UNIVERSITÀ DEGLI STUDI DI MILANO

SCUOLA DI DOTTORATO:
Scienze biomediche, cliniche e sperimentali
Dipartimento di Scienze Cliniche e di Comunità
Sezione di Statistica Medica e Biometria "G. A. Maccacaro"

CORSO DI DOTTORATO
Statistica Biomedica - XXVI Ciclo

TESI DI DOTTORATO DI RICERCA:

Approccio Frequentista e Bayesiano, Due Modi Diversi di Vedere La Stessa Realtà: Applicazioni alla Modulazione del Dolore a Misure Ripetute

Settore scientifico disciplinare MED/01

NOME DEL DOTTORANDO

Manuela Giangreco

TUTOR: Adriano Decarli

COORDINATORE: Adriano Decarli

A.A.2012/2013

Indice

A mio marito

Ai miei genitori

Introduzione

Nello studio di un fenomeno in generale si cerca di formalizzare la realtà nella quale ci si trova ad operare tramite un modello che permetta la piena conoscenza del fenomeno stesso. Per fare ciò si ricorre in ambito statistico all'inferenza cioè ad una particolare procedura che permetta di ottenere, da dati raccolti tramite un campione, informazioni generalizzabili alla popolazione dalla quale questi dati sono stati estratti. L'inferenza, però, dipende dal campione estratto, dal modello probabilistico sottostante che genera tale campione ma anche dall'approccio che si sceglie per ricondurre alla popolazione incognita ciò che si viene a conoscere dal campione noto. Si distinguono infatti due importanti approcci quello classico frequentista e quello bayesiano, entrambi caratterizzati da procedure inferenziali che rispettano loro specifici criteri sottostanti. Entrambi gli approcci, che verranno presentati nel prosieguo, hanno i loro meriti e punti di forza e non è escluso che possano essere usati in coppia. Inoltre, se le conclusioni alle quali si giunge differiscono non vorrà dire che un metodo è migliore dell'altro; scoprire invece le ragioni che portano alla discrepanza sarà utile ad essere sicuri di non fare assunzioni inappropriate o a comprendere che magari fondamentali informazioni sono state tralasciate a favore di altre meno importanti nell'uno o nell'altro approccio. In sintesi quello che caratterizza l'approccio classico è che la probabilità è vista come un valore oggettivo, i parametri sono fissi e incogniti e le procedure inferenziali sono basate su un campionamento ripetuto nelle stesse condizioni. Quello che invece sottostà all'approccio bayesiano è la probabilità intesa come valore soggettivo, parametri come variabili casuali e la procedura inferenziale è basata sulla distribuzione di probabilità dei

parametri osservando il campione estratto e avendo a disposizione ulteriori informazioni.

La struttura che si impiegherà di seguito riguarderà in primo luogo la presentazione dei due approcci con le loro specifiche caratteristiche, prima quello frequentista e poi quello bayesiano. In seguito le specificità dei due approcci verranno esplicate in applicazione ad un caso pratico di misure ripetute nell'ambito della modulazione del dolore. Infine si discuteranno i risultati applicativi dei due approcci utilizzando il software SAS 9.3.

Materiali e metodi

Parlare di statistica frequentista e di statistica bayesiana significa affrontare un determinato problema attraverso due approcci che intendono la teoria della probabilità e quindi l'inferenza statistica in modo differente.

1. L'approccio frequentista

Quando si sceglie di seguire l'approccio classico, il primo passo da compiere è quello di capire il concetto di probabilità e di comprendere come avviene la sua misurazione.

La probabilità è infatti un concetto primitivo, cioè innato nell'uomo, che obbedisce a criteri logici traducibili in assiomi tramite i quali poi si possono dimostrare teoremi. Inoltre la probabilità è misurabile in quanto al concetto primitivo viene di fatto associata una valutazione numerica.

Negli anni si sono succedute diverse definizioni di probabilità nessuna delle quali completamente soddisfacente. Qui si presenta quella definita 'frequentista' che concepisce la probabilità come una caratteristica intrinseca di un dato evento rappresentandola da un punto di vista puramente oggettivo.

La probabilità di un evento risulta quindi, secondo la definizione tradizionale, '*il rapporto fra il numero di esperimenti in cui esso si è verificato e il numero totale di esperimenti eseguiti nelle stesse condizioni, essendo tale numero opportunamente grande'*. In termini matematici si scriverà:

$$P(E) = \lim_{n\to\infty} \frac{n_E}{n}$$

dove E è l'evento di interesse n_E è il numero di esperimenti in cui si è verificato l'evento E ed n è il numero totale degli esperimenti.

Tale definizione e tutto l'approccio frequentista che ne segue, poggiano su alcuni postulati che ne permettono quindi tutta la formulazione matematica per associarne una valutazione numerica.

Siano E_i, i=1,2… eventi di Ω. La probabilità di un evento E_i è una funzione di insieme a valori reali, che si indicherà con P(E_i) che soddisfa le seguenti leggi:

1) P(E_i)≥0, $\forall E_i \subset \Omega$
2) P(Ω)=1
3) $E_i \cap E_j = \emptyset$, $\forall$ i≠j ⇒ P($\bigcup_{i=1}^{\infty} E_i$)=$\sum_{i=1}^{\infty} P(E_i)$.

Quello che si intende è che la probabilità è una funzione di insieme che assegna ad ogni evento un numero reale non negativo, l'evento certo ha probabilità 1 e la probabilità dell'unione di una infinità numerabile di eventi a due a due incompatibili è la somma delle singole probabilità.

In questo modo quindi si elimina l'arbitrarietà nella scelta dello sperimentatore in quanto ad ogni evento è associata una ed una sola funzione di probabilità, costante per ogni decisore. Il problema che si pone è ora assegnare alla probabilità un valore numerico in base alle caratteristiche dell'insieme Ω senza contraddire nessuno dei postulati.

Se Ω è costituito da un numero finito di eventi $E_1, E_2, \dots E_n$ tali che questi siano equiprobabili, quindi la P(E_i) è costante al variare di i=1,2…n applicando i postulati visti sopra si ha che 1 = P(Ω) = P($E_1 \cup E_2 \cup \dots \cup E_n$) = P($E_1$) + P($E_2$) +…+ P($E_n$) = P($E_i$) + P($E_i$) +…+ P($E_i$) = nP($E_i$). Da questo si deduce che 1 = nP(E_i) cioè P(E_i) = $\frac{1}{n} \forall\, i = 1,2, \dots, n$.

In particolare se A ⊂ Ω è composto da m eventi elementari, allora si ottiene che P(A) = P($E_1 \cup E_2 \cup \dots \cup E_m$) = P($E_1$) + P($E_2$) +…+ P($E_m$) = $\frac{1}{n} + \frac{1}{n} + \dots + \frac{1}{n} = \frac{m}{n}$ cioè se $E_1, E_2, \dots, E_m$ formano una partizione di A ⊂ Ω allora la probabilità di A ⊂ Ω è il rapporto tra il numero degli eventi elementari di cui A si compone (casi favorevoli ad A) e il numero totale degli eventi elementari di Ω (casi ugualmente possibili).

Se Ω è costituito da un numero infinito di eventi rappresentato quindi per semplicità da $\mathbb{R} = [-\infty < x < +\infty]$, si definirà l'evento E come qualsiasi evento del tipo E_x = (si verifica un numero reale X ≤ x) per cui Ω è l'insieme di tutti gli intervalli aperti a sinistra e chiusi a destra del tipo $(-\infty, x]$. Allora la P(E_x) = P (X ≤ x) =F(x). Attraverso tale funzione, detta funzione di ripartizione, è possibile assegnare la probabilità a qualunque evento $E_x \subset \Omega$ con x ∈ ℝ. La funzione di ripartizione è una funzione che ha per dominio ℝ e per codominio l'intervallo [0,1] essendo una probabilità. Inoltre è funzione non decrescente tra 0 e 1 ed è continua a destra.

Una volta definita la probabilità di evento è necessario introdurre sullo spazio degli eventi una relazione algebrica e una d'ordine mediante la definizione di variabile casuale (v.c) in quanto si è interessati ad una funzione dell'evento

piuttosto che all'evento stesso. Questa non è altro che una funzione misurabile a valori reali definita sullo spazio Ω e quindi per ogni evento E ⊂ Ω la variabile casuale X assume un valore reale x creando di fatto una corrispondenza tra l'insieme Ω degli eventi e l'insieme ℝ dei numeri reali. Allo stesso modo di quanto fatto in precedenza anche qui si possono distinguere variabili casuali discrete e variabili casuali continue. Nel primo caso i valori che la variabile casuale può assumere sono in numero finito o al più numerabile, nel secondo caso, invece, i valori che può assumere sono in numero infinito non numerabile. Le variabili casuali discrete sono caratterizzate da P(X=x_1) = p_1, P(X=x_2) = p_2,…, P(X=x_i) = p_i e sono ben definite se e solo se vale che $p_i \geq 0 \; \forall i = 1,2, \ldots$ e $\sum_{i=1}^{\infty} p_i = 1$. Le variabili casuali continue sono correttamente note se, per ogni x reale, sono conosciute la F(x) o la f(x) legate dalle seguenti relazioni F(x) = P($-\infty <$X ≤ x) = P (X ≤ x) = $\int_{-\infty}^{x} f(w)dw$ o f(x) = $\frac{d}{dx} F(x)$. La F(x) è detta funzione di ripartizione mentre la f(x) è la funzione di densità e la variabile casuale X è ben definita se e solo se f(x) ≥ 0 e $\int_{-\infty}^{+\infty} f(x)dx = 1$.

In generale quindi si può dire che:

$$\text{F(x)} = \text{P}(X \leq x) = \begin{cases} \int_{-\infty}^{x} f(w)dw & \text{se } X \text{ è una variabile casuale continua} \\ \sum_{x_j \leq x} p_j & \text{se } X \text{ è una variabile casuale discreta} \end{cases}$$

Da questo discendono le seguenti proprietà:

1) F(x) è non decrescente cioè $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$
2) $\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to +\infty} F(x) = 1$

3) F(x) è continua da destra, cioè $\lim_{x \to x_0^+} F(x) = F(x_0)$

1.1 Inferenza classica

Il problema che l'inferenza statistica si pone è quello di capire, sulla base di eventi osservati, quale possa essere la popolazione che li ha generati. Quest'ultima sarà una variabile casuale X definita completamente dalla sua funzione di ripartizione F(x;$\underline{\vartheta}$) e si indicherà X~ f(x; $\underline{\vartheta}$) dove $\underline{\vartheta}$ è un vettore di parametri m ≥ 1, fissi e incogniti. L'insieme dei valori che i parametri $\underline{\vartheta}$ possono assumere sarà definito spazio parametrico e indicato con $\Omega(\underline{\vartheta})$. Dalla popolazione X viene quindi estratto casualmente un sottoinsieme di n unità statistiche $(X_1, X_2, \dots, X_n)$ la cui realizzazione numerica $(x_1, x_2, \dots, x_n)$ è detta campione casuale. Le $X_1, X_2, \dots, X_n$ sono anch'esse variabili casuali e non sono altro che repliche della variabile casuale X e ne hanno la stessa distribuzione.

L'impostazione classica del problema dell'inferenza si avvale del principio che tutte le procedure statistiche vengano svolte sotto la condizione di un ipotetico e ripetuto campionamento dei dati. Il campione quindi, su cui si baseranno tutte le analisi, sarà considerato come ottenuto alla stregua di tutti i possibili campioni che si sarebbero potuti ricavare ripetendo un gran numero di volte e nelle stesse condizioni l'esperimento.

Le tecniche utilizzate nell'inferenza classica per riportare l'informazione del campione alla popolazione sono sostanzialmente tre e sono tra loro interconnesse

- Teoria della stima con la quale si cerca di determinare un valore numerico per il vettore di parametri $\underline{\vartheta}$ che caratterizza la popolazione sulla base delle informazioni contenute nel campione.
- Teoria della verifica delle ipotesi statistiche con la quale si cerca di controllare quale tra due affermazioni complementari circa il vettore di parametri possa essere più verosimile sulla base delle informazioni campionarie.
- Teoria degli intervalli di confidenza con la quale si cerca di determinare una regione di valori reali in cui riporre una prefissata fiducia per il vettore di parametri.

1.1.1 Teoria della stima

Tutte le informazioni statisticamente rilevanti e ricavabili dal campione riguardanti la popolazione di interesse sono contenute all'interno di quella che è detta la funzione di verosimiglianza.

Se prima di estrarre un campione si sa che la v.c. X~ f(x; $\underline{\vartheta}$), allora si può dire che con probabilità f(x; $\underline{\vartheta}$) si verificherà il vettore numerico $\underline{x} = (x_1, x_2, \dots, x_n)'$. Se invece il campione $\underline{x} = (x_1, x_2, \dots, x_n)'$ è stato già estratto e $\underline{\vartheta}$ è sconosciuto allora la distribuzione congiunta $\mathcal{L}(\underline{\vartheta}, x)$ = f($x_1, x_2, \dots, x_n$; $\underline{\vartheta}$) = f(x_1; $\underline{\vartheta}$) f(x_2; $\underline{\vartheta}$) … f(x_n; $\underline{\vartheta}$) = $\prod_{i=1}^{n}$ f(x_i; $\underline{\vartheta}$) = $\mathcal{L}(\underline{\vartheta})$ è funzione solo del vettore di parametri fissi e incogniti $\underline{\vartheta} \in \Omega(\vartheta)$. Questa è la funzione di verosimiglianza che esprime quanto è verosimile che il

campione estratto provenga da una popolazione di parametri $\underline{\vartheta}$. Essa non è una distribuzione di probabilità e risulta essere una quantità sempre non negativa. Per tale ragione spesso si è più soliti utilizzare la log-verosimiglianza data da $\ln \mathcal{L}(\underline{\vartheta}, x) = \ln[\mathrm{f}(x_1;\ \underline{\vartheta})\ \mathrm{f}(x_2;\ \underline{\vartheta})\ \ldots\ \mathrm{f}(x_n;\ \underline{\vartheta})] = \ln[\prod_{i=1}^{n} \mathrm{f}(x_i;\ \underline{\vartheta})] = \sum_{i=1}^{n} \ln \mathrm{f}(x_i;\ \underline{\vartheta}) = \ell(\underline{\vartheta})$ con l'eventuale convenzione che $\ell(\underline{\vartheta}) = -\infty$ se $\mathcal{L}(\underline{\vartheta}) = 0$.

L'obiettivo principale della teoria della stima è determinare i valori più plausibili per un determinato $\underline{\vartheta}$ incognito e fisso che caratterizza una popolazione X. Innanzitutto si supponga la v.c. X con funzione di densità $\mathrm{f}(x;\ \underline{\vartheta})$ chiaramente nota a meno del vettore di parametri $\underline{\vartheta}, \underline{\vartheta} \in \Omega(\vartheta)$. Da questa popolazione si estrae un campione X = $(X_1, X_2, \ldots, X_n)$ la cui determinazione numerica è $(x_1, x_2, \ldots, x_n)$. Si definisce stimatore qualunque funzione T() nota e a valori reali definita su $(X_1, X_2, \ldots, X_n)$ e si indicherà $T_n = T(X_1, X_2, \ldots, X_n)$ mentre si definisce stima qualunque funzione T() nota e a valori reali definita su $(x_1, x_2, \ldots, x_n)$ cioè $t_n = t(x_1, x_2, \ldots, x_n)$. Per come sono stati definiti quindi lo stimatore è una variabile casuale mentre la stima è un valore reale.

Da quanto detto finora si sa quindi che la funzione di verosimiglianza ci permette di racchiudere tutte le informazioni riguardo alla popolazione e lo stimatore è una sintesi del campione che si è estratto. Allora quello che si vuole è cercare di non disperdere l'informazione contenuta nella funzione di verosimiglianza cercando di ottenere determinate caratteristiche dagli stimatori che si costruiscono. Si supponga X=$(X_1, X_2, \ldots, X_n)$ un campione generato dalla v.c. X~ f(x; ϑ) con ϑ parametro di interesse oggetto di stima

$\vartheta \in \Omega(\vartheta)$. La prima caratteristica che si considera è la sufficienza, in particolare si dirà che uno stimatore $T_n = T(X_1, X_2, \dots, X_n)$ è sufficiente per ϑ se questo assume lo stesso valore in due punti distinti dello spazio campionario solo se questi due punti hanno verosimiglianze equivalenti cioè $T(y_1, y_2, \dots, y_n) = T(z_1, z_2, \dots, z_n) \Rightarrow \mathcal{L}(\vartheta, y) \propto \mathcal{L}(\vartheta, z)$.

Se $T(X_1, X_2, \dots, X_n)$ è una statistica sufficiente, allora $\mathcal{L}(\vartheta)$ dipende da $(X_1, X_2, \dots, X_n)$ solo tramite $T(X_1, X_2, \dots, X_n)$ cioè esiste una funzione g tale che $\mathcal{L}(\vartheta) \propto g(T(x_1, x_2, \dots, x_n); \vartheta)$. Siccome $\mathcal{L}(\vartheta) \propto$ f(x; ϑ) ne segue che

$$\mathrm{f}(\mathrm{x};\vartheta) \Big/ g(T(x_1, x_2, \dots, x_n); \vartheta)$$

non dipende da ϑ. Chiamando tale rapporto h$(x_1, x_2, \dots, x_n)$ si ha che se $T(X_1, X_2, \dots, X_n)$ è una statistica sufficiente allora vale la relazione f(x; ϑ) = h$(x_1, x_2, \dots, x_n)$ $g(T(x_1, x_2, \dots, x_n); \vartheta)$ per delle opportune funzioni g e h non negative. Se vale tutto questo allora $\mathcal{L}(\vartheta)$ è funzione di $(x_1, x_2, \dots, x_n)$ solo tramite $T(X_1, X_2, \dots, X_n)$ e quindi $T(X_1, X_2, \dots, X_n)$ è sufficiente, perciò tutte le informazioni contenute nella funzione di verosimiglianza sono trasferite sullo stimatore. Si può quindi enunciare il teorema di fattorizzazione di Neyman che si esprime nel seguente modo: la statistica $T(X_1, X_2, \dots, X_n)$ è sufficiente per ϑ se e solo se f(x; ϑ) = h$(x_1, x_2, \dots, x_n)$ $g(T(x_1, x_2, \dots, x_n); \vartheta)$.

Ora l'obiettivo è quello di ottenere la massima sintesi dei dati conservando tutta 'informazione campionaria e quindi trovare lo stimatore di sufficienza minimale. Formalmente uno stimatore $T_n = T(X_1, X_2, \dots, X_n)$ è sufficiente

minimale per ϑ se per qualsiasi altro stimatore sufficiente T_n^*, la statistica T_n è una funzione di T_n^*.

Altre proprietà auspicabili per gli stimatori sono:

- Correttezza se $\mathrm{E}(T_n)=\vartheta$, indica la bontà dello stimatore T_n richiedendo che la distribuzione di T_n sia centrata su ϑ ;
- Efficienza ; tra tutti gli stimatori corretti per ϑ si preferisce quello con varianza $[\mathrm{E}(T_n) - \vartheta]^2$ più bassa.
- Consistenza in media quadratica se $\lim_{n\to\infty}[\mathrm{E}(T_n) - \vartheta]^2=0$; questa garantisce che asintoticamente lo stimatore risulti non distorto e quindi informa su quanto velocemente T_n converge a ϑ.
- Consistenza in probabilità se $\lim_{n\to\infty} \mathrm{P}[|(T_n) - \vartheta| < \varepsilon]=1$; questa assicura che all'aumentare della dimensione del campione aumenta la probabilità che T_n stimi ϑ con un errore ε infinitamente piccolo.

Una volta indicate le proprietà che uno stimatore dovrebbe avere è importante capire come si costruisce uno stimatore nell'inferenza classica. Due sono i metodi principalmente adottati: metodo dei minimi quadrati e metodo della massima verosimiglianza anche se ne esistono molti altri.

Nel metodo dei minimi quadrati ordinari si suppone che ciascuna delle v.c. che compongono il campione $(X_1, X_2, \dots, X_n)$ possano esprimersi nella forma $X_i = g_i(\vartheta) + \varepsilon_i$ dove $g_i(\vartheta)$ è una quantità deterministica a valori reali del parametro ϑ e ε_i sono v.c. che esprimono errori accidentali e hanno media zero, varianza uno e sono tra loro incorrelati. Si supponga che sul campione si osservino i valori $(x_1, x_2, \dots, x_n)$ e quindi risulti che $x_i = g_i(\vartheta) + e_i$ con

e_i realizzazioni delle v.c. ε_i. Allora il parametro ϑ può essere stimato in modo tale che lo stimatore $G(\vartheta) = \sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}[x_i - g_i(\vartheta)]^2 = min.$

Normalmente la soluzione è quella che rende nulla la $G'(\vartheta)$ rispetto a ϑ e verificando che $G''(\vartheta)$ sempre rispetto a ϑ sia positiva.

Se la funzione g() è lineare nel parametro ϑ allora lo stimatore dei minimi quadrati è BLUE cioè il migliore stimatore lineare non distorto ed efficiente. Inoltre è anche consistente e asintoticamente normale.

Nel metodo della massima verosimiglianza sia $(X_1, X_2, \dots, X_n)$ un campione generato da X~ f(x; ϑ) con $\vartheta \in \Omega(\vartheta)$. Si propone come stima per il parametro il valore t=T$(x_1, x_2, \dots, x_n)$ che massimizza la funzione di verosimiglianza $\mathcal{L}(\vartheta, \underline{x})$ per il quale risulti quindi che $\mathcal{L}(t, \underline{x}) \geq \mathcal{L}(t^*, \underline{x})$ per qualunque $t \neq t^*$. La logica sottostante è quella che si cerca di preferire il valore del parametro ϑ che corrisponde alla massima probabilità di generare i dati osservati.

Allora il parametro ϑ può essere stimato con il valore che rende nulla $\mathcal{L}'(\vartheta, \underline{x})$ rispetto a ϑ e verificando che $\mathcal{L}''(\vartheta, \underline{x}) < 0$. Sotto condizioni di regolarità, gli stimatori di massima verosimiglianza godono di tutte le proprietà auspicabili per uno stimatore e inoltre godono anche della proprietà di invarianza cioè se $T_n = T(X_1, X_2, \dots, X_n)$ è lo stimatore di massima verosimiglianza per ϑ e se $\tau(\vartheta)$ è una funzione biunivoca di ϑ allora lo stimatore di massima verosimiglianza di $\tau(\vartheta)$ è $\tau(T_n)$.

1.1.2 Teoria della verifica delle ipotesi statistiche

Si è interessati qui a capire come prendere decisioni riguardo al parametro ϑ sulla popolazione servendosi solo di dati campionari. Questo avviene facendo ricorso a un test, regola sullo spazio campionario, utilizzando la quale, in funzione del campione osservato, si sceglie se rifiutare o meno una ipotesi statistica H_0 riguardante la popolazione.

Tale ipotesi H_0, detta ipotesi nulla, è l'insieme di un particolare sottogruppo di valori $w_0 \subset \Omega(\vartheta)$ che è lo spazio parametrico che definisce completamente la v.c. X~ f(x; ϑ). Se H_0 non è vera allora risulterà vera H_1, l'ipotesi alternativa e quindi $\vartheta \notin w_0$. Le ipotesi H_0 e H_1 sono sempre esaustive e disgiunte e se la decisione è quella di rifiutare H_0 la conseguenza è univoca in quanto si attesta la non validità di una affermazione precisa al contrario la decisione di non rifiutare H_0 non implica che H_0 sia vera ma solo che dal campione estratto non si evincono elementi tali da rifiutare H_0.

La verifica di quale ipotesi sia da rifiutare avviene mediante la formulazione di una statistica-test, ossia di uno stimatore T_n che fa corrispondere ad ogni campione $(x_1, x_2, \dots, x_n)$ dell'universo campionario $\mathbb{R}^n$, un valore del parametro di interesse classificato secondo due diverse possibilità: coerente o non coerente con H_0. La statistica test si utilizza quindi per misurare la differenza tra il valore del parametro stimato dal campione e il valore che si osserverebbe se l'ipotesi nulla fosse vera. Ogni test da quindi luogo alla partizione dello spazio campionario in due regioni complementari: A_0 di campioni $(x_1, x_2, \dots, x_n)$ in cui i valori della statistica test sono coerenti con

H_0 e C_0 di campioni $(x_1, x_2, \dots, x_n)$ in cui i valori della statistica-test non sono coerenti con H_0. La regione A_0 è definita regione di accettazione o di non rifiuto mentre C_0 è la regione critica o regione di rifiuto e risulta quindi che $A_0 \cup C_0 = \mathbb{R}^n$. Nell'effettuare un test però si possono commettere due tipi di errori nel giudicare le ipotesi:

- ✓ Si potrebbe rifiutare H_0 quando in realtà questa è vera, si parla di errore di prima specie
- ✓ Si potrebbe accettare H_0 quando in realtà questa è falsa, si parla di errore di seconda specie.

Ovviamente non è possibile sapere se si sta compiendo uno di questi errori perché non si conosce se H_0 sia vera o falsa. Per cui le probabilità associate agli errori di cui sopra si indicano nel seguente modo:

- ✓ $\alpha = P(rifiutare\ H_0 | H_0 \text{ è vera}) = P[(x_1, x_2, \dots, x_n) \in C_0 | H_0 : \vartheta \in w_0]$
- ✓ $\beta = P(non\ rifiutare\ H_0 | H_0 \text{ è falsa}) = P[(x_1, x_2, \dots, x_n) \notin C_0 | H_0 : \vartheta \notin w_0]$

α è la probabilità dell'errore di primo tipo conosciuta come livello di significatività del test o ampiezza della regione critica, β è la probabilità dell'errore di secondo tipo. Inoltre si introduce $\gamma = 1 - \beta$ che è la probabilità di rifiutare correttamente H_0 che è nota come potenza del test. Come si può facilmente notare tutte le probabilità sopra descritte dipendono dalla regione di rifiuto e può essere ragionevole richiedere che la regione critica sia tale che sia la probabilità α che β siano entrambe sufficientemente

piccole, ma non è possibile farle tendere contemporaneamente a 0 per una stessa dimensione campionaria. Infatti tra queste due quantità esiste relazione esprimibile tramite una funzione decrescente, per cui è opportuno cercare un compromesso in base all'importanza che riveste ciascun errore all'interno del particolare studio che si sta conducendo. In genere si ritiene più grave commettere un errore di primo tipo rispetto ad uno del secondo tipo.

Per la costruzione di un test appropriato bisogna far riferimento innanzitutto a quella che è definita regione critica ottimale di ampiezza α, RCO(α), che risulta essere una regione critica C_0 per H_0 tale che la $P[(x_1, x_2, \dots, x_n) \in C_0 \,|\mathrm{H}_0] = \alpha$ e che per qualunque altra regione critica C_0' di uguale ampiezza α risulti che $\gamma(C_0) = P[(x_1, x_2, \dots, x_n) \in C_0 \,|\mathrm{H}_1] > P[(x_1, x_2, \dots, x_n) \in C_0' |\mathrm{H}_1] = \gamma(C_0')$. Quindi è ottimale per un prefissato livello α, la regione critica che tra tutte quelle di pari ampiezza α possiede la potenza più elevata.

Per costruire tale regione bisogna, nel caso in cui le ipotesi che si testano sono semplici o unidirezionali, ricorrere al Lemma di Neyman-Pearson che connette la teoria del test a quella della stima utilizzando il concetto di funzione di verosimiglianza. Sia $\underline{X} = (X_1, X_2, \dots, X_n)$ un campione di v.c. generato da $\mathrm{X} \sim \mathrm{f}(\mathrm{x}; \vartheta)$ dove $\vartheta \in \Omega(\vartheta)$ e si voglia verificare $\mathrm{H}_0: \vartheta = \vartheta_0$ contro $\mathrm{H}_1: \vartheta = \vartheta_1$. Se $\mathcal{L}(\vartheta, \underline{X})$ è la funzione di verosimiglianza di $\underline{X}$ allora la RCO(α) di H_0 contro H_1 è quella regione C_0 dello spazio campionario $\mathbb{R}^n$ tale che soddisfa:

- $\mathcal{L}(\vartheta_1, \underline{X}) / \mathcal{L}(\vartheta_0, \underline{X}) \geq c$
- $P[\underline{X} \in C_0 \,|\mathrm{H}_0] = \alpha$

Il Lemma permette di costruire una delle possibili RCO(α) la cui forma è determinata dal primo vincolo e l'ampiezza dal secondo vincolo. Esso favorisce la scelta dell'ipotesi che risulta più plausibile in termini di verosimiglianza determinando la costante c in modo che il rischio di commettere un errore del primo tipo sia fissato ad α.

Se invece le ipotesi testate non rientrano in quelle viste in precedenza, per costruire una RCO(α) si ricorre ad una generalizzazione data dal test del rapporto di verosimiglianza.

Sia $\underline{X} = (X_1, X_2, \dots, X_n)$ un campione di v.c. generato da X~ f(x; ϑ) dove $\vartheta \in \Omega(\vartheta)$ e si voglia verificare $H_0: \vartheta \in w_0$ contro $H_1: \vartheta \notin w_0$. Si definisce rapporto di verosimiglianza $\lambda(\underline{X})$ il seguente rapporto $\max_{\vartheta \in w_0} \mathcal{L}(\vartheta, \underline{X}) / \max_{\vartheta \in \Omega(\vartheta)} \mathcal{L}(\vartheta, \underline{X})$ e si costruisce una regione critica di ampiezza α in modo tale che sia $P[\lambda(\underline{X}) \leq c_\alpha | H_0] = \alpha$. La regione critica così definita $C_0 = [\underline{X}: \lambda(\underline{X}) \leq c_\alpha]$ è una regione critica di ampiezza α cotruita con il metodo del test di massima verosimiglianza.

1.1.3 Teoria degli intervalli di confidenza

Quando si stima un parametro ϑ, individuare un singolo valore non basta, è sempre necessario associare a questo un insieme di valori plausibili per il parametro di interesse in modo da trovare un grado di affidabilità riguardo la sua collocazione più probabile. Nella stima puntuale, infatti, non si ha alcuna informazione circa l'errore che si commette nello stimare il parametro

incognito della popolazione, questo perché non si tiene conto della distribuzione dello stimatore nello spazio campionario.

Ora sia $\underline{X} = (X_1, X_2, \dots, X_n)$ un campione di v.c. generato da X~ f(x; ϑ), si definisce intervallo casuale per il parametro ϑ con coefficiente di confidenza $1 - \alpha$ qualsiasi sottoinsieme $\mathcal{S}(\underline{X})$ di $\mathbb{R}^n$ tale che P($\vartheta \in \mathcal{S}(\underline{X})$) = $1 - \alpha$. Quando alla regione $\vartheta \in \mathcal{S}(\underline{X})$ si sostituisce il campione osservato, si ottiene un intervallo di confidenza per ϑ con coefficiente di confidenza $1 - \alpha$ e lo si indicherà $IC_\alpha(\vartheta)$. Più in generale per un vettore di parametri si parlerà di regione di confidenza. L'idea che sta alla base quindi è trovare per il parametro ϑ, che caratterizza una v.c X~ f(x; ϑ) e per il quale sia possibile costruire uno stimatore con una certa distribuzione di probabilità, i due estremi dell'intervallo casuale L($\underline{X}$) e U($\underline{X}$) con probabilità $1 - \alpha$, che siano funzione del campione di variabili casuali $\underline{X}$ estratto dalla v.c X. Una volta osservato il campione $\underline{x} = (x_1, x_2, \dots, x_n)$ si troverà l'intervallo di confidenza per ϑ $\left[L(\underline{x}), U(\underline{x})\right]$ con coefficiente di confidenza $1 - \alpha$. Il significato che si dà è che se la procedura si ripete per tutti i possibili campioni casuali che è possibile estrarre, quindi per un numero elevato di volte, si otterrà una successione di intervalli di confidenza per cui la frazione di quelli che contengono il valore vero del parametro ϑ tende a $1 - \alpha$. Non si può dire con certezza che uno specifico intervallo contiene il valore vero del parametro incognito.

Nella scelta di un intervallo di confidenza occorre ricercare un compromesso tra il coefficiente di confidenza, che si vorrebbe elevato, e l'ampiezza

dell'intervallo che si vorrebbe invece piccola. Un modo di costruire l'intervallo di confidenza è quello della variabile casuale pivot.

Dato $\underline{X} = (X_1, X_2, \dots, X_n)$ campione casuale estratto da X~ f(x; ϑ), si definisce quantità pivotale una v.c $v(\underline{X}, \vartheta)$ che è funzione del campione casuale e del parametro ϑ, la cui distribuzione è nota e non dipende dal parametro ϑ. La quantità pivotale perciò è una v.c. in quanto dipendente dai dati campionari ma non è una statistica in quanto dipende da ϑ incognito. Se questa v.c. esiste per il parametro si ha:

$$1 - \alpha = P\big(v_L \leq v\big(\underline{X}, \vartheta\big) \leq v_U\big) = P(v^{-1}(v_L, \underline{X}) \leq \vartheta \leq v^{-1}(v_U, \underline{X}))$$

per cui si ha che a livello di confidenza $1 - \alpha$ l'intervallo casuale per il parametro ϑ sarà dato da $L\big(\underline{X}\big) = v^{-1}(v_L, \underline{X})$; $U\big(\underline{X}\big) = v^{-1}(v_U, \underline{X})$ mentre l'intervallo di confidenza relativo sarà $IC_\alpha(\vartheta) = \big[L\big(\underline{x}\big) = v^{-1}\big(v_L, \underline{x}\big);\ U(x) = v^{-1}\big(v_U, \underline{x}\big)\big]$.

2. L'approccio bayesiano

Il contesto bayesiano ha le sue fondamenta in un concetto di probabilità che è diverso da quello visto in precedenza. La probabilità di un evento assume una concezione del tutto soggettivista, in quanto viene ad esprimere il *grado di fiducia* che lo sperimentatore ripone nel verificarsi di un dato evento e dipende quindi dallo stato di conoscenza, o di ignoranza di tale evento, che fa parte dell'esperienza acquisita in precedenza da ciascuno.

La teoria bayesiana, quindi, fa ricorso ad una concezione non legata strettamente e unicamente all'evento ma anche e soprattutto al soggetto assegnante la funzione di probabilità. Quest'ultima, non è caratteristica dell'evento ma si viene a trovare tra individuo e mondo esterno, seguendo criteri razionali nell'assegnazione, i quali però non imporranno che la probabilità dell'evento sia uguale per ogni decisore ma potrà invece variare in base a tutto il bagaglio culturale acquisito e alle informazioni a disposizione. (J.Bernoulli, XVIII secolo). Tutto quanto detto finora fa si che anche in questo caso la probabilità così formulata debba comunque rispettare gli assiomi del sistema probabilistico. Quindi, in particolare, supponendo E_i, i=1,2… eventi di Ω deve risultare:

1) $0 \leq P(E_i) \leq 1, \forall E_i \subset \Omega$
2) P(Ω)=1
3) $E_i \cap E_j = \emptyset, \forall\, i \neq j \Rightarrow P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.
4) $P(\bar{E}_i) = 1 - P(E_i),\ \forall E_i \subset \Omega$, supponendo $\bar{E}_i \subset \Omega\ \forall i \in N$ insieme dei numeri naturali.

2.1 Il teorema di Bayes e l'inferenza bayesiana

La differenza sostanziale fra approccio soggettivo ed oggettivo riguarda l'utilizzo dell'evidenza empirica a scopo inferenziale ed il trattamento dei parametri da stimare nel modello statistico. Nell'inferenza bayesiana, infatti, la stima dei parametri avviene basandosi anche sulla eventuale disponibilità di una distribuzione di probabilità *a priori*, cioè di informazioni sul fenomeno che si

analizza disponibili prima rispetto ai dati del campione e che dipendono dall'esperienza dello sperimentatore. I parametri quindi, non sono più fissi e incogniti ma vengono considerati come variabili casuali e i dati provenienti dal campione risulteranno delle costanti.

Mentre l'approccio frequentista basa l'inferenza sulla ripetibilità degli eventi analizzati, l'approccio bayesiano fa principalmente ricorso al teorema di Bayes per ricavare nuove informazioni dai dati. Quest'ultimo si esplica, in maniera generale, nel seguente modo:

Sia $(E_i)_{i\geq 1}$ una partizione dell'evento certo Ω tale che

1) $\bigcup_{i=1}^{\infty} E_i = \Omega$
2) $E_i \cap E_j = \emptyset, i \neq j$
3) $P(E_i) > 0, i = 1 \dots \infty$

Sia H ⊂ Ω un evento tale che P(H) >0, allora per i=1…∞:

$$P(E_i/H) = \frac{P(H/E_i)P(E_i)}{\sum_{j=1}^{\infty} P(H/E_j)P(E_j)}$$

Dal punto di vista computazionale, si supponga di avere una v.c. X~ f(x; ϑ) dove $\vartheta \in \Omega(\vartheta)$ è il parametro incognito. Dalla popolazione X viene quindi estratto casualmente un sottoinsieme di n unità statistiche X=$(X_1, X_2, \dots, X_n)$ la cui realizzazione numerica è x=$(x_1, x_2, \dots, x_n)$. Il teorema di Bayes si semplifica nella seguente formula:

$$P(\vartheta/x) = \frac{P(x/\vartheta)P(\vartheta)}{P(x)} = \frac{P(x/\vartheta)P(\vartheta)}{\int P(x/\vartheta)P(\vartheta)d\vartheta} \qquad (1)$$

dove si evidenziano le tre distribuzioni fondamentali nell'analisi di tipo bayesiano.

$P(\vartheta)$, *probabilità a priori*, è la distribuzione di probabilità del parametro di interesse nello studio che esprime quanto il ricercatore "scommetterebbe" sui possibili valori del parametro senza considerare i dati campionari dello studio. Tale distribuzione a priori riflette l'informazione che il ricercatore ha sul fenomeno in studio prima ancora di aver raccolto i dati campionari, e generalmente si basa sulle evidenze di letteratura o su informazioni già note allo sperimentatore in base alla sua esperienza precedente. Nell'analisi bayesiana quindi il parametro non è più un valore fisso ed incognito ma diventa una variabile casuale con una propria distribuzione di probabilità.

$P(x/\vartheta)$, *funzione di verosimiglianza*, è la distribuzione di probabilità che il ricercatore assegnerebbe ai dati campionari osservati in corrispondenza di ogni specifico valore che il parametro di interesse può assumere. Da un punto di vista bayesiano tale distribuzione rappresenta un altro set di "scommesse": il modello $P(x/\vartheta)$ quantifica quanto il ricercatore "scommetterebbe" sui dati osservati se conoscesse il parametro di interesse.

$P(\vartheta/x)$, *la distribuzione a posteriori*, è la distribuzione di probabilità del parametro di interesse nello studio che esprime quanto il ricercatore "scommetterebbe" sui possibili valori del parametro dopo aver esaminato i dati campionari dello studio. In altri termini, dopo aver osservato i dati in studio il ricercatore desidera aggiornare le proprie aspettative sul fenomeno di indagine,

ovvero desidera combinare le proprie conoscenze a priori con le nuove conoscenze derivate dall'osservazione dei dati.

Si può inoltre essere sicuri che se sia la distribuzione *a priori* che la funzione di verosimiglianza supportano un particolare valore del parametro ϑ allora anche la distribuzione *a posteriori* lo supporterà. Ma se un determinato valore del parametro ϑ non è supportato dalla distribuzione *a priori* o dalla funzione di verosimiglianza o da entrambe, allora il parametro ϑ non sarà supportato neanche dalla distribuzione *a posteriori*.

Nella formula (1) precedentemente esposta, $P(x)$, dipendendo esclusivamente dal campione ed essendo quindi una quantità fissa in ambito bayesiano, ha il solo scopo di assicurare che la $P(\vartheta/x)$ sia ancora una distribuzione.

Tutto ciò che riguarda l'inferenza si ottiene basandosi sulla distribuzione *a posteriori* in quanto essa contiene tutte le informazioni circa il parametro di interesse.

2.1.1 Teoria della stima in campo bayesiano

Per ciò che riguarda la stima puntuale del parametro ϑ, il fatto di avere a disposizione una distribuzione di probabilità *a posteriori* per il parametro di interesse fa si che si opti per un valore puntuale che sintetizzi tale distribuzione, per cui si è soliti scegliere tra moda, mediana o media a seconda del tipo di distribuzione con cui si ha a che fare e della sua forma.

La moda della distribuzione è quella maggiormente in uso in quanto è interpretata come il valore più credibile (probabile) nella situazione considerata. Questa è definita come:

$$\hat{\vartheta} = max_{\vartheta} P(\vartheta/x)$$

Ed è quindi il valore per il parametro ϑ che massimizza la distribuzione *a posteriori.*

La seconda misura che si può scegliere come stima puntuale è la media della distribuzione *a posteriori* data da

$$\bar{\vartheta} = \int \vartheta \, P(\vartheta/x) d\vartheta$$

e infine la terza misura di locazione che può essere considerata è la mediana *a posteriori* secondo la seguente formula

$$0.5 = \int_{\bar{\vartheta}_M} P(\vartheta/x) d\vartheta$$

Quale tra queste tre misure scegliere dipende comunque da questioni computazionali e dalla forma della distribuzione *a posteriori*.

La misura di variabilità per il parametro è invece la varianza *a posteriori* data da

$$\bar{\sigma}^2 = \int (\vartheta - \bar{\vartheta})^2 P(\vartheta/x) d\vartheta$$

La radice quadrata di quest'ultima è la deviazione standard *a posteriori* che fornisce l'incertezza *a posteriori* circa il parametro di interesse.

2.1.2 Teoria della verifica delle ipotesi in campo bayesiano

Si supponga di voler testare l'ipotesi nulla H_0 contro l'ipotesi alternativa H_1, ipotesi che devono essere definite in modo esplicito e completamente probabilizzate. In ambito bayesiano si può calcolare la distribuzione *a posteriori* che H_0 sia vera applicando il teorema di Bayes, per cui si ha:

$$P(H_0/x) = \frac{P(x/H_0)P(H_0)}{P(x/H_0)P(H_0) + P(x/H_1)P(H_1)}$$

Visto che H_1 è il complemento dell'ipotesi nulla, cioè $P(H_1) = 1 - P(H_0)$ e che $P(H_1|x) = 1 - P(H_0|x)$ si ottiene

$$\frac{P(H_0/x)}{1 - P(H_0|x)} = \frac{P(x/H_0)}{P(x/H_1)} * \frac{P(H_0)}{1 - P(H_0)}$$

Questa espressione afferma che l'odds *a priori* per l'ipotesi nulla H_0 è trasformato nell'odds *a posteriori* per H_0 tramite la moltiplicazione per il fattore $\frac{P(x/H_0)}{P(x/H_1)}$ che viene definito Fattore di Bayes, indicato come BF_{01}. Esso è il corrispettivo del test del rapporto di verosimiglianza dell'inferenza classica ed indica la veridicità di H_0 rispetto ad H_1 basandosi solo sui dati osservati e non sulle conoscenze soggettive. Il Fattore di Bayes è il rapporto tra le verosimiglianze sotto H_0 e H_1 e varia tra 0 e ∞. Esso è simmetrico in

quanto non predilige nessuna delle due ipotesi formulate e inoltre l'ipotesi nulla H_0 viene rifiutata se tale rapporto è inferiore ad una certa soglia critica fissata. Questa soglia può essere fissata nel seguente modo ma non è detto che sia quello più opportuno.

Se $BF_{01} > 1$ allora l'ipotesi nulla è maggiormente supportata dai dati campionari rispetto all'ipotesi alternativa.

Se $BF_{01} < 1$ allora l'ipotesi alternativa è maggiormente supportata dai dati campionari rispetto all'ipotesi nulla.

Se $BF_{01} = 1$ le due ipotesi sono supportate allo stesso modo dai dati campionari.

Jeffreys (1961) fornisce delle linee guida su scala logaritmica per l'interpretazione del fattore di Bayes, per esempio se $1 < logBF_{01} \leq 2$ allora c'è una forte evidenza contro l'ipotesi alternativa. Kass & Raftery (1995) invece preferiscono la trasformazione $2\ln BF_{01}$ in quanto in questo modo il Fattore di Bayes può essere letto sulla stessa base del rapporto di verosimiglianza dell'approccio frequentista. In questo caso una forte evidenza contro H_1 si ha se $6 \leq 2\ln BF_{01} \leq 10$. Queste soglie possono essere utili in molte situazioni ma in generale è il contesto nel quale si trova ad operare che guida la scelta.

2.1.3 Teoria degli intervalli di credibilità

Per ciò che riguarda gli intervalli di confidenza nel caso bayesiano si parla più propriamente di intervalli di credibilità intendendo associare al campione

osservato x=$(x_1, x_2, \dots, x_n)$ un intervallo che con una certa probabilità *a posteriori* conterrà il vero valore del parametro. L'interpretazione è molto più semplice in quanto il parametro, che è una variabile aleatoria, si troverà nell'intervallo con probabilità (1-α) e questo permetterà anche di capire direttamente quali siano i valori del parametro che hanno più probabilità. Infatti il 100(1- α)% dell'intervallo di credibilità contiene sicuramente il 100(1- α)% dei più plausibili valori del parametro incognito.

Formalmente [a,b] ⊂ $\Omega(\vartheta)$ è un 100(1- α)% intervallo di credibilità per ϑ, con α scelto di solito uguale al 0.05, se $P(a \leq \vartheta \leq b|x) = \int_a^b P(\vartheta|x)d\vartheta = 1 - \alpha$ con $1 - \alpha$ livello di confidenza bayesiano. La definizione data è molto generica e lascia libertà di scelta sul modo più appropriato per costruire tale intervallo. I più popolari e usati metodi di costruzione degli intervalli di credibilità di tipo bayesiano sono a) *a code uguali* e b) *a densità a posteriori più alta*.

L'intervallo di credibilità a code uguali si ottiene se l'intervallo di credibilità [a,b] è tale per cui $P(\vartheta \leq a/x) = \frac{\alpha}{2}$ e $P(\vartheta \geq b/x) = \frac{\alpha}{2}$. In questo caso alcuni valori del parametro ϑ inclusi nell'intervallo di credibilità hanno più bassa probabilità *a posteriori* rispetto a quelli fuori dell'intervallo.

L'intervallo di credibilità a densità *a posteriori* più alta si ha se per tutti $\vartheta_1 \in [a, b]$ e per tutti i $\vartheta_2 \notin [a, b]$, la $P(\vartheta_1/x) \geq P(\vartheta_2/x)$. Questo assicura quindi di avere all'interno dell'intervallo di credibilità tutti i valori di ϑ che sono *a posteriori* più plausibili. Graficamente questo intervallo può essere ricavato tracciando una linea orizzontale sulla rappresentazione della distribuzione a posteriori regolando l'altezza della linea in modo tale che

l'area sotto la curva sia pari a $1-\alpha$. Questo tipo di intervallo è il meno ampio che si possa determinare e inoltre se la distribuzione *a posteriori* è simmetrica unimodale i due intervalli sopra presentati corrispondono.

2.2 Scelta appropriata della distribuzione a priori

La specificazione della distribuzione *a priori* è chiaramente necessaria e cruciale nell'approccio Bayesiano e per capire questo basta fare l'esempio dell'assurdo che per tutti i parametri incogniti ϑ per cui la P(ϑ)=0 la loro distribuzione *a posteriori* $P(\vartheta|x)$=0. Questo dimostra la necessaria cura che bisognerebbe avere nello scegliere la distribuzione *a priori*.

Riguardo a questo aspetto si potrebbe essere in una delle seguenti condizioni riguardo al parametro incognito ϑ *a priori*:

- Nessuna informazione
- Conoscenza sostanziale

2.2.1 Nessuna conoscenza a priori per il parametro ϑ

Se non si dispone di informazioni utili riguardo al parametro incognito di interesse la distribuzione *a priori* deve comunque essere esplicitata e a tal proposito Jeffreys propone di utilizzare il principio di ragione insufficiente, il quale prevede che entro un insieme discreto di alternative in condizioni di ignoranza non c'è ragione di assegnare a qualcuna di queste, una probabilità diversa. Le distribuzioni *a priori* alle quali si ricorre in questo caso vengono

dette improprie e vengono scelte in base allo spazio parametrico. Più in particolare:

- ✓ Se $\Omega(\vartheta) = (-\infty, +\infty)$ allora si potrebbe scegliere una *a priori* non informativa del tipo $P(\vartheta) \propto 1$, distribuzione uniforme cosicché la distribuzione *a posteriori* sia proporzionale alla funzione di verosimiglianza $P(x/\vartheta)$
- ✓ Se $\Omega(\vartheta) = (0, +\infty)$ la scelta potrebbe ricadere su una *a priori* $P(\vartheta) \propto \frac{1}{\vartheta}$
- ✓ Se $\Omega(\vartheta) = (0,1)$ la *a priori* non informativa può essere scelta in diversi modi quali ad esempio $P(\vartheta) \propto \vartheta^{-1}(1-\vartheta)^{-1}$ proposta da Haldane che risulta ottimale per le sue proprietà di invarianza o $P(\vartheta) \propto \vartheta^{-1/2}(1-\vartheta)^{-1/2}$ proposta da Jeffreys che rispetto alla precedente ha meno peso negli estremi. Quello che quest'ultimo autore consiglia è comunque di dare probabilità discrete a certi valori del parametro ϑ e distribuzione uniforme altrove.

Grande cura dovrebbe essere comunque riposta nell'assicurare che la distribuzione *a posteriori* derivante da una *a priori* impropria sia appropriata (propria). E' difficile dare delle linee guida riguardo a questo infatti ad esempio nei modelli non lineari le distribuzioni *a priori* improprie non dovrebbero mai essere usate mentre nei modelli lineari generalizzati le distribuzioni improprie per i parametri di regressione spesso, ma non sempre, portano ad una distribuzione *a posteriori* propria.

2.2.2 Conoscenza sostanziale a priori per il parametro ϑ

Una conoscenza chiara riguardo al parametro di interesse ϑ porta ad esplicitare la distribuzione *a priori* in maniera forte tale da far in modo che la distribuzione *a posteriori* sia piuttosto diversa dalla funzione di verosimiglianza. In questo caso, se il fenomeno in studio lo permette, è possibile utilizzare una distribuzione *a priori* con proprietà matematiche buone quali sono le cosiddette coniugate. Una *a priori* coniugata è quella per cui $P(\vartheta/x)$ e $\mathrm{P}(\vartheta)$ appartengono alla stessa famiglia. Quindi una distribuzione *a priori* $\mathrm{P}(\vartheta)$ appartenente ad una famiglia di distribuzioni $\mathbb{Q} = \{\mathrm{P_k}(\vartheta): k \in \mathcal{K}\}$ è detta coniugata se la distribuzione *a posteriori* $P(\vartheta/x)$ appartiene ancora alla stessa famiglia $\mathbb{Q}$. In questo caso se si suppone che k_0 racchiude l'informazione *a priori* sul parametro ϑ, il campione osservato lo trasforma in k_1 che invece sintetizza l'informazione *a posteriori* sul parametro. I vantaggi sono quindi la possibilità di scomporre l'informazione proveniente *a priori* con quella proveniente dai dati campionari in quanto vale che $k_1 = k_0 + (k_1 - k_0)$. Per cui k_0 è il contributo della *a priori* e $k_1 - k_0$ è la variazione indotta dai dati campionari.

Un ulteriore vantaggio è che la *a posteriori* si ottiene con facili passaggi computazionali in quanto questa fa parte della stessa famiglia di distribuzioni della *a priori* a meno del parametro k. E' da sottolineare però che le coniugate esistono solo in pochi casi e uno di questi è quello della famiglia esponenziale. E' necessario, infatti, che esista una statistica

sufficiente per il parametro ϑ, la *a priori* appartenga alla stessa famiglia di distribuzione della statistica sufficiente e il parametro k possa essere espresso come funzione della dimensione campionaria e della statistica sufficiente. Anche qui importante la funzione della statistica sufficiente come nel caso dell'approccio frequentista.

2.3 MCMC: Markov Chain Monte Carlo

Il calcolo della distribuzione *a posteriori* tramite metodi analitici non è sempre facile visto che il teorema di Bayes mette in gioco al denominatore un integrale. Questo può risultare più semplice se l'inferenza riguarda un solo parametro ma nelle applicazioni della vita reale diventa proibitivo. Quello che si propone in questi casi come soluzione sono dei metodi che permettano di approssimare e simulare una distribuzione di probabilità, nell'approccio bayesiano la distribuzione a posteriori. Si parla di Markov Chain Monte Carlo (MCMC) metodi di simulazione che fanno uso dell'integrazione Monte Carlo attraverso Catene di Markov. Più in particolare i metodi di simulazione Monte Carlo sono quelli nei quali si ottiene una serie di campioni casuali indipendenti ϑ^t da una distribuzione P(ϑ) scelta e si stima il valore atteso di una funzione $\hbar(\vartheta)$ nel seguente modo: $E\big(\hbar(\vartheta|x)\big) = \int \hbar(\vartheta)P(\vartheta|x)d\vartheta \approx \frac{1}{T}\sum_{t=1}^{T}\hbar(\vartheta^t)$. L'accuratezza è ottenuta generando un numero elevato di campioni mentre ciascun elemento del campione è campionato sequenzialmente come elemento di una Catena di Markov.

Sia $\mathcal{S} = [s_1, s_2, \dots s_k]$ un insieme finito o numerabile. Una famiglia $(X_n)_n$ di variabili aleatorie a valori in $\mathcal{S}$ definite tutte sullo stesso spazio di probabilità è una catena di Markov se qualsiasi sia n si ha: $P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j | X_n = i)$. Perciò il valore di X_{n+1} dipende esclusivamente dalla conoscenza del valore X_n. Gli elementi di $\mathcal{S}$ sono stati possibili della catena e si pone che $P_{ij}(n) = P(X_{n+1} = j | X_n = i)$, in cui n è il tempo e il valore $P_{ij}(n)$ è la probabilità di transizione dallo stato i al tempo n allo stato j al tempo n+1. Si dicono catene di Markov omogenee quelle per cui vale $P_{ij}(n) = P_{ij}$. La matrice $\mathcal{P}=(P_{ij})_{(i,j)}$ è detta matrice delle probabilità d transizione e soddisfa le seguenti proprietà:

- $P_{ij} \geq 0$
- $\sum_{j=1}^{k} P_{ij} = 1 \quad per\ qualsiasi\ i$

Una catena di Markov si dice aperiodica se gli stati non vengono osservati mai in modo periodico, mentre si dice irriducibile se tutti gli stati comunicano con gli altri ed è sempre possibile muoversi da uno stato all'altro. Ci si può chiedere cosa succede alla catena di Markov per n crescente. Questa è di natura stocastica quindi fluttuerà nel tempo e non si può sperare in una sua convergenza per n→ ∞, però si può pensare che giunga ad una situazione di equilibrio non dipendente dal tempo, si parla allora di distribuzione stazionaria. Sia X_n una catena di Markov con $\mathcal{S}$ spazio di stati e $\mathcal{P}$ matrice di transizione. Il vettore $\pi = (\pi_1, \pi_2, \dots \pi_k)$ è una distribuzione stazionaria se:

- $\pi_i \geq 0 \ \ \forall i = 1 \dots k$

- $\sum_{i=1}^{k} \pi_i = 1$
- $\pi = \pi\mathcal{P}$ *cioè* $\pi_j = \sum_{i=1}^{k} P_{ij}\pi_i \quad \forall j = 1 \dots k$

Da sottolineare infine che ogni catena irriducibile e aperiodica ammette una ed una sola distribuzione stazionaria se è finto il numero degli stati.

Ora data una catena di Markov finita, aperiodica e irriducibile possiamo determinare la distribuzione stazionaria associata oppure data una distribuzione stazionaria possiamo costruire una catena di Markov che converga a tale distribuzione stazionaria. Ed è questo che fanno i MCMC.

La strategia di campionamento MCMC consiste:

1) Nella costruzione di una catena di Markov X_n aperiodica e irriducibile per la quale la distribuzione stazionaria sia esattamente la distribuzione target π cioè in campo bayesiano la distribuzione *a posteriori* del parametro ϑ di interesse.
2) Simulare l'andamento della catena X_n per un tempo sufficientemente lungo
3) Restituire il valore finale della catena.

Per fare ciò ci serviamo di due algoritmi fondamentali: quello Metropolis-Hastings e quello Gibbs sampler. Si precisa che le procedure inferenziali che si metteranno in atto dopo aver applicato gli algoritmi, e quindi aver generato catene di Markov per la distribuzione *a posteriori* $P(\vartheta|x)$, sono esattamente le stesse che sono state già discusse in precedenza in campo bayesiano.

2.3.1 Algoritmo di Metropolis-Hastings

Supponiamo di voler giungere ad una distribuzione stazionaria P($\vartheta|x$), la nostra distribuzione *a posteriori*, l'algoritmo si esplica nei seguenti passi:

1. Si sceglie un valore di partenza ϑ^0 dalla distribuzione P($\vartheta|x$)
2. All'iterazione t-esima si campiona un valore ϑ^* dalla matrice di transizione $\mathcal{J}_t(\vartheta^*|\vartheta^{t-1})$
3. Si costruisce un rapporto di accettazione del tipo

$$r = \frac{\mathrm{P}(\vartheta^*|x)/\mathcal{J}_t(\vartheta^*|\vartheta^{t-1})}{\mathrm{P}(\vartheta^{t-1}|x)/\mathcal{J}_t(\vartheta^{t-1}|\vartheta^*)}$$

4. Si accetta ϑ^* come ϑ^t con probabilità min[r,1]. Se ϑ^* non è accettato allora ϑ^t=ϑ^{t-1}
5. Si ripetono i passi 2-4, M volte per ottenere M campioni da P($\vartheta|x$).

Nel passo 1. si campiona un valore ϑ^0 da una distribuzione P($\vartheta|x$) che si suppone essere la distribuzione stazionaria della catena di Markov che andiamo a generare. E' importante che questo valore abbia probabilità positiva cioè sia P($\vartheta^0|x$)> 0 altrimenti si inizierebbe da un valore che non può essere campionato.

Nel passo 2. $\mathcal{J}_t(\vartheta^*|\vartheta^{t-1})$ determina solo una delle possibili mosse effettuabili nella successiva iterazione della catena di Markov quindi tale distribuzione dovrebbe avere come supporto quello della *a posteriori*. L'algoritmo originale di Metropolis dichiarava che la $\mathcal{J}_t(\vartheta^*|\vartheta^{t-1})$ dovesse

essere simmetrica cioè $\mathcal{J}_t(\vartheta^*|\vartheta^{t-1}) = \mathcal{J}_t(\vartheta^{t-1}|\vartheta^*)$ ma questo non è strettamente necessario. Se fosse però simmetrica ricadremmo nell'algoritmo della passeggiata aleatoria di Metropolis.

Nel passo 4. di fatto si sceglie un valore u da una $\mathcal{U}[0,1]$ e se $u \leq r$ si accetta ϑ^* come ϑ^t, altrimenti ϑ^t=ϑ^{t-1}; per cui si ha sempre un valore da scegliere.

E' importante monitorare il tasso di accettazione cioè la frazione di ϑ^* accettatti, in quanto se il tasso di accettazione è troppo alto non si riesce ad indagare tutto lo spazio parametrico abbastanza velocemente mentre se il tasso di accettazione è troppo basso l'algoritmo risulta inefficiente, rifiuta troppi ϑ^* nonostante visiti più velocemente lo spazio campionario. Questo algoritmo ha delle proprietà che lo rendono utilizzabile nelle più svariate applicazioni. Infatti l'algoritmo Metropolis – Hastings è facilmente applicabile quando è conosciuta la distribuzione *a posteriori*. Inoltre il ricercatore non ha bisogno di conoscere le distribuzioni condizionate, come nel caso dell'algoritmo che si presenterà nel prossimo paragrafo, quello dell'algoritmo Gibbs Sampler.

Algoritmo Gibbs Sampler

Un caso particolare di algoritmo Metropolis – Hastings è quello che viene presentato di seguito e noto con il nome di Gibbs sampler. Due sono le differenze rispetto al Metropolis – Hastings:

1) Si accetta sempre il valore campionato
2) Le distribuzioni condizionate devono essere sempre note.

Proprio questo ultimo punto rende tale algoritmo meno applicabile rispetto al precedente.

Si assuma che la distribuzione stazionaria di interesse sia $P(\underline{\vartheta}|x)$, con $\underline{\vartheta} = (\vartheta_1, \vartheta_2, \dots, \vartheta_k)$ e che le distribuzioni full conditional $P_i(\vartheta_i) = P(\vartheta_i|\vartheta_{-i})$ per i=1,2,…,k siano note e facilmente campionabili. L'algoritmo segue i seguenti passi:

1. Si sceglie un vettore di partenza $\underline{\vartheta}^0 = (\vartheta_1^0, \vartheta_2^0, \dots, \vartheta_k^0)$
2. All'iterazione t-esima si ottiene un nuovo valore $\underline{\vartheta}^t = (\vartheta_1^t, \vartheta_2^t, \dots, \vartheta_k^t)$ a partire da $\underline{\vartheta}^{t-1}$ attraverso successive generazioni di valori tramite le distribuzioni full conditional nel seguente modo:

$$\vartheta_1^t \sim P(\vartheta_1|\vartheta_2^{t-1}, \dots, \vartheta_k^{t-1})$$

$$\vartheta_2^t \sim P(\vartheta_2|\vartheta_1^t, \dots, \vartheta_k^{t-1})$$

$$\vdots$$

$$\vdots$$

$$\vartheta_k^t \sim P(\vartheta_k|\vartheta_1^t, \dots, \vartheta_{k-1}^t)$$

3. Si incrementa ogni volta da t a t+1 e si ripercorre il passo 2. fino a giungere a convergenza.

Una iterazione dell'algoritmo di Gibbs Sampler è completa dopo k passi e genera una catena di Markov omogenea.

2.3.2 Convergenza

Generalmente le procedure di campionamento MCMC hanno un costo tra cui la valutazione della convergenza in quanto si ha bisogno di monitorare la catena per assicurare che le misure di sintesi sulla distribuzione *a posteriori* siano abbastanza accurate. Nonostante si sia sicuri che la catena generata dal MCMC converga alla distribuzione *a posteriori* in quanto la catena generata è aperiodica, irriducibile e finita, questo potrebbe non essere subito chiaramente visibile per i fenomeni che normalmente si riscontrano nella realtà e tra l'altro non si sa quando si giunge a convergenza. Diversi sono i metodi per valutare la convergenza e tra questi il test di Geweke che suddivide la catena di Markov in due parti non uguali tra loro, dopo aver eliminato la prima parte della catena. Se la catena è in uno stato stazionario allora le medie dei due campioni così ottenuti dovrebbero essere uguali. Egli ha costruito un test basato sulla differenza standardizzata tra le due medie campionarie, e, se la catena converge, tale test tende ad avere una distribuzione normale standard. Uno dei punti di forza di questo test è che è facile da capire, ma un punto di negatività è che dipende fortemente dalla scelta delle due parti della catena. Un altro test, il più popolare attualmente utilizzato, è quello di Gelman-Rubin. Il test considera $m > 1$ catene di Markov campionate in modo tale che i loro valori iniziali siano dispersi lungo la distribuzione *a posteriori*. Dopo aver eliminato la prima parte delle catene, si applica un'analisi della varianza analizzando la varianza tra le catene e la varianza entro catena. La varianza della distribuzione stazionaria

sarà data da $\hat{\sigma}^2 = \frac{(n-1)W}{n} + \frac{B}{n}$ dove W è la varianza empirica entro catena, $\frac{B}{n}$ è la varianza empirica tra le catene e n è il numero di iterazioni. Se la catena converge allora i due addendi sono non distorti altrimenti il primo sarà sottostimato mentre il secondo sarà sovrastimato. Il test di Gelman-Rubin si basa sull'assunzione che la distribuzione *a posteriori* sia normale e calcola un intervallo di credibilità bayesiano usando una t di student con media $\hat{\mu}$ media delle catene combinate, varianza $\hat{V} = \hat{\sigma}^2 + \frac{B}{mn}$ e gradi di libertà $d = \frac{2*\hat{V}^2}{Var(\hat{V})}$. Il test sarà dato dalla quantità $R = \sqrt{(d+3)\hat{V}/(d+1)W}$ che altro non è che il rapporto tra la varianza stimata tra catene, $\hat{V}$, e la varianza entro catena, W, moltiplicato per un fattore che tiene in considerazione la extra varianza della ditrbuzione t di Student. Valori vicino a 1 indicano sostanzialmente una buona convergenza.

Ci sono però dei metodi che permettono un'accelerazione della convergenza. Quali ad esempio la scelta appropriata dei valori iniziali della catena, attuare una trasformazione dei dati soprattutto se questi differiscono in magnitudine, attuare una riparametrizzazione dei parametri.

3. Il disegno per misure ripetute

Quando si parla di misure ripetute ci si riferisce alla condizione in cui dati riguardanti un certo soggetto vengono raccolti in istanti temporali diversi o sotto diverse condizioni. Questo può avere il vantaggio di poter valutare l'evoluzione della variabile risposta nel tempo su ogni singolo individuo, inoltre permette di

considerare un campione di numerosità inferiore rispetto ad altri tipi di disegni con un'unica rilevazione. Nonostante ciò questo tipo di disegno produce degli stimatori più efficienti per i parametri di interesse, in quanto si riesce a controllare la variabilità tra i soggetti separandola dall'errore sperimentale. Tra svantaggi si rileva il fatto che poiché la variabile risposta è osservata più volte sulla stessa unità, le diverse misurazioni non possono essere considerate indipendenti, ma sono tra loro correlate. Inoltre il ricercatore non sempre ha il controllo di come i dati vengono raccolti, per cui ci si può trovare di fronte a dati incompleti e quindi non bilanciati.

In genere i dati sono raccolti nella seguente forma e notazione:

GRUPPO	SOGGETTO	MISURA 1	....	MISURA h	....	MISURA t
1	1	Y_{111}	…	Y_{11h}	…	Y_{11t}
	…	…	…	…	…	…
	j	Y_{1j1}	…	Y_{1jh}	…	Y_{1jt}
	…	…	…	…	…	…
	n_1	Y_{1n_11}	…	Y_{1n_1h}		Y_{1n_1t}
			……			
i	1	Y_{i11}	…	Y_{i1h}	…	Y_{i1t}
	…	…	…	…	…	…
	j	Y_{ij1}	…	Y_{ijh}	…	Y_{ijt}
	…	…	…	…	…	…
	n_i	Y_{in_i1}	…	Y_{in_ih}		Y_{in_it}
			…			
M	1	Y_{M11}	…	Y_{M1h}	…	Y_{M1t}
	…	…	…	…	…	…
	j	Y_{Mh1}	…	Y_{Mjh}	…	Y_{Mjt}
	…	…	…	…	…	…
	n_M	Y_{Mn_M1}	…	Y_{Mn_Mh}		Y_{Mn_Mt}

dove:

n_i è il numero di soggetti nel gruppo i-mo

$n = \sum_{i=1}^{M} n_i$

Y_{ijh} risposta al tempo h dell'j-mo soggetto nel gruppo i, con h=1…t, i=1…M e j=1…n_M

$\overline{Y_{...}} = \frac{\sum_{h=1}^{t}\sum_{i=1}^{M}\sum_{j=1}^{n_M} Y_{ijh}}{nt}$ media generale

$\overline{Y_{i..}} = \frac{\sum_{h=1}^{t}\sum_{j=1}^{n_M} Y_{ijh}}{n_i t}$ media nel gruppo i-mo

$\overline{Y_{..h}} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{n_M} Y_{ijh}}{n}$ media al tempo h

$\overline{Y_{ij.}} = \frac{\sum_{h=1}^{t} Y_{ijh}}{t}$ media del j-mo soggetto nel gruppo i

3.1 Il modello

Il modello che in genere viene utilizzato per descrivere il disegno a misure ripetute è quello detto lineare ad effetti misti. Si assuma un disegno completamente randomizzato di soggetti in M gruppi con n_i soggetti assegnati al gruppo i-simo. Si suppone che i dati sui differenti soggetti siano indipendenti e che ci siano t misurazioni equamente spaziate nel tempo per ogni soggetto. Si ponga che il vettore Y_{ijh} è la risposta misurata al tempo h per il soggetto j del gruppo i per i=1…M, h=1…t e j=1…n_M. La parte fissa del modello darà che

$E[Y_i] = \mu_{ijh}$ ed è modellata come funzione del gruppo, del tempo e di altre covariate fisse. La parte casuale del modello specifica la struttura di covarianza delle osservazioni e si assume che le osservazioni su diversi soggetti siano indipendenti perciò $Cov(Y_{ijh}, Y_{i'j'l}) = 0$ se $i \neq i'$ e $j \neq j'$. Inoltre si pone che la matrice di varianze e covarianze delle misurazioni di uno stesso soggetto sono le stesse all'interno di ciascun gruppo. Tuttavia si può pensare che le varianze non siano omogenee in tutti i tempi e che le covarianze tra osservazioni a tempi diversi nello stesso soggetto non siano le stesse tra coppie di tempi soprattutto se i tempi sono ravvicinati. Una struttura generale per la matrice di varianza e covarianza è data da $Cov(Y_{ijh}, Y_{ijl}) = \sigma_{h,l}$ dove $\sigma_{h,l}$ è la covarianza tra misurazioni ai tempi h ed l sullo stesso soggetto e $\sigma_{h,h} = \sigma_h^2$ è la varianza al tempo h. Come strutture di covarianza si dovrà scegliere quella che maggiormente caratterizza i dati tra quelle possibili.

Sia $\boldsymbol{Y}_{ij}=(Y_{ij1}, \dots, Y_{ijt})^T$ il vettore risposta ai tempi 1,2…,t sul soggetto j nel gruppo i. Quindi si ha

$$\boldsymbol{Y}_{ij} = \boldsymbol{\mu}_{ij} + \boldsymbol{\varepsilon}_{ij}$$

dove $\boldsymbol{\mu}_{ij}=(\mu_{ij1}, \dots, \mu_{ijt})^T$ è il vettore delle medie ed $\boldsymbol{\varepsilon}_{ij}=(\varepsilon_{ij1}, \dots, \varepsilon_{ijt})^T$ è il vettore degli errori, rispettivamente per il soggetto j nel gruppo i. Per cui si ha che $E[Y_{ij}] = \mu_{ij}$ e $Var[Y_{ij}] = V_{ij}$ dove V_{ij} è la matrice t*t con $\sigma_{h,l}$ nella riga h e colonna l e si assuma che questa sia la stessa per tutti i soggetti cioè per ogni i e per ogni j. La matrice dei dati sarà $Y = (Y'_{11}, \dots, Y'_{1n}, Y'_{21}, \dots, Y'_{2n}, \dots, Y'_{g1}, \dots, Y'_{gn})$ e ugualmente il vettore dei valori attesi e

degli errori $E[\boldsymbol{Y}] = \boldsymbol{\mu} = \left(\mu'_{11}, \dots, \mu'_{1n}, \mu'_{21}, \dots, \mu'_{2n}, \dots, \mu'_{g1}, \dots, \mu'_{gn}\right)$ e $\boldsymbol{\varepsilon} = \left(\varepsilon'_{11}, \dots, \varepsilon'_{1n}, \varepsilon'_{21}, \dots, \varepsilon'_{2n}, \dots, \varepsilon'_{g1}, \dots, \varepsilon'_{gn}\right)$. Per cui il modello può ancora essere scritto nel seguente modo:

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \qquad (2)$$

con $Var[\boldsymbol{Y}] = diag\left\{V_{ij}\right\}$. Il modello applicato alle misure ripetute si esplicita nella seguente forma:

$$Y_{ijh} = \mu + \lambda x_{ij} + \alpha_i + d_{ij} + \tau_h + (\tau\alpha)_{ih} + e_{ijh} \qquad (3)$$

dove μ è una costante comune a tutte le osservazioni, λ è il coefficiente fisso delle covariate x_{ij} per il paziente j nel gruppo i, α_i è il parametro corrispondente al gruppo i, τ_h è il parametro riferito al tempo h, $(\tau\alpha)_{ih}$ è il parametro dell'interazione tra il tempo h e il gruppo i, d_{ij} è una variabile normalmente distribuita con media 0 e varianza σ_d^2corrispondente al soggetto j nel gruppo i e e_{ijh} sono variabili casuali normalmente distribuite con media 0 e varianza σ_e^2, indipendente da d_{ij} e riferita al soggetto j nel gruppo i al tempo h. Segue che $E\left[Y_{ijh}\right] = \mu_{ijh} = \mu + \lambda x_{ij} + \alpha_i + \tau_h + (\tau\alpha)_{ih}$, $Var\left[Y_{ijh}\right] = \sigma_d^2 + \sigma_e^2$ e $Cov\left(Y_{ijh}, Y_{ijl}\right) = \sigma_d^2 + cov(e_{ijh}, e_{ijl})$. Il modello (3) in forma matriciale si può ricondurre a

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{ZU} + \boldsymbol{e} \qquad (4)$$

dove **X** è la matrice dei coefficienti noti dei parametri degli effetti fissi $\mu, \lambda, \alpha_i, \tau_h, (\tau\alpha)_{ih}$, $\boldsymbol{\beta}$ è il vettore di questi parametri degli effetti fissi, $\boldsymbol{Z}$ è la matrice di coefficienti, 0 e 1, dell'effetto random dei soggetti d_{ij}, $\boldsymbol{U}$ è il vettore di tali effetti random d_{ij} e infine $\boldsymbol{e}$ è il vettore degli errori e_{ijh}. Assumendo soltanto che $\boldsymbol{U}$ ed $\boldsymbol{e}$ siano indipendenti si ottiene che $Var[\boldsymbol{Y}] = ZV(U)Z' + V(\boldsymbol{e})$. Per cui nelle misure ripetute la varianza totale del fenomeno è scomposta rispettivamente nella varianza tra i soggetti $ZV(U)Z'$ e varianza entro soggetto $V(\boldsymbol{e})$.

3.2 Inferenza classica

Considerando il modello scritto nella forma (4), si sa che E(**U**) = 0, E(**e**) = 0 e per semplicità si scriva Var(**U**) = **G** e Var(**e**) = **R**. Quindi E(**Y**)= $\boldsymbol{X\beta}$. Assumendo che **U** e **e** siano indipendenti si ottiene che $Var[\boldsymbol{Y}] = \boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$ dalla quale si deduce che $Y \sim N(\boldsymbol{X\beta}, \boldsymbol{ZGZ'} + \boldsymbol{R})$. La migliore stima per i parametri $\boldsymbol{\beta}$ è ottenuta attraverso il metodo dei minimi quadrati generalizzati ed è data da

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}Y}$$

e

$$Var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X'V^{-1}X})^{-1}.$$

Entrambe le stime per i parametri $\boldsymbol{\beta}$, sono funzioni di $Var[\boldsymbol{Y}] = \boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$ ed in molti casi questa sarà dipendente da parametri incogniti, per cui in realtà $\boldsymbol{V}$

sarà sostituita con una stima $\widehat{\boldsymbol{V}}$**.** La matrice **G** in genere dipende da un set di parametri mentre **R** da un altro gruppo di parametri che vengono stimati attraverso il metodo della massima verosimiglianza. Lo stimatore $\widehat{\boldsymbol{\beta}}$ è funzione lineare di **Y** per cui sotto l'ipotesi di normalità dei dati, anche $\widehat{\boldsymbol{\beta}}$ sarà distribuito normalmente. Per cui si ha che $(\boldsymbol{X'V^{-1}X})^{1/2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \rightarrow_d \boldsymbol{N(0, I)}$**.**

Se si considera una famiglia di funzioni lineari $\boldsymbol{\beta}$ del tipo $\boldsymbol{A'\beta}$, la stima di tale famiglia sarà $\boldsymbol{A'}\widehat{\boldsymbol{\beta}}$ e la stima della sua varianza sarà $\boldsymbol{A'(X'V^{-1}X)^{-1}A}$**.** Per cui $\boldsymbol{A'}\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{A'\beta}, \boldsymbol{A'(X'V^{-1}X)^{-1}A})$

La statistica utilizzata per testare gli effetti fissi è data da

$$F = \widehat{\boldsymbol{\beta}}'\boldsymbol{A}\left[\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right)\right]^{-1}\boldsymbol{A'}\widehat{\boldsymbol{\beta}}/rango(\boldsymbol{A})$$

Tale statistica si distribuisce come una F di Fisher con gradi di libertà dati da $rango(\boldsymbol{A})$ ma molto spesso il calcolo dei gradi di libertà è molto macchinoso per cui una possibile scelta è quella proposta con l'approssimazione di Satterwhite.

Per quanto riguarda gli effetti casuali rifacendosi sempre al modello nella forma (4) e assumendo che $\mathbf{U} \sim N(\mathbf{0},\mathbf{G})$ e $\mathbf{e} \sim N(\mathbf{0},\mathbf{R})$ con **U** ed **e** indipendenti, si pone $\boldsymbol{Y^* = Y - X\beta}$. La stima dei parametri ad effetti casuali sarà

$$\widehat{\boldsymbol{U}} = \widehat{\boldsymbol{G}}^{-1}\boldsymbol{Z}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{Y}}^*$$

e

$$Var(\widehat{\boldsymbol{U}}) = \widehat{\boldsymbol{G}} - \widehat{\boldsymbol{G}}^{-1}\boldsymbol{Z}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{G}}$$

andando a stimare i parametri di ciascuna matrice attraverso la massima verosimiglianza.

3.3 Inferenza bayesiana

Si consideri il modello nella forma (4) e si scelga $\mathbf{U}\sim N(\mathbf{0},\mathbf{G})$ e $\mathbf{e}\sim N(\mathbf{0},\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}\boldsymbol{I})$ per semplicità.

Il metodo di inferenza bayesiana inizia stabilendo:

- ✓ La Funzione di verosimiglianza: $P(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{U},\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}})$
- ✓ La *a priori* per gli effetti casuali: $\mathrm{P}(\boldsymbol{U}|\boldsymbol{G})$
- ✓ Le *a priori* per i parametri $\mathrm{P}(\boldsymbol{\beta},\boldsymbol{G},\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}})$

Per quanto riguarda quest'ultimo punto è di comune norma assumere che siano indipendenti cioè $\mathrm{P}(\boldsymbol{\beta},\boldsymbol{G},\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}) = P(\boldsymbol{\beta})P(\boldsymbol{G})P(\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}})$. In genere la scelta per $P(\boldsymbol{\beta})$ e per $P(\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}})$ ricade rispettivamente su una distribuzione normale multivariata e su una distribuzione inversa Gamma in quanto sono flessibili nel coprire un insieme di informazioni abbastanza ampio. La scelta per $P(\boldsymbol{G})$ è meno facile anche se in genere anche per questa si sceglie una distribuzione inversa Gamma con parametri specificati.

A questi passaggi segue poi la determinazione via Teorema di Bayes o attraverso l'utilizzo dei metodi MCMC della distribuzione *a posteriori*

$$P(\boldsymbol{\beta},\boldsymbol{U},\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{U},\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}})\mathrm{P}(\boldsymbol{U}|\boldsymbol{G})\mathrm{P}(\boldsymbol{\beta},\boldsymbol{G},\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}})$$

4. Caso in studio

Otto soggetti volontari sani sono stati reclutati per lo studio previo consenso informato. I dati sono stati raccolti in due fasi – fase di condizionamento e fase di esperimento - in cui i soggetti sono stati stimolati sulla cute del piede con impulsi di due intensità, alta e bassa. L'intensità degli stimoli usati è stata scelta in base alla sensibilità del soggetti, in modo da risultare dolorosi ma sopportabili senza pena. Su di un monitor, prima della stimolazione e per 5 secondi, apparivano cue visivi indicanti l'intensità dello stimolo, rosso=alto e verde=basso. Nel condizionamento i soggetti hanno imparato l'associazione tra intensità dello stimolo e colore del cue (12 stimoli alti e 13 bassi, in ordine casuale). Nell'esperimento effettivo, la procedura non è cambiata agli occhi del soggetto, ma a 21 dei cue verdi è stato associato uno stimolo alto (condizione placebo), a 16 cue verdi uno stimolo basso (rinforzo) e ai cue rossi è seguito sempre uno stimolo alto; per un totale alla fine di 483 osservazioni escludendo le osservazioni di rinforzo. Lo studio si è svolto nel rispetto della convenzione di Helsinki. Il dolore è stato rilevato utilizzando la scala VAS. Obiettivo dello studio è capire se possa esistere un effetto placebo nella determinazione della scala di dolore da parte del soggetto.

Risultati

I dati che sono stati raccolti possono essere sintetizzati nella seguente tabella:

Soggetto	Cue	Misurazione1	…	…	…	Misurazione h
1	r	VAS_{1r1}	…	…	…	VAS_{1rh}
	v	VAS_{1v1}	…	…	…	VAS_{1vh}
…	r	…	…	…	…	…
	v	…	…	…	…	…
…	r	…	…	…	…	…
	v	…	…	…	…	…
…	r	…	…	…	…	…
	v	…	…	…	…	…
8	r	VAS_{8r1}	…	…	…	VAS_{8rh}
	v	VAS_{8v1}	…	…	…	VAS_{8vh}

e il modello utilizzato per l'analisi è il seguente: $VAS_{ijh} = \mu + \lambda CUE_{ij} + d_{ij} + \tau_h + +e_{ijh}$ dove μ è una costante comune a tutte le osservazioni, λ è il coefficiente fisso per la covariata CUE_{ij} per il CUE j nel soggetto i, τ_h è il parametro riferito alla misurazione h, d_{ij} è una variabile normalmente distribuita con media 0 e varianza σ_d^2 corrispondente al CUE j nel soggetto i e e_{ijh} sono variabili casuali normalmente distribuite con media 0 e varianza σ_e^2, indipendente da d_{ij} e riferita al CUE j nel soggetto i alla misurazione h. Nell'approccio bayesiano sono state fatte diverse prove con scelta di *a priori* diverse per i parametri λ e scegliendo invece come *a priori* per i parametri τ una distribuzione normale standardizzata, per i parametri casuali una distribuzione normale di media 0 e varianza stimata attraverso una distribuzione *a priori* inversa gamma. La stessa distribuzione *a priori* inversa gamma è stata scelta per tutti i parametri di varianza. I modelli quindi si distinguono nel seguente modo:

1) $\lambda_r \sim N(\mu_r, \sigma_r^2)$, $\lambda_v \sim N(\mu_v, \sigma_v^2)$, $\tau \sim N(0,1)$ dove μ_r, μ_v sono le medie e σ_r^2, σ_v^2 sono le varianze della VAS rispettivamente per il cue rosso e per il cue verde calcolate sul campione.

2) $\lambda_r \sim U(min_r, max_r)$, $\lambda_v \sim U(min_v, max_v)$, $\tau \sim N(0,1)$, dove min_r, min_v e max_r, max_v sono rispettivamente il valore minimo e il valore massimo della VAS per il cue rosso e il cue verde

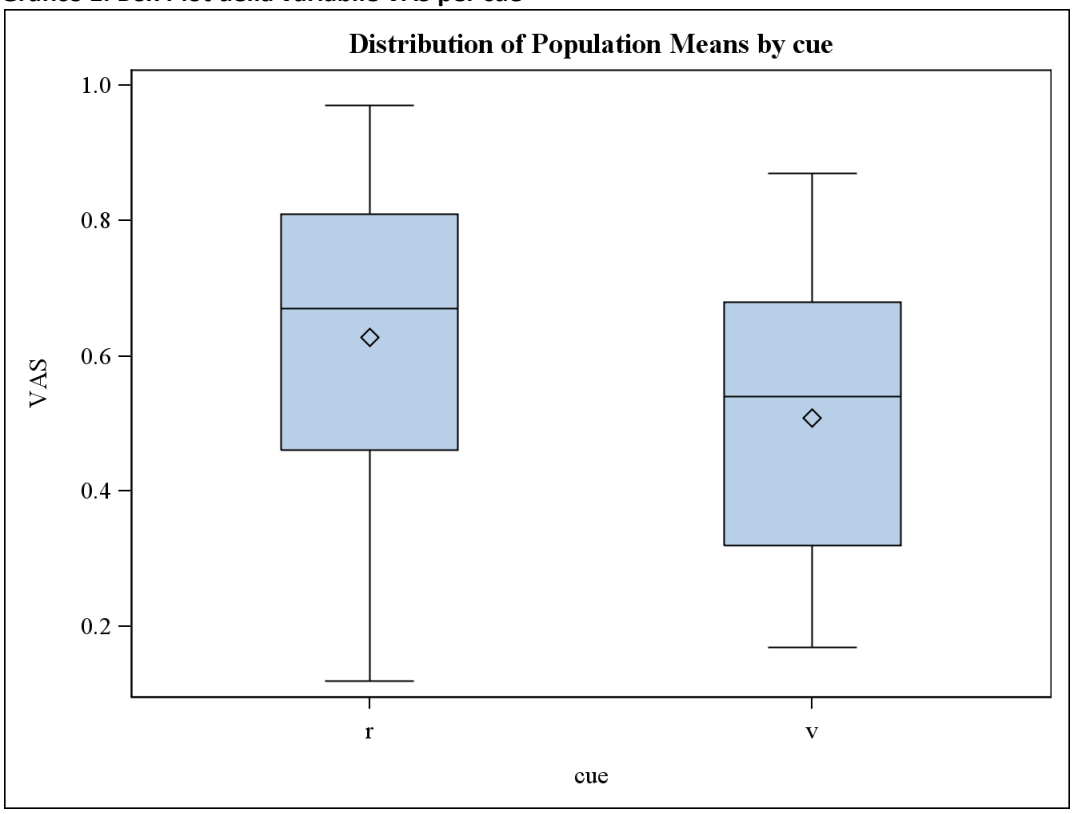
3) $\lambda_r \sim Beta(\alpha_r, \beta_r)$, $\lambda_v \sim Beta(\alpha_v, \beta_v)$, $\tau \sim N(0,1)$ dove α_r, α_v e β_r, β_v sono i parametri della distribuzione beta calcolati basandosi sui valori della VAS sul campione rispettivamente per il cue rosso e per il cue verde.

4) $\lambda_r \sim Beta(\alpha_r, \beta_r)$, $\lambda_v \sim U(min_v, max_v)$, $\tau \sim N(0,1)$ dove α_r, β_r sono i parametri della distribuzione beta calcolati basandosi sui valori della VAS sul campione per il cue rosso e min_v, max_v sono il valore minimo e il valore massimo della VAS sul campione per il cue verde.

L'algoritmo per campionare dalla distribuzione a posteriori e quindi dalla distribuzione di stazionarietà è l'algoritmo Metropolis – Hastings.

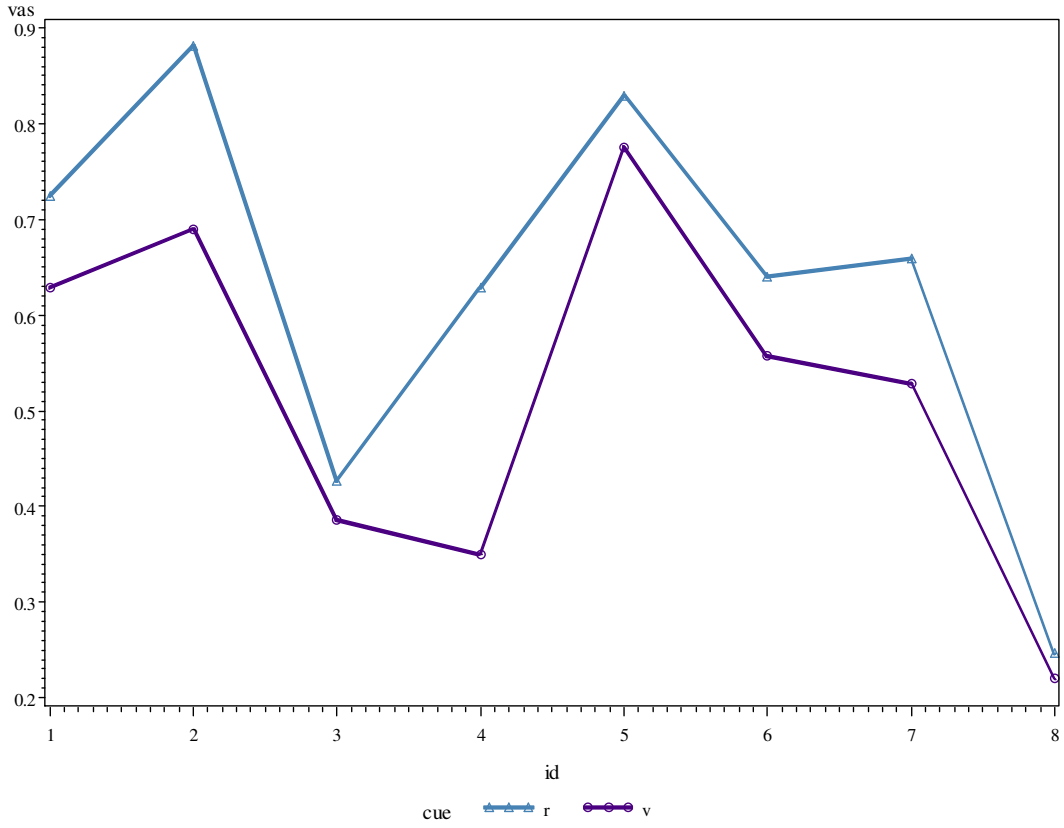
I grafici box plot dell'esperimento sono i seguenti:

Grafico 1. Box Plot della variabile VAS per cue



e rivelano come la VAS sia in media più bassa nel cue verde rispetto al cue rosso nonostante lo stimolo inviato al soggetto sia sempre alto. In ciascun soggetto la situazione non cambia come si può vedere dal seguente grafico:

Grafico 2. Distribuzione della media per la variabile VAS in ciascun soggetto per cue



Per cui in media la VAS per il CUE rosso è in media sempre più elevata che per il CUE verde nonostante lo stimolo inviato sia sempre elevato.

I risultati che si sono ottenuti con i modelli vengono sintetizzati nella seguente tabella premettendo che l'interazione tra misurazione e cue non è stata inserita in quanto risultata non significativa.

Tabella 1. Confronto di Stima e Errore standard tra modello classico e modelli bayesiani

Variabili	Mod. classico		Mod. bayesiano 1		Mod. bayesiano 2		Mod. bayesiano 3		Mod. bayesiano 4	
	Stima	Err.St	Stima	Err.St	Stima	Err.St	Stima	Err.St	Stima	Err.St
Cue Rosso	0.63*	0.07	0.62	0.009	0.61	0.02	0.61	0.01	0.65	0.02
Cue Verde	0.51*	0.07	0.49	0.009	0.48	0.02	0.49	0.01	0.52	0.02
Differenza Cue	0.12*	0.03	0.13	0.0002	0.13	0.0001	0.12	0.0002	0.13	0.0001
Misurazione	-0.002*	0.0005	-0.002	0.000009	-0.002	0.000006	-0.002	0.000009	-0.002	0.000005

*La stima è significativa ($p<0.05$)

Da questa tabella si evince che le stime per i due approcci sono molto similari anche se l'ultimo modello bayesiano si discosta dagli altri modelli bayesiani e dal modello classico dando stime più alte per le variabili cue rosso e cue verde. Quello che però cambia è che sembra esserci una stima più precisa per quanto riguarda le stime dell'approccio bayesiano specie per il primo modello. In particolare si noti anche come gli errori standard tendano ad essere più alti per il modello bayesiano in cui le *a priori* sono uniformi per i parametri λ

Se invece si va considerare l'intervallo di confidenza per il modello classico e l'intervallo di credibilità per i modelli bayesiani si ottiene quanto segue:

Tabella 2. Confonto di intervallo di confidenza e intervalli di credibilità tra modello classico e modelli bayesiani

Variabili	Modello classico		Mod. bayesiano 1		Mod. bayesiano 2		Mod. bayesiano 3		Mod. bayesiano 4	
	Intervallo di confidenza		Intervallo di credibilità HPD[2]		Intervallo di credibilità HPD		Intervallo di credibilità HPD		Intervallo di credibilità HPD	
	LI[1]	LS[1]	LI[1]	LS[1]	LI[1]	LS[1]	LI[1]	LS[1]	LI[1]	LS[1]
Cue Rosso	0.49	0.78	0.44	0.81	0.36	0.85	0.37	0.85	0.40	0.85
Cue Verde	0.36	0.66	0.31	0.68	0.24	0.72	0.23	0.72	0.28	0.72
Differenza Cue	0.06	0.19	0.10	0.15	0.10	0.15	0.10	0.15	0.10	0.15
Misurazione	-0.0024	-0.0006	-0.0027	-0.0004	-0.0027	-0.0004	-0.0027	-0.0005	-0.003	-0.0005

[1] LI =Limite inferiore; LS=limite superiore

[2] HPD= Higher Probability Density

Da qui si evince che in genere l'intervallo bayesiano è più ampio di quello dell'approccio classico, soprattutto se le distribuzioni a priori che si danno ai parametri sono uniformi ma con l'intervallo bayesiano si è sicuri che al 95% il valore vero del parametro cade in tale intervallo. Inoltre per la differenza tra le medie, che è la variabile che più interessa, nel modello bayesiano, l'ampiezza dell'intervallo è più piccola rispetto a quella del modello classico e non comprende lo 0 quindi la differenza tra la VAS per il cue verde e quella per il cue rosso esiste ed è evidente e potrebbe oscillare tra 0.10 e 0.15.

Per quanto riguarda il solo approccio bayesiano di seguito viene presentata una tabella per valutare la convergenza della catena di Markov.

Tabella 3. Test di Geweke di convergenza per i modelli bayesiani

Variabili	Mod. bayesiano 1		Mod. bayesiano 2		Mod. bayesiano 3		Mod. bayesiano 4	
	Test Geweke		Test Geweke		Test Geweke		Test Geweke	
	z	Pr > |z|	z	Pr > |z|	z	Pr > |z|	z	Pr > |z|
Cue Rosso	-0.35	0.73	-1.17	0.24	0.69	0.49	-0.63	0.53
Cue Verde	-0.34	0.74	-1.13	0.26	0.63	0.53	-0.62	0.53
Differenza Cue	0.42	0.68	1.56	0.12	-1.65	0.10	0.44	0.66
Misurazione	0.51	0.61	1.46	0.15	-1.58	0.11	-0.86	0.39

Tabella 4. Test di Gelman – Rubin di convergenza per i modelli bayesiani

Variabili	Modello bayesiano 1		Modello bayesiano 2		Modello bayesiano 3		Modello bayesiano 4	
	Test Gelman - Rubin		Test Gelman - Rubin		Test Gelman - Rubin		Test Gelman - Rubin	
	Stima	97.5% LS[1]	Stima	97.5% LS[1]	Stima	97.5% LS[1]	Stima	97.5% LS[1]
Cue Rosso	1.0218	1.0061	1.1418	1.0298	1.0221	1.0024	1.0733	1.0199
Cue Verde	1.0210	1.0054	1.1423	1.0302	1.0227	1.0024	1.0744	1.0203
Differenza Cue	1.0004	1.0010	1.0001	1.0008	1.0005	1.0010	1.0005	1.0008
Misurazione	1.0007	1.0010	1.0000	1.0009	1.0004	1.0008	1.0001	1.0010

[1] LS=limite superiore

Da entrambi i test si evince che la stima è corretta in quanto le catene di Markov generate giungono a convergenza. Questo perché il primo test è sempre non

significativo e quindi non rifiuta l'ipotesi nulla di convergenza della catena di Markov generata e il secondo sempre molto vicino a 1, che è il valore che ci si aspetta per la convergenza.

Ultimo cenno ai tassi di accettazione dell'algoritmo di campionamento Metropolis – Hastings

Tabella 5. Tassi di accettazione per l'algoritmo di Metropolis – Hastings nei modelli bayesiani

Modello bayesiano 1		Modello bayesiano 2		Modello bayesiano 3		Modello bayesiano 4	
Tasso di accettazione		Tasso di accettazione		Tasso di accettazione		Tasso di accettazione	
Basso	Alto	Basso	Alto	Basso	Alto	Basso	Alto
0.19	0.28	0.19	0.28	0.18	0.26	0.16	0.29

Il range, infine, dei tassi di accettazione è buono. E' stato dimostrato, infatti, che se questo è compreso nel range 0.15 - 0.50 il metodo di campionamento è almeno all'80% efficiente. Per cui in questo caso essendo i valori trovati compresi in 0.15 - 0.50 per tutti e quattro i modelli si è sicuri che l'algoritmo ha funzionato bene andando a campionare correttamente in tutto lo spazio parametrico i valori per i parametri di interesse.

Discussione

I risultati ottenuti mostrano come è evidente l'effetto placebo in questo esperimento. La differenza nel valore di VAS riportato dai soggetti quando vedono l'immagine rossa e quando vedono l'immagine verde è in media esistente e netta. Per cui nonostante l'impulso sia alto sia in corrispondenza del cue rosso sia in corrispondenza di quello verde, i soggetti in media sentono meno dolore in corrispondenza di quest'ultimo. E questo tipo di risultato si raggiunge più o meno simile, sia con l'approccio classico che con quello bayesiano anche se questo produce stime più precise. Soprattutto nel quarto modello bayesiano, quello con le distribuzioni *a priori* beta e uniforme rispettivamente per il cue rosso e il cue verde, le stime per quanto riguarda la VAS nei due CUE iniziano a diversificarsi dal modello classico quindi sembra che la distribuzione *a priori* inizi a pesare di più sui dati osservati. Infatti in genere, quanto più il peso dei dati osservati è grande rispetto alla distribuzione *a priori*, per numerosità del campione molto elevato, tanto più le inferenze ottenute con i due approcci sono uguali (principio della misura precisa). Al contrario quando il campione è molto piccolo o quando la distribuzione *a priori* è molto forte, i due approcci tendono a dare risposte molto diverse. Però allo stesso modo dare distribuzioni *a priori* non corrette può portare a dati fuorvianti e distribuzioni *a posteriori* che dipendono fortemente dalla distribuzione *a priori* scelta. Quindi bisognerebbe prestare molta attenzione quando si prendono decisioni sulla distribuzione *a priori*. Uno dei punti di maggior critica dell'approccio bayesiano è infatti proprio la scelta della distribuzione *a priori*. Se da una parte, tramite la scelta della distribuzione *a priori*, l'approccio bayesiano tende ad essere molto affascinante in quanto imita il processo di conoscenza del fenomeno, dall'altro mette in

gioco la soggettività del ricercatore in quanto tale metodo di fatto non dà indicazioni su come scegliere correttamente la distribuzione *a priori*. A vantaggio dell'approccio bayesiano, però in questo caso, va la possibilità di effettuare delle prove di sensibilità nella scelta della distribuzione *a priori*.

Un altro punto di critica dell'approccio bayesiano rispetto a quello classico è il fatto di non prevedere delle possibilità di verifica delle metodiche inferenziali, aspetto che invece viene affrontato nell'ambito classico attraverso i concetