco; avremo così

- il fattore di Bayes intrinseco aritmetico, ottenuto come media aritmetica dei vari $B_{hj}(\mathbf{x}_{(-t)})$

$$B_{hj}^A = \frac{1}{T} \sum_{t=1}^{T} B_{hj}(\mathbf{x}_{(-t)}) = B_{hj}^N(\mathbf{x}) \frac{1}{T} \sum_{t=1}^{T} B_{jh}^N(\mathbf{x}_{(t)}) \tag{8.14}$$

  dove T è il numero dei diversi campioni di prova.
- il fattore di Bayes intrinseco geometrico, ottenuto come media geometrica dei vari $B_{hj}(\mathbf{x}_{(-t)})$

$$B_{hj}^G = \left[\prod_{t=1}^{T} B_{hj}(\mathbf{x}_{(-t)}) \right]^{1/T} = B_{hj}^N(\mathbf{x}) \left[\prod_{t=1}^{T} B_{jh}^N(\mathbf{x}_{(t)}) \right]^{1/T} \tag{8.15}$$

- il fattore di Bayes intrinseco mediano, ovvero la mediana dei $B_{hj}(\mathbf{x}_{(t)})$, $t = 1, \cdots, T$.

$$B_{hj}^{Me} = \underset{t=1,\cdots,T}{\text{Med}} \left[B_{hj}(\mathbf{x}_{(-t)}) \right] = B_{hj}^N(\mathbf{x}) \underset{t=1,\cdots,T}{\text{Med}} \left[B_{jh}^N(\mathbf{x}_{(t)}) \right] \tag{8.16}$$

Esempio 8.5 (continua).
Assumiamo per θ, nel modello M_2, la legge a priori di riferimento $\pi^N(\theta) \propto \theta^{-1}$. Allora

$$m_2^N(\mathbf{x}) = \int_0^\infty \theta^{n-1} e^{-n\bar{x}\theta} d\theta = \frac{\Gamma(n)}{(n\bar{x})^n}$$

e

$$B_{21}^N(\mathbf{x}) = \frac{m_2^N(\mathbf{x})}{L(\theta_0; \mathbf{x})} = \frac{\Gamma(n) e^{n\bar{x}\theta_0}}{(n\bar{x}\theta_0)^n}.$$

La dimensione minimale del campione di prova è qui pari a 1 e il generico fattore di Bayes $B_{12}^N(x_j)$ vale allora

$$B_{12}^N(x_j) = x_j \theta_0 e^{-x_j \theta_0}.$$

Avremo così

$$B_{21}^A = \frac{\Gamma(n) e^{n\bar{x}\theta_0}}{(n\bar{x})^n \theta_0^{n-1}} \frac{1}{n} \sum_{j=1}^{n} \left(x_j e^{-x_j \theta_0} \right); \tag{8.17}$$

$$B_{21}^G = \frac{\Gamma(n) e^{(n-1)\bar{x}\theta_0}}{(n\bar{x})^n \theta_0^{n-1}} \prod_{j=1}^{n} x_j^{1/n}; \tag{8.18}$$

$$B_{21}^{Me} = \frac{\Gamma(n)e^{n\bar{x}\theta_0}}{(n\bar{x})^n\theta_0^{n-1}} \underset{j=1,\cdots,n}{\text{Med}} \left[x_j e^{-x_j\theta_0}\right] \tag{8.19}$$

Il fatto che i fattori di Bayes intrinseci siano calcolati come medie su tutti i possibili campioni di prova potrebbe fornire, in presenza di piccole dimensioni campionarie, valori instabili, soprattutto per B_{hj}^A e B_{hj}^G. In questi casi è possibile definirne una versione *media*, che qui illustriamo solo per il caso aritmetico. L'idea intuitiva è quella di sostituire la media che appare nella (8.14) con il suo valore atteso calcolato nell'ipotesi che $\boldsymbol{\theta}_h$ sia pari alla sua stima di massima verosimiglianza. Definiremo così il fattore di Bayes intrinseco *aritmetico atteso* come

$$B_{hj}^{EA} = B_{hj}^N \frac{1}{T}\sum_{t=1}^{T} \mathbf{E}_{\hat{\theta}_h}^{M_h}\left[B_{jh}^N(\mathbf{x}_{(t)})\right] \tag{8.20}$$

Quando le osservazioni sono scambiabili, come abbiamo sempre supposto finora, i T valori attesi presenti nella formula precedente sono ovviamente tutti uguali e la (8.20) diventa

$$B_{hj}^{EA} = B_{hj}^N \mathbf{E}_{\hat{\theta}_h}^{M_h}\left[B_{jh}^N(\mathbf{x}_{(l)})\right],$$

calcolata su un generico campione di prova $\mathbf{x}_{(l)}$.

Esempio 8.5 (continua).

In questo caso la stima di massima di massima verosimiglianza sotto il modello più grande è $\hat{\theta}_2 = 1/\bar{x}$, da cui

$$\mathbf{E}_{\hat{\theta}_2}^{M_2}\left[B_{12}^N(x_i)\right] = \int_0^\infty \frac{\theta_0 x_i}{\bar{x}} e^{-x_i(\theta_0+1/\bar{x}}dx_i = \frac{\theta_0\bar{x}}{(1+\theta_0\bar{x})^2},$$

e dunque

$$B_{21}^{EA} = \frac{\Gamma(n)e^{n\bar{x}\theta_0}}{(n\bar{x})^n\theta_0^{n-1}} \frac{\bar{x}}{(1+\theta_0\bar{x})^2}. \tag{8.21}$$

Una condizione necessaria per l'utilizzo di B_{hj}^{EA} è che il modello M_j sia annidato in M_h [8]. In assenza di tale assunzione, la consistenza di B_{hj}^{EA}, quando è vero il modello M_j, non è più assicurata.

CONFRONTO CON IL RISULTATO CLASSICO

Il fattore di Bayes frazionario

Un modo alternativo di suddividere l'informazione campionaria in due parti separate è discussa in [65], il quale propone di una *potenza frazionaria* $0 < b < 1$ della funzione di verosimiglianza per rendere propria la legge a priori. Si sceglie in pratica un valore di b tale che, per ogni $j = 1, \cdots, K$,

$$m_j^b(\mathbf{x}) = \int_{\boldsymbol{\Omega}_j} p_j(\mathbf{x};\boldsymbol{\theta}_j)^b \pi_j^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j < \infty,$$

dove $\pi_j^N(\boldsymbol{\theta}_j)$ è la legge a priori, eventualmente impropria, relativa ai parametri del modello M_j. In questo modo le leggi a priori sui diversi modelli vengono aggiornate attraverso il teorema di Bayes in

$$\pi^b(\boldsymbol{\theta}_j \mid \mathbf{x}) = \frac{\pi_j^N(\boldsymbol{\theta}_j)\, p_j(\mathbf{x};\boldsymbol{\theta}_j)^b}{m_j^b(\mathbf{x})}. \tag{8.22}$$

La restante frazione della funzione di verosimiglianza, $p_j(\mathbf{x};\boldsymbol{\theta}_j)^{(1-b)}$, viene utilizzata per la discriminazione tra i modelli, cosicch il generico fattore di Bayes per confrontare i modelli M_h e M_j, prende il nome di *frazionario* e si scrive come

$$B_{hj}^F(\mathbf{x}) = \frac{\int_{\boldsymbol{\Omega}_h} p_h(\mathbf{x};\boldsymbol{\theta}_h)^{(1-b)}\pi^b(\boldsymbol{\theta}_h \mid \mathbf{x})d\boldsymbol{\theta}_h}{\int_{\boldsymbol{\Omega}_j} p_j(\mathbf{x};\boldsymbol{\theta}_j)^{(1-b)}\pi^b(\boldsymbol{\theta}_j \mid \mathbf{x})d\boldsymbol{\theta}_j}$$

Utilizzando la (8.22), si può riesprimere $B_{hj}^F(\mathbf{x})$ nella forma computazionalmente più comoda

$$B_{hj}^F(\mathbf{x}) = B_{hj}^N(\mathbf{x})\frac{\int_{\boldsymbol{\Omega}_h} p_h(\mathbf{x};\boldsymbol{\theta}_h)^b\pi^N(\boldsymbol{\theta}_h)d\boldsymbol{\theta}_h}{\int_{\boldsymbol{\Omega}_j} p_j(\mathbf{x};\boldsymbol{\theta}_j)^b\pi^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j} \tag{8.23}$$

Il metodo di O'Hagan risolve il problema della dipendenza dal particolare campione di prova ma ne introduce un'altra, relativa al valore di b. Se i modelli a confronto appartengono tutti alla famiglia esponenziale è facile stabilire che b va posto pari a m/n, dove m è la dimensione minimale del campione di prova. In contesti più generali non è semplice stabilire il valore di b. Si vedano [65] e [13] per un'ampia discussione sul problema. Va inoltre ricordato che, nelle intenzioni del proponente, il fattore di Bayes intrinseco non aveva come scopo primario quello di rendere utilizzabili le distribuzioni improprie per la scelta tra modelli, quanto quello di dosare in modo opportuno l'informazione campionaria tra le due esigenze contrastanti, ovvero quella di un irrobustimento delle leggi a priori e quella di conservare sufficiente potere discriminatorio tra modelli alternativi.

Esempio 8.5 (continua).

Per b generico si ottiene facilmente che

$$B_{21}^F(\mathbf{x}) = \frac{\Gamma(n)e^{n\bar{x}\theta_0}}{(n\bar{x}\theta_0)^n}\frac{(nb\bar{x}\theta_0)^{nb}}{\Gamma(nb)e^{nb\bar{x}\theta_0}}$$

che, per $b = 1/n$, vale

$$B_{21}^F(\mathbf{x}) = \frac{\Gamma(n)e^{(n-1)\bar{x}\theta_0}}{(n\bar{x})^n\theta_0^{n-1}}\,\bar{x}, \tag{8.24}$$

molto simile alla versione geometrica del fattore di Bayes intrinseco (8.18), dove la media aritmetica delle osservazioni è sostituita dalla media geometrica. B_{21}^F può anche essere confrontato con B_{21}^{EA} nel quale il fattore $\exp\{-\bar{x}\theta_0\}$ è stato sostituito da $(1+\theta_0\bar{x})^{-2}$.

FARE I CALCOLI ESPLICITI

Una legge a priori convenzionale

I fattori di Bayes alternativi illustrati in questa sezione sono senza dubbio uno strumento efficace per la discriminazione tra modelli secondo un'impostazione bayesiana in assenza di specifiche informazioni a priori. Vanno però sempre considerati esattamente per quello che sono e non bisogna dimenticare i loro punti deboli, come riportati ade esempio in [76]

- Coerenza: per fattori di Bayes derivati da leggi di probabilit proprie vale sempre la relazione

  $$B_{jk}(\mathbf{x}) = \frac{B_{jh}(\mathbf{x})}{B_{kh}(\mathbf{x})} = B_{jh}(\mathbf{x})\,B_{hk}(\mathbf{x}), \tag{8.25}$$

  ovvero il confronto tra i modelli M_j e M_K può sempre essere espresso in modo relativo rispetto ad un terzo modello M_j. Questa propriet non si estende direttamente al caso del fattore di Bayes intrinseco.
- Abbiamo visto come i fattori di Bayes alternativi possano spesso essere considerati come veri fattori di Bayes, calcolati rispetto a particolari leggi iniziali, le cosiddette *intrinsic priors*. Può accadere tuttavia che la intrinsic prior relativa ad un particolare confronto tra modelli non sia completamente convincente. Si veda a tal proposito [13].

- Per quanto in modo debole, i fattori di Bayes alternativi dipendono dalla distribuzione non informativa inizialmente adottata: in pratica può accadere che il fattore di Bayes intrinseco o frazionario ottenuto a partire da una legge iniziale di Jeffreys sia diverso da quello ottenibile mediante l'uso della reference prior.
- Calibrazione: può accadere che la distribuzione a priori intrinseca nell'uso di uno specifico fattore di Bayes non dia probabilit uguale alle due ipotesi a confronto. Questo fenomeno è importante soprattutto in problemi di verifica di ipotesi unilaterali. **Esempio 8.6** *Ipotesi unilaterale*

♢

8.4 Metodi MC e MCMC

Da un punto di vista applicativo il problema formale della scelta del modello da un punto di vista bayesiano è ancora difficile da affrontare, soprattutto per le difficolt computazionali. Esistono due approcci alternativi: il primo mira ad una stima diretta della distribuzione marginale dei dati sotto ciascun modello; in pratica, per ogni M_j, $j = 1, \cdots, K$, si calcola una stima, di tipo Monte Carlo o Monte Carlo Markov Chain

$$\hat{p}_j(\mathbf{x}; M_j)$$

della grandezza 8.1.Questo approccio è discusso nella § 8.4.1. Alternativamente, è possibile considerare i K modelli alternativi come componenti di un meta-modello, che verr analizzato con un metodo di tipo MCMC. In questo caso, saranno i tempi di soggiorno della catena di Merkov nelle varie componenti del meta-modello a fornire, stavolta, una stima diretta delle probabilit a posteriori dei diversi modelli. Tale approccio verr discusso brevemente nella § 8.4.2.

8.4.1 Stima diretta della distribuzione marginale

Abbiamo gi sottolineato il ruolo centrale del fattore di Bayes nel calcolo delle probabilit finali dei modelli in competizione. In situazioni regolari, poi, il fattore di Bayes non è altro che il rapporto tra le distribuzioni marginali dei dati sotto due modelli alternativi. Proprio il calcolo della distribuzione marginale diviene allora importante e, il più delle volte, tale distribuzione non è disponibile in forma esplicita e si deve ricorrere ad approssimazioni di tipo Monte Carlo. In questa sezione, poich si analizza un metodo di calcolo diretto della 8.1, si ometter nelle formule la dipendenza dal particolare modello in esame.

Il metodo di Chib

Il metodo di [24] si basa inizialmente sulla elementare osservazione che, esprimendo in versione logaritmica il teorema di Bayes,

$$\log \pi(\boldsymbol{\theta} \mid \mathbf{x}) = \log \pi(\boldsymbol{\theta}) + \log p(\mathbf{x}; \boldsymbol{\theta}) - \log m(\mathbf{x}),$$

si ottiene la seguente relazione

$$\log m(\mathbf{x}) = \log p(\mathbf{x}; \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta} \mid \mathbf{x}), \tag{8.26}$$

valida qualunque sia il valore $\boldsymbol{\theta} \in \boldsymbol{\Omega}$. Nella formula precedente, i primi due addendi del membro di destra sono in genere disponibili in forma esplicita: per ottenere una stima di $\pi_(\mathbf{x})$, occorre dunque una stima di $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ per un valore $\boldsymbol{\theta}$ scelto a piacere. È ragionevole però scegliere un punto $\tilde{\boldsymbol{\theta}}$ di alta densit a posteriori, come la media oppure il valore che massimizza la funzione di verosimiglianza. Descriviamo ora in dettaglio il metodo di Chib, che si applica solo quando il modello è gestibile attraverso l'uso del Gibbs sampler. Assumiamo, per semplicit, che il parametro sia suddivisibile in due soli blocchi, ovvero

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

La quantit da stimare può essere allora scritta come

$$\pi(\tilde{\boldsymbol{\theta}} \mid \mathbf{x}) = \pi(\tilde{\boldsymbol{\theta}_2} \mid \tilde{\boldsymbol{\theta}_1}, \mathbf{x})\pi(\tilde{\boldsymbol{\theta}_1} \mid \mathbf{x}). \tag{8.27}$$

L'utilizzo del Gibbs sampler implica che le cosiddette "full conditionals" sono disponibili in modo esplicito; quindi il primo fattore della (8.27) è ottenibile in forma chiusa, mentre il secondo fattore può essere stimato (vedi Capitolo 7) utilizzando il vettore $(\boldsymbol{\theta}_2^{(1)}, \cdots \boldsymbol{\theta}_2^{(M)})$dei valori di output del Gibbs sampler

$$\hat{\pi}(\tilde{\boldsymbol{\theta}_1} \mid \mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \pi(\tilde{\boldsymbol{\theta}_1} \mid \boldsymbol{\theta}_2^{(m)}, \mathbf{x}) \tag{8.28}$$

Introducendo questa stima nella (8.26) ed eliminando i logaritmi, si ottiene allora la stima di Chib

$$\hat{\pi}(\mathbf{x}) = \frac{p(\mathbf{x}; \tilde{\boldsymbol{\theta}_1}, \tilde{\boldsymbol{\theta}_2})\pi(\tilde{\boldsymbol{\theta}_1}, \tilde{\boldsymbol{\theta}_2})}{\hat{\pi}(\tilde{\boldsymbol{\theta}_1} \mid \mathbf{x})p(\tilde{\boldsymbol{\theta}_2} \mid \tilde{\boldsymbol{\theta}_1}, \mathbf{x})} \tag{8.29}$$

Il metodo appena descritto è facilmente generalizzabile al caso in cui il vettore dei parametri sia suddiviso in più di due blocchi ma resta vincolato all'uso del Gibbs sampler. Esiste però una generalizzazione del metodo [23] utilizzabile anche in presenza di output derivanti dall'uso dell'algoritmo di Metropolis - Hastings.

8.4.2 Il meta-modello

Invece di stimare le distribuzioni marginali una alla volta, è possibile considerare un meta-modello in cui i singoli modelli alternativi $M_1, \cdots, M_k$ giocano il ruolo di iperparametri e la cui analisi viene effettuata attraverso un algoritmo MCMC che esplori lo spazio parametrico prodotto, definito come

$$\boldsymbol{\Omega} = \mathcal{M} \times \boldsymbol{\Omega}_1 \times \boldsymbol{\Omega}_2 \times \cdots \times \boldsymbol{\Omega}_K,$$

dove $\mathcal{M} = \{M_1, \cdots, M_k\}$ è l'insieme dei possibili modelli. Siano $(\gamma_1, \cdots, \gamma_K)$, come nella (8.2), le probabilit a priori dei K modelli e assumiamo, in accordo con l'intuizione, che le distribuzioni a priori su $(\boldsymbol{\Omega}_1 \times \cdots \boldsymbol{\Omega}_K)$ condizionate ai vari modelli M_j, siano a componenti indipendenti, ovvero per $j = 1, \cdots, K$,

$$\pi\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_K \mid M_j\right) = \prod_{k=1}^{K} \pi\left(\boldsymbol{\theta}_k \mid M_j\right).$$

Questa assunzione, in associazione al fatto che i modelli alternativi sono stati definiti in modo indipendente uno dall'altro, ognuno coi propri parametri specifici, implica che le $\pi\left(\boldsymbol{\theta}_k \mid M_j\right)$, o in modo più compatto le $\pi_j(\boldsymbol{\theta}_k)$, per $k \neq j$, siano del tutto irrilevanti dal punto di vista inferenziale, e possano essere scelte in modo da semplificare l'utilizzo pratico dell'algoritmo.

Assumiamo quindi che sia possibile adottare un algoritmo di tipo Gibbs sullo spazio prodotto Ω. Le *full conditionals* saranno allora così ottenute [48]. Per $k = 1, \cdots, K$,

$$\begin{aligned}
\pi(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{j\neq k}, M_k, \mathbf{x}) &\propto p_k(\mathbf{x}; \boldsymbol{\theta}_k, M_k)\pi_k(\boldsymbol{\theta}_k) \\
\pi(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{j\neq k}, M_h, \mathbf{x}) &\propto \pi_h(\boldsymbol{\theta}_k), \qquad h \neq k \\
\Pr(M_k \mid \boldsymbol{\theta}, \mathbf{x}) &= \frac{\gamma_k\, p_k(\mathbf{x}; \boldsymbol{\theta}_k, M_k)\, \prod_{j=1}^{K} \pi_k(\boldsymbol{\theta}_j}{\sum_{h=1}^{K} \left(\gamma_h\, p_h(\mathbf{x}; \boldsymbol{\theta}_h, M_h)\, \prod_{j=1}^{K} \pi_h(\boldsymbol{\theta}_j\right)}
\end{aligned}$$

Con le accortezze necessarie quando si utilizza un algoritmo MCMC (verifica delle condizioni di di regolarità, monitoraggio della convergenza della catena di Markov, eliminazione della fase di burn-in, etc.), si possono così stimare le probabilità a posteriori dei vari modelli attraverso la frequenza relativa di visite della catena al modello in questione. Avremo così

$$\hat{Pr}(M_j \mid \mathbf{x}) = \frac{1}{G} \sum_{g=1}^{G} I_{M_j}(M^{(g)}), \tag{8.30}$$

dove G è il numero delle iterazioni utilizzate della catena di Markov ed $M^{(g)}$ rappresenta il modello visitato dalla catena nella g-esima iterazione. Analogamente è possibile ottenere una approssimazione MCMC dei fattori di Bayes mediante la formula

$$\hat{B}_{hj}(\mathbf{x}) = \frac{\hat{Pr}(M_h \mid \mathbf{x})}{\hat{Pr}(M_j \mid \mathbf{x})} \times \frac{\gamma_j}{\gamma_h}. \tag{8.31}$$

Questo approccio è stato generalizzato da [39], che introducono una versione Metropolis dell'algoritmo stesso.

8.4.3 L'algoritmo Reversible Jump

La tecnica del Reversible Jump (RJ) è stata introdotta da [45]. Seguendo tale approccio, non è necessario definire lo spazio prodotto $\boldsymbol{\Omega}$ come in precedenza, n tantomeno elicitare le altrimenti inutili pseudo a priori. Piuttosto, si costruisce uno spazio parametrico globale attraverso l'*unione* degli spazi parametrici relativi ai singoli modelli in esame. L'indubbio vantaggio logico in termini di formalizzazione del problema viene tuttavia pagato in termini di complessità di costruzione della relativa catena di Markov, che deve stavolta essere in grado di "saltare" (da qui il nome dell'algoritmo) tra spazi di diversa dimensione e, per mantenere valida la condizione di reversibilità, occorrerà introdurre dei fattori di correzione nel calcolo delle probabilità di accettazione. Una discussione dettagliata dell'algoritmo *Reversible Jump* va oltre gli scopi di questo testo. Qui di seguito delineiamo brevemente lo spirito dell'algoritmo, rimandando a [45] e [21] per una discussione adeguatamente approfondita. Sia allora, per $j = 1, \cdots, k$, d_j la dimensione del vettore dei parametri presenti nel modello M_j.

Pseudo codice per l'algoritmo RJ

Supponiamo di trovarci, al passo t in $(M_j^{(t)}, \boldsymbol{\theta}_j^{(t)})$, ovvero, stiamo visitando il modello M_j nel punto $\boldsymbol{\theta}_j$ dello spazio parametrico.

1. Proponi un nuovo modello $M_{j'}$ con probabilità $h(j, j')$, $j' = 1, \cdots, k$.

2. Genera un valore del parametrico $\mathbf{u}$ dalla densità di proposta $q(\mathbf{u} \mid \boldsymbol{\theta}_j^{(t)}, M_j, M_{j'})$
3. Poni $(\boldsymbol{\theta}'_j, \mathbf{u}') = f_{j,j'}(\boldsymbol{\theta}_j, \mathbf{u})$, dove $f_{j,j'}$ è una funzione deterministica che "mappa" il valore attuale $(\boldsymbol{\theta}'_j, \mathbf{u}')$ in un punto dello spazio parametrico $\boldsymbol{\Omega}_{j'}$. Tale funzione va costruita in modo da uguagliare le dimensioni degli spazi, ovvero $d_j + \dim(\mathbf{u}) = d_{j'} + \dim(\mathbf{u}')$.
4. Accetta il nuovo valore proposto $M_{j'}, \boldsymbol{\theta}_{j'}$ con probabilità $\beta = \min(1, \alpha_{j,j'})$, dove

$$\alpha_{j,j'} = \frac{p'_j(\mathbf{x}; \theta'_{j'}, M_{j'})\pi_j(\theta'_{j'})\gamma_{j'}, h(j', j)q(\mathbf{u}' \mid \boldsymbol{\theta}_{j'}^{(t)}, M_j, M_{j'})}{p_j(\mathbf{x}; \theta_j, M_j)\pi_j(\theta_j)\gamma_j, h(j, j')q(\mathbf{u} \mid \boldsymbol{\theta}_j^{(t)}, M_{j'}, M_j)} \mid \frac{\partial f(\boldsymbol{\theta}_j, \mathbf{u})}{\partial(\boldsymbol{\theta}_j, \mathbf{u})} \mid$$

Se il nuovo modello proposto coincide con quello dove ci si trova, l'algoritmo si semplifica e diventa in pratica un semplice Metropolis- Hastings.

L'algoritmo RJ è piuttosto generale in termini di applicabilità ma non semplice da implementare, soprattutto in fase di costruzione delle varie funzioni $f_{j,j'}$ e dei corrispondenti Jacobiani, il cui calcolo è necessario per calcolare correttamente le probabilità di accettazione delle proposte. Esso rappresenta uno strumento essenziale in tutti quei problemi dove, per usare un'espressione di Green

> *...il numero di cose da stimare è una delle cose da stimare.*

Analizzeremo situazioni di questo genere nel capitolo 12.1 a proposito dei modelli mistura.

8.5 Altre impostazioni

8.5.1 Cross Validation

8.6 Esercizi

9

Il modello lineare

Il modello lineare rappresenta uno dei capisaldi della teoria dell'inferenza statistica e viene regolarmente utilizzato in diversi contesti applicativi, dalla biostatistica all'economia. Assumiamo qui che il lettore abbia una conoscenza, sia pure elementare, almeno della teoria della regressione lineare in un contesto classico.
La situazione pi elementare è quella in cui il modello è definito dalla relazione

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{9.1}$$

dove $\mathbf{Y}$ è un vettore di n osservazioni indipendenti, la matrice $\mathbf{X}$ ha dimensione $n \times p$, mentre il vettore $\boldsymbol{\beta}$ ha dimensione p e denota i coefficienti della relazione lineare (9.1); infine $\boldsymbol{\varepsilon}$ è un vettore aleatorio n-dimensionale che rappresenta gli errori associati alle osservazioni. Per semplicità espositiva assumeremo che

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n).$$

Questo equivale a dire che gli errori di osservazione che si commettono su diverse unità sono indipendenti tra loro e e la loro "variabilità" attesa, rappresentata dalla varianza σ^2, è costante (ipotesi di *omoschedasticità*); in genere σ^2 è un ulteriore parametro incognito del modello. La matrice $\mathbf{X}$ viene spesso definita come "matrice del disegno" oppure "matrice delle covariate": essa è assunta, in una trattazione elementare, indipendente dal vettore $\boldsymbol{\varepsilon}$; senza perdere troppo in generalità nel prosieguo di questo capitolo, $\mathbf{X}$ verrà assunta nota.

Le assunzioni precedenti sono ovviamente restrittive e, soprattutto in ambito econometrico, poco realistiche: il lettore interessato ad approfondimenti in tal senso, secondo una logica bayesiana, può consultare ad esempio [53]. Il modello appena introdotto può essere adattato a diverse e tutte importanti situazioni applicative a seconda della natura della matrice delle covariate $\mathbf{X}$.

ESEMPI: ANOVA Regressione

Le assunzioni sopra descritte, assieme alle note proprietà sulla distribuzione di trasformate lineari di vettori aleatori normali, conducono a dire che

$$\mathbf{Y} \mid (\boldsymbol{\beta}, \sigma^2) \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n\right), \tag{9.2}$$

cosicch la funzione di verosimiglianza associata al vettore osservato $\mathbf{y}$ e alla matrice di covariate $\mathbf{X}$ (il condizionamento alla quale non riporteremo nelle formule seguenti) risulta essere

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) \propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{9.3}$$

Dalla teoria classica del modello lineare è noto che, se $\mathbf{X}$ ha rango pieno, lo stimatore di massima verosimiglianza, nonch dei minimi quadrati, per il parametro $\boldsymbol{\beta}$ è dato da $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ e i valori "teorici" delle osservazioni prodotti dal modello mediante il valore stimato $\hat{\boldsymbol{\beta}}$ possono essere espressi come

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y},$$

dove $\mathbf{H} = \mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'$ rappresenta la cosiddetta matrice di proiezione [2]; in questo senso, il vettore $\hat{\mathbf{y}}$ rappresenta la proiezione del vettore $\mathbf{y}$ sul sottospazio generato dalle colonne della matrice $\mathbf{X}$. Possiamo allora scrivere, con ovvie semplificazioni,

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) &\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\hat{\mathbf{y}})'\,(\mathbf{y}-\hat{\mathbf{y}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\} \\ &= \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left((n-p)S^2 + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right\}, \end{aligned} \tag{9.4}$$

dove $(n-p)S^2 = (\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}}) = \sum_{j=1}^{n}(y_i - \hat{y}_i)^2$ rappresenta la cosiddetta devianza residua, mentre $S^2 = (\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}})/(n-p)$ viene indicato con il nome di varianza residua corretta[1].

9.1 Analisi bayesiana coniugata

La conduzione di un'inferenza bayesiana in questo contesto prevede l'elicitazione di una distribuzione a priori sui parametri $(\boldsymbol{\beta}, \sigma^2)$. Ferma restando la soggettiva libertà di scegliere la forma distribuzionale che meglio formalizzi le informazioni a priori note al ricercatore, considerazioni pratiche suggeriscono di utilizzare una legge a priori coniugata alla (9.4). Una possibile scelta, simile a quanto già visto §4.3.2 è allora la seguente:

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta} \mid \sigma^2)\,\pi(\sigma^2) \tag{9.5}$$

dove

$$\begin{aligned} \boldsymbol{\beta} \mid \sigma^2 &\sim N_p(\beta_0, \sigma^2 V_0) \\ \sigma^2 &\sim GI(c_0/2, d_0/2). \end{aligned}$$

In altri termini, assumiamo che il parametro $(\boldsymbol{\beta}, \sigma^2)$ segua a priori, una distribuzione di tipo Normale-Gamma Inversa con iperparametri $(\boldsymbol{\beta}_0, V_0, c_0/2, d_0/2)$. La conseguente densità a posteriori congiunta è dunque

$$\pi(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}+\frac{c_0}{2}+\frac{p}{2}+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-p)S^2 + d_0 + Q(\boldsymbol{\beta})\right]\right\}.$$

dove

$$Q(\boldsymbol{\beta}) = (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'V_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0).$$

[1] Il motivo per cui la varianza residua si ottenga dividendo la devianza per $n-p$ e non per n è tutto nella statistica classica; in questo modo lo "stimatore" varianza residua risulta non distorto ovvero la media su tutti i possibili valori osservabili della varianza residua è esattamente pari al valore del parametro incognito σ^2: qui si è deciso di mantenere l'uso della varianza residua corretta per facilitare il confronto tra la metodologia bayesiana qui descritta e l'approccio basato sulla funzione di verosimiglianza

Le due forme quadratiche che compongono $Q(\boldsymbol{\beta})$ possono essere elaborate secondo il Lemma C.1; avremo così

$$Q(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_\star)' V_\star^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_\star) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \left(\mathbf{X}'\mathbf{X}\right) V_\star V_0^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

dove $V_\star = (\mathbf{X}'\mathbf{X} + V_0^{-1})^{-1}$ e $\boldsymbol{\beta}_\star = V_\star(\mathbf{X}'\mathbf{y} + V_0^{-1}\boldsymbol{\beta}_0)$ La distribuzione finale si può scrivere allora come

$$\begin{aligned}\pi(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) \propto & \frac{1}{(\sigma^2)^{\frac{n+c_0}{2}+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-p)S^2 + d_0 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \left(\mathbf{X}'\mathbf{X}\right) V_\star V_0^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right]\right\} \\ & \times \frac{1}{(\sigma^2)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta} - \boldsymbol{\beta}_\star)' V_\star^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_\star)\right]\right\}.\end{aligned}$$

Ne segue che la distribuzione finale del parametro $(\boldsymbol{\beta}, \sigma^2)$ è ancora di tipo Normale-Gamma Inversa con parametri

$$\left(\beta_\star, V_\star, \frac{c_\star}{2}, \frac{d_\star}{2}\right) \tag{9.6}$$

dove

$$c_\star = c_0 + n, \qquad d_\star = d_0 + (n-p)S^2 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \mathbf{X}'\mathbf{X} V_\star V_0^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

In particolare, quindi, risulterà che la distribuzione finale di $\boldsymbol{\beta}$ condizionatamente al valore di σ^2 è ancora di tipo $N\left(\boldsymbol{\beta}_\star, \sigma^2 V_\star\right)$, mentre la legge marginale finale di σ^2 è ancora di tipo $\mathrm{GI}(c_\star/2, d_\star/2)$. Utilizzando il Teorema E.1 si può invece dimostrare che la distribuzione finale marginale del vettore $\boldsymbol{\beta}$ è di tipo t di Student multivariata ovvero

$$\boldsymbol{\beta} \mid \mathbf{y} \sim St_p\left(c_\star, \boldsymbol{\beta}_\star, \frac{d_\star}{c_\star} V_\star\right). \tag{9.7}$$

Pur non essendo in questo contesto molto interessante, è opportuno determinare anche la distribuzione condizionata finale $\pi(\sigma^2 \mid \boldsymbol{\beta}, \mathbf{y})$. Essa risulterà particolarmente utile in un approccio di calcolo basato sulla simulazione. Dalla (9.6) si vede facilmente che

$$\sigma^2 \mid \boldsymbol{\beta}, \mathbf{y} \sim GI(\frac{n + c_0 + p}{2}, \frac{d_0 + (n-p)S^2 + Q(\boldsymbol{\beta})}{2}) \tag{9.8}$$

9.2 Il caso non informativo

Come già visto nel capitolo 5 esistono diversi criteri per stabilire quale sia la distribuzione di riferimento per un particolare modello. Nel caso della regressione lineare, è possibile dimostrare (vedi §C.5) che la distribuzione iniziale di Jeffreys e la reference prior quando il parametro di principale interesse è $\boldsymbol{\beta}$, valgono rispettivamente

$$\pi_J(\beta, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{p+2}{2}}}, \qquad \pi_R(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

Nei calcoli che seguono indicheremo genericamente la distribuzione iniziale non informativa con il simbolo $\pi_\eta(\beta, \sigma^2) \propto 1/(\sigma^2)^{\eta+1}$: valori di η pari a $p/2$ e a 0 corrispondono alle scelte sopra descritte.

La distribuzione π_η può essere considerata un caso limite della distribuzione coniugata Normale-Gamma Inversa, ottenibile facendo crescere ad infinito le varianze a priori degli elementi del vettore β (in pratica facendo tendere la matrice V_0^{-1} ad una matrice $\mathbf{0}$, composta di tutti zeri e ponendo

$d_0 = 0$ e $c_0 = 2\eta$). Le formule del precedente paragrafo possono allora essere facilmente adattate per ottenere la distribuzione finale non informativa per (β, σ^2)

$$\pi_\eta(\beta, \sigma^2 \mid \mathbf{y}) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}+\eta+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-p)S^2 + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]\right\},$$

da cui discende, tra l'altro, che

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{y} \sim N_p\left(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right), \quad \sigma^2 \mid \mathbf{y} \sim GI(\eta + \frac{(n-p)}{2}, \frac{(n-p)S^2}{2}). \tag{9.9}$$

Il valore di η agisce soltanto sulla legge di σ^2. Utilizzando ancora il Teorema E.1 è possibile ottenere facilmente la legge marginale finale di $\boldsymbol{\beta}$ che sarà del tipo

$$\boldsymbol{\beta} \mid \mathbf{y} \sim t_p\left(2\eta + n - p, \hat{\boldsymbol{\beta}}, \frac{(n-p)S^2}{2\eta + n - p}(\mathbf{X}'\mathbf{X})^{-1}\right)$$

Va inoltre notato che i gradi di libertà della distribuzione marginale di $\boldsymbol{\beta}$ sono $2\eta + (n-p)$ e non $2\eta + n$ come l'adattamento della (9.7) potrebbe suggerire: questo si spiega con l'uso di una legge a priori costante su $\boldsymbol{\beta}$ in cui non compare il fattore $1/(\sigma^2)^{p/2}$.

Le formule ottenute meritano alcuni commenti. È evidente come, soprattutto nel caso $\eta = 0$, le formule ottenute, ricordino i risultati ottenibili in ambito classico. Anche in questo caso la stima puntuale di $\boldsymbol{\beta}$ è data dalla stima dei minimi quadrati e la varianza a posteriori di $\boldsymbol{\beta}$ equivale alla varianza calcolata, in modo frequentista, per $\hat{\boldsymbol{\beta}}$. Per quanto concerne la stima di σ^2, tenendo conto della (9.9) si ha

$$\mathbf{E}\left(\sigma^2 \mid \mathbf{y}\right) = \frac{(n-p)S^2}{2\eta + (n-p) - 2}, \qquad \mathrm{Var}\left(\sigma^2\right) = \frac{2(n-p)^2S^4}{(2\eta + (n-p) - 2)^2(2\eta + (n-p) - 4)}.$$

FARE ANCHE ESEMPIO ANOVA **Esempio 9.1** [*Regressione lineare con tre covariate (CEMENT).*]

I dati sono tratti dal testo di [31]. Consideriamo il seguente data set in cui $\mathbf{y}$ rappresenta un vettore di.... mentre le covariate X_1 e X_2 fanno riferimento a ...

♢

Difficoltà di elicitazione: le *g*-prior

L'analisi coniugata precedentemente condotta non presenta difficoltà eccessive, ma spesso non esistono sufficienti informazioni per l'elicitazione della matrice delle covarianze a priori $\mathbf{V}_0$, e la soluzione non informativa che, in qualche modo, rende le componenti indipendenti, potrebbe non rispecchiare bene l'intuizione. Una proposta computazionalmente efficace per queste situazioni è stata proposta da [87] ed è basata su un approccio bayesiano empirico: Zellner ha proposto di utilizzare una distribuzione non informativa per il parametro σ^2 ma, per quanto riguarda il vettore $\boldsymbol{\beta}$, la distribuzione iniziale proposta è

$$\boldsymbol{\beta} \mid \sigma^2 \sim N_p\left(\boldsymbol{\beta}_0, c\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right).$$

In pratica si evita di elicitare la matrice di varianze e covarianze a priori assumendo che essa risulti pari ad un multiplo c della matrice di varianze e covarianze empirica. La quantità c viene

scelta in modo da calibrare il peso relativo della funzione di verosimiglianza e della distribuzione a priori. In una versione standard delle g-prior, inoltre, si pone $\boldsymbol{\beta}_0 = \mathbf{0}$, e qui ci si adeguerà a tale scelta. La trattazione analitica del modello lineare quando si utilizzano le g-prior è ovviamente un caso particolare di quanto visto a proposito dell'analisi coniugata, sebbene i calcoli risultino particolarmente semplificati. La g-prior è infatti un caso limite della distribuzione Normale-Gamma Inversa con le seguenti scelte degli iperparametri:

$$\boldsymbol{\beta}_0 = \mathbf{0}; \quad V_0 = c\,(\mathbf{X}'\mathbf{X})^{-1}; \quad c_0 = 2\eta; \quad d_0 = 0, \tag{9.10}$$

Con questa scelta, dunque, c resta l'unico iperparametro della distribuzione a priori da elicitare. Segue allora facilmente che, in questo caso la media a posteriori di $\boldsymbol{\beta}$ (marginalmente o non) vale

$$\frac{c}{c+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \frac{c}{c+1}\hat{\boldsymbol{\beta}},$$

mentre la matrice di varianza e covarianza della legge condizionata di $\boldsymbol{\beta} \mid (\sigma^2, \mathbf{y})$ vale

$$\sigma^2 \left[\mathbf{X}'\mathbf{X}\left(1 + \frac{1}{c}\right)\right]^{-1} = \sigma^2 \frac{c}{c+1}\,(\mathbf{X}'\mathbf{X})^{-1}$$

La legge marginale a posteriori di σ^2 è invece di tipo $\mathrm{GI}(a/2, b/2)$, con

$$a = 2\eta + n; \quad b = (n-p)S^2 + \frac{1}{c+1}\hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = (n-p)S^2 + \frac{1}{c+1}\mathbf{y}'H\mathbf{y}.$$

Ricordando che $(n-p)S^2 = (\mathbf{y} - H\mathbf{y})'(\mathbf{y} - H\mathbf{y}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'H\mathbf{y}$, e la proprietà di idempotenza della matrice H, il parametro di forma della Gamma Inversa si può anche scrivere come

$$\frac{1}{c+1}\left(c(n-p)S^2 + \mathbf{y}'\mathbf{y}.\right)$$

Di conseguenza, la legge marginale finale di $\boldsymbol{\beta}$ risulta essere di tipo

$$St_p\left(n-p, \boldsymbol{\beta}_\star, \frac{c(n-p)S^2 + \mathbf{y}'\mathbf{y}}{(1+c)(2\eta+n)}(\mathbf{X}'\mathbf{X})^{-1}\right).$$

9.3 Regioni di credibilità.

La costruzione delle regioni di credibilità per il vettore dei coefficienti $\boldsymbol{\beta}$ o per il parametro σ^2 non comporta eccessive difficoltà, almeno sotto le assunzioni iniziali esplicitate nella §9.1. In situazioni non trattabili analiticamente si può ricorrere invece alle tecniche illustrate nel paragrafo §6.2. Illustreremo qui di seguito il caso in cui si adotti una legge iniziale di tipo coniugato: le formule relative al caso non informativo e delle g-prior possono essere dedotte come casi particolari, e verranno elencate alla fine del paragrafo.

Poich $\boldsymbol{\beta} \mid \mathbf{y} \sim St_p\,(c_\star, \boldsymbol{\beta}_\star d_\star V_\star / c_\star)$, la densità marginale finale di $\boldsymbol{\beta}$ è funzione decrescente della forma quadratica

$$G(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_\star)'\,V_\star^{-1}\,(\boldsymbol{\beta} - \boldsymbol{\beta}_\star)\,.$$

Inoltre, le curve di livello di della densità di $\boldsymbol{\beta}$ sono di tipo ellissoidale e la regione HPD di livello generico $1-\alpha$ si ottiene semplicemente come quel sottoinsieme di $\mathbb{R}^p$ tale che

$$\{\boldsymbol{\beta} \in \mathbb{R}^p : G(\boldsymbol{\beta}) \leq z(\alpha)\}$$

per un qualche z funzione del livello di credibilità α. Ricordando che

$$\sigma^2 \mid \mathbf{y} \sim GI\left(c_\star/2, d_\star/2\right),$$

il teorema E.2.a garantisce che

$$W = \frac{d_\star}{\sigma^2} \sim \chi^2_{c_\star},$$

Inoltre, ricorrendo al teorema E.2 , in Appendice, si ottiene

$$\frac{(\boldsymbol{\beta} - \boldsymbol{\beta}_\star)' V_\star^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_\star)}{d_\star p} \sim F_{p,c_\star},$$

e dunque l'espressione finale della regione di credibilità di livello $1-\alpha$ per il vettore $\boldsymbol{\beta}$ può essere ottenuta come

$$\left\{\boldsymbol{\beta} \in \mathbb{R}^p : G(\boldsymbol{\beta}) \leq d_\star\, p\, F^{-1}_{p,c_\star}(1-\alpha)\right\}, \tag{9.11}$$

dove il simbolo $F^{-1}_{m,n}(\gamma)$ indica il quantile di ordine γ per una v.a. di Fisher con m ed n gradi di libertà. Nel caso non informativo, utilizzando la reference prior $\pi(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$, le regione (9.11) diventa

$$\left\{\boldsymbol{\beta} \in \mathbf{R}^p : \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \leq (n-p)\, p\, S^2\, F^{-1}_{p,n-p}(1-\alpha)\right\}.$$

Formula sopra è da verificare se e' (n-p) p oppure solo p ESEMPIO CON I DATI GIA' INTRODOTTI

AGGIUNGERE FORMULE CASI NONINFO E gPRIOR

9.4 Regressione lineare attraverso metodi di simulazione

Nelle §9.1 abbiamo illustrato in modo dettagliato le procedure di calcolo analitico necessarie ad ottenere la distribuzione finale dei parametri di un modello di regressione lineare, peraltro in un contesto molto semplice, con assunzioni molto comode, come la normalità del vettore degli errori e una distribuzione a priori di tipo coniugato. Non appena qualcuna di queste assunzioni venga a mancare la strada analitica diventa molto complicata e occorre utilizzare un approccio diverso, basato sulla simulazione di un congruo numero di realizzazioni della distribuzione a posteriori dei parametri, da utilizzare poi come campione rappresentativo di tale distribuzioni per tutte le inferenze necessarie, secondo la logica descritta nel Capitolo 7. Gli stessi risultati analitici delle §§9.1 e 9.2 sono riottenibili secondo questo approccio, non appena si osservi che la distribuzione congiunta finale di $(\boldsymbol{\beta}, \sigma^2)$ è del tipo Normale-Gamma Inversa con parametri $(\boldsymbol{\beta}_\star, \mathbf{V}_\star, c_\star, d_\star)$, ed è dunque possibile effettuare direttamente simulazioni della legge a posteriori. Il codice in **R** che segue consente di produrre, a partire da un vettore di osservazioni $\mathbf{y}$ e una matrice di dati $\mathbf{X}$ con $p-1$ colonne, stime puntuali e intervallari per i parametri di interesse del modello, basate sulla distribuzione a priori di $\pi_\eta(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$.

```
function(x, y, beta0=rep(0,p),nsim=10) {
x<-cbind(1,x)
 p<-dim(x)[2]
 n<-length(y)
 k<-n-p
 simbeta<-numeric()
 simss<-numeric()
 simsd<-numeric()
```

```
sigma<-1
var<-as.matrix(solve(t(x)%*%x))
ivar<-as.matrix(t(x)%*%x)
hbeta<-as.vector(var%*%t(x)%*%y)
dev<-as.vector(t(y-x%*%hbeta)%*%(y-x%*%hbeta))
 for (i in 1:nsim)
  {
    beta  <- as.vector(rmvnorm (1, hbeta, sigma*var))
sigma<-1/rgamma(1, shape=k/2, rate = dev/2)
 sdd<-sqrt(sigma)
    simbeta <- rbind(simbeta,beta)
    row.names(simbeta)<-NULL
    simss <- c(simss,sigma)
    simsd <- c(simsd,sdd)}
stimabeta<-apply(simbeta,2,mean)
stimasig<-mean(simss)
par(mfrow=c(2,2))
d1<-density(simbeta[,1],bw="sj")
plot(d1,xlab=expression(theta[1]),main="Legge finale per theta1")
hist(simbeta[,1],prob=T,add=T)
d2<-density(simbeta[,2],bw="sj")
plot(d2,xlab=expression(theta[2]),main="Legge finale per theta2")
hist(simbeta[,2],prob=T,add=T)
d3<-density(simbeta[,3],bw="sj")
plot(d3,xlab=expression(theta[3]),main="Legge finale per theta3")
hist(simbeta[,3],prob=T,add=T)
d4<-density(simss,bw="sj")
plot(d4,xlab=expression(sigma^2),main="Legge finale per sigma2")
hist(simss,prob=T,add=T)
cat("Press <Enter> to continue...")
readline()
h2d <- hist2d(simbeta[,3],simbeta[,4],show=FALSE, same.scale=TRUE, nbins=c(20,30))
#filled.contour( h2d$x, h2d$y, xlab=expression(theta[3]), ylab=expression(theta[4]),
my2d( h2d$x, h2d$y, xlab=expression(theta[3]), ylab=expression(theta[4]),
h2d$counts, nlevels=10, col=gray((4:0)/4),main="Legge congiunta per theta3 e theta4" )
list(beta=stimabeta, stimasig=stimasig)
 }
```

Esempio 9.1 (continua).
Riprendiamo in considerazione i dati su.... e perseguiamo ora gli stessi obiettivi attraverso un approccio basato sulla simulazione.

METTERE LE DISTRIBUZIONI DI BETA1 E BETA2 E FORNIRE GLI INTERVALLI DI CREDIBILITA PER I QUATTRO PARAMETRI ♢

Esempio 9.2 [*Analisi della varianza.*]
L'analisi della varianza è un caso particolare del modello lineare che si ha in corrispondenza di una particolare struttura della matrice $\mathbf{X}$. Ne discuteremo qui soltanto un esempio introduttivo rimandando, ad esempio, a [20] per approfondimenti.

ESEMPIO DI ANOVA con due fattori ortogonali inserire anche la distribuzione di $\beta_1 - \beta_2$ ♢
Come già sottolineato nel Capitolo 7 anche nell'esempio precedente abbiamo utilizzato una delle maggiori potenzialità dell'approccio basato su simulazione, ovvero la possibilità di utilizzare il campione dalle distribuzioni a posteriori dei parametri di interesse per ottenere un'altrettanto adeguata rappresentazione della legge finale di un qualsiasi altro parametro d'interesse funzione dei parametri originari; nell'esempio precedente, la distribuzione a posteriori di $\delta = \beta_1 - \beta_2$ è stata ottenuta mediante il campione simulato $(\delta^{(1)}, \delta^{(i)}, \cdots \delta^{(n)})$ dove, per $i = 1, \cdots, n$, $\delta^{(i)} = \beta_1^{(i)} - \beta_2^{(i)}$.

9.4.1 Regressione lineare con errori a code pesanti

L'altro enorme vantaggio legato all'approccio basato su dati simulati è la possibilità di utilizzare modelli più sofisticati, più aderenti alla natura dei dati utilizzati, ma che non risulterebbero tratta-

bili per via analitica. È questa semplice constatazione che ha radicalmente cambiato l'atteggiamento degli statistici applicati nei confronti della metodologia bayesiana che, da approccio sofisticato e di interesse accademico, si è trasformata in uno strumento di lavoro essenziale. In questa sezione illustriamo brevemente come trattare il modello di regressione lineare bayesiano nel caso in cui la distribuzione degli errori $\boldsymbol{\varepsilon}$ sia di tipo St con ν gradi di libertà, dove ν, per semplicità è fornito dal ricercatore. Tale generalizzazione del modello è particolarmente utile quando si ha il sospetto che i dati presentino una curtosi superiore a quella catturabile da un modello gaussiano, o più semplicemente, quando l'analisi dei residui effettuata mediante la regressione con errori gaussiani, segnali tale extra-curtosi. Questo fenomeno avviene con particolare frequenza nelle applicazioni finanziarie [53]. Assumiamo dunque che

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim St_n\left(\nu, 0, \sigma^2 I_n\right).$$

La funzione di verosimiglianza associata al modello è dunque

$$L(\boldsymbol{\beta}, \sigma^2) \propto \prod_{i=1}^{n} \left[\frac{1}{\sigma} \left(1 + \frac{(y_i - x_i'\boldsymbol{\beta})^2}{\nu\sigma^2} \right)^{-\frac{\nu+1}{2}} \right] \tag{9.12}$$

assolutamente intrattabile dal punto di vista analitico, tanto più che non esistono distribuzioni a priori di tipo "direttamente" coniugato alla (9.12). Uno stratagemma per risolvere tale difficoltà consiste nell'utilizzo di tecniche di *data augmentation*: si veda il Capitolo 7 o, per approfondimenti, [77]. Invece di considerare la singola osservazione Y_i, $i = 1 \cdots, n$, assumiamo, artificialmente, che essa rappresenti una componente di una struttura gerarchica (Y_i, R_i), dove

$$Y_i \mid R_i \sim N(x_i'\boldsymbol{\beta}, \sigma^2 R_i), \qquad R_i \sim GI(\frac{\nu}{2}, \frac{\nu}{2})$$

e le coppie (Y_i, R_i) sono considerate mutuamente indipendenti. In altri termini si sta utilizzando ancora la rappresentazione (E.1) di Dickey per una distribuzione $\mathrm{St}(\nu, \mu, \sigma)$. Naturalmente, le R_i sono variabili non osservabili (latenti, nella terminologia statistica classica) che qui utilizzeremo esclusivamente per ottenere una rappresentazione della funzione di verosimiglianza più facile da trattare. La funzione di verosimiglianza associata al dato completo $(\mathbf{y}, \mathbf{r})$ è allora

$$L(\boldsymbol{\beta}, \sigma^2 \mathbf{r}; \mathbf{y}) \propto \prod_{i=1}^{n} p(y_i \mid r_i, \beta, \sigma^2) p(r_i)$$

$$= \frac{1}{(\sigma^2)^{n/2} \prod_{i=1}^{n} \left(r_i^{1+\frac{1+\nu}{2}} \right)} \exp\left\{ -\frac{1}{2} \left[\sum_{i=1}^{n} \frac{1}{r_i} \left(\frac{(y_i - x_i'\boldsymbol{\beta})^2}{\sigma^2} + \nu \right) \right] \right\}.$$

Adottando, per comodità di esposizione, una distribuzione a priori coniugata di tipo Normale-Gamma Inversa dove $\beta \mid \sigma^2 \sim N_p(\boldsymbol{\beta}_0, \sigma^2 \mathbf{V}_0)$ e $\sigma^2 \sim GI(c_0/2, d_0/2)$, e utilizzando calcoli simili a quelli illustrati nella §9.1 si ottiene la seguente distribuzione a posteriori congiunta[2] per i parametri di interesse e per il vettore $\mathbf{r}$:

$$\pi(\boldsymbol{\beta}, \sigma^2, \mathbf{r} \mid \mathbf{y}) \propto \frac{1}{(\sigma^2)^{\frac{n+c_0+p}{2}+1}} \frac{1}{\prod_{i=1}^{n} \left(r_i^{1+\frac{1+\nu}{2}} \right)} \exp\left\{ -\frac{1}{2\sigma^2} \left[d_0 + Q_{\boldsymbol{\Psi}}(\boldsymbol{\beta}) + (n-p) S_{\boldsymbol{\Psi}}^2 \right] - \sum_{i=1}^{n} \left(\frac{\nu}{2r_i} \right) \right\}, \tag{9.13}$$

[2] occorre soltanto tener conto della presenza delle r_i che cambia leggermente l'espressione della stima dei minimi quadrati e della devianza

dove

$$\boldsymbol{\Psi} = \text{ diag } \left(\frac{1}{r_1}, \frac{1}{r_2}, \ldots, \frac{1}{r_n}\right),$$

$$Q_{\boldsymbol{\Psi}}(\boldsymbol{\beta}) = \left(\boldsymbol{\beta} - \boldsymbol{\beta}_{\star,\boldsymbol{\Psi}}\right)' V_{\star,\boldsymbol{\Psi}}^{-1} \left(\boldsymbol{\beta} - \boldsymbol{\beta}_{\star,\boldsymbol{\Psi}}\right) + \left(\hat{\boldsymbol{\beta}}_{\boldsymbol{\Psi}} - \boldsymbol{\beta}_0\right)' \left(\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X} + V_0^{-1}\right) \left(\hat{\boldsymbol{\beta}}_{\boldsymbol{\Psi}} - \boldsymbol{\beta}_0\right),$$

$$V_{\star,\boldsymbol{\Psi}}^{-1} = (\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X} + V_0^{-1}), \quad \boldsymbol{\beta}_{\star,\boldsymbol{\Psi}} = V_{\star,\boldsymbol{\Psi}}(\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{y} + V_0^{-1}\boldsymbol{\beta}_0),$$

e infine

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\Psi}} = (\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{y}, \quad (n-p)S_{\boldsymbol{\Psi}}^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{\Psi}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{\Psi}}).$$

CONTROLLARE LA FORMA QUADRATICA!

La (9.13) ha ancora una forma analitica complessa, ma è facile verificare che le distribuzioni a posteriori di ciascun parametro e di ciascuna variabili artificiale $(r_1, \cdots, r_n)$ condizionatamente ai dati e al resto dei parametri e delle variabili artificiali assume una forma nota. Infatti

- $\boldsymbol{\beta} \mid \sigma^2, r_1, \cdots r_n \sim N_p\left(\boldsymbol{\beta}_{\star,\boldsymbol{\Psi}}, \sigma^2\mathbf{V}_{\star,\boldsymbol{\Psi}}\right)$
- $\sigma^2 \mid \boldsymbol{\beta}, r_1, \cdots r_n \sim \text{GI}\left(\frac{n+c_0+p}{2}, \frac{d_0+(n-p)S_{\boldsymbol{\Psi}}^2+Q_{\boldsymbol{\Psi}}(\boldsymbol{\beta})}{2}\right)$
- $r_i \mid \beta, \sigma^2, r_1, \cdots r_{i-1}, r_{i+1}, \cdots r_n \sim \text{GI}\left(\frac{\nu+1}{2}, \frac{1}{2}\left(\nu + \frac{(y_i - x_i'\boldsymbol{\beta})^2}{\sigma^2}\right)\right) \qquad (i = 1, \ldots, n)$

In appendice è riportato il codice in **R** per utilizzare questo modello. La versione in Bugs è reperibile in [53].

9.5 Confronto tra modelli di regressione alternativi

9.5.1 Il fattore di Bayes per modelli lineari

RUOLO DEL FATTORE DI BAYES NEL TEST DI IPOTESI E NELLA SCELTA TRA MODELLI

9.5.2 Il calcolo della marginale di y

Poich il fattore di Bayes è esprimibile attraverso il rapporto delle distribuzioni marginali del vettore dai dati **y** sotto le due ipotesi alternative, o sotto i due modelli lineari alternativi, è importante saper esprimere in forma analitica la quantità

$$p(\mathbf{y}) = \int_{\mathbf{R}^p} \int_0^\infty p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta} \mid \sigma^2)\pi(\sigma^2)d\boldsymbol{\beta}d\sigma^2.$$

Un modo semplice per ottenere tale distribuzione quando la legge a priori è di tipo Normale-Gamma Inversa è il seguente: dalle assunzioni iniziali (9.2) e (9.5) si deduce che $\mathbf{X}\boldsymbol{\beta} \mid \sigma^2 \sim N_n(\mathbf{X}\beta_0, \sigma^2\mathbf{X}\mathbf{V}_0\mathbf{X}')$ cosicch per il Lemma 6.1 applicato al caso multidimensionale, risulta

$$\mathbf{Y} \mid \sigma^2 \sim N_n\left(X\boldsymbol{\beta}_0, \sigma^2(I_n + \mathbf{X}\mathbf{V}_0\mathbf{X}')\right).$$

Utilizzando allora il Teorema E.1 si deduce che, marginalmente, **Y** ha distribuzione di tipo

$$St_n(c_0, \mathbf{X}\boldsymbol{\beta}_0, \frac{d_0}{c_0}\left(I_n + \mathbf{X}\mathbf{V}_0\mathbf{X}'\right),$$

ovvero

$$p(\mathbf{y}) = \frac{d_0^{c_0/2}\Gamma((c_0+n)/2)/\Gamma(c_0/2)}{(\pi)^{n/2} \mid I_n + \mathbf{X}\mathbf{V}_0\mathbf{X}' \mid^{1/2}} \left[d_0 + G(\mathbf{y}, \boldsymbol{\beta}_0, \mathbf{X})\right]^{-\frac{n+c_0}{2}}, \tag{9.14}$$

dove

$$G(\mathbf{y}, \boldsymbol{\beta}_0, \mathbf{X}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)'(I_n + \mathbf{X}\mathbf{V}_0\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0). \tag{9.15}$$

La 9.15 può essere espressa in diversi modi per mettere in risalto diverse caratteristiche della distribuzione stessa. Seguendo [64] si può dimostrare (vedi Appendice C.6) che

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) = (I_n + \mathbf{X}\mathbf{V}_0\mathbf{X}')\,(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\star),$$

e quindi, con facili semplificazioni, si ottiene una forma alternativa per G,

$$G(\mathbf{y}, \boldsymbol{\beta}_\star, \mathbf{X}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\star)'(I_n + \mathbf{X}\mathbf{V}_0\mathbf{X}')(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\star),$$

dove questa volta la forma quadratica G è espressa in funzione della media finale $\boldsymbol{\beta}_\star$ piuttosto che della media iniziale $\boldsymbol{\beta}_0$.

9.5.3 Uso delle *g*-priors

Come già visto in precedenza, le g-priors sono un caso particolare delle leggi a priori coniugate (9.2) e (9.5). In questo caso particolare si avrebbe

$$\boldsymbol{\beta}_0 = 0, \qquad \mathbf{V}_0 = c(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1}$$

mentre la legge a priori su σ^2 è la legge non informativa $\pi(\sigma^2) \propto \sigma^{-2}$. (DIRE PERCHE SI PUO USRAE)

da cui

$$\mathbf{X}_\gamma\mathbf{V}_0\mathbf{X}_\gamma' = \mathbf{X}_\gamma\left(\mathbf{X}_\gamma'\mathbf{X}_\gamma\right)^{-1}\mathbf{X}_\gamma' = \mathbf{H}_\gamma$$

e

$$m_\gamma(\mathbf{y}) = \frac{\Gamma(n/2)}{\pi^{n/2} \mid \mathbf{I}_n + c\mathbf{H}_\gamma \mid^{1/2}} \left(y'(\mathbf{I}_n + c\mathbf{H}_\gamma)^{-1}y\right)^{-\frac{n}{2}}$$

È possibile poi verificare che

$$(\mathbf{I}_n + c\mathbf{H}_{\boldsymbol{\gamma}})^{-1} = \mathbf{I}_n - \frac{c}{c+1}\mathbf{H}_{\boldsymbol{\gamma}}$$

; inoltre si ha[3]

$$\det(\mathbf{I}_n + c\mathbf{H}_{\boldsymbol{\gamma}}) = (c+1)^{q_\gamma+1}$$

avremo che

$$m_{\boldsymbol{\gamma}}(y) = \frac{\Gamma(n/2)}{\pi^{n/2}(c+1)^{\frac{q_\gamma+1}{2}}} \left[\mathbf{y}'\mathbf{y} - \frac{c}{c+1}\mathbf{y}'\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y}\right]^{-\frac{n}{2}}$$

Alcuni Commenti.

- L'uso delle g-priors consente di ottenere, in forma esplicita, per ogni possibile modello M_γ, un numero ($m_\gamma(\mathbf{y})$) che si può interpretare, di fatto, come una probabilità a posteriori non normalizzata.
- Tale numero dipende solo dal numero delle variabili considerate q_γ e dalla matrice di proiezione associata al modello stesso $\mathbf{H}_\gamma$.

[3] questo risultato si dimostra mediante l'uso la teoria spettrale per matrici definite positive, ricordando che le matrici idempotenti hanno tutti gli autovalori uguali a 0 oppure ad 1.)

- Quando il numero delle covariate a disposizione è relativamente piccolo (circa 10-15) è possibile calcolare $m_\gamma(\mathbf{y})$ per ogni possibile modello in competizione e scegliere quello che fornisce il più alto valore di $m_\gamma(\mathbf{y})$.
- Quando $(n-p)$ è più grande, non è possibile occorre utilizzare un algoritmo di tipo Gibbs sampler.

Esempio 9.3 *cement data*

Si hanno quattro covariate, ovvero 2^4 modelli, tutti includenti il termine noto. Ponendo $c = 5$ e assumendo una legge impropria per σ^2 si ottengono le seguenti probabilità finali per i 16 modelli

mod	c0	c12	c13	14	c15	c123	c124	c125
prob	0.0364	0.0522	0.0573	0.0441	0.0576	0.0716	0.0527	0.0713

mod	c134	c135	c145	c1234	c1235	c1245	c1345	ctutte
prob	0.0713	0.0578	0.0694	0.0718	0.0718	0.0717	0.0713	0.0718

♢

Esistono però limitazioni all'uso di questo approccio. Ad esempio, supponiamo di voler confrontare il modello "pieno, M_1, con tutte le covariate presenti, contro il modello senza covariate, M_0 (ovvero $\boldsymbol{\beta} = \mathbf{0}$).

Si può dimostrare che in questo caso, al tendere del vettore delle stime dei minimi quadrati a più infinito (ovvero, quando l'evidenza contro l'ipotesi nulla è inconfutabile), il fattore di Bayes B_{01} converge alla costante

$$(1+c)^{\frac{p-n}{2}}.$$

Questo risultato è , in sostanza, un'altra manifestazione del paradosso di Lindley, e sconsiglia, in caso di confronto tra due modelli l'utilizzo delle g-prior. In alternativa, esistono approcci basati sulle cosiddette distribuzioni iniziali intrinseche o frazionarie: si veda ad esempio, O'Hagan e Forster (2004).

9.6 Previsioni

Discutiamo ora brevemente il problema della previsione: una volta calibrato il modello di regressione sulla base di un campione di osservazioni $\mathbf{y}$ a cui è associata la matrice delle covariate $\mathbf{X}$, accade spesso di dover utilizzare tale modello per effettuare previsioni del valore di un futuro vettore di r nuove osservazioni, diciamo $\mathbf{Y}_0$, sulla base dell'informazione fornita dalla matrice di covariate associata $\mathbf{X}_0$, di dimensioni (r, p).

Il modo in cui fino ad ora abbiamo utilizzato l'informazione ci consente di rispondere facilmente a questa domanda. Da un punto di vista formale, infatti si tratta di costruire, la funzione di densità marginale di $\mathbf{Y}_0$ sulla base di tutte le informazioni a nostra disposizione, ovvero $p(\mathbf{y}_0 \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_0)$. Si può allora scrivere che

$$p(\mathbf{y}_0 \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_0) = \int_{\mathbf{R}^p} \int_0^\infty p(\mathbf{y}_0, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_0) d\boldsymbol{\beta} d\sigma^2$$

$$= \int_{\mathbf{R}^p} \int_0^\infty p(\mathbf{y}_0 \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_0, \boldsymbol{\beta}, \sigma^2) \pi(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{X}_0, \mathbf{y}) d\boldsymbol{\beta} d\sigma^2$$

Il primo fattore nell'ultima espressione è la funzione di verosimiglianza associata alle nuove osservazioni $\mathbf{y}_0$, nella quale non opera il condizionamento rispetto a $\mathbf{y}$ e $\mathbf{X}$; il secondo termine è invece la distribuzione finale dei parametri del modello nella quale non opera il condizionamento rispetto a $\mathbf{X}_0$. Poich $Y_0 \mid (\mathbf{x}_0, \boldsymbol{\beta}, \sigma^2) \sim N_r(\mathbf{X}_0\boldsymbol{\beta}, \sigma^2 I_r)$ mentre la distribuzione finale di $\boldsymbol{\beta}$ e σ^2 è di tipo Normale-Gamma Inversa con parametri dati dalla (9.6), il problema è assolutamente identico a quello del calcolo della distribuzione marginale di $\mathbf{y}$, discusso nella §9.5.2. Vale dunque il seguente risultato

Teorema 9.1 *La distribuzione predittiva finale di* $\mathbf{Y}_0$ *è di tipo*

$$St_r\left(2c_\star, \mathbf{X}_0\boldsymbol{\beta}_\star, \frac{d_\star}{c\star}(\mathbf{I}_r + \mathbf{X}_0(\mathbf{X}'\mathbf{X} + \mathbf{V}_0^{-1})^{-1}\mathbf{X}_0')\right),$$

Dimostrazione 9.1 *Si tratta di un semplice adattamento dei calcoli già effettuati per ottenere la (9.14).*

DEVO SPECIFICARE CHE SI TRATTA SOLO DEL CASO CONIUGATO???

Il modello appena descritto è sufficientemente semplice da consentire una distribuzione predittiva con forma analitica nota. È sufficiente però modificare solo qualche ingrediente del modello per complicare le espressioni analitiche e dover ricorrere per forza a simulazioni. Qui descriviamo brevemente come ottenere un campione simulato dalla legge finale di $\mathbf{Y}_0$ sotto le ipotesi precedenti: modelli più complessi richiederanno aggiustamenti specifici dell'algoritmo che resta però identico nella filosofia.

Per $t = 1, \cdots, M$,

- genera $\boldsymbol{\beta}_t \sim \pi(\boldsymbol{\beta} \mid \sigma_{(t-1)}, \mathbf{y})$ (gaussiana)
- genera $\sigma_t \sim \pi(\sigma \mid \mathbf{y})$ (Gamma Inversa)
- genera $\mathbf{y}_t \sim \pi(\mathbf{y} \mid \boldsymbol{\beta}_t, \sigma_t)$ (gaussiana)

9.7 Esercizi

10

Modelli lineari generalizzati

10.1 Introduzione ed esempi

SALVAN, IBRAHIM, O'HAGAN

10.2 Distribuzioni a priori

10.3 Tecniche di calcolo

10.4 Alcune esemplificazioni

10.4.1 Dati dicotomici

10.4.2 Dati di conteggio

10.4.3 sopravvivenza

10.5 Esercizi

11

I modelli gerarchici

11.1 Introduzione

In molte applicazioni reali le osservazioni effettuate su diverse unità statistiche possono variare non solo per effetto del caso: anche il loro valore atteso può differire in funzione di peculiari caratteristiche delle unità stesse.

Esempio 11.1 [*Prove cliniche multicentriche*]
Si vuole condurre uno studio clinico sull'efficacia di un nuovo trattamento cardiaco; il piano sperimentale prevede il coinvolgimento di più ospedali, diciamo un numero K, diversi per qualità, tipologia di utente. La quantità di interesse è la probabilità di sopravvivenza, denotata con θ_j, relativa all'ospedale j-esimo, per i pazienti sottoposti al trattamento. In un tale contesto è ragionevole supporre che i diversi θ_j, relativi a un campione di ospedali siano in qualche modo "legati" tra loro, pur rappresentando realtà differenti. ♢

Esempio 11.2 [*Modello gerarchico gaussiano*]
La tavola che segue riassume il contenuto in mg. di AAA in $n = 36$ campioni di vino relativi a $k = 6$ vitigni che indicheremo con **CA**= Cabernet, **CH**= Chianti, **SG**= Sangiovese, **SH**= Shiraz, **ME**= Merlot e **NE**= Nebbiolo.

Vitigno	Rilevazioni								$\bar{x}, s$	
CA	62	60	63	59					61	1.82
CH	63	67	71	64	65	66			66	2.83
SG	68	66	71	67	68	68			68	1.67
SH	68	66	71	67	68	68			68	1.67
ME	56	62	60	61	63	64	63	59	61	2.62
NE	56	62	60	61	63	64	63	59	61	2.62

La peculiarità di questo insieme di dati è costituita dal fatto che osservazioni relative allo stesso vitigno tendono ad essere più omogenee rispetto ad osservazioni su vitigni differenti e tali relazioni devono essere specificate nel modello statistico ♢

In queste situazioni non è più ragionevole considerare le osservazioni come n realizzazioni indipendenti di una stessa variabile aleatoria "madre"; uno strumento statistico più adeguato è la modellizzazione gerarchica. Esistono molti altri settori applicativi dove l'interesse verte su un numero spesso congruo di parametri, che possono essere ritenuti "collegati" in un senso da definire,

dalla struttura del problema; il modello statistico utilizzato in questi contesti dovrebbe, in qualche modo, incorporare la interdipendenza dei parametri θ_j.

Vedremo che un modo naturale (ma non necessario) di modellare le relazioni tra i θ_j, $j = 1, \cdots, K$, è quello di ritenerli scambiabili, ovvero considerarli come diverse realizzazioni indipendenti di un'unica variabile aleatoria.

11.2 Modelli gerarchici

Dal punto di vista modellistico, l'aspetto essenziale di una modellizzazione gerarchica è che la generica osservazione y_{ij} effettuata sull' i-esima unità del j-gruppo (ospedale, vitigno) può essere utilizzata per stimare alcuni aspetti della distribuzione dei θ_j anche se questi non sono direttamente osservabili. Da un punto di vista pratico, la modellizzazione gerarchica semplifica non poco le strategie computazionali. Esiste ormai una letteratura sterminata sulle applicazioni dei modelli gerarchici. Qui ricordiamo

- **Small Area Estimation** (Stima per piccole aree) (dire qualcosa)
- **Controllo di qualità** proporzioni di pezzi difettosi in una serie di diversi lotti di produzione provenienti dalla stessa ditta produttrice
- **Sicurezza sul lavoro** Tassi di infortuni sul lavoro osservati su un campione di ditte operanti nello stesso settore industriale
- **Disease mapping**
  Analisi epidemiologica "spaziale".
- **Meta-analysis** Utilizzo di studi precedenti simili per migliorare l'inferenza in un nuovo contesto

11.2.1 Strategie per l'analisi dei modelli gerarchici

Esistono in letteratura principalmente due approcci all'analisi dei modelli gerarchici; essi vengono definiti come

- *Empirical Bayes* (EB) Approccio bayesiano empirico
- *Hierarchical Bayes* (HB) Approccio bayesiano gerarchico.

Nel contesto EB si assume che i diversi θ_j seguano una legge *(detta a priori, ma in realtà parte integrante del modello)* del tipo

$$p(\boldsymbol{\theta} \mid \lambda)$$

considerata nota a meno del parametro (magari vettoriale) λ.

Nell'approccio EB parametrico occorre calcolare la distribuzione marginale del vettore dei dati condizionatamente a λ,

$$\begin{aligned} p(\mathbf{y} \mid \lambda) = \int_{\boldsymbol{\Omega}_1} \cdots \int_{\boldsymbol{\Omega}_K} & p(\mathbf{y} \mid \theta_1, \cdots, \theta_K) \\ & \times p(\theta_1, \cdots, \theta_K \mid \lambda) d\theta_1 \cdots d\theta_K. \end{aligned} \tag{11.1}$$

In questo modo, i parametri specifici a particolari sottogruppi della popolazione, nella fattispecie i θ_j, vengono eliminati e la (11.1) viene considerata una vera e propria funzione di verosimiglianza per il vettore λ.

Di contro, nell'impostazione gerarchica, tutte le quantità in gioco sono considerate aleatorie, e la loro distribuzione viene specificata in diversi stadi. Al primo stadio, come nel caso EB si specifica la legge delle y_{ij} condizionate al relativo θ_j

$$p(\underline{\mathbf{y}}_j \mid \theta_j)$$

dove $\underline{\mathbf{y}}_j = (y_{1j}, y_{2j}, \cdots, y_{n_j,j})$ è il vettore delle n_j osservazioni relative al gruppo j-esimo.

Al secondo stadio, come in EB, si specifica la

$$p(\theta_1, \cdots, \theta_K \mid \lambda)$$

. In genere essa risulta pari a

$$\prod_{j=1}^{K} p(\theta_j \mid \lambda),$$

ovvero i θ_j vengono assunti condizionatamente (a λ) indipendenti; in molte applicazioni, inoltre, la costruzione del modello giustifica l'ulteriore assunzione che i θ_j possano essere considerati condizionatamente somiglianti, e quindi marginalmente scambiabili. Infine al terzo stadio, si specifica anche una distribuzione di probabilità per gli iperparametri λ. È utile tenere conto del fatto che, spesso, λ è un vettore (λ_1, λ_2). Per motivi computazionali, sarà bene specificare la distribuzione di terzo stadio nel modo seguente

$$p(\lambda_1, \lambda_2) = p_1(\lambda_1 \mid \lambda_2) p_2(\lambda_2)$$

Infatti può accadere, come ad esempio nei modelli gaussiani, che alcuni dei calcoli necessari (integrazione rispetto a λ_1) siano effettuabili in modo esplicito mentre l'integrazione rispetto a λ_2 richiederà l'uso di metodi di approssimazione, o numerici, o di tipo MCMC. Entrambi gli approcci riconoscono l'incertezza (variabilità) intrinseca dei θ_j attraverso una legge governata da λ, ma va sottolineato che

- nell'approccio EB il vettore dei parametri λ viene "stimato" solo attraverso i dati, in genere attraverso la massimizzazione della funzione di verosimiglianza (11.1), oppure utilizzando il metodo dei momenti (rif ???)
- nell'approccio HB il vettore λ ha una sua distribuzione a priori, spesso non informativa e impropria, aggiornabile attraverso la $p(\mathbf{y} \mid \lambda)$.

I due metodi forniscono spesso, in termini di stime puntuali, risultati comparabili, soprattutto nel caso gaussiano (nei primi due stadi). Tuttavia, nel calcolo degli errori standard delle stime puntuali i risultati possono differire notevolmente, in genere a vantaggio del HB. Col metodo EB l'incertezza su λ è sostituita da una stima puntuale che non riesce a introdurre nella (11.1) il grado di incertezza relativo a tale operazione di stima. Per giunta, il calcolo dell'errore standard di tali stime è frequentemente e si basa su accurate tecniche di approssimazione.

Di contro, col metodo HB, la stima naturale di λ è la media a posteriori, corredata dal suo naturale indicatore di variabilità, la deviazione standard a posteriori di λ, che, per quanto complicata da ottenere in modo esplicito ha un significato naturale e può essere calcolata attraverso metodi MCMC.

Esempio 11.2 (continua).
Il modello da utilizzare è allora il seguente

Primo stadio: Dati i parametri $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$, , e σ^2, quest'ultimo considerato noto, poniamo

$$y_{ij} \overset{ind.}{\sim} N(\theta_j, \sigma^2), \qquad i = 1, \cdots, n_j; \;\; j = 1, \cdots, K$$

Si può esprimere l'informazione campionaria attraverso le statistiche sufficienti

$$\bar{y}_j \overset{ind.}{\sim} N(\theta_j, \sigma_j^2), \qquad j = 1, \cdots, K,$$

dove $\sigma_j^2 = \sigma^2/n_j$.

Secondo stadio: Oltre all'informazione campionaria non si hanno altre informazioni sui diversi θ_j ed è quindi ragionevole assumere una distribuzione a priori scambiabile per i θ_j stessi. Per convenienza computazionale assumeremo una densità gaussiana (coniugata). Condizionatamente a due iperparametri $\lambda = (\mu, \tau^2)$,

$$\theta_j \mid \mu, \tau^2 \overset{i.i.d.}{\sim} N(\mu, \tau^2), \qquad j = 1, \cdots, K.$$

Terzo stadio: nel caso HB viene specificata una distribuzione a priori sugli iperparametri (μ, τ^2), che per ora lasciamo indeterminata, $p(\mu, \tau^2)$. Per semplicità σ^2 verrà considerato noto: il caso in cui σ^2 è incognito comporta per lo più complicazioni di carattere computazionale che rimandiamo a testi più avanzati (si veda ad esempio [28] dove modelli gerarchici di diversi gradi di complessità vengono descritti e affrontati mediante il software BUGS (reference). ◇

11.3 Il modello gerarchico gaussiano

In questo paragrafo considereremo in dettaglio le elaborazioni analitiche possibili nel caso di modelli che adottano, come nell'esempio 11.2, distribuzioni gaussiane sui primi due stadi della gerarchia. In quanto segue considereremo nota la varianza di primo stadio σ^2: tale assunzione può essere rimossa ricorrendo a metodi computazionali oggi implementati in diversi software, a cominciare dal già ricordato BUGS.

11.3.1 Il caso EB

Nell'impostazione bayesiana empirica, i parametri di terzo livello $\lambda = (\mu, \tau^2)$ vengono considerati incogniti ma non aleatori, secondo una filosofia classica; ne consegue che la (11.1) diventa una vera e propria funzione di verosimiglianza, che denoteremo con $L_{EB}(\mu, \tau^2)$, per i due parametri incogniti. Qui di seguito otteniamo la forma esplicita di $L_{EB}(\mu, \tau^2)$; per agilità di notazione indicheremo con il simbolo $N(x; a; b^2)$ il valore nel punto x della densità gaussiana di media a e varianza b^2. Avremo quindi

$$L_{EB}(\mu, \tau^2) = p(\bar{\mathbf{y}} \mid \mu, \tau^2) = \int_{\mathbb{R}^k} \prod_{j=1}^{K} \left[p(\bar{\mathbf{y}} \mid \theta_j) p(\theta_j \mid \mu, \tau^2) \right] d\theta_j =$$

$$\prod_{j=1}^{K} \left[\int_{\mathrm{R}} N(\bar{y}_j, \theta_j, \sigma_j^2) N(\theta_j, \mu, \tau^2) d\theta_j \right].$$

Per il Lemma 6.1, gli integrali possono essere risolti analiticamente, per ottenere

$$L_{EB}(\mu, \tau^2) = \prod_{j=1}^{K} \left[N(\bar{y}_j, \mu, \sigma_j^2 + \tau^2) \right] \propto \frac{1}{\prod_j \sqrt{\sigma_j^2 + \tau^2}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^{K} \frac{(\bar{y}_j - \mu)^2}{\sigma_j^2 + \tau^2} \right\} \tag{11.2}$$

L'ultima espressione viene poi di solito massimizzata rispetto a μ e τ^2 per ottenere le stime EB dei parametri di terzo stadio. Tali parametri, tuttavia, potrebbero non rappresentare le quantità di diretto interesse inferenziale; è possibile ad esempio che l'interesse verta su alcuni (o tutti) i parametri θ_j che rappresentano gli effetti specifici di alcune delle nostre macro unità (ospedali e vitigni, nei due esempi). In questo caso l'impostazione EB si riduce ad una analisi bayesiana standard in cui le distribuzioni iniziali sul vettore $\boldsymbol{\theta}$ dipende dagli iperparametri $\hat{\mu}$ e $\hat{\tau}^2$ stimati mediante massimizzazione della $L_{EB}(\mu, \tau^2)$. Illustreremo tali calcoli nella §11.3.2.

11.3.2 L'approccio HB

Nell'impostazione gerarchica HB, tutte le quantità presenti nel modello sono considerate aleatorie e dotate di legge di probabilità. La distribuzione a posteriori congiunta del vettore dei parametri $(\boldsymbol{\theta}, \mu, \tau^2)$ risulta così proporzionale a

$$p(\underline{\boldsymbol{\theta}}, \mu, \tau^2 \mid \mathbf{y}) \propto p(\mu, \tau^2) \prod_{j=1}^{K} N(\theta_j \mid \mu, \tau^2) \prod_{j=1}^{K} N(\bar{y}_j \mid \theta_j, \sigma_j^2) \tag{11.3}$$

Da tale espressione è possibile ottenere in via esplicita o numerica diverse distribuzioni (o sintesi di queste) di interesse; nei prossimi sotto paragrafi presenteremo in un certo dettaglio alcuni dei calcoli necessari.

La legge finale di $\boldsymbol{\theta}$ per $(\boldsymbol{\mu}, \boldsymbol{\tau^2})$ fissati

Iniziamo con il considerare la legge finale dei vari θ_j per valori fissati degli iperparametri fissati: questo tipo di analisi è importante soprattutto in applicazioni, come quella nel campo delle piccole aree, dove è importante stimare le grandezze relative alle singole aree geografiche, o come negli esempi precedenti, ai singoli ospedali o vitigni. Va inoltre notato che tale analisi è importante anche nel caso EB laddove i valori condizionanti degli iperparametri siano quelli relativi alle loro stime EB, $\hat{\mu}$ e $\hat{\tau}^2$, ottenute come nella §11.3.1.

Condizionatamente agli iperparametri, i vari θ_j risultano indipendenti e appaiono in differenti fattori della funzione di verosimiglianza, cosicch $p(\boldsymbol{\theta} \mid \mu, \tau^2, \mathbf{y})$ *fattorizza* in K diverse componenti, in ognuna delle quali è possibile utilizzare il fatto che la distribuzione *iniziale* dei θ_j è coniugata al corrispondente fattore della funzione di verosimiglianza. Si ha quindi, per la (4.6,

$$\theta_1, \cdots \theta_K \mid \tau, \mu, \mathbf{y} \overset{ind.}{\sim} N(\hat{\theta}_j, V_j),$$

dove

$$\hat{\theta}_j = \frac{\frac{\bar{y}_j}{\sigma_j^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \qquad V_j = \frac{\sigma_j^2 \tau^2}{\sigma_j^2 + \tau^2}, \tag{11.4}$$

Da notare fin d'ora che le medie finali dei vari θ_j sono combinazioni convesse delle medie campionarie specifiche e di una componente comune, che per ora è rappresentata da μ, la media generale.

La La legge finale di (μ, τ^2)

Consideriamo qui l'espressione della distribuzione finale degli iperparametri. In generale si ha

$$p(\mu, \tau^2 \mid \mathbf{y}) \propto p(\mu \mid \tau^2) p(\tau^2) p(\mathbf{y} \mid \mu, \tau^2),$$

ma questa formula raramente è di qualche aiuto perch non è facile esprimere in forma esplicita il secondo fattore. Nel caso del modello gerarchico gaussiano, però, questo è possibile, ricordando il Lemma 6.1, e le elaborazioni già viste nella §11.3.1.

Nella stragrande maggioranza delle applicazioni si pone $p(\mu \mid \tau^2) \propto 1$, sebbene sia possibile ottenere risultati analitici anche ponendo una ulteriore distribuzione gaussiana al terzo stadio: la scelta di non informatività sul parametro μ è giustificata con il fatto che l'informazione campionaria su tale parametro sarà in genere sufficiente, in quanto tutte le osservazioni campionarie concorreranno a fornire informazioni su μ. Con questa scelta, e lasciando per ora indeterminata la scelta di $p(\tau^2)$, si avrà che

$$p(\mu, \tau^2 \mid \mathbf{y}) \propto p(\tau^2) L_{EB}(\mu, \tau^2) \propto \frac{1}{\prod_j \sqrt{\sigma_j^2 + \tau^2}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{K} \frac{(\bar{y}_j - \mu)^2}{\sigma_j^2 + \tau^2} \right\}$$

La La legge finale di

$$\boldsymbol{\mu \mid \tau^2}$$

Per le assunzioni fatte al punto precedente, la quantità di interesse $p(\mu \mid \tau^2, \mathbf{y})$ si può fattorizzare in

$$p(\mu \mid \tau^2, \mathbf{y}) \propto p(\mu \mid \tau^2) p(\mathbf{y} \mid \tau^2 \mu) \propto L_{EB}(\mu, \tau^2) \tag{11.5}$$

È possibile elaborare ulteriormente la (11.2) (vedi Appendice ????) per ottenere

$$\begin{aligned} L_{EB}(\mu, \tau^2) &\propto \frac{1}{\prod_j \sqrt{\sigma_j^2 + \tau^2}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{K} \frac{(\bar{y}_j - \mu)^2}{\sigma_j^2 + \tau^2} \right\} \\ &\propto \sqrt{V_\mu} \exp\left\{ -\frac{1}{2} V_\mu (\mu - \hat{\mu})^2 \right\}, \end{aligned}$$

dove

$$\hat{\mu} = \frac{\sum_{j=1}^{K} \frac{\bar{y}_j}{\sigma_j^2 + \tau^2}}{\sum_{j=1}^{K} \frac{1}{\sigma_j^2 + \tau^2}} \tag{11.6}$$

rappresenta la media delle medie osservate sui singoli campioni, ponderate con le rispettive precisioni (ovvero i reciproci delle varianze); inoltre

$$V_\mu^{-1} = \sum_{j=1}^{K} \frac{1}{\sigma_j^2 + \tau^2} \tag{11.7}$$

rappresenta la somma delle precisioni. In pratica la legge finale di $\mu \mid \tau^2$ è gaussiana con media fornita dalla (11.6) e varianza espressa dal reciproco della (11.7)

In tal modo, qualora il parametro d'interesse sia la media generale, e i singoli θ_j rappresentano parametri incidentali, relativi ad esempio ad ospedali che sono stati osservati per effetto di

estrazione casuale, ma che non rappresentano il fine ultimo della nostra analisi, la stima bayesiana di μ avviene attraverso l'integrazione di diverse fonti informative, provenienti dalle medie osservate nei singoli sottogruppi, opportunamente ponderate attraverso le rispettive precisioni.

Fino a qui

La legge finale di τ^2.

Si può ottenere in modo esplicito dalla (11.5) mediante un semplice trucco. Poich

$$p(\tau^2 \mid \mathbf{y}) = \frac{p(\tau^2, \mu \mid \mathbf{y})}{p(\mu \mid \tau^2, \mathbf{y})} \propto \frac{p(\tau^2) \prod_{j=1}^{K} N(\bar{y}_j \mid \mu, \sigma_j^2 + \tau^2)}{N(\mu \mid \hat{\mu}, V_\mu)},$$

tale identità deve valere per ogni valore di μ (la cui presenza nella formula è ... *virtuale*). In particolare deve valere per $\mu = \hat{\mu}$. Quindi,

$$p(\tau^2 \mid \mathbf{y}) \propto \frac{p(\tau^2) \prod_{j=1}^{K} N(\bar{y}_j \mid \hat{\mu}, \sigma_j^2 + \tau^2)}{N(\hat{\mu} \mid \hat{\mu}, V_\mu)} \propto$$

$$p(\tau^2) \sqrt{V_\mu} \prod_{j=1}^{K} \frac{1}{\sqrt{\sigma_j^2 + \tau^2}} \exp\left\{ -\frac{1}{2} \frac{(\bar{y}_j - \hat{\mu})^2}{\sigma_j^2 + \tau^2} \right\}$$

Tale espressione è abbastanza complessa perch $\hat{\mu}$ e V_μ dipendono da τ^2 ma comunque è possibile disegnare la distribuzione a posteriori (non normalizzata) di τ^2. Fin qui i calcoli sono stati effettuati per una generica distribuzione iniziale per τ^2: nella prossima sezione vengono discussi alcuni aspetti legati a tale scelta.

11.3.3 Sulla scelta della distribuzione a priori di τ^2

Se $p(\tau^2)$ è impropria dobbiamo verificare che la distribuzione a posteriori sia propria. Questo problema non si pone se utilizziamo una distribuzione propria. Ne proveremo 3:

- $p(\tau) \propto 1$
- $p(\tau^2) \propto 1$
- $p(\tau^2) \sim IG(3, 20)$
  (Gamma inversa, propria)

  Nota: Se $X \sim IG(\alpha, \beta)$ la funzione di densità è

$$f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp\left\{ -\frac{\beta}{x} \right\} \frac{1}{x^{\alpha+1}}$$

Con tutte e tre le a priori utilizzate si arriva ad una distribuzione a posteriori propria. Questo non accade però utilizzando la consueta distribuzione non informativa per parametri di scala $p(\tau^2) \propto \tau^{-2}$

Infatti,

$$p(\tau^2 \mid \mathbf{y}) \propto p(\tau^2) \prod_{j=1}^{k} \left(\frac{1}{\sqrt{\sigma_j^2 + \tau^2}} \right) V_\mu^{1/2} \exp\left\{ -\frac{1}{2} \sum_j \frac{(y_j - \hat{\mu})^2}{\sigma_j^2 + \tau^2} \right\}.$$

È facile allora verificare che, per $\tau^2 \to 0$,

$$\prod_{j=1}^{k}\left(\frac{1}{\sqrt{\sigma_j^2+\tau^2}}\right) \to \prod_{j=1}^{k}\frac{1}{\sigma_j} \qquad \text{costante}$$

$$\hat{\mu} = \frac{\sum_{j=1}^{k}\frac{y_j}{\sigma_j^2+\tau^2}}{\sum_{j=1}^{k}\frac{1}{\sigma_j^2+\tau^2}} \to \frac{\sum_{j=1}^{k}\frac{y_j}{\sigma_j^2}}{\sum_{j=1}^{k}\frac{1}{\sigma_j^2}} \qquad \text{costante}$$

$$\exp\left\{-\frac{1}{2}\sum_j \frac{(y_j-\hat{\mu})^2}{\sigma_j^2+\tau^2}\right\} \to \text{costante}$$

Ne segue che, in un intorno di zero,

$$p(\tau^2 \mid \mathbf{y}) \approx p(\tau^2)$$

ed è allora necessario utilizzare una distribuzione a priori integrabile per $\tau^2 \to 0$.

11.4 Il calcolo dei momenti a posteriori

In questa sezione

11.4.1 Media e varianza dei θ_j

.

I momenti non possono essere ottenuti in forma esplicita. Usando le (11.4), la (11.6) e la (11.7) possiamo calcolare medie e varianze dei parametri d'interesse attraverso integrali numerici unidimensionali.

1^o passo μ e τ^2 fissati, (utile nel caso HB e EB)

utilizzando la (11.4) si ha

$$\hat{\theta}_j^{(1)} = \hat{\theta}_j(\tau^2, \mu, \mathbf{y}) = \frac{\tau^2\bar{y}_j + \sigma_j^2\mu}{\tau^2+\sigma_j^2}$$

e

$$V_j^{(1)} = V_j^{(1)}(\theta_j \mid \tau^2, \mu, \mathbf{y}) = V_j^{(1)}(\tau^2) = \frac{\sigma_j^2\tau^2}{\sigma_j^2+\tau^2}$$

2^o passo τ^2 fissato, e integrando rispetto a μ

(la legge a priori è $p(\mu \mid \tau^2) \propto 1$) (utile nel caso HB)

Integrando rispetto a μ si ottengono le stime HB per τ^2 noto:

$$\hat{\theta}_j^{(2)} = \hat{\theta}_j^{(2)}(\tau^2, \mathbf{y}) = E^{\mu}\left(\theta_j \mid \tau^2, \mathbf{y}\right)$$

$$\begin{aligned}
&= \int_{\boldsymbol{\Omega}_j} \theta_j p(\theta_j \mid \tau^2, \mathbf{y}) \\
&= \int_{\mu}\int_{\boldsymbol{\Omega}_j} \theta_j p(\theta_j \mid \mu, \tau^2, \mathbf{y}) p(\mu \mid \tau^2, \mathbf{y}) \\
&= \int_{\mu} \hat{\theta}_j^{(1)} p(\mu \mid \tau^2, \mathbf{y}) \\
&= \frac{\tau^2\bar{y}_j}{\tau^2+\sigma_j^2} + \frac{\sigma_j^2}{\tau^2+\sigma_j^2} E(\mu \mid \tau^2, \mathbf{y}) \\
&= \frac{\tau^2\bar{y}_j}{\tau^2+\sigma_j^2} + \frac{\sigma_j^2}{\tau^2+\sigma_j^2}\hat{\mu}
\end{aligned}$$

La varianza a posteriori per τ^2 fissato è con

$$\begin{aligned} V_j^{(2)} &= V_j^{(2)}(\theta_j \mid \tau^2, \mathbf{y}) = V_j^{(2)}(\tau^2) \\ &= E[V(\theta_j \mid \mu, \tau^2, \mathbf{y}) \mid \tau^2, \mathbf{y}] + V[E(\theta_j \mid \mu, \tau^2, \mathbf{y}) \mid \tau^2, \mathbf{y}] \\ &= E^{\mu}(V(\theta_j \mid \mu, \tau^2, \mathbf{y})) + V^{\mu}E(\theta_j \mid \mu, \tau^2, \mathbf{y})) \\ &= V_j^{(1)}(\tau^2) + V^{\mu}\left(\frac{\tau^2\bar{y}_j + \sigma_j^2\mu}{\sigma_j^2 + \tau^2}\right) \\ &= V_j^{(1)}(\tau^2) + \frac{\sigma_j^4}{(\sigma_j^2 + \tau^2)^2}V(\mu \mid \tau^2, \mathbf{y}) \\ &= V_j^{(1)}(\tau^2) + \frac{\sigma_j^4}{(\sigma_j^2 + \tau^2)^2}V_{\mu}(\tau^2) \end{aligned}$$

11.5 Le stime finali

Se poi si integra rispetto a τ^2, si ottengono le stime HB conclusive dei parametri e relativi errori standard.

$$\hat{\theta}_{jHB} = \int_{\tau^2} \hat{\theta}_j^{(2)}(\tau^2, \mathbf{y}) p(\tau^2 \mid \mathbf{y}) d\tau^2$$

da risolvere numericamente.

Per quanto riguarda la varianza

$$\begin{aligned} V_{jHB} &= V(\theta_j \mid \mathbf{y}) = E\left(V(\theta_j \mid \tau^2, \mathbf{y})\right) + V\left(E(\theta_j \mid \tau^2, \mathbf{y})\right) \\ &= E\left(V_j^{(2)}(\tau^2) \mid \mathbf{y}\right) + V\left(\hat{\theta}_j^{(2)}(\tau^2, \mathbf{y}) \mid \mathbf{y}\right) \\ &= E\left(V_j^{(1)}(\tau^2) \mid \mathbf{y}\right) + E\left(\frac{\sigma_j^4}{(\sigma_j^2 + \tau^2)^2}V_{\mu}(\tau^2) \mid \mathbf{y}\right) + V\left(\hat{\theta}_j^{(2)}(\tau^2, \mathbf{y}) \mid \mathbf{y}\right) \\ &= V_{jHB}^{(I)} + V_{jHB}^{(II)} + V_{jHB}^{(III)} \end{aligned}$$

- $V_{jHB}^{(I)}$ misura l'incertezza relativa alla stima di θ_j per μ e τ^2 noti
- $V_{jHB}^{(II)}$ misura l'incertezza relativa alla stima di θ_j per il solo τ^2 noto (si inserisce la legge $p(\mu \mid \tau^2)$)
- $V_{jHB}^{(III)}$ misura l'incertezza aggiuntiva relativa alla stima di θ_j fornita da $p(\tau^2)$

11.5.1 La Strategia EB

Invece di specificare una distribuzione a priori per μ e τ^2 si possono ottenere stime EB dei vari θ_j sostituendo a μ e τ^2 delle opportune stime. Denoteremo con

$$\hat{\theta}_{jEB} = \hat{\theta}_j^{(1)}(\hat{\tau}^2, \hat{\mu}(\hat{\tau}^2), \mathbf{y})$$

la stima EB di θ_j, dove le stima $\hat{\tau}^2$ è in genere una stima di tipo ML, REML, oppure ANOVA.

Pe il calcolo degli errori standard si possono produrre due diverse stime

- $V_{jEB}^{(I)} = V_j^{(1)}(\hat{\tau}^2)$
- $V_{jEB}^{(II)} = V_j^{(1)}(\hat{\tau}^2) + \frac{\sigma_j^4}{(\sigma_j^2+\tau^2)^2}V_{\mu}(\hat{\tau}^2)$

La prima stima è piuttosto *naive* perch, non tenendo conto della variabilità rispetto a μ e τ^2, sottostima seriamente gli errori standard, con conseguente cattiva copertura degli intervalli di confidenza

Nella seconda si introduce la variabilità rispetto a μ ma non rispetto a τ^2

La difficoltà nel riportare "buone"' misure della variabilità associata alle stime è il limite principale dell'approccio EB

Esempio: continua I risultati relativi a diverse combinazioni di stime EB e di distribuzioni a priori per μ e τ^2.

$\bar{y}_j$	$\hat{\theta}_j^{HB}$	$\hat{\theta}_j^{HB}$	$V_{jHB}^{(I)}$	$V_{jHB}^{(II)}$	$V_{jHB}^{(III)}$	$V_{jEB}^{(I)}$	$V_{jEB}^{(II)}$
61	61.22	61.41	1.30	0.024	0.044	1.21	1.255
66	65.90	65.81	0.88	0.012	0.010	0.84	0.867
68	67.80	67.62	0.88	0.012	0.041	0.84	0.867
61	61.12	61.22	0.67	0.007	0.015	0.65	0.662

Tabella 11.1. Tavola 2: Stime EB di tipo ANOVA e $\mathbf{p}(\mu, \tau) \propto 1$.

$\bar{y}_j$	$\hat{\theta}_j^{HB}$	$\hat{\theta}_j^{HB}$	$V_{jHB}^{(I)}$	$V_{jHB}^{(II)}$	$V_{jHB}^{(III)}$	$V_{jEB}^{(I)}$	$V_{jEB}^{(II)}$
61	61.44	61.32	1.19	0.048	0.033	1.25	1.286
66	65.80	65.85	0.84	0.024	0.008	0.86	0.881
68	67.59	67.70	0.84	0.024	0.032	0.86	0.881
61	61.24	61.17	0.64	0.014	0.012	0.66	0.670

Tabella 11.2. Stime EB di tipo REML e $\mathbf{p}(\mu, \tau^2) = IG(\tau^2; 3, 20)$

$\bar{y}_j$	$\hat{\theta}_j^{HB}$	$\hat{\theta}_j^{HB}$	$V_{jHB}^{(I)}$	$V_{jHB}^{(II)}$	$V_{jHB}^{(III)}$	$V_{jEB}^{(I)}$	$V_{jEB}^{(II)}$
61	61.11	61.32	1.35	0.012	0.022	1.25	1.286
66	65.95	65.85	0.91	0.006	0.005	0.86	0.881
68	67.90	67.70	0.91	0.006	0.019	0.86	0.881
61	61.06	61.17	0.69	0.003	0.007	0.66	0.670

Tabella 11.3. Stime EB di tipo REML e $\mathbf{p}(\mu, \tau^2) \propto 1$

.

Le analisi succitate possono essere facilmente implementate con BUGS. Il codice che segue si riferisce a ...

```
model
```

```
{
for( j in 1 : T ) {
for( i in 1 : N ) {
Y[i , j] ~ dnorm(te[ j],sig.n)}
te[ j] ~ dnorm(mu,tau.n)
}
  tau.n~ dgamma(0.001,0.001)
  mu~ dnorm(0.0,1.0E-6)
}

#Dati
list(N=8, T=4, sig.n=1,
Y = structure(
        .Data = c(62,63,68,56,
                  60,67,66,62,
                  63,71,71,60,
                  59,64,67,61,
                  NA,65,68,63,
                  NA,66,68,64,
                  NA,NA,NA,63,
                  NA,NA,NA,59),
        .Dim = c(8,4)))

#Dati iniziali
list(te = c(0,0,0,0)
     tau.n = 1,mu= 0)
```

Previsioni. Vengono di solito considerate due tipi di previsioni nell'ambito dei modelli gerarchici. Nel linguaggio dell'esempio precedente, esse corrispondono a

- una futura osservazione y_j^* relativa ad uno specifico θ_j già considerato
- una futura osservazione y^* relativa ad un futuro θ^*, non ancora considerato

Nel primo caso si tratterebbe di una nuova rilevazione effettuata sotto una delle 4 diete oggetto di studio.

Nel secondo caso si tratterebbe di una nuova rilevazione effettuata sotto una nuova dieta ("scambiabile" con le altre")

Caso A

$$p(y_j^* \mid \underline{\mathbf{y}}) = \int_{\boldsymbol{\Omega}_j} p(\theta_j, y_j^* \mid \underline{\mathbf{y}}) d\theta_j$$

$$= \int_{\boldsymbol{\Omega}_j} p(y_j^* \mid \theta_j) p(y_j^* \mid \underline{\mathbf{y}}) d\theta_j$$

Dal punto di vista computazionale,

- Si generano $\theta_j^{(1)}, \theta_j^{(2)}, \cdots, \theta_j^{(M)}$ da $\theta_j \mid \underline{\mathbf{y}}$. (a posteriori di θ_j)
- $\forall h = 1, \cdots, M$ si genera $y_j^{(h)}$ da $y \mid \theta_j^{(h)}$
- Si considera l'istogramma delle $y_j^{(h)}$ generate come stima della densità predittiva a posteriori $p(y_j^* \mid \underline{\mathbf{y}})$

$$p(y^* \mid \underline{\mathbf{y}}) = \int_{\Lambda} \int_{\boldsymbol{\Omega}} p(y^*, \theta, \lambda \mid \underline{\mathbf{y}}) d\theta d\lambda$$
$$= \int_{\Lambda} \int_{\boldsymbol{\Omega}} p(y^* \mid \theta) p(\theta \mid \lambda, \underline{\mathbf{y}}) p(\lambda \mid \underline{\mathbf{y}})$$

e un'approssimazione di Monte Carlo dell'integrale si ottiene in modo simile al caso A:

-
-
-

Ci sono $K = 12$ ospedali che effettuano un particolare intervento chirurgico al cuore su neonati. Per ognuno di essi si registra il numero di interventi n_j e il numero di decessi r_j.

		Ospedali										
	A	B	C	D	E	F	G	H	I	J	K	L
n	47	148	119	114	150	200	79	187	313	97	256	360
r	0	18	8	12	19	21	4	21	22	8	29	24

Analizzeremo questo problema mediante un modello bayesiano di regressione logistica con effetti casuali (modello gerarchico) Formalmente,

- Verosimiglianza:
  Risposte binomiali con effetti random

$$r_j \sim \mathbf{Binomial}(\pi_j, n_j)$$
$$\mathbf{logit}(\pi_j) = \alpha_j$$
$$\alpha_j \sim N(\mu, \frac{1}{\tau})$$

- Priori

$$\mu \sim N(0, 10^6)$$
$$\tau \sim G(10^{-3}, 10^{-3})$$
$$\sigma = \frac{1}{\sqrt{\tau}}$$

Programma BUGS

```
model {
# Likelihood
for (i in 1:K) {
r[i]   dbin(p[i], n[i]);
logit(p[i]) ¡-alpha[i];
alpha[i] dnorm(mu,tau)
}
# Priors:
```

```
mu   dnorm(0.0, 1.0E-6); # (in termini di precisione)
tau   dgamma(1.0E-3, 1.0E-3);
sigma ¡- 1/sqrt(tau);
pop.mean ¡-exp(mu/(1+mu))
}
   } # Dati
list(K=12,
n=c(47,148,119,114,150,200,79,187,313,97,256,360),
r=c(0,18,8,12,19,21,4,21,22,8,29,24))

   list(mu=0,tau=1)
```

(i valori iniziali possono essere scelti oppure generati da WinBugs

Alcuni consigli pratici

- Parametrizzazione
- Scelta delle a priori

Parametrizzazione

I campioni MCMC mostrano a volte uno scarso "mixing, ovvero il "sampler" non si muove rapidamente nel supporto della distribuzione. Questo causa

- Lenta convergenza e una maggiore varianza Monte Carlo
- Le catene tendono ad essere correlate

La causa per il secondo problema è spesso la forte correlazione a posteriori tra i parametri d'interesse.

Un semplice Esempio

Consideriamo un semplice modello di Regressione lineare.

$$
\begin{aligned}
y_i &\sim N(\mu, \sigma^2) \\
\mu_i &= \alpha + \beta x_i \\
p(\alpha, \beta) &\propto 1
\end{aligned}
$$

La correlazione a posteriori tra α e β è

$$\rho_{\alpha,\beta} = -\frac{\bar{x}}{\sqrt{\bar{x}^2 + \frac{1}{n}\sum_i (x_i - \bar{x})^2}}$$

- $\bar{x} >> sd(x) \Rightarrow \rho_{\alpha,\beta} \to \pm 1$
- Un rimedio: standardizzare le x_i cosicch

$$\mu_i = \alpha' + \beta'(x_i - \bar{x})$$

e

$$\rho_{\alpha',\beta'} = 0$$

Scelta delle distribuzioni a priori C'è spesso l'esigenza di utilizzare distribuzioni a priori "poco" informative. Va subito chiarito che BUGS **pretende** distribuzioni a priori proprie. Occorre poi distinguere tra

- Parametri *primari*: i parametri d'interesse
- Parametri *secondari* o di disturbo, in genere inseriti per una maggiore flessibilità del modello

Alcuni casi particolari

- **parametri di posizione**
  in genere si utilizza una distribuzione quasi uniforme, ad esempio per un coefficiente di regressione β,
  $$\beta \sim N(0, 10^6)$$
- **parametri di scala** Le a priori standard sono
  $$p(\tau) \propto \frac{1}{\tau}, \qquad p(\sigma^2) \propto \frac{1}{\sigma^2}$$
  Come già visto, al secondo livello gerarchico queste a priori producono distribuzioni a posteriori improprie poich nell'intorno di $\sigma^2 = 0$ la funzione di verosimiglianza è non trascurabile.

Possibili opzioni

- Utilizzare una distribuzione propria del tipo $\tau \sim G(\varepsilon, \varepsilon)$
- Utilizzare una distribuzione propria del tipo $\sigma \sim U(0, r)$
- Processo di elicitazione
- Scegliere un'altra distribuzione a priori ($p(\tau) \propto 1$)

11.6 Approccio basato sulla simulazione

11.7 Conclusioni

Il non uso di modelli gerarchici quando i dati hanno una struttura gerarchica (come quella dell'esempio) crea non pochi problemi

- Se usati con pochi parametri essi producono un cattivo adattamento per data set molto grandi
- Se usati con molti parametri, essi producono il fenomeno dell'*overfitting* che produce buon adattamento ai dati esistenti ma pessime capacità previsive.

11.8 Esercizi

12

Approfondimenti

12.1 Modelli a struttura latente

12.1.1 Mistura finita di distribuzioni gaussiane

12.1.2 Frontiera stocastica

12.2 Il problema della stima della numerosità di una popolazione

12.3 Scelta della numerosità campionaria

12.4 Esercizi

A

Alcune nozioni di algebra lineare

Definizioni preliminari

Una matrice $\mathbf{A}$ si dice quadrata se ha lo stesso numero d di righe e colonne. Relativamente ad una matrice quadrata $\mathbf{A}$ si dice che

- A è *simmetrica* se $a_{ij} = a_{ji}, \qquad i, j = 1, \cdots d$;
- A è *definita positiva* (*definita non-negativa*) se

$$\mathbf{t}^{'}\mathbf{A}\mathbf{t} > 0; \qquad (\mathbf{t}^{'}A\mathbf{t} \geq 0)$$

  per ogni vettore $\mathbf{t} \in \mathbb{R}^d$, $\mathbf{t} \neq \mathbf{0}_d$, dove $\mathbf{0}_d$ è un vettore composto da d zeri.
- A è *ortogonale* se $A^{'} = A^{-1}$ (ovvero se, $A^{'}A = AA^{'} = \mathbf{I}_d$).
- Si dice che $\lambda \in \mathbb{C}$ è un *autovalore* di A se $A - \lambda I$ è singolare (o, in modo equivalente se $\det(A - \lambda I) = 0$). Se $\mathbf{x} \in \mathbb{C}^d \backslash \{0\}$ soddisfa la relazione $A\mathbf{x} = \lambda\mathbf{x}$, allora $\mathbf{x}$ è un *autovettore* associato all'autovalore λ.

  Si definisce *traccia* di $\mathbf{A}$ la somma degli elementi sulla diagonale principale, ovvero

$$\mathrm{tr}(\mathbf{A}) = \sum_{j=1}^{k} a_{jj}.$$

La traccia gode di alcune proprietà. Se $\mathbf{A}$ e $\mathbf{B}$ sono due matrici quadrate di dimensione d, allora

$$\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B}) \tag{A.1}$$

e

$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}). \tag{A.2}$$

La proprietà (A.1) garantisce che la traccia è un operatore lineare. La proprietà (A.2) continua a valere anche in caso di matrici non quadrate ma tali che lo siano i loro prodotti, ovvero nel caso in cui $\mathbf{A}$ ha dimensione $k \times h$ e $\mathbf{B}$ ha dimensione $h \times k$.
Si può dimostrare che se una matrice simmetrica è definita positiva allora i suoi k autovalori sono tutti strettamente positivi.

Teorema A.1 [*della decomposizione spettrale*].
Se $\mathbf{A}$ *è una matrice simmetrica definita positiva, esiste una matrice ortonormale*[1] $\mathbf{Q}$ *della stessa dimensione di* $\mathbf{A}$ *tale che*

[1] tale cioè, che le colonne hanno norma pari a 1, sono tra loro ortogonali, e vale la relazione $\mathbf{Q}' = \mathbf{Q}^{-1}$.

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}',$$

dove $\boldsymbol{\Lambda} = \text{diag}\,(\lambda_1, \cdots, \lambda_k)$ *è una matrice diagonale costruita con gli autovalori di* $\mathbf{A}$*, mentre le colonne di* $\mathbf{Q}$ *sono gli autovalori di* $\mathbf{A}$*.*

Da quanto sopra si deduce anche, ricordando proprietà elementari del determinante e dell'inversa di una matrice quadrata, che

$$\mid \mathbf{A} \mid = \prod_{j=1}^{k} \lambda_j$$

e

$$\mathbf{A}^{-1} = \mathbf{Q}\boldsymbol{\Lambda}^{-1}\mathbf{Q}'.$$

Data una matrice simmetrica di dimensione k definita positiva, è possibile definire la matrice *radice quadrata* di $\mathbf{A}$ ovvero quella matrice $\mathbf{R}$ tale che $\mathbf{A} = \mathbf{R}\mathbf{R}'$; in virtù del teorema precedente basta porre

$$\mathbf{R} = \mathbf{Q}\boldsymbol{\Lambda}^{1/2} = \mathbf{Q}\,\text{diag}\left(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_k}\right).$$

Da questo si deduce, inoltre, che

$$\mid \mathbf{R} \mid = \mid \mathbf{A} \mid^{1/2}; \quad \mathbf{R}^{-1} = \boldsymbol{\Lambda}^{-1/2}\mathbf{Q}'.$$

B

Nozioni di probabilità

Sia Ω lo spazio di tutti i possibili risultati di un esperimento. Valutare in modo probabilistico un esperimento significa, sia in termini tecnici che sostanziali, saper calcolare la probabilità che il risultato dell'esperimento appartenga ad un sottoinsieme B di Ω, qualunque sia B appartenente ad una famiglia $\mathcal{F}$ di sottoinsiemi. Ragioni tecniche impongono che la famiglia $\mathcal{F}$ sia una σ-algebra, ovvero soddisfi le seguenti proprietà:

- $\emptyset \in \mathcal{F}$ (l'insieme vuoto è in $\mathcal{F}$)
- $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ ($\mathcal{F}$ è chiusa rispetto all'unione numerabile)
- $A \in \mathcal{F} \Rightarrow A^c = \Omega \backslash A \in \mathcal{F}$ ($\mathcal{F}$ è chiusa rispetto alla complementazione)

Una misura di probabilità $\mathbb{P}$ è una funzione d'insieme che associa, ad ogni elemento di $\mathcal{F}$, un valore reale compreso nell'intervallo $[0, 1]$. In altri termini

$$\mathbb{P} : \mathcal{F} \to [0, 1];$$

La funzione $\mathbb{P}$ soddisfa le seguenti proprietà assiomatiche:

- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(\Omega) = 1$
- $A_1, A_2, \ldots \in \mathcal{F}$, mutuamente disgiunti, $\Rightarrow \mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Quando l'insieme dei risultati possibili dell'esperimento è un sottoinsieme di $\mathbb{R}^d$ per qualche intero d, la σ-algebra naturale da utilizzare è quella di Borel, in genere indicata come $\mathcal{B}(\mathbb{R}^d)$); essa è la più piccola σ-algebra che contiene tutti gli insiemi aperti di $\mathbb{R}^d$; questo garantisce che qualunque sottoinsieme "ragionevole" di $\mathbb{R}^d$ appartenga a $\mathcal{B}(\mathbb{R}^d)$.

Variabili e vettori aleatori

Spesso, di un dato esperimento, ci interessano soltanto uno o più specifici aspetti numerici. Ad esempio, quando si estrae un campione casuale di 10 soggetti da una popolazione di studenti, su questi vengono poi rilevate alcune grandezze numeriche importanti per l'indagine in questione (come il peso, l'altezza o il numero di esami già sostenuti) mentre ne vengono trascurate tantissime altre. Allo stesso modo, quando si lancia un dado, il più delle volte ci interessa sapere quale numero da 1 a 6 mostrerà sulla faccia superiore, meno frequentemente saremo interessati al tempo che il dado ha impiegato per arrestarsi! Questo significa che, ad ogni possibile realizzazione $\omega \in \Omega$, è

possibile associare un valore $X(\omega)$ che rappresenta il valore numerico associato a quella particolare realizzazione.

Definizione B.1 *Dato uno spazio Ω, si chiama variabile aleatoria una funzione $X : \Omega \to \mathbb{R}$ tale che, $\forall B \in \mathcal{B}(\mathbb{R})$,*

$$\mathbb{P}(X \in B) = \mathbb{P}(\omega : X^{-1}(\omega) \in B). \tag{B.1}$$

La formula (B.1) stabilisce una condizione di misurabilità della funzione X e afferma che è possibile calcolare la probabilità di un evento B solo quando l'immagine inversa di B appartiene alla σ-algebra $\mathcal{F}$. È per questo motivo che ci limitiamo a calcolare la probabilità di eventi B appartenenti a $\mathcal{B}(\mathbb{R})$. Tra le variabili aleatorie (v.a.), grande importanza rivestono due famiglie:

- v.a. **discrete**;
- v.a. **assolutamente continue**.

Una v.a. si dice *discreta* se l'insieme dei valori assumibili è finito o al più numerabile. In questo caso si definisce la *distribuzione di probabilità* della v.a. X elencando i valori assumibili dalla X e le probabilità con cui questi valori vengono assunti. Avremo così

valori di X	x_1	x_2	x_3	$\cdots$	x_j	$\cdots$	x_k
$\mathbb{P}(X = x_i)$	p_1	p_2	p_3	$\cdots$	p_j	$\cdots$	p_k

dove le $p_j = \mathbb{P}(X = x_j)$ debbono soddisfare alcuni vincoli imposti dagli assiomi della probabilità; in particolare

- $0 \le p_j \le 1, \quad j = 1, \ldots, k$;
- $\sum_j p_j = 1$.

Come già detto, l'insieme dei valori assunti dalla X può avere cardinalità finita o numerabile; in entrambi i casi esso prende il nome di spettro e verrà indicato con il simbolo S.

Esempio B.1 [*Distribuzione binomiale.*] Si lancia tre volte una moneta che dà testa (T) con probabilità p, e croce (C) con probabilità $q = 1 - p$. I tre lanci possono essere considerati indipendenti e siamo interessati allo studio della v.a. $X = \{$ numero di T nei tre lanci $\}$.

Al generico lancio i-esimo associamo la v.a. Y_i che può assumere i due 0 e 1 abbinati rispettivamente agli eventi C e T. Ne consegue che

$$X = Y_1 + Y_2 + Y_3;$$

inoltre è presto visto (mediante elencazione di tutti i $2^3 = 8$ possibili risultati) che la X può assumere solo i valori interi da 0 a 3 compresi. Inoltre, per l'indipendenza dei lanci e per semplici ragionamenti di carattere combinatorio si ha che, per $j = 0, 1, 2, 3$,

$$\mathbb{P}(X = j) = \binom{3}{j} p^j q^{3-j}.$$

La formula appena scritta è un caso particolare della legge binomiale che stabilisce, in presenza di un generico numero n di prove indipendenti e dicotomiche e tali che la probabilità di successo in ciascuna prova è costante e vale p, le probabilità di osservare k successi ed $n - k$ insuccessi. $\diamondsuit$

Una v.a.reale X si dice assolutamente continua se esiste una funzione non negativa f tale che

$$\int_{\mathbb{R}} f(x)dx = 1$$

e, $\forall B \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}(X \in B) = \int_B f(x)dx.$$

In pratica, una v.a. si dice assolutamente continua quando l'insieme dei valori assumibili risulta più che numerabile, e nessuno dei valori assumibili ha probabilità positiva di essere osservato. Parleremo allora di *densità di probabilità*, che è proprio il nome con cui viene indicata la funzione f.

Esempio B.2 [*Distribuzione uniforme.*] In una versione super tecnologica della ruota della fortuna, supponiamo di azionare una lancetta che può fermarsi in un qualunque punto di una circonferenza di lunghezza $2\pi r$, dove r è il raggio; assumiamo inoltre che nessun punto possa considerarsi più probabile di un altro. In altri termini, tutti i valori da 0 a $2\pi r$ hanno la stessa densità di probabilità, e questo implica che la funzione f debba essere costante, ovvero

$$f(x) = \frac{1}{2\pi r}, \quad 0 \leq x \leq 2\pi r.$$

♢

Esistono poi v.a. che non sono classificabili n come discrete n come assolutamente continue. Esempi di questo tipo sono forniti da esperimenti che, con una certa probabilità p forniscono un valore specifico, mentre con probabilità complementare generano un valore casuale appartenente ad un intervallo specifico. Una situazione concreta di questo tipo è il nostro tempo di attesa aleatorio al semaforo: se arriviamo con il verde il nostro tempo di attesa è zero, mentre se arriviamo con il rosso attenderemo un tempo aleatorio che dipende da quando il segnale di rosso è iniziato.

Le definizioni appena date si estendono in modo immediato al caso dei vettori aleatori.

Definizione B.2 *Un* vettore aleatorio $\mathbf{X} = (X_1, \ldots, X_d)$ *è una funzione* $\mathbf{X} : \Omega \to \mathbb{R}^d$ *dove*

$$\mathbf{X}^{-1}(B) = \{\omega \in \Omega : \mathbf{X}(\omega) \in B\} \ \in \mathcal{F}$$

qualunque sia $B \in \mathcal{B}(\mathbb{R}^d)$. *Anche in questo caso diremo che il vettore* $\mathbf{X}$ *è* misurabile.

La distribuzione del vettore aleatorio $\mathbf{X}$ è anche in questo caso fornita dalla distribuzione di probabilità su $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, cioè

$$\Pr(\mathbf{X} \in B) = \mathbb{P}(\mathbf{X}^{-1}(B)).$$

Se le componenti $X_1, \ldots, X_d$ del vettore $\mathbf{X}$ sono discrete, allora la distribuzione di probabilità del vettore $\mathbf{X}$ è data dalla funzione

$$p(x_1, x_2, \ldots, x_d) = \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d)$$

dove ogni valore x_j varia nello spettro S_j della v.a. X_j. La formula sopra scritta viene in genere definita distribuzione congiunta del vettore $\mathbf{X}$.

Analogamente diremo che la distribuzione di $\mathbf{X}$ è *assolutamente continua* se

$$\Pr(\mathbf{X} \in B) = \int_B f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}$$

per qualche funzione non negativa $f : \mathbb{R}^d \to \mathbb{R}_+$ che prende il nome di densità del vettore $\mathbf{X}$.

Valore atteso.

Sia $\mathbf{X}$ un vettore aleatorio d-dimensionale con funzione di densità f. Sia inoltre $\phi : \mathbb{R}^d \to \mathbb{R}$, una funzione a valori reali di $\mathbf{X}$; allora il valore atteso o media della variabile aleatoria $\phi(\mathbf{X})$ è

$$\mathbf{E}\left(\phi(X)\right) = \int_{\mathbb{R}^d} \phi(\mathbf{x}) f(\mathbf{x}) d\mathbf{x},$$

purch esista finito l'integrale

$$\int_{\mathbb{R}^d} |\phi(x)|\, f_{\mathbf{X}}(x) dx < \infty.$$

Densità marginali e condizionate

Supponiamo che il vettore aleatorio d-dimensionale X abbia densità $f_{\mathbf{X}}(\cdot)$. Siano inoltre $\mathbf{Y} = (X_1, \ldots, X_k)$ e $\mathbf{Z} = (X_{k+1}, \ldots, X_d)$ per qualche $1 \leq k \leq d-1$, due sub-vettori di X cosicch $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$, e poniamo $f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y}, \mathbf{z}) = f_{\mathbf{X}}(\mathbf{x})$. Volendo calcolare la distribuzione del vettore $\mathbf{Y}$, avremo

$$\begin{aligned}
\Pr\left(Y \in B\right) &= \Pr\left((\mathbf{Y}, \mathbf{Z}) \in B \times \mathbb{R}^{d-k}\right) \\
&= \int_{B \times \mathbb{R}^{d-k}} f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y}, \mathbf{z}) d\mathbf{y} d\mathbf{z} \\
&= \int_B \left[\int_{\mathbb{R}^{d-k}} f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y}, \mathbf{z}) d\mathbf{z} \right] d\mathbf{y} \\
&= \int_B f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y},
\end{aligned}$$

dove si è posto

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_{\mathbb{R}^{d-k}} f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y}, \mathbf{z}) d\mathbf{z}.$$

Abbiamo così ottenuto la distribuzione marginale di $\mathbf{Y}$ che, essendo espressa mediante un integrale risulta anch'essa assolutamente continua con densità $f_{\mathbf{Y}}$. La densità del vettore $\mathbf{Z}$ condizionata al fatto che $\mathbf{Y} = \mathbf{y}$, si esprime attraverso la seguente formula avremo

$$f_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y}) = \frac{f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y}, \mathbf{z})}{f_{\mathbf{Y}}(\mathbf{y})},$$

definibile per ogni valore $\mathbf{y}$ tale che $f_{\mathbf{Y}}(\mathbf{y}) \neq 0$. La $f_{\mathbf{Z}|\mathbf{Y}}$ è una densità in $\mathbf{z}$, appunto la *densità condizionata di* $\mathbf{Z}$ *dato* $Y = y$.

Indipendenza

Diremo che le v.a. X e Y sono *indipendenti*, e lo indicheremo con il simbolo

$$X \perp\!\!\!\perp Y$$

se, per ogni coppia di insiemi A, $B \in \mathcal{B}(\mathbb{R})$ risulta:

$$\Pr\left(\mathbf{X} \in A, Y \in B\right) = \Pr\left((X, Y) \in A \times B\right) = \Pr\left(X \in A\right) \Pr\left(Y \in B\right).$$

Nel caso in cui il vettore aleatorio (X, Y) è assolutamente continuo con densità $f_{X,Y}(x, y)$, e di conseguenza, le densità marginali sono fornite dalle

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy$$

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx,$$

allora avremo che le seguenti affermazioni sono equivalenti:

$$\begin{aligned} X \perp\!\!\!\perp Y &\iff f_{X,Y}(x,y) = f_X(x) f_Y(y) \\ &\iff f_{X|Y}(x|y) = f_X(x) \\ &\iff f_{Y|X}(y|x) = f_Y(y). \end{aligned}$$

Trasformazioni di variabili aleatorie

Consideriamo il vettore aleatorio $\mathbf{X}$ con densità $f_{\mathbf{X}}$ e supponiamo che esista un insieme aperto $S \subseteq \mathbb{R}^d$ tale che $\Pr(\mathbf{X} \in S) = 1$. Venga definita poi su S la funzione $g : \mathbb{R}^d \to \mathbb{R}^d$ invertibile e continuamente differenziabile, con determinante dello Jacobiano strettamente diverso da zero per ogni punto di S. Allora il vettore aleatorio $\mathbf{Y} = g(\mathbf{X})$ ha a sua volta una densità assolutamente continua con densità $f_{\mathbf{Y}}$ data da

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y})) \left| \frac{d\mathbf{x}}{d\mathbf{y}} \right| \mathbf{1}_{g(S)}(y),$$

dove $|d\mathbf{x}/d\mathbf{y}|$ rappresenta il modulo del determinante dell'inverso dello Jacobiano della trasformazione:

$$\det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_d} \\ \vdots & \vdots & \\ \frac{\partial x_d}{\partial y_1} & \cdots & \frac{\partial x_d}{\partial y_d} \end{pmatrix},$$

e $g(S)$ è l'immagine $\{y = g(x) : x \in S\}$ di S mediante g.

Esempio B.3 [*Somma e rapporto di v.a. di tipo Gamma.*] Siano X e Y v.a. indipendenti; sia $X \sim \mathrm{Ga}(\alpha_1, 1/\beta)$ e $Y \sim \mathrm{Ga}(\alpha_2, 1/\beta)$. Vogliamo determinare la densità congiunta delle v.a. (U, V), dove

$$U = X + Y, \qquad V = X/Y.$$

L'applicazione $(x, y) \to (u, v)$ è definita su tutto il quadrante positivo tranne che sull'asse delle x. Tuttavia $\Pr(Y = 0) = 0$ e possiamo applicare il risultato precedente. La funzione inversa è data da

$$x = \frac{uv}{1+v}, \quad y = \frac{u}{1+v}.$$

Inoltre anche le variabili U e V sono strettamente positive perch lo sono x e y. Il modulo del determinante dello Jacobiano vale

$$\begin{vmatrix} v/(1+v) & u/(1+v)^2 \\ 1/(1+v) & -u/(1+v)^2 \end{vmatrix} = \frac{u}{(1+v)^2}$$

Ne segue che, per $u, v, > 0$

$$
\begin{aligned}
f_{U,V} &= f_X\left(\frac{uv}{1+v}\right) f_Y\left(\frac{u}{1+v}\right)\frac{u}{(1+v)^2} \\
&= \frac{1}{\Gamma(\alpha_1)\,\Gamma(\alpha_2)\,\beta^{\alpha_1+\alpha_2}} \exp\left(-\frac{uv}{\beta(1+v)} + \frac{u}{(\beta(1+v)}\right) \times (\frac{uv}{1+v})^{\alpha_1-1}(\frac{u}{1+v})^{\alpha_2-1}\frac{u}{(1+v)^2} \\
&= \frac{1}{\Gamma(\alpha_1)\,\Gamma(\alpha_2)\,\beta^{\alpha_1+\alpha_2}} \exp\left(-\frac{u}{\beta}\right) u^{\alpha_1+\alpha_2-1} \times (\frac{v}{1+v})^{\alpha_1-1}(\frac{1}{1+v})^{\alpha_2+1} \\
&= \frac{1}{\Gamma(\alpha_1+\alpha_2)\,\beta^{\alpha_1+\alpha_2}} \exp\left(-\frac{u}{\beta}\right) u^{\alpha_1+\alpha_2-1} \times \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\,\Gamma(\alpha_2)}(\frac{v}{1+v})^{\alpha_1-1}(\frac{1}{1+v})^{\alpha_2+1} \\
&= f_U(u) \times f_V(v)
\end{aligned}
$$

Abbiamo così dimostrato che U e V sono indipendenti. Inoltre

- $U \sim \text{Ga}(\alpha_1+\alpha_2, 1/\beta)$:
- la legge di V è simile ad una F di Fisher. In particolare, utilizzando ancora la regola di trasformazione di v.a., si può dimostrare che

$$\frac{V}{1+V} \sim \text{Beta}(\alpha_1, \alpha_2);$$

da notare che la distribuzione di V non dipende da β. ♢

B.1 Funzione generatrice dei momenti

B.2 Convergenza di variabili aleatorie

C

Alcuni risultati e dimostrazioni

C.1 Statistiche d'ordine

Siano $(X_1, \cdots, X_n)$ indipendenti e somiglianti con funzione di ripartizione F_X. Definiamo con $Y_{(j)}$ la j-esima più piccola realizzazione di $(X_1, \cdots, X_n)$ cosicch, ad esempio,

$$Y_{(1)} = \min_{i=1,\cdots,n} (X_1, \cdots, X_n), \quad Y_{(n)} = \max_{i=1,\cdots,n} (X_1, \cdots, X_n).$$

Vogliamo calcolare la funzione di ripartizione $F^{(j)}$ della generica statistica d'ordine $Y_{(j)}$, $j = 1, \cdots, n$. L'evento $\{Y_{(j)} \leq t\}$ si verifica se e solo se almeno j delle X_i sono minori o uguali di t; tenendo conto che ciascuna delle X_i risulta minore o uguale a t con probabilità $F_X(t)$, risulta

$$F^{(j)}(t) = \Pr\left(Y_{(j)} \leq t\right) = \sum_{h=j}^{n} \binom{n}{h} F_X(t)^h \left[1 - F_X(t)\right]^{n-h}. \tag{C.1}$$

Quando le X_i sono assolutamente continue con densità f_X, lo saranno anche le $Y_{(j)}$ e la densità della generica variabile aleatoria $Y_{(j)}$ si ottiene per derivazione della (C.1). Con calcoli semplici [[30], pag. 114] si verifica che

$$f^{(j)}(t) = \binom{n}{j} j\, F_X(t)^{j-1} \left[1 - F_X(t)\right]^{n-j} f_X(t). \tag{C.2}$$

Con argomenti simili si possono calcolare anche le distribuzioni congiunte di più statistiche d'ordine. Ad esempio, la funzione di ripartizione della variabile aleatoria doppia $(Y_{(h)}, Y_{(j)})$, si ottiene dal seguente teorema (per la dimostrazione si veda [82]):

Teorema C.1 *Siano $(X_1, \cdots, X_n)$ indipendenti e somiglianti con funzione di ripartizione F_X; siano inoltre $(Y_{(1)}, \cdots, Y_{(n)})$ le corrispondenti statistiche d'ordine. Allora, per $h < j$, si ha*

$$\begin{aligned} &F^{(h,j)}(s,t) = \Pr\left(Y_{(h)} \leq s, Y_{(j)} \leq t\right) \\ &= \begin{cases} \sum_{i=h}^{n} \sum_{k=\max(0,j-i)}^{n-i} \binom{n}{i,k,n-i-k} F_X(s)^i [F_X(t) - F_X(s)]^k [1 - F_X(t)]^{n-i-k} & s < t \\ \Pr\left(Y_{(j)} \leq t\right) & s \geq t \end{cases} \end{aligned}$$

Anche in questo caso, quando le v.a. $(X_1, \cdots, X_n)$ hanno densità $f_X(\cdot)$, si vede facilmente che anche le v.a. doppie $(Y_{(h)}, Y_{(j)})$ sono assolutamente continue e la densità è data, per $h < j$, e per $s < t$,

$$\begin{aligned} f^{(h,j)}(s,t) &= \frac{n!}{(h-1)!(j-h-1)!(n-j)!} \\ &\times F_X(s)^{h-1} [F_X(t) - F_X(s)]^{j-h-1} [1 - F_X(t)]^{n-j} f_X(s) f_X(t). \end{aligned}$$

C.2 Alcuni approfondimenti matematici

C.2.1 Derivazione della distribuzione di Jeffreys

Dato il consueto modello statistico $\mathcal{E} = \{\mathcal{X}, \mathcal{P}, \boldsymbol{\Theta}\}$, con $\mathcal{P}\{p(\cdot|\theta), \quad \theta \in \boldsymbol{\Theta}\}$, consideriamo la distanza "simmetrica" di Kullback-Leibler (vedi §2.6) tra due elementi di $\mathcal{E}$, indicizzati da θ e θ':

$$J_{KL}(\theta', \theta) = \int_{\mathcal{X}} \log \frac{p(x|\theta')}{p(x|\theta)} \left[p(x|\theta') - p(x|\theta)\right] dx.$$

Si dimostra che, per $\theta' \to \theta$, la distanza simmetrica di Kullback e Leibler può essere approssimata dalla metrica di Fisher, cosicch

$$\lim_{\theta' \to \theta} \frac{J_{KL}(\theta', \theta)}{(\theta' - \theta)^2} = I(\theta).$$

La dimostrazione dell'assunto è basata sul fatto che, scrivendo $\theta' = \theta + \delta$, e ponendo

$$\frac{J_{KL}(\theta', \theta)}{(\theta' - \theta)^2} = \frac{r(\theta, \delta)}{\delta^2},$$

si ha

$$\frac{r(\theta, \delta)}{\delta^2} = \int_{\mathcal{X}} \frac{\log p(x \mid \theta + \delta) - \log p(x \mid \theta)}{\delta} \; \frac{p(x \mid \theta + \delta) - p(x \mid \theta)}{\delta} dx.$$

Quindi, assumendo scambiabili l'operatore di limite e quello di integrazione, e assumendo la derivabilità di $p(x \mid \theta)$ e la continuità dell'informazione attesa di Fisher $I(\theta)$,

$$\lim_{\delta \to 0} \frac{r(\theta, \delta)}{\delta^2} = \int_{\mathcal{X}} \frac{\partial \log p(x \mid \theta)}{\partial \theta} \frac{\partial p(x \mid \theta)}{\partial \theta} dx$$

$$= \int_{\mathcal{X}} \frac{\partial \log p(x \mid \theta)}{\partial \theta} \frac{p'(x \mid \theta)}{p(x \mid \theta)} p(x \mid \theta)) dx = \int_{\mathcal{X}} \left(\frac{\partial \log p(x|\theta}{\partial \theta} \right)^2 dx = I(\theta).$$

Inoltre, se λ rappresenta una qualunque riparametrizzazione biunivoca di θ, per la (5.6) risulta

$$I(\theta) = I(\lambda(\theta)) \left(\frac{d\lambda}{d\theta} \right)^2,$$

ovvero scritto in una forma alternativa,

$$\sqrt{I(\theta)} d\theta = \sqrt{I(\lambda)} d\lambda.$$

La distribuzione di Jeffreys

$$\pi^J(\theta) \propto \sqrt{I(\theta)},$$

dunque, "pesa" sottoinsiemi di $\mathcal{E}$ in modo proporzionale al valore della matrice I, indipendentemente dalla parametrizzazione adottata. In altri termini, la metrica naturale per il modello $\mathcal{E}$ non è quella euclidea bensì quella indotta dalla matrice di informazione I, e la distribuzione iniziale di Jeffreys può essere interpretata come la distribuzione uniforme su $\mathcal{E}$, secondo tale metrica.

C.3 Sulla scambiabilità

C.3.1 Dimostrazione del Teorema 6.1

La dimostrazione è tratta da [15]. Per n fissato S_n può assumere i valori $0 \leq s \leq n$. Per l'ipotesi di scambiabilità si ha

$$\Pr(S_n = s) = \binom{n}{s} \Pr(x_{i_1}, x_{i_2}, \cdots, x_{i_n}),$$

dove $x_{i_1}, x_{i_2}, \cdots, x_{i_n}$ rappresenta una generica permutazione delle n realizzazioni delle variabili tale che $x_{i_1} + x_{i_2} + \cdots + x_{i_n} = s$. Inoltre, per un qualunque $N \geq n$, si ha che l'evento $\{S_n = s\}$ è compatibile con un numero di successi su N prove che va da un minimo di s (già realizzati nelle prime n prove) ad un massimo di $N - (n - s)$, in quanto si sono già verificati $n - s$ insuccessi: ne segue che

$$\Pr(S_n = s) = \sum_{r=s}^{N-(n-s)} \Pr\left([S_n = s] \bigcup [S_N = r]\right). \tag{C.3}$$

La (C.3) viene poi trasformata nel modo seguente:

$$= \sum_{r=s}^{N-(n-s)} \Pr(S_n = s \mid S_N = r) \Pr(S_N = r)$$

$$= \sum_{r=s}^{N-(n-s)} \frac{\binom{n}{s}\binom{N-n}{r-s}}{\binom{N}{r}} \Pr(S_N = r);$$

l'ultimo passaggio è dovuto al fatto che la probabilità di avere s successi su n prove quando se ne sono osservati r su N prove, equivale alla probabilità ipergeometrica di estrarre, da un'urna che contiene n palline bianche e $N - n$ palline nere, esattamente s palline bianche ed $r - s$ palline nere quando si effettuino r estrazioni senza ripetizione. Semplificando alcuni fattoriali, la (C.3) vale

$$= \binom{n}{s} \sum_{r=s}^{N-(n-s)} \frac{(r)_s (N-r)_{(n-s)}}{(N)_n} \Pr(S_N = r),$$

dove il simbolo $(A)_t$ rappresenta il fattoriale troncato al termine t-esimo

$$(A)_t = A(A-1)(A-2)\cdots(A-t+1).$$

Per un N grande arbitrario, tale che $N \geq n \geq s \geq 0$, definiamo la funzione di ripartizione $Q_N(\theta)$ nel modo seguente

$$Q_N(\theta) = \begin{cases} 0 & \theta < 0 \\ \Pr(S_N = 0) & 0 \leq \theta < \frac{1}{N} \\ \Pr(S_N \leq 1) & \frac{1}{N} \leq \theta < \frac{2}{N} \\ \cdots & \cdots \\ \Pr(S_N \leq k) & \frac{k-1}{N} \leq \theta < \frac{k}{N} \\ \cdots & \cdots \\ 1 & \theta \geq 1 \end{cases}$$

In pratica $Q_N(\theta)$ "salta" sugli $N + 1$ punti k/N, per $k = 0, 1, \cdots, N$ e, per ogni k, il salto vale esattamente $\Pr(S_N = k)$. Ponendo $N\theta = r$, si ha

$$\Pr(S_n = s) = \binom{n}{s} \int_0^1 \frac{(N\theta)_s [N(1-\theta)]_{n-s}}{(N)_n} dQ_N(\theta),$$

dove l'ultima espressione è un integrale rispetto alla funzione di ripartizione a salti $Q_N(\theta)$, e come tale, una somma. Per $N \to \infty$, è facile vedere che

$$\frac{(N\theta)_s[N(1-\theta)]_{n-s}}{(N)_n} \to \theta^s(1-\theta)^{n-s},$$

uniformemente in θ. Inoltre, per il teorema C.2, deve esistere una sottosuccessione $Q_{N_1}, Q_{N_2}, \cdots$ della successione delle Q_N per cui vale

$$\lim_{N_j \to \infty} Q_{N_j}(\theta) = Q(\theta),$$

dove $Q(\theta)$ è una funzione di ripartizione, mentre

$$\theta = \lim_{N \to \infty} \frac{S_N}{N}.$$

Teorema C.2 (Helly) *Per ogni successione $F_1, F_2, \cdots$, di funzioni di ripartizione, tali che*

$$\forall \epsilon > 0, \exists a \text{ tale che } \forall i > i_0, F_i(a) - F_i(-a) > 1 - \epsilon,$$

esiste una funzione di ripartizione F e una sottosuccessione $F_{i1}, F_{i2}, \cdots$ tali che

$$\lim_{i_k \to \infty} F_{i_k}(x) = F(x),$$

per ogni x punto di continuità per F.

Dimostrazione C.1 *Si veda, ad esempio, [1].*

C.4 Sulle forme quadratiche

C.4.1 Combinazione di due forme quadratiche

Lemma C.1. *Siano $\mathbf{x}, \mathbf{a}, \mathbf{b}$ vettori di $\mathbb{R}_k$ e $\mathbf{A}, \mathbf{B}$ matrici simmetriche $k \times k$ tali che $(\mathbf{A}+\mathbf{B})^{-1}$ esiste. Allora,*

$$(\mathbf{x}-\mathbf{a})'\mathbf{A}(\mathbf{x}-\mathbf{a}) + (\mathbf{x}-\mathbf{b})'\mathbf{B}(\mathbf{x}-\mathbf{b})$$
$$= (\mathbf{x}-\mathbf{c})'(\mathbf{A}+\mathbf{B})(\mathbf{x}-\mathbf{c}) + (\mathbf{a}-\mathbf{b})'\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}(\mathbf{a}-\mathbf{b})$$

dove

$$\mathbf{c} = (\mathbf{A}+\mathbf{B})^{-1}(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b})$$

Nel caso in cui $x \in \mathbb{R}$ la formula si semplifica nella (4.7)

Dimostrazione C.2 *[20].*

$$(\mathbf{x}-\mathbf{a})'\mathbf{A}(\mathbf{x}-\mathbf{a}) + (\mathbf{x}-\mathbf{b})'\mathbf{B}(\mathbf{x}-\mathbf{b})$$
$$= \mathbf{x}'(\mathbf{A}+\mathbf{B})\mathbf{x} - 2\mathbf{x}(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b}) + \mathbf{a}'\mathbf{A}\mathbf{a} + \mathbf{b}'\mathbf{B}\mathbf{b}$$

Si aggiunge e si sottrae la quantità $\mathbf{c}'(\mathbf{A}+\mathbf{B})\mathbf{c}$ e quindi la quantità sopra scritta è pari a

$$(\mathbf{x}-\mathbf{c})'(\mathbf{A}+\mathbf{B})(\mathbf{x}-\mathbf{c}) + \mathbf{G},$$

dove $\mathbf{G} = \mathbf{a}'\mathbf{A}\mathbf{a} + \mathbf{b}'\mathbf{B}\mathbf{b} - \mathbf{c}'(\mathbf{A}+\mathbf{B})\mathbf{c}$. Inoltre

$$\mathbf{c}'(\mathbf{A}+\mathbf{B})\mathbf{c} = (\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b})'(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b}) =$$

(si aggiunge e si sottrae $\mathbf{A}\mathbf{b}$ nel primo fattore e $\mathbf{B}\mathbf{a}$ nel terzo)

$$[\mathbf{A}(\mathbf{a}-\mathbf{b}) + (\mathbf{A}+\mathbf{B})\mathbf{b}]'(\mathbf{A}+\mathbf{B})^{-1}[(\mathbf{A}+\mathbf{B})\mathbf{a} - \mathbf{B}(\mathbf{a}-\mathbf{b})]$$
$$= -(\mathbf{a}-\mathbf{b})'\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}(\mathbf{a}-\mathbf{b}) + \mathbf{a}'\mathbf{A}\mathbf{a} + \mathbf{b}'\mathbf{B}\mathbf{b};$$

Quindi

$$\mathbf{G} = (\mathbf{a}-\mathbf{b})'\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}(\mathbf{a}-\mathbf{b}).$$

C.5 Sul calcolo delle distribuzioni non informative nel modello lineare

Per la definizione della distribuzione di Jeffreys, è necessario calcolare la matrice d'informazione attesa associata all'esperimento. Dalla (9.4) si ha che la funzione di log-verosimiglianza vale

$$\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) \propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(nS^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right). \tag{C.4}$$

Occorre ora calcolare le derivate prime e seconde della (C.4). La derivate prime, rispetto al vettore $\boldsymbol{\beta}$ e rispetto a σ^2, valgono

$$\frac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}'} =,$$

$$\frac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4}\left(nS^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right).$$

Infine,

$$-\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}'} = \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}:$$

$$-\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}' \partial \sigma^2} =;$$

$$-\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2 \partial \sigma^2} =$$

Calcolando il valore atteso delle precedenti derivate si ottiene

$$I_{\boldsymbol{\beta},\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}; \quad I_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{\beta}, \sigma^2) = \mathbf{0}; \quad I_{\sigma^2,\sigma^2} = \frac{n}{2\sigma^4},$$

ovvero, la matrice d'informazione vale

$$I(\boldsymbol{\beta}, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}; & \mathbf{0} \\ \mathbf{0}' & \frac{n}{2\sigma^4} \end{pmatrix}. \tag{C.5}$$

La legge a priori di Jeffreys è allora pari a

$$\pi^J(\boldsymbol{\beta}, \sigma^2) \propto \sqrt{\det\left(I(\boldsymbol{\beta}, \sigma^2)\right)} = \frac{1}{(\sigma^2)^{p/2+1}}$$

C.6 Sul calcolo della marginale per un modello lineare

Consideriamo la quantità $(\mathbf{y} - X\boldsymbol{\beta}_0)$. Aggiungendo e sottraendo la quantità $\mathbf{X}\boldsymbol{\beta}_\star$, e successivamente usando l'espressione di $\boldsymbol{\beta}_\star$, si ottiene

$$\begin{aligned}
\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0 &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\star) + \mathbf{X}(\boldsymbol{\beta}_\star - \boldsymbol{\beta}_0) \\
&= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\star) + \mathbf{X}\left[(\mathbf{V}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{V}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}'\mathbf{y}) - \boldsymbol{\beta}_0\right] \\
&= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\star) + \mathbf{X}(\mathbf{V}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + \mathbf{X}(\mathbf{V}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{V}_0^{-1}\boldsymbol{\beta}_0 - \mathbf{X}\boldsymbol{\beta}_0 \\
&= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\star) + \mathbf{X}(\mathbf{V}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{X}(I_p - (\mathbf{V}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{V}_0^{-1})\boldsymbol{\beta}_0
\end{aligned}$$

Aggiungendo e sottraendo la quantità $(\mathbf{V}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$ nella parentesi quadra, si ottiene

$$= (\mathbf{y} - \mathbf{X}\beta_\star) + \mathbf{X}(\mathbf{V}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{X}(\mathbf{V}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_0$$

$$= (\mathbf{y} - \mathbf{X}\beta_\star) + \mathbf{X}(\mathbf{V}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(y - \mathbf{X}\boldsymbol{\beta}_0).$$

Perciò,

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0 = \left(I_n - \mathbf{X}(\mathbf{V}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\star),$$

che per un noto lemma sull' inversione di matrici [2, pag. 299], è equivalente a

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0 = (I_n + \mathbf{X}\mathbf{V}_0\mathbf{X}')(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\star)$$

D

Catene di Markov

Ci limiteremo in questa appendice ad elencare i risultati principali sulle catene di Markov in tempo discreto, privilegiando gli aspetti più importanti nell'ottica delle utilizzazioni di tipo computazionale che di questi processi si fanno nella statistica bayesiana. Il lettore interessato ad approfondire lo studio delle catene di Markov può fare riferimento, tra i testi in italiano, a [30] o [4]. Tra i testi in inglese si suggeriscono i classici [46], [62], [61]; per le applicazioni in ambito bayesiano si veda [77].

D.1 Catene in tempo discreto

Sia $\mathbb{S}$ un insieme finito o numerabile, detto *spazio degli stati*. Gli elementi di $\mathbb{S}$ verranno indicati genericamente con le lettere latine $\{\dots, i, j, k, \dots\}$. sia inoltre $\boldsymbol{\pi} = (\pi_i; i \in \S)$ una generica legge di probabilità sull'insieme $\mathbb{S}$. Sia infine $\mathbf{P} = (p_{ij}; i, j \in \S)$ una matrice quadrata con numero di righe e colonne pari alla cardinalità di $\mathbb{S}$. Tale matrice si dice *stocastica* se

- $0 \leq p_{ij} \leq 1, \qquad i, j \in \mathbb{S}$
- $\forall i \in \mathbb{S}, \quad \sum_{j \in \mathbb{S}} p_{ij} = 1$ (ogni riga di $\mathbf{P}$ ha somma 1).

Consideriamo ora un processo stocastico a parametro discreto $\mathbf{X} = (X_n; n \geq 0)$. Ad ogni istante $n \geq 0$ la v.a. X_n rappresenta *lo stato occupato dal processo al tempo* n ed ha come spettro l'insieme $\mathbb{S}$. Diremo che $(X_n; n \geq 0)$ è una catena di Markov omogenea nel tempo con distribuzione iniziale $\boldsymbol{\pi}^{(0)}$ e matrice di transizione $\mathbf{P}$ se

- la v.a. X_0 ha distribuzione $\boldsymbol{\pi}^{(0)}$.
- per ogni $n \geq 1$, la distribuzione della v.a. X_{n+1} condizionata all'informazione $X_n = i$, è fornita dalla riga i-esima della matrice $\mathbf{P}$.

Indicheremo il suddetto processo con il simbolo $CM(\boldsymbol{\pi}^{(0)}, \mathbf{P})$. Un modo equivalente per definire $CM(\boldsymbol{\pi}^{(0)}, \mathbf{P})$ è il seguente: $\forall n \geq 0$ e per ogni scelta degli stati $i_0, i_1, \dots, i_{n+1} \in \mathbb{S}$, si ha $\mathbb{P}(X_0 = i_0) = \pi^{(0)}_{i_0}$ e

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) = p_{i_n, i_{n+1}}.$$

Detto a parole, una catena di Markov è un processo per il quale la probabilità di trovarsi in un certo stato al tempo n non dipende da tutta la storia passata bensì solo dalla posizione del processo nell'ultimo istante osservato. Si suole anche dire che, in una catena di Markov, il futuro è

indipendente dal passato, condizionatamente al presente.

Proprietà di Markov

Sia $(X_n; n \geq 0)$ un processo $CM(\boldsymbol{\pi}^{(0)}, \mathbf{P})$. Allora, condizionatamente all'informazione $\{X_m = i\}$, il processo $(Y_n; n \geq 0) = (X_{n+m}; n \geq 0)$ è ancora una catena di Markov con la stessa matrice di transizione $\mathbf{P}$ e distribuzione iniziale concentrata sullo stato i, ovvero $\mathbb{P}(Y_0 = i = 1)$. La proprietà di Markov afferma in pratica che, nota la posizione al tempo n, la catena si rigenera in modo indipendente dal passato secondo lo stesso meccanismo aleatorio di transizione, governato dalla matrice $\mathbf{P}$.

D.1.1 Distribuzione del processo ad un tempo prefissato

Calcolare la distribuzione della v.a. X_1 è semplice. Per $j \in \mathbb{S}$,

$$\mathbb{P}(X_1 = j) = \sum_{i \in \mathbb{S}} \mathbb{P}(X_1 = j \mid X_0 = i)\mathbb{P}(X_0 = i) = \sum_{i \in \mathbb{S}} p_{ij}\pi_i^{(0)}.$$

In forma matriciale si può allora scrivere

$$\boldsymbol{\pi}^{(1)} = \boldsymbol{\pi}^{(0)}\mathbf{P}. \tag{D.1}$$

Allo stesso modo, per X_2 avremo

$$\mathbb{P}(X_2 = j) = \sum_{i \in \mathbb{S}} \mathbb{P}(X_2 = j \mid X_1 = i)\mathbb{P}(X_1 = i) = \sum_{i \in \mathbb{S}} p_{ij}\pi_i^{(1)}.$$

In forma matriciale, $\boldsymbol{\pi}^{(2)} = \boldsymbol{\pi}^{(1)}\mathbf{P}$, oppure per la (D.1),

$$\boldsymbol{\pi}^{(2)} = \boldsymbol{\pi}^{(0)}\mathbf{P}\mathbf{P} = \boldsymbol{\pi}^{(0)}\mathbf{P}^2.$$

Più in generale avremo allora che la distribuzione del processo al tempo n si ottiene moltiplicando in senso matriciale la legge al tempo precedente per la matrice di transizione $\mathbf{P}$, oppure la distribuzione iniziale per la potenza ennesima di $\mathbf{P}$, ovvero

$$\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(n-1)}\mathbf{P} = \boldsymbol{\pi}^{(0)}\mathbf{P}^n. \tag{D.2}$$

Da quanto sopra si deduce inoltre che, per ogni n, la matrice di transizione in n passi $\mathbf{P}^{(n)}$ con elemento generico

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j \mid X_0 = i)$$

coincide esattamente con $\mathbf{P}^n$. Generalizzando quanto appena dimostrato, è possibile scrivere anche la distribuzione al tempo n in funzione di quanto accaduto ad un tempo $m < n$.

$$\mathbb{P}(X_n = j) = \pi_j^{(n)} = \sum_{i \in \mathbb{S}} \mathbb{P}(X_n = j \mid X_m = i)\mathbb{P}(X_m = i) = \sum_{i \in \mathbb{S}} \pi_i^{(m)} p_{ij}^{(n-m)};$$

la relazione precedente prende il nome di equazione di Chapman e Kolmogorov.

Esempio D.1 [*Catena a due stati*]

?????????????????????????????????? ♢

D.1.2 Probabilità di assorbimento

In quello che segue sarà utile abbreviare la quantità $\mathbb{P}(X_n = j \mid X_0 = i)$ attraverso la notazione $\mathbb{P}_i(X_n = j)$.

Sia $A \in \mathbb{S}$ un sottoinsieme di stati. Si definisce *Tempo di prima visita in A* la v.a. $V_A : \Omega \to \mathbb{N} \cup \{\infty\}$ definita da

$$V_A(\omega) = \inf \{n \geq 0; X_n(\omega) \in A\}.$$

La probabilità, partendo dallo stato i, che la catena raggiunga prima o poi l'insieme A è indicata da

$$v_i^A = \mathbb{P}_i(V_A < \infty).$$

Se A è una classe chiusa, la quantità v_i^A prende il nome di probabilità di assorbimento. Quando $v_i^A = 1$, è possibile definire anche il tempo medio di assorbimento, indicato col simbolo

$$E_i^A = \mathbf{E}_\mathrm{i}[V_A] = \sum_{n \geq 1} n \Pr(V_A = n).$$

La quantità E_i^A non necessariamente esiste finita; le quantità v_i^A e E_i^A possono essere spesso calcolate in modo semplice attraverso un sistema di equazioni lineari associate alla matrice $\mathbf{P}$.

Esempio D.2 Consideriamo la catena di Markov con matrice di transizione definita dal diagramma in Figura D.1. Supponendo che $X_0 = B$, si vuole calcolare la probabilità di assorbimento in D ed il tempo medio di assorbimento in A oppure D.

Figura D.1. Diagramma sostitutivo della matrice di transizione

Si tratta allora di calcolare v_B^D e $E_B^{(A;D)}$. Per il calcolo di v_B^D si condiziona a quanto avviene al tempo 1, ovvero

$$v_B^D = \mathbb{P}_B(X_1 = A)v_A^D + \mathbb{P}_B(X_1 = C)v_C^D,$$

e, essendo impossibile finire in D partendo da A,

$$v_B^D = (1-p)v_C^D$$

Allo stesso modo avremo

$$v_C^D = \mathbb{P}_C(X_1 = B)v_B^D + \mathbb{P}_C(X_1 = D)v_D^D = qv_B^D + (1-q).$$

Combinando le due relazioni si ottiene

$$v_B^D = 1 - \frac{p}{1-q+qp}.$$

Il calcolo di $E_B^{(A,D)}$ si effettua usando la stessa tecnica. L'unica differenza è che, condizionando a quanto avviene al tempo 1, occorre considera che, appunto, un'unità di tempo è già trascorsa. Avremo allora

$$E_B^{(A,D)} = 1 + \mathbb{P}_B(X_1 = A)E_A^{(A,D)} + \mathbb{P}_B(X_1 = C)E_C^{(A,D)} = 1 + (1-p)E_C^{(A,D)}.$$

Inoltre, $E_C^{(A,D)} = 1 + qE_B^{(A,D)}$, da cui

$$E_B^{(A,D)} = \frac{2-p}{1-q+qp}.$$

♢

Altri esempi notevoli di applicazione delle tecniche appena descritte sono il classico problema della rovina del giocatore e l'analisi dei processi di nascite e morti; si veda ad esempio [62].

D.1.3 Tempi di arresto e proprietà forte di Markov

La proprietà di Markov asserisce tra l'altro che, condizionatamente all'informazione $X_m = i$, il processo $(X_{n+m}; n \geq 0)$ si comporta in modo indipendente da quanto avvenuto prima del tempo m. La proprietà di Markov forte fa riferimento allo stesso contesto, con la differenza tuttavia che il condizionamento opera in un tempo aleatorio M. In altri termini, se M è il tempo aleatorio tale che $X_M = i$, è possibile ancora asserire che il processo $(X_{n+M}; n \geq 0)$ è indipendente da quanto avvenuto fino al tempo M? Si può dimostrare che una condizione sufficiente affinch questo sia vero è che la v.a. M sia un *tempo di arresto*, ovvero una v.a. $M : \Omega \to \mathbb{N} \cup \{\infty\}$ tale che, $\forall n \in \mathbb{N}$, l'evento $\{M = n\}$ dipende soltanto dalle v.a. $X_0, X_1, \ldots, X_n$.

D.1.4 Classificazioni degli stati

Può accadere che lo spazio degli stati $\mathbb{S}$ di una catena possa essere logicamente suddiviso in sottoclassi di stati tra loro più "simili". Questo si effettua mediante il concetto di *comunicabilità* tra stati.

Si dice che uno stato i comunica con lo stato j, e si scrive $i \to j$, se $p_{ij}^{(n)} > 0$ per almeno un n. Due stati i e j *intercomunicano*, e si scrive $j \leftrightarrow i$, se $i \to j$ e $i \to j$.

Una classe di stati $C \subset \mathbb{S}$ si dice chiusa se

$$i \in C, \quad i \to j \Rightarrow j \in C;$$

In altre parole, se la catena si trova nella classe C, non ne può più uscire. Se la classe C è formata da un solo stato questo si dice *assorbente*: lo stato $j \in \mathbb{S}$ è dunque assorbente se

$$p_{jj} = 1, \quad \text{e} \quad p_{jr} = 0, \forall r \neq j.$$

In ogni catena di Markov è possibile suddividere l'insieme degli stati S in sottoinsiemi disgiunti di stati $\mathbb{S}_1, \mathbb{S}_2, \cdots$ tali che

- All'interno di ogni $\mathbb{S}_j$, $j = 1, 2, \cdots$, ogni coppia di stati è intercomunicante.
- Stati appartenenti a diversi sottoinsiemi non possono essere intercomunicanti.

Quando non è possibile partizionare in tal maniera l'insieme $\mathbb{S}$ si dice che la catena è *irriducibile*.

Un ulteriore tipo di classificazione degli stati è basata sulla probabilità che la catena ha di tornare in uno stato già visitato.

Definizione D.1 *Sia i uno stato di $\mathbb{S}$. Si chiama probabilità di ritorno la quantità*

$$f_i = \mathbb{P}_i(X_n = i, \text{ per qualche } n \geq 1). \tag{D.3}$$

Si dice allora che

- Uno stato $i \in \mathbb{S}$ è *transitorio* se $f_i < 1$.
- Uno stato $i \in \mathbb{S}$ è *ricorrente* se $f_i = 1$.

È inoltre possibile classificare gli stati sulla base del tempo di primo ritorno.

Definizione D.2 *Sia i un generico stato di $\mathbb{S}$, e definiamo la v.a.*

$$T_i = \inf \{n \geq 1 : X_n = i | X_0 = i\}.$$

T_i rappresenta, quindi, il tempo (aleatorio) di primo ritorno nello stato i: sia inoltre $\mu_i = \mathbb{E}(T_i)$. Diremo allora che

- *Lo stato i è ricorrente positivo se $\mu_i < \infty$*
- *Lo stato i è ricorrente nullo se $\mu_i = \infty$*

La ricorrenza (positiva o nulla) e la transitorietà di uno stato sono proprietà di classe, ovvero due stati intercomunicanti sono entrambi transitori oppure entrambi ricorrenti. Inoltre, ogni classe ricorrente è per forza chiusa, mentre ogni classe finita e chiusa è per forza ricorrente. Se infine la classe C è finita e la catena è irriducibile, allora la catena deve essere ricorrente positiva.

Un'altra classificazione degli stati è basata sul periodo.

Definizione D.3 *Il periodo d_i di uno stato $i \in S$è il massimo comune divisore degli interi n tali che $p_{ii}^{(n)} > 0$, ovvero*

$$d_i = MCD\left\{n \geq 1 : p_{ii}^{(n)} > 0\right\}.$$

Se $d_i = 1$ lo stato si dice aperiodico; per convenzione, se uno stato i è tale che l'insieme su cui calcolare il MCD è vuoto, si assume che $d_i = 1$.

Anche la periodicità è una proprietà di classe. In definitiva lo spazio degli stati $\mathbb{S}$ può essere decomposto nel modo seguente

$$S = T \cup C_1 \cup C_2 \cup \cdots$$

dove T è l'insieme degli stati transitori mentre ogni singola classe C_i è una classe irriducibile di stati ricorrenti. Ogni singola C_i può essere ricorrente positiva o ricorrente nulla; inoltre, tutti gli stati appartenenti ad una classe hanno il medesimo periodo. Ne consegue che se la catena parte da uno stato della classe C_i per qualche i allora resterà in C_i per sempre. Se invece la catena parte da uno stato di T il processo si può evolvere in due modi: o la catena resta sempre in T oppure ne esce per entrare in una delle classi C_i per poi rimanerci.

Esempio D.3 [*Classificazione degli stati*]
Consideriamo il processo $\mathbf{X} = \{X_0, X_1, \cdots\}$ con spazio degli stati $\mathbb{S} = \{A, B, C\}$ e matrice di transizione

$$\mathbf{P} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix} \end{array}.$$

In questo caso l'insieme degli stati ricorrenti C_1 è costituito dal solo stato A mentre l'insieme degli stati transitori t è costituito dagli stati B e C. ♢

Esempio D.4 [*Classificazione degli stati*]
Consideriamo il processo $\mathbf{X} = \{X_0, X_1, \cdots\}$ con spazio degli stati $\mathbb{S} = \{A, B, C, D\}$ e matrice di transizione

$$\mathbf{P} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array}.$$

Questo è un esempio di catena irriducibile in quanto tutti gli stati sono ricorrenti. ♢

Definizione D.4 *Uno stato $i \in S$, si dice ergodico se è ricorrente positivo e aperiodico.*

D.1.5 Distribuzioni invarianti

Una legge di probabilità $\boldsymbol{\pi}^* = (\pi_j, j \in \mathbb{S})$ si dice invariante per una catena di Markov con matrice di transizione P se

$$\boldsymbol{\pi}^* = \boldsymbol{\pi}^* \mathbf{P}, \quad \text{ovvero} \quad \boldsymbol{\pi}^* (\mathbf{I} - \mathbf{P}) = 0. \tag{D.4}$$

Ovviamente, se $\boldsymbol{\pi}$ è anche la distribuzione iniziale della catena, allora il processo risulta stazionario, ovvero

$$\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}^{(1)} = \boldsymbol{\pi}^{(2)} = \cdots = \boldsymbol{\pi}^*.$$

L'esistenza (ma non l'unicità) di una distribuzione invariante è garantita solo quando S è finito.

Esiste anche un legame tra distribuzioni invarianti e distribuzioni limite. Quando $\mathbb{S}$ è finito, infatti, se accade che, per qualche $i \in \mathbb{S}$ e per $n \to \infty$,

$$p_{ij}^{(n)} \to \pi_j, \qquad \forall j \in \mathbb{S},$$

allora la legge "limite" $\boldsymbol{\pi} = (\pi_j, j \in \mathbb{S})$ è invariante. Va ricordato tuttavia che, quando $\mathbb{S}$ è numerabile, può accadere che tutte le $p_{ij}^{(n)}$ convergano a delle costanti (ad esempio 0) che però non costituiscono una distribuzione di probabilità.

Quando la catena è irriducibile, è possibile caratterizzare la distribuzione invariante nel modo seguente

Teorema D.1 *Sia $(X_n; n \geq 0)$ una catena irriducibile. Allora le tre affermazioni seguenti sono equivalenti*

- *Esiste uno stato ricorrente.*
- *Tutti gli stati sono ricorrenti.*
- *Esiste una misura invariante $\boldsymbol{\pi}^*$ tale che, $\forall j \in \mathbb{S}$*

$$\pi_j^* = \frac{1}{\mu_j}$$

La terza affermazione del Teorema D.1 ci dice che, se una catena è irriducibile ricorrente positiva, la distribuzione invariante è data dai reciproci dei tempi medi di ritorno nei singoli stati. Se poi la catena è ricorrente nulla allora la misura invariante **non è** una legge di probabilità in quanto $\pi_j^* = 0$, $\forall j \in \mathbb{S}$.

D.1.6 Equilibrio di una catena

Teorema D.2 *Consideriamo una catena di Markov con matrice di transizione bP irriducibile e aperiodica con distribuzione invariante* $\boldsymbol{\pi}^*$*. Qualunque sia la legge iniziale* $\boldsymbol{\pi}^{(0)}$ *del processo, avremo che*

- *La distribuzione al tempo n, al crescere di n converge alla distribuzione invariante, ovvero*

$$\Pr(X_n = j) \to \pi_j, \text{ per } n \to \infty \text{ e } \forall j \in \mathbb{S}$$

- *Al crescere di n le probabilità di transizione in n passi convergono alla distribuzione invariante, indipendentemente dallo stato di partenza, ovvero*

$$p_{ij}^{(n)} \to \pi_j, \text{ per } n \to \infty \text{ e } \forall i, j \in \mathbb{S}.$$

Il secondo punto del teorema precedente implica che, se ad esempio $\mathbb{S} = \{0, 1, 2, \cdots\}$ la matrice di transizione in n passi $\mathbf{P}^{(n)}$ converge verso la matrice, con righe tutte uguali,

$$\begin{vmatrix} \pi_0 & \pi_1 & \pi_2 & \cdots & \cdots \\ \pi_0 & \pi_1 & \pi_2 & \cdots & \cdots \\ \pi_0 & \pi_1 & \pi_2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{vmatrix}$$

Lo stesso secondo punto implica che $\boldsymbol{\pi}$ rappresenta la distribuzione di equilibrio della catena; essa rappresenta anche la proporzione di tempo, nel lungo periodo, che la catena trascorrerà nei vari stati.

Concludiamo questa sezione con due (contro-)esempi. È possibile che una catena abbia una distribuzione invariante ma non abbia distribuzione di equilibrio.

Esempio D.5

Consideriamo la catena a due stati, ovvero $\mathbb{S} = \{A, B\}$, e matrice di transizione

$$\mathbf{P} = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}$$

in questo caso la distribuzione $\boldsymbol{\pi} = (1/2, 1/2)$ è invariante per la catena ma non esiste alcuna distribuzione di equilibrio, in quanto la catena oscilla in modo deterministico tra gli stati A e B senza convergere ad alcun limite. ♢

È altresì possibile che una catena di Markov sia dotata di più di una distribuzione invariante ma non abbia distribuzione di equilibrio.

Esempio D.6

Consideriamo la catena a due stati, ovvero $\mathbb{S} = \{A, B\}$, e matrice di transizione

$$\mathbf{P} = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}$$

in questo caso la catena resta per sempre nello stato in cui si trova all'inizio. Qualunque distribuzione $\boldsymbol{\pi}$ risulterà invariante ma non esistono distribuzioni di equilibrio. ♢

D.1.7 Reversibilità

Come già detto, l'esistenza di una distribuzione invariante garantisce che la catena abbia un comportamento *stazionario*, nel senso che le distribuzioni marginali ai vari istanti sono identiche. È importante allora determinare condizioni sufficienti a garantire la stazionarietà di una catena, ovvero l'esistenza di una distribuzione invariante.

Definizione D.5 *Una distribuzione di probabilità* $\boldsymbol{\pi}$ *sullo spazio degli stati* $\mathbb{S}$ *soddisfa il principio di bilanciamento locale se, per ogni coppia di stati* $(i, j) \in S$*, vale la relazione*

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{D.5}$$

Il principo di bilanciamento locale, in sostanza, afferma che il flusso dallo stato i allo stato j equivale, nel lungo periodo, a quello inverso tra lo stato j e lo stato i. È facile vedere che, se una catena soddisfa il principio di bilanciamento locale, allora soddisfa anche una sorta di bilanciamento globale, ovvero l'equazione di invarianza $\pi = \pi\mathbf{P}$. Esiste un legame stretto tra le nozioni di bilanciamento locale e il concetto di reversibilità nel tempo. Data una catena di Markov $\{X_n, n = \dots, -1, 0, 1, \dots\}$, con matrice di transizione $\mathbf{P}$, si chiama catena *rovesciata* la successione

$$\mathbf{Y} = \dots, Y_{-1}, Y_0, Y_1, \dots,$$

tale che $Y_n = X_{-n}$ per ogni n. È facile vedere che $\mathbf{Y}$ ancora una catena di Markov. In generale, tuttavia, la catena rovesciata $\mathbf{Y}$ avr una matrice di transizione $\mathbf{Q}$ differente da $\mathbf{P}$.

Definizione D.6 *Una catena di Markov stazionaria* $\{X_n, n = \dots, -1, 0, 1, \dots\}$*, con matrice di transizione* $\mathbf{P}$*, si dice reversibile se la matrice di transizione* $\mathbf{Q}$ *della catena rovesciata è uguale a* $\mathbf{P}$*.*

È poi facile verificare che una catena reversibile se e solo se la sua matrice di transizione sodddifa il principio di bilanciamento in dettaglio, per ogni coppia di stati $(i, j) \in S$.

È noto che, per una catena di Markov, passato e futuro sono indipendenti condizionatamente al presente; è naturale allora supporre che la catena anche se osservata in senso temporale inverso sia ancora markoviana; in questo senso la catena ha un comportamento simmetrico. D'altro canto la convergenza ad una distribuzione di equilibrio è un esempio di comportamento asimmetrico, in quanto la catena può passare da una situazione molto specifica (ad esempio, l'informazione che X_0 assuma uno stato ben preciso) ad una situazione di incertezza al limite (distribuzione di equilibrio). La condizione di reversibilità (D.5) serve proprio a garantire che la catena, quando osservata all'indietro nel tempo sia ancora markoviana.

Se una distribuzione $\boldsymbol{\pi}$ soddisfa la condizione di reversibilità allora essa è anche invariante per la catena: infatti

$$\sum_{i \in S} \pi_i p_{ij} = \sum_{i \in S} \pi_j p_{ji} = \pi_j \sum_{i \in S} p_{ij} = \pi_j.$$

La condizione di reversibilità è più semplice da verificare rispetto all'invarianza: va però ricordato che essa rappresenta una condizione solo sufficiente, cosicch una distribuzione potrebbe essere stazionaria senza soddisfare la (D.5).

D.2 Catene continue

Vedi oulu2x2.pdf

E

Le principali distribuzioni di probabilità

In questa appendice, dopo un breve cenno ad alcune speciali funzioni matematiche, molto frequenti nei calcoli statistici, sono elencate le principali famiglie di leggi di probabilità utilizzate nei problemi di inferenza, bayesiana e non. Per ognuna di esse, oltre ad una breve introduzione sulla genesi della famiglia, vengono fornite le principali caratteristiche. Per comodità di lettura elenchiamo qui la notazione utilizzata:

- funzione di ripartizione $F(x) = P(X \leq x)$;
- funzione di densità (nel caso assolutamente continuo) $f(x; \boldsymbol{\theta})$, dove $\boldsymbol{\theta}$ rappresenta il generico vettore dei parametri (nel caso in cui il parametro è scalare, verrà indicato con θ) oppure funzione di probabilità (nel caso discreto) $p(x; \theta) = P(X = x; \boldsymbol{\theta})$;
- media: $\mathbf{E}(X; \boldsymbol{\theta})$; varianza: $\mathrm{Var}(X; \boldsymbol{\theta})$.
- si indica con $S(X; \boldsymbol{\theta})$ il supporto della variabile aleatoria X, ovvero l'insieme dei valori x per i quali $P(X = x; \boldsymbol{\theta}) > 0$ (caso discreto) oppure $f(x; \boldsymbol{\theta}) > 0$ (caso assolutamente continuo).

Funzione Gamma di Eulero

Si chiama funzione Gamma e si indica con $\Gamma(t)$ la funzione definita, per $t > 0$, come

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx.$$

Si può facilmente dimostrare, mediante un'integrazione per parti, che vale la relazione ricorrente

$$\Gamma(t+1) = t\,\Gamma(t), \qquad t > 0. \tag{E.1}$$

Dalla (E.1) e dal fatto che

$$\Gamma(1) = \int_0^\infty e^{-x} dx = 1,$$

discende che, se t è un numero intero,

$$\Gamma(t) = (t-1)!$$

Inoltre vale la formula $\Gamma(1/2) = \sqrt{\pi}$, che si può dimostrare attraverso il cambio di variabile $x = y^2/2$ e ricordando l'espressione della densità della curva normale standardizzata. Per valori elevati dell'argomento t, $\Gamma(t)$ può essere approssimata mediante la formula di Stirling

$$\Gamma(t+1) = \sqrt{2\pi t}\; t^t \;\exp\{-t + \frac{\epsilon}{12t}\} \left(1 + O(\frac{1}{t})\right), \quad 0 \leq \epsilon \leq 1. \tag{E.2}$$

Funzione Beta di Eulero

Si definisce funzione Beta e si indica con $B(s,t)$ la funzione definita, per $s>0, t>0$, come

$$B(s,t)=\int_0^1 x^{s-1}(1-x)^{t-1}dx;$$

Si può dimostrare, anche in questo caso attraverso un semplice cambio di variabile, che $B(s,t)$ è esprimibile in termini della funzione Gamma attraverso la relazione

$$B(s,t)=\frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}.$$

E.1 Distribuzioni discrete

Bernoulliana [Be(θ)]

Si dice che la v.a. $X\sim Be(\theta)$ quando $S(X;\theta)=\{0,1\}$ per ogni $\theta\in[0,1]$, e

$$p(x;\theta)=\theta^x(1-\theta)^{1-x},\quad x=\,0,\,1.$$

Inoltre, si calcola facilmente che

$$\mathbf{E}\,(X;\theta)=\theta,\qquad \mathrm{Var}\,(X;\theta)=\theta(1-\theta).$$

Binomiale [Bin(n,θ)]

Si dice che la v.a. $X\sim \mathrm{Bin}(n,\theta)$ quando $S(X;\theta)=\{0,1,2,\cdots,n\}$ per ogni $\theta\in[0,1]$ e

$$P(k;n,\theta)=P(X=k;n,\theta)=\binom{n}{k}\theta^k(1-\theta)^{n-k},\qquad k=0,1,2,\cdots,n.$$

Si può facilmente dimostrare che, se $Y_1,Y_2,\cdots,Y_n$ sono n v.a. indipendenti, tutte con distribuzione $Be(\theta)$, allora la variabile somma

$$X=Y_1+Y_2+\cdots Y_n\sim \mathrm{Bin}(n,\theta); \tag{E.3}$$

dalla (E.3) è immediato dedurre che, se $X\sim\mathrm{Bin}(n,\theta)$, allora

$$\mathbf{E}\,(X;n,\theta)=n\theta,\qquad \mathrm{Var}\,(X;n,\theta)=n\,\theta\,(1-\theta).$$

Geometrica [Geo(θ)]

Si dice che la v.a. $X\sim\mathrm{Geo}(\theta)$ quando $S(X;\theta)=\{0,1,2,\cdots\}$ per ogni $\theta\in[0,1]$ e

$$P(k;\theta)=P(X=k;\theta)=\theta\,(1-\theta)^k,\qquad k=0,1,2,\cdots.$$

La v.a. geometrica ha una naturale interpretazione come il numero di insuccessi che precedono il verificarsi del primo successo in una successione di prove bernoulliane, ovvero indipendenti e tutte con la stessa probabilità di successo. Il calcolo della media si effettua con un piccolo trucco

$$\mathbf{E}\,(X;\theta)=\sum_{j=0}^{\infty}j\theta(1-\theta)^j=\sum_{j=1}^{\infty}j\theta(1-\theta)^j$$

$$= \sum_{k=0}^{\infty}(k+1)\theta(1-\theta)^{k+1} = -\theta\,(1-\theta)\sum_{k=0}^{\infty}\frac{\partial}{\partial\theta}(1-\theta)^{k+1};$$

Assumendo la possibilità di invertire il simbolo di integrazione e quello di serie, si ottiene che la quantità precedente è pari a

$$-\theta(1-\theta)\frac{\partial}{\partial\theta}\sum_{k=0}^{\infty}(1-\theta)^{k+1} =$$

$$= -\theta\,(1-\theta)\frac{\partial}{\partial\theta}\frac{(1-\theta)}{\theta} = \frac{\theta(1-\theta)}{\theta^2} = \frac{1-\theta}{\theta},$$

la quale suggerisce come il numero atteso di insuccessi che precedono il primo successo è inversamente proporzionale alla probabilità di successo nella singola prova. Con calcoli simili si ottiene che $\mathrm{Var}\,(X;\theta) = (1-\theta)/\theta^2$. A volte, la v.a. geometrica viene definita come il numero Z di prove necessarie per ottenere il primo successo. È ovvio che risulta $Z = X + 1$ e che

$$P(k;\theta) = \Pr\,(Z = k;\theta) = \theta\,(1-\theta)^{k-1}, \qquad k = 1, 2, 3, \cdots.$$

Binomiale negativa [BiNeg(n, θ)]

Si dice che la v.a. $X \sim \mathrm{BiNeg}(n,\theta)$ quando $S(X;\theta) = \{0, 1, 2, \cdots\}$ per ogni $\theta \in [0,1]$ e

$$P(k;n,\theta) = \Pr\,(X = k;n,\theta) = \binom{n+k-1}{k-1}\theta^n\,(1-\theta)^k, \qquad k = 0, 1, 2, \cdots.$$

La v.a. binomiale negativa, in analogia con quanto detto a proposito della v.a. geometrica, ha una naturale interpretazione come il numero di insuccessi che precedono il verificarsi dell'n-esimo successo in una successione di prove bernoulliane, ovvero indipendenti e tutte con la stessa probabilità di successo.
Una v.a. $X \sim \mathrm{BiNeg}(n,\theta)$ può essere vista come la somma di n v.a. $Y_1, Y_2, \cdots, Y_n$, indipendenti e somiglianti, con distribuzione $\mathrm{Geo}(\theta)$. Ne segue che

$$\mathbb{E}\,(X;\theta) = \sum_{j=1}^{n}\mathbb{E}\,(Y_j;\theta) = n\frac{1-\theta}{\theta}; \quad \mathrm{Var}\,(X;\theta) = \sum_{j=1}^{n}\mathrm{Var}\,(Y_j;\theta) = n\frac{1-\theta}{\theta^2};$$

Ipergeometrica [IpGeo(N, n, θ)]

Si dice che la v.a. $X \sim \mathrm{IpGeo}(N,n,\theta)$ quando $S(X;N,n,\theta) = \{0, 1, 2, \cdots, n\}$ per ogni $0 < \theta < 1, n < N$ ed $N\theta \in \mathbb{N}$ e

$$P(k;N,n,\theta) = \binom{N\theta}{k}\binom{(1-\theta)N}{n-k}\Big/\binom{N}{n},$$

per $n-(1-\theta)N \le k \le N\theta$. La distribuzione ipergeometrica emerge in modo naturale negli schemi di estrazione senza ripetizione o in blocco. Consideriamo un'urna contenente N palline di cui $N\theta$ di colore rosso, e effettuiamo l'estrazione in blocco di n palline. Allora il numero di palline rosse tra le n estratte avrà distribuzione di tipo $\mathrm{IpGeo}(N,n,\theta)$. Con semplici calcoli si dimostra che

$$\mathbb{E}\,(X;N,n,\theta) = n\theta; \qquad \mathrm{Var}\,(X;N,n,\theta) = \frac{N-n}{N-1}n\theta(1-\theta)$$

Beta-binomiale [BeBi(n, α, β)]

Si dice che la v.a. $X \sim \text{BeBi}(n, \alpha, \beta)$ quando $S(X;\theta) = \{0, 1, 2, \cdots, n\}$ per ogni $\alpha > 0, \beta > 0, n \in \mathbb{N}$ e

$$P(k; n, \alpha, \beta) = \binom{n}{k} \frac{B(\alpha + k, \beta + n - k)}{Beta(\alpha, \beta)}$$

L'interpretazione più naturale di una v.a. beta-binomiale è quella di una mistura di distribuzioni binomiali con parametro n fissato e parametro θ aleatorio con distribuzione di tipo Beta(α, β). Si vede infatti facilmente che

$$P(k; n, \alpha, \beta) = \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} \frac{1}{Beta(\alpha, \beta)} \theta^{\alpha-1} d\theta (1-\theta)^{\beta-1}.$$

Calcoli semplicissimi (utilizzando la definizione e le proprietà della funzione Beta di Eulero) conducono a

$$\mathbb{E}(X; n, \alpha, \beta) = \frac{n\alpha}{\alpha + \beta}, \quad \text{Var}(X; n, \alpha, \beta) = \frac{n\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)}.$$

Poisson [Po(θ)]

Si dice che $X \sim Po(\theta)$ quando $S(X;\theta) = \{0, 1, 2, \cdots\}$ per ogni $\theta \in [0, 1]$, e

$$P(k;\theta) = \Pr(X = k;\theta) = e^{-\theta} \frac{\theta^k}{k!}$$

Inoltre

$$\mathbb{E}(X;\theta) = \text{Var}(X;\theta) = \theta.$$

E.2 Distribuzioni assolutamente continue

Beta [Beta(α, β)]

Si dice che $X \sim \text{Beta}(\alpha, \beta)$ quando $S(x, \alpha, \beta) = [0, 1]$ e, per ogni $\alpha > 0$, e $\beta > 0$, la funzione di densità vale

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \qquad x \in [0, 1].$$

Dall'espressione della densità si ottiene la seguente uguaglianza, utile per il calcolo dei momenti:

$$\int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

Da questo si deduce immediatamente che, per ogni k positivo,

$$\mathbb{E}(X^k) = \frac{\Gamma(\alpha + k)\Gamma(\beta)}{\Gamma(\alpha + \beta + k)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{\alpha(\alpha+1)\cdots(\alpha+k-1)}{(\alpha+\beta)(\alpha+\beta+1)\cdots(\alpha+\beta+k-1)}.$$

Così, ad esempio,

$$\mathbb{E}(X) = \frac{\alpha}{\alpha + \beta}; \quad \text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{E.4}$$

Caso particolare. Per $\alpha = \beta = 1$ si ottiene la distribuzione uniforme nell'intervallo $[0, 1]$. Una v.a. con tale distribuzione si denota con il simbolo $X \sim \text{U}(0, 1)$.

Esponenziale [Exp(θ)]

Si dice che $X \sim \text{Exp}(\theta)$ quando $S(x, \theta) = (0, \infty)$ e, per ogni $\theta > 0$, la funzione di densità vale

$$f(x; \theta) = \theta\, e^{-\theta x}, \qquad x > 0$$

I momenti si ottengono come casi particolari dei momenti di una distribuzione di tipo Gamma.

Gamma [Gamma(α, θ)]

Si dice che $X \sim \text{Gamma}(\alpha, \theta)$ quando $S(x, \alpha, \theta) = (0, \infty)$ e, per ogni α e θ positivi, la funzione di densità vale

$$f(x; \alpha, \theta) = \frac{\theta^\alpha}{\Gamma(\alpha)} e^{-\theta x} x^{\alpha-1}, \qquad x > 0 \tag{E.5}$$

Il parametro θ prende il nome di parametro di scala mentre α è detto parametro di forma. Dalla forma della densità (E.5) si deduce l'uguaglianza

$$\int_0^\infty e^{-\theta x} x^{\alpha-1} dx = \frac{\Gamma(\alpha)}{\theta^\alpha},$$

utile per il calcolo dei momenti. Infatti,

$$\mathbb{E}\left(X^k\right) = \int_0^\infty \frac{\theta^\alpha}{\Gamma(\alpha)} e^{-\theta x} x^{\alpha+k-1} dx = \frac{\theta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\theta^{\alpha+k}} = \frac{\alpha(\alpha+1)\cdot(\alpha+k-1)}{\theta^k},$$

da cui, ad esempio

$$\mathbb{E}\left(X\right) = \frac{\alpha}{\theta}, \qquad \text{Var}\left(X\right) = \frac{\alpha(\alpha+1)}{\theta^2} - \frac{\alpha^2}{\theta^2} = \frac{\alpha}{\theta^2}.$$

Casi particolari.

- Se $\alpha = 1$, $X \sim \text{Exp}(\theta)$.
- Se $\alpha = \nu/2$ e $\theta = 1/2$,
$$f(x; \nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \exp\left\{-\frac{1}{2}x\right\} x^{\frac{\nu}{2}-1},$$
e la distribuzione prende il nome di Chi quadrato con ν gradi di libertà: in simboli $X \sim \chi^2_\nu$.
- Una variabile $X \sim \chi^2_2$ è equivalente ad una $X \sim \text{Exp}(\frac{1}{2})$.
- Se $X \sim \text{Gamma}(\alpha, \theta)$, la trasformazione lineare $Y = 2\theta X$ ha distribuzione $\chi^2_{2\alpha}$.

Gamma inversa [GI(α, θ)]

Si dice che $X \sim GI(\alpha, \theta)$ quando $S(X, \alpha, \theta) = \mathbb{R}^+$ per ogni $\alpha, \theta > 0$, e la funzione di densità vale

$$f(x; \alpha, \theta) = \frac{\theta^\alpha}{\Gamma(\alpha)} \frac{1}{x^{\alpha+1}} e^{-\theta/x}, \quad x > 0. \tag{E.6}$$

La densità (E.6) deve il suo nome al fatto che

$$X \sim GI(\alpha, \theta) \longrightarrow 1/X \sim \text{Gamma}(\alpha, \theta).$$

Dall'espressione della (E.6) si deduce la seguente identità, utile per il calcolo dei momenti della X:

$$\int_0^\infty e^{-\theta/x}\frac{1}{x^{\alpha+1}}dx = \frac{\Gamma(\alpha)}{\theta^\alpha} \tag{E.7}$$

Utilizzando la (E.7) si ottiene facilmente che, ad esempio,

$$\mathbb{E}(X) = \frac{\theta}{\alpha-1}; \quad \mathrm{Var}(X) = \frac{\theta^2}{(\alpha-1)^2(\alpha-2)}$$

.

Pareto [Pa(γ, β)]

Si dice che $X \sim \mathrm{Pa}(\gamma,\beta)$ quando $S(X,\gamma,\beta) = (\beta, +\infty)$ per ogni γ , e la funzione di densit vale

$$f(x;\gamma,\beta) = \gamma\frac{\beta^\gamma}{x^{\gamma+1}}, \quad x > \beta,\ \beta > 0.$$

Inoltre

$$\mathbb{E}(X;\gamma,\beta) = \frac{\gamma}{\gamma-1}\beta; \quad \mathrm{Var}(X;\gamma,\beta) = \frac{\gamma}{(\gamma-1)^2(\gamma-2)}\beta^2.$$

Normale o Gaussiana [N(μ, σ^2)]

Si dice che $X \sim \mathrm{N}(\mu,\sigma)$ quando $S(x,\mu,\sigma) = \mathbb{R}$ e, per ogni $\mu \in \mathbb{R}$ e $\sigma > 0$, la funzione di densità vale

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \qquad x \in \mathbb{R}.$$

Con semplici integrazioni per parti si ottengono i momenti di X. Elenchiamo di seguito i più importanti dal punto di vista statistico;

$$\mathbb{E}(X) = \mu; \quad \mathrm{Var}(X) = \sigma^2;$$

$$\mathbb{E}(X-\mu)^{2k-1} = 0, \forall k \in \mathbb{N}, \quad \mathbb{E}(X-\mu)^{2k} = \frac{(2k)!\sigma^{2k}}{k!2^k}.$$

Quando $\mu = 0$ e $\sigma = 1$, la v.a. prende il nome di normale standardizzata e la densit viene in genere indicata con il simbolo $\varphi(\cdot)$.

La funzione di ripartizione non ha una espressione esplicita. Nel caso standardizzato, per approssimare

$$\Phi(x) = \int_{-\infty}^x \varphi(t)dt,$$

si utilizza, per $x \to +\infty$, il seguente risultato:

$$1 - \Phi(x) \approx \frac{\varphi(x)}{x}.$$

Cauchy [Ca(μ, σ)]

Si dice che $X \sim \mathrm{Ca}(\mu,\sigma)$ quando $S(x,\mu,\sigma) = \mathbb{R}$ e, per ogni $\mu \in \mathbb{R}$ and $\sigma > 0$, la funzione di densità vale

$$f(x;\mu,\sigma) = \frac{\sigma}{\pi\left(\sigma^2 + (1+(x-\mu)^2\right)}, \qquad x \in \mathbb{R}.$$

Si dimostra facilmente che la media (e ovviamente tutti i momenti di ordine superiore) di una v.a. di Cauchy non esiste. La v.a. di Cauchy può essere ottenuta mediante trasformazioni elementari di altre v.a. note. Ad esempio,

- se $X \sim \mathrm{U}(-\frac{\pi}{2}, \frac{\pi}{2})$, allora $Y = \tan(X) \sim \mathrm{Ca}(0,1)$ (questa relazione può essere utile per generare valori pseudo-aleatori da una legge di Cauchy).
- se X_1 e X_2 sono indipendenti con distribuzione $N(0,1)$, allora $Y = X_1/X_2 \sim \mathrm{Ca}(0,1)$

Logistica [Lo(μ, σ)]

Si dice che $X \sim \text{Lo}(\mu, \sigma)$ quando $S(x, \mu, \sigma) = \mathbb{R}$ e, per ogni $\mu \in \mathbb{R}$ e $\sigma > 0$, la funzione di densità vale

$$f(x; \mu, \sigma) = \frac{1}{\sigma} \frac{\exp\left\{\frac{x-\mu}{\sigma}\right\}}{\left(1 + \exp\left\{\frac{x-\mu}{\sigma}\right\}\right)^2}, \qquad x \in \mathbb{R}.$$

Si dimostra facilmente che

$$V \sim \text{U}(0,1) \longrightarrow X = \log \frac{V}{1-V} \sim \text{Lo}(0,1) \tag{E.8}$$

Laplace o doppia esponenziale [La(μ, λ)]

Si dice che $X \sim \text{La}(\mu, \lambda)$ quando $S(x, \mu, \lambda) = \mathbb{R}$ e, per ogni $\mu \in \mathbb{R}$ e $\lambda > 0$, la funzione di densità vale

$$f(x; \mu, \lambda) = \frac{\lambda}{2} \exp\left\{-\lambda|x - \mu|\right\}, \qquad x \in \mathbb{R}.$$

La densità è simmetrica rispetto al parametro di posizione μ che ne rappresenta quindi la media e la mediana. Il parametro λ, o meglio $1/\lambda$, è il parametro di scala. Inoltre $\text{Var}\,(X; \lambda) = 2/\lambda^2$.

Student [St(ν, μ, σ)]

Si dice che $X \sim \text{St}(\nu, \mu, \sigma)$ quando $S(x, \mu, \sigma, \nu) = \mathbb{R}$ e, per ogni $\mu \in \mathbb{R}$, $\nu > 0$, e $\sigma > 0$, la funzione di densità vale

$$f(x; \mu, \sigma) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\Gamma(1/2)\sigma\sqrt{\nu}} \left(1 + \frac{(x-\mu)^2}{\sigma^2 \nu}\right)^{-(\nu+1)/2}, \qquad x \in \mathbb{R}\,.$$

Il parametro ν prende il nome di *gradi di libertà*. Si può dimostrare che il momento k-esimo di una v.a. t di Student esiste solo quando $\nu > k$.
Caso particolare: per $k = 1$ si riottiene una distribuzione di Cauchy.
È importante ricordare che se $X \sim \text{N}(\mu, \sigma^2)$ e $Y \sim \chi^2_\nu$, con X e Y indipendenti, allora

$$\frac{(X-\mu)\sqrt{\nu}}{\sigma\sqrt{Y}} \sim \text{St}(\nu, \mu, \sigma)$$

Fisher [Fis(ν, ξ)]

Si dice che $X \sim \text{Fis}(\nu, \xi)$ quando $S(x, \nu, \xi) = \mathbb{R}$ e, per ogni ν,$\xi > 0$, la funzione di densità vale

$$f(x; \nu, \xi) = \frac{\Gamma((\nu+\xi)/2)}{\Gamma(\nu/2)\Gamma(\xi/2)} \nu^{\xi/2}\, \xi^{\nu/2} \frac{x^{(\nu-2)/2}}{(\nu + \xi x)^{(\nu+\xi)/2}}, \qquad x > 0\,.$$

I parametri ν e ξ prendono il nome di *gradi di libertà*. Questa legge appare in molti sviluppi della teoria del campionamento da popolazioni gaussiane. Ad esempio se $X \perp\!\!\!\perp Y$, $X \sim \chi^2_\nu$ e $Y \sim \chi^2_\xi$, allora

$$\frac{X\,\xi}{Y\,\nu} \sim \text{Fis}(\nu, \xi).$$

Inoltre

$$X \sim \text{Fis}(\nu, \xi) \Rightarrow \frac{\xi X}{\nu + \xi X} \sim \text{Beta}(\nu, \xi).$$

E.3 Distribuzioni multivariate

Multinomiale $[\mathbf{MNom_k(n, p)}]$

Si dice che il vettore k-dimensionale ha distribuzione multinomiale e si indica con il simbolo $\mathbf{X} \sim \text{MNom}_k(n, \mathbf{p})$, dove n è un intero e $\mathbf{p} = (p_1, p_2, \cdots, p_k)$, con $p_j \geq 0$ e $p_1 + p_2 + \ldots p_k = 1$, quando

$$S(\mathbf{x}, n, \mathbf{p},) = \left\{ (n_1, \cdots, n_k) \text{ interi } : n_i \geq 0, \sum_{i=1}^{k} n_i = n \right\}$$

e la funzione di probabilità vale

$$\Pr(X_1 = n_1, \cdots, X_k = n_k) = \frac{n!}{n_1!\, n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}.$$

La distribuzione multinomiale rappresenta la versione multidimensionale della legge binomiale; per questo motivo, se $\mathbf{X} \sim \text{MNom}_k(n, \mathbf{p})$, ogni coordinata del vettore $\mathbf{X}$ ha distribuzione binomiale. Più precisamente,

$$\mathbf{X} \sim \text{MNom}_k(n, \mathbf{p}) \Longrightarrow X_j \sim \text{Bin}(n, p_j), \qquad j = 1, \cdots, k.$$

Inoltre,

$$\mathbb{E}(\mathbf{X}) = \begin{pmatrix} np_1 \\ np_2 \\ \cdots \\ np_k \end{pmatrix},$$

e

$$\text{Var}(X_j) = np_j(1 - p_j),\ j = 1, \cdots, k; \quad \text{Cov}(X_i, X_j) = -np_i p_j, \forall i \neq j.$$

Dirichlet $[\mathbf{Dir_k(\boldsymbol{\alpha}, \boldsymbol{\gamma})}]$

Si dice che il vettore k-dimensionale $\mathbf{X} \sim \text{Dir}_k(\boldsymbol{\alpha}, \gamma)$ quando

$$S(\mathbf{x}, \boldsymbol{\alpha}, \gamma) = \left\{ \mathbf{x} \in \mathbb{R}^k : x_i > 0, \sum_{i=1}^{k} x_i < 1 \right\}$$

e la funzione di densità vale

$$f(\mathbf{x}; \mathbf{p}, \gamma) = \frac{\Gamma(\alpha_1 + \cdots \alpha_k + \gamma)}{\gamma \prod_{i=1}^{k} \Gamma(\alpha_j)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{k} x_i\right)^{\gamma - 1}. \tag{E.9}$$

La distribuzione di Dirichlet si dice anche "Beta multivariata"; infatti, per $k = 1$, la (E.9) si riduce alla densità di una Beta(α_1, γ).
Un modo costruttivo per ottenere una v.a. con legge di Dirichlet è il seguente: siano $Z_1, Z_2, \cdots, Z_{k+1}$ v.a. indipendenti tali che $Z_i \sim \text{Gamma}(\alpha_i, 1)$, $i = 1, \cdots, k+1$, e sia $T = \sum_{i=1}^{k+1} Z_i$. Allora il vettore

$$\mathbf{X} = \frac{1}{T}(Z_1, \cdots, Z_k)$$

ha distribuzione di Dirichlet con parametri $\mathbf{p} = (\alpha_1, \cdots, \alpha_k)$ e $\gamma = \alpha_{k+1}$. Da questa rappresentazione si deduce che $\mathbf{Z} \stackrel{d}{=} T\,\mathbf{X}$; si può inoltre dimostrare facilmente che T è indipendente da $\mathbf{X}$:

quindi, moltiplicando tra loro le coordinate dei due vettori, si ottiene che, per ogni k-pla di interi $(r_1, \cdots, r_k)$, si ha

$$\prod_{i=1}^{k} Z_i^{r_i} \stackrel{d}{=} T^r \prod_{i=1}^{k} X_i^{r_i},$$

con $r = \sum_{i=1}^{k} r_i$, ovvero

$$\mathbb{E}\left(\prod_{i=1}^{k} X_i^{r_i}\right) = \frac{\prod_{i=1}^{k} \mathbb{E}\left(Z_i^{r_i}\right)}{\mathbb{E}\left(T^r\right)}$$

Da questa formula generale, ricordando che $T \sim \text{Gamma}(\sum_{i=1}^{k+1} \alpha_i, 1)$, si deduce facilmente che

$$\begin{aligned}
\mathbb{E}\left(X_i\right) &= \frac{\mathbb{E}\left(Z_i\right)}{\mathbb{E}\left(T\right)} = \frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i + \gamma}; \\
\mathbb{E}\left(X_i^2\right) &= \frac{\mathbb{E}\left(Z_i^2\right)}{\mathbb{E}\left(T^2\right)} = \frac{\alpha_i(\alpha_i + 1)}{(\sum_{i=1}^{k} \alpha_i + \gamma)(\sum_{i=1}^{k} \alpha_i + \gamma + 1)}; \\
\mathbb{E}\left(X_i X_j\right) &= \frac{\mathbb{E}\left(Z_i\right)\mathbb{E}\left(Z_j\right)}{\mathbb{E}\left(T^2\right)} = \frac{\alpha_i \alpha_j}{(\sum_{i=1}^{k} \alpha_i + \gamma)(\sum_{i=1}^{k} \alpha_i + \gamma + 1)}.
\end{aligned}$$

Normale multivariata $[\mathbf{N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})}]$

Si dice che il vettore $\mathbf{X} = (X_1, \ldots, X_k)$ ha distribuzione Normale multivariata con parametri di posizione e scala pari a $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, matrice definita positiva, e si indica col simbolo $N_k(\boldsymbol{\mu}, \Sigma)$, se la densità vale

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}\right)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}.$$

Per $k = 1$ si ottiene la distribuzione normale univariata. Si può verificare facilmente che

$$\mathbb{E}\left(\mathbf{X}\right) = \boldsymbol{\mu}, \qquad \text{Var}\left(\mathbf{X}\right) = \boldsymbol{\Sigma}.$$

Una delle proprietà più importanti della distribuzione $N_k(\boldsymbol{\mu}, \Sigma)$ è la seguente, che stabilisce che ogni trasformazione lineare di $\mathbf{X}$ ha ancora legge normale.

Proposizione E.1 *Sia $\mathbf{A}_{p,k}$ una matrice di rango $p \leq k$: allora*

$$\mathbf{AX} \sim N_p(\mathbf{A}\boldsymbol{\mu}; \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}').$$

Distribuzioni marginali e condizionate.
Se $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \Sigma)$ allora ogni sottoinsieme delle coordinate di $\mathbf{X}$ ha ancora legge normale; anche la distribuzione di un sottoinsieme delle coordinate di $\mathbf{X}$ condizionatamente al resto delle coordinate ha legge normale; più precisamente, consideriamo la partizione di $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, con $\mathbf{X}_1$ di dimensione $p < k$ e $\mathbf{X}_2$ di dimensione $k - p$.

Proposizione E.2 *Sia*

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right);$$

allora

$$\mathbf{X}_1 \sim N_k\left(\boldsymbol{\mu}_1, \Sigma_{11}\right) \tag{E.10}$$

e

$$[\mathbf{X}_2 \mid \mathbf{X}_1 = x_1] \sim N_{k-p}\left(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right) \tag{E.11}$$

Normale Gamma [NoGa($\mu, \sigma, \alpha, \theta$)]

Si dice che il vettore (X, Y) ha distribuzione normale-gamma, e si indica col simbolo $\text{NoGa}(\mu, \sigma, \alpha, \theta)$, se la densità vale, per $x \in \mathbb{R}$ e $y > 0$,

$$f(x, y; \mu, \sigma, \alpha, \theta) = f(x; y, \mu, \sigma) f(y; \alpha, \theta) = \frac{\sqrt{y}}{\sigma\sqrt{2\pi}} e^{-\frac{y}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2} \frac{\theta^\alpha}{\Gamma(\alpha)} e^{-\theta y} y^{\alpha-1}.$$

In pratica, la distribuzione della variabile doppia viene definita in termini della legge marginale di Y, di tipo $\text{Gamma}(\alpha, \theta)$, e della legge condizionata di $X \mid Y = y$, di tipo $\text{N}(\mu, \sigma^2/y)$. L'importanza di questa legge è dovuta al fatto che essa rappresenta la distribuzione a priori coniugata nel modello gaussiano con parametro di posizione e scala entrambi incogniti.

Student multivariata [$\text{St}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu})$]

Si dice che il vettore $\mathbf{X} = (X_1, \ldots, X_k)$ ha distribuzione t con ν gradi di libertà e parametri di posizione e scala pari a $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, matrice definita positiva, e si indica col simbolo $St_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, se la densità vale

$$f(\mathbf{x}) = \frac{\Gamma((\nu + k)/2)}{|\Sigma|^{1/2} \Gamma(\nu/2)(\nu\,\pi)^{k/2}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+k}{2}}.$$

Per $k = 1$ si ottiene la t di Student univariata. La distribuzione t multivariata può essere ottenuta come mistura di scala di distribuzioni normali multivariate come mostra il seguente risultato [41].

Teorema E.1 *Siano* $\mathbf{X}$ *un vettore aleatorio* k*-dimensionale e* Y *una variabile aleatoria positiva tali che*

$$\mathbf{X} \mid Y \sim N_k(\boldsymbol{\mu}, Y\boldsymbol{\Psi}), \qquad Y \sim GI(a, b);$$

allora la legge marginale di $\mathbf{X}$ *è del tipo*

$$\mathbf{X} \sim St_k\left(2a, \boldsymbol{\mu}, \frac{b}{a}\boldsymbol{\Psi}\right).$$

Se poi, come caso particolare, si pone $a = \nu/2$ e $b = 1/2$, allora $Y^{-1} \sim \chi^2_\nu$ e $\mathbf{X} \sim St_k(\nu, \boldsymbol{\mu}, \Psi/\nu)$.

Dimostrazione E.1 *Lasciata per esercizio.*

Da sottolineare anche il seguente noto risultato.

Teorema E.2 *Nelle condizioni del teorema precedente si ha che*

(a)

$$\mathbf{W} = \frac{2b}{Y} \sim \chi^2_{2a};$$

(b) La v.a.

$$\mathbf{V} = \frac{(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Psi}^{-1} (\mathbf{X} - \boldsymbol{\mu})}{Y} \sim \chi^2_k$$

è indipendente da Y*.*

(c)

$$\frac{2a\,(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Psi}^{-1} (\mathbf{X} - \boldsymbol{\mu})}{2bk} \sim \text{Fis}(k, 2a)$$

Dimostrazione E.2 *Si tratta di un risultato classico della teoria del campionamento da popolazioni normali. Si veda ad esempio, [2].*

Wishart $[\mathbf{W}_k(m, \boldsymbol{\Sigma})]$

Si dice che la matrice quadrata k-dimensionale $\mathbf{V}$, definita positiva, ha distribuzione di Wishart con m gradi di libertà e parametro di scala pari a $\boldsymbol{\Sigma}$, matrice definita positiva, e si indica col simbolo $W_k(m, \boldsymbol{\Sigma})$, se la densità vale

$$f(\mathbf{V}) = \frac{1}{2^{mk/2}\Psi_k(m/2)|\boldsymbol{\Sigma}|^{m/2}}|\mathbf{V}|^{(m-k-1)/2}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\mathbf{V}\right)\right\},$$

dove

$$\Psi_k(u) = \pi^{k(k-1)/4}\prod_{i=1}^{k}\Gamma\left(u - \frac{1}{2}(i-1)\right), \quad u > \frac{k-1}{2}.$$

Un modo di costruire una matrice aleatoria di Wishart è il seguente: siano $(\mathbf{Z}_1, \cdots, \mathbf{Z}_m) \overset{\text{iid}}{\sim} N_k(\mathbf{0}, \mathbf{I})$; allora la quantità

$$W = \sum_{i=1}^{m}\mathbf{Z}_i\mathbf{Z}_i'$$

si distribuisce secondo una legge di Wishart $W_k(m, \mathbf{I})$.
Alcune proprietà della distribuzione $W_k(m, \Sigma)$:

Proposizione E.3 *Sia* $\mathbf{V} \sim W_k(m, \Sigma)$*:*

- *se* $\mathbf{A}$ *è una matrice* $q \times k$*, allora* $\mathbf{Y} = \mathbf{AVA}' \sim W_q(m, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$
- $\operatorname{tr}(\mathbf{V}) \sim \chi^2_{mk}$.
- $\mathbf{E}(V) = m\boldsymbol{\Sigma}$

Wishart inversa $[\mathbf{W}_{\boldsymbol{k}}^{-\mathbf{1}}(\boldsymbol{m}, \boldsymbol{\Sigma})]$

Sia $\mathbf{V} \sim W_k(m, \Sigma)$. Poich $\mathbf{V}$ è definita positiva con probabilità 1, è possibile calcolare la funzione di densità della matrice aleatoria inversa $\mathbf{Z} = \mathbf{V}^{-1}$:

$$f(\mathbf{Z}) = \frac{|\mathbf{Z}|^{-(m+k+1)/2}}{2^{mk/2}\Psi_k(m/2)|\boldsymbol{\Sigma}|^{m/2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\mathbf{Z}^{-1}\right)\right\}.$$

Inoltre

$$\mathbf{E}(\mathbf{Z}) = \frac{\boldsymbol{\Sigma}^{-1}}{m-k-1}.$$

La distribuzione è particolarmente utile nell'analisi coniugata del modello normale multivariato.

Soluzioni

Problemi del capitolo 1

1.1 La soluzione è presentata qui.

1.2 Titolo del problema

(a) La soluzione della prima parte del problema è presentata qui.

(b) La soluzione della seconda parte del problema è presentata qui.

Riferimenti bibliografici

[1] R. B. Ash. *Real Analysis and Probability*, volume 11. Academic Press, New York, 1972.

[2] A. Azzalini. *Inferenza Statistica. Una presentazione basata sul concetto di verosimiglianza*. Springer Italia, Milano, 2000.

[3] Adelchi Azzalini. A class of distributions which includes the normal ones. *Scand. J. Statist.*, 12:171–178, 1985.

[4] P. Baldi. *Calcolo delle probabilità e statistica*. Mc Graw-Hill, Italia, 1993.

[5] T. Bayes. An essays towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, 53:370–418, 1763.

[6] J.0. Berger. The frequentist viewpoint and conditioning. In *Proceedings ot the Berkeley Conference in honor of J. Neyman and J. Kiefer (L. LeCam and R. Olsen (eds))*, volume I, pages 197–222. Wadsworth, Belmont, Canada, 1985.

[7] James O. Berger, Jos M. Bernardo, and Manuel Mendoza. On priors that maximize expected information. In *Recent Developments in Statistics and their Applications (J. P. Klein and J. C. Lee, eds.)*, pages 1–20, Seoul, 1989. Freedom Academy Publishing.

[8] James O. Berger and Luis R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.*, 91(433):109–122, 1996.

[9] James O. Berger and Robert L. Wolpert. *The likelihood principle*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 6. Institute of Mathematical Statistics, Hayward, CA, 1984.

[10] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 1985.

[11] J.O. Berger, J.M. Bernardo, and D. Sun. *Objective Bayesian Analysis*. NON LO SO,, 2006.

[12] J.O. Berger and M. Delampady. Testing precise hypotheses (con discussione). *Statistical Science*, 2:317–352, 1987.

[13] J.O. Berger and L.R. Pericchi. Objective bayesian methods for model selection: introduction and comparison [with discussion]. *In 'Model Selection' (P.Lahiri, editor), IMS Lecture Notes, Monograph Series, vol. 38*, pages 135–207, 2001.

[14] J.M. Bernardo. Reference posterior distributions for bayesian inference (with discussion). *JRSS*, 2:113–147, 1979.

[15] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, New York, 1996.

[16] Jos Miguel Bernardo. Reference analysis. In *Bayesian Thinking: Modeling and Computation (D. Dey and C.R. Rao (eds))*, volume 25 of *Handbook of Statistics*, pages 17–90. Elsevier, The Netherlands, 2005.

[17] Stephen Bernstein and Ruth Bernstein. *Statistica inferenziale*. Collana Schaum. McGraw-Hill Italia, 2003.

[18] Allan Birnbaum. On the foundations of statistical inference. *J. Amer. Statist. Assoc.*, 57:269–326, 1962.

[19] A.A. Borovkov. *Estadistica Matematica (in spagnolo)*. Mosca, Russia, 1984.

[20] G.E.P. Box and R.E. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, New York, 1973.

[21] G. Casella and C. Robert. *Monte Carlo statistical methods, 2nd ed.* Springer, New York, 2004.

[22] George Casella and Roger L. Berger. *Statistical Inference, 2nd ed.* Wadsworth, New York, 1998.

[23] S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96:270–281, 2001.

[24] Siddhartha Chib. Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.*, 90(432):1313–1321, 1995.

[25] D.M. Cifarelli and P. Muliere. *Statistica bayesiana*. Giappichelli, Torino, 1989.

[26] Donato M. Cifarelli. *Calcolo delle probabilit*. Mc Graw-Hill Italia, Milano, 1996.

[27] Donato M. Cifarelli. *Introduzione alla teoria della stima*. EGEA, Milano, 2001.

[28] P. Congdon. *Bayesian Statistical Modelling*. Wiley, New York, 2001.

[29] D.R Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman & Hall, London, 1974.

[30] G. Dall'Aglio. *Calcolo delle Probabilit (II ed.)*. Zanichelli, Bologna, 2000.

[31] A. C. Davison. *Statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2003.

[32] B. de Finetti. La prévision: ses lois logiques, ses sources sujectives. *Ann. Inst. H. Poincaré*, 7:1–68, 1937.

[33] Bruno de Finetti. *Teoria delle probabilità: sintesi introduttiva con appendice critica. Volumi primo e secondo*. Giulio Einaudi Editore, Turin, 1970. Nuova Biblioteca Scientifica Einaudi, 25* et 25**.

[34] Bruno de Finetti. *Theory of probability: a critical introductory treatment. Vol. 1*. John Wiley & Sons, London-New York-Sydney, 1974. Translated by Antonio Machì and Adrian Smith, With a foreword by D. V. Lindley, Wiley Series in Probability and Mathematical Statistics.

[35] Bruno de Finetti. *Theory of probability: a critical introductory treatment. Vol. 2*. John Wiley & Sons, London-New York-Sydney, 1975. Translated from the Italian by Antonio Machì and Adrian Smith, Wiley Series in Probability and Mathematical Statistics.

[36] Bruno de Finetti. *Scritti (1926–1930)*. Casa Editrice Dott. Antonio Milani (CEDAM), Padua, 1981. With a preface by Massimo de Felice.

[37] F. De Santis. *Fattori di Bayes per la scelta del modello statistico*. Dip. di Statistica, Probabilita' e Statistiche Applicate, Univ. di Roma La Sapienza. Tesi di dottorato, 1997.

[38] F. De Santis and F. Spezzaferri. Methods for default and robust bayesian model comparison: The fractional bayes factor approach. *International Statistical Review*, 67:267–286, 1999.

[39] P. Dellaportas, J.J. Forster, and I. Ntzoufras. On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 417:???–207, 1999.

[40] Arthur P. Dempster. The direct use of likelihood for significance testing. In O. Barndorff-Nielsen, P. Blaesild, and G. Sihon, editors, *Proceedings of the Conference on Foundational Questions in Statistical Inference*, pages 335–352. University of Aarhus, 1974.

[41] James M. Dickey. Three multidimensional-integral identities with Bayesian applications. *Ann. Math. Statist.*, 39:1615–1628, 1968.

[42] H. Gamerman, D.and Lopes. *Markov Chain Monte Carlo: stochastic simulation and Bayesian inference*. Chapman & Hall/CRC, Boca Raton, FL, 2006.

[43] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis (II ed.)*. CRC Press, 2003.

[44] J.K. Ghosh and R.V. Raamamoorthy. *Bayesian Nonparametrics*. Springer Text in Statistics. Springer, New York, 2003.

[45] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[46] G. Grimmett and D. Stirzaker. *Probability and Random Processes (II ed.)*. Oxford Univ. Press, 1992.

[47] J Haldane. The precision of observed values of small frequencies. *Biometrika*, 35:297–303, 1948.

[48] C. Han and B.M. Carlin. MCMC methods for computing Bayes factors: a comparative review. *J. Amer. Statist. Assoc.*, 417:2??–207, 2001.

[49] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[50] H. Jeffreys. *Theory of Probability*. Oxford University Press, 1961.

[51] N.L. Johnson and S. Kotz. *Continuous Univariate Distributions*. Wiley, New York, 1970.

[52] R.E. Kass and A.E. Raftery. Bayes factor and model uncertainty. *Journal of the American Statistical Association*, 90:773–795, 1995.

[53] T. Lancaster. *An introduction to Bayesian Econometrics*. NON LO SO, 2004.

[54] P.S. Laplace. *Theorie Analitique des Probabilities*. Courcier, Paris, 1812.

[55] Chap T. Le. *Introductory biostatistics*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2003.

[56] E.L. Lehmann and G Casella. *Theory of point estimation, 2nd ed.* Springer, New York, 1998.

[57] D.V. Lindley. On a measure of information provided by an experiment. *Ann. Math. Statist.*, 27:986–1005, 1956.

[58] D.V. Lindley. The use of prior probability distributions in statistical inference and decision. *Proc. Fourth Berkeley Symp. Math. Statist. Probab. University of California Press*, 1961.

[59] B. Liseo and N. Loperfido. Default Bayesian analysis of the skew normal distribution. *J. Statist. Plann. Inference*, (136):373–389, 2006.

[60] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[61] P. Meyn and G Tweedie. *Monte Carlo statistical methods, 2nd ed.* Springer Texts in Statistics. Springer-Verlag, New York, 2004.

[62] J. R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, UK, 1997.

[63] A. O'Hagan. A moment of indecision. *Biometrika*, 68(1):329–330, 1981.

[64] A. O'Hagan and J. Forster. *Kendall's Advanced Theory of Statistics, Vol. 2B. Bayesian Inference (2nd ed.)*. Arnold, London, 2004.

[65] Anthony O'Hagan. Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B*, 57(1):99–138, 1995. With discussion and a reply by the author.

[66] L. Pace and A. Salvan. *Teoria della Statistica*. Cedam, Padova, 1996.

[67] Y. Pawitan. *In All Likelihood*. Oxford Univ. Press, 2001.

[68] L. Piccinato. *Metodi per le decisioni statistiche*. Springer Italia, Milano, 1996.

[69] L. Piccinato. Il concetto statistico di evidenza. *Pubblicazioni del Dip. di Statist. Probab. e Stat. Appl., Universita' di Roma La Sapienza*, 1999.

[70] G. Pompilj. Teorie statistiche della significativit e conformit dei risultati sperimentali agli schemi teorici. *Statistica*, 8:7–42, 1948.

[71] E. Regazzini. *Sulle probabilità coerenti*. CLUP, Bologna, 1983.

[72] N. Reid. Saddlepoint methods and statistical inference (with discussion). *Statistical Science*, 3:213–238, 1988.

[73] N. Reid. The role of conditioning. *Statistical Science*, 11:12–15, 1996.

[74] B.D. Ripley. *Stochastic Simulation*. Wiley, New York, 1987.

[75] Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 157–163, Berkeley and Los Angeles, 1956. University of California Press.

[76] C. Robert. *The Bayesian choice*. Springer, New York, 2001.

[77] Christian P. Robert and George Casella. *Monte Carlo statistical methods, 2nd ed.* Springer Texts in Statistics. Springer-Verlag, New York, 2004.

[78] R. Royall. *Statistical Evidence*. CRC press, 1996.

[79] L. Schmidt, L. Jahn, and L. Rodin. Esempio di psicocinesi. *Chiedere a Berger*, 2:non so, 1987.

[80] G. Schwarz. Estimating the dimension of model. *Annals of Statistics*, 6:461–64, 1978.

[81] Thomas A. Severini. *Likelihood methods in statistics*. Oxford University Press, Oxford, 2000.

[82] Thomas A. Severini. *Elements of Distribution Theory*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2005.

[83] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *J. Roy. Statist. Soc. Ser. B*, 64(3):583–639, 2002.

[84] D. Stirzaker. *Stochastic Processes and Models*. Oxford Univ. Press, 2005.

[85] L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86, 1986.

[86] Larry Wasserman. *All of statistics*. Springer Texts in Statistics. Springer-Verlag, New York, 2004. A concise course in statistical inference.

[87] Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques. Essay in honor of Bruno de Finetti. P.K.Goel and A. Zellner (eds.)*, pages 233–243. Elsevier, Amsterdam, 1986.

Indice analitico

de Finetti, 110
distribuzione di Cauchy, 136
scambiabilità, 110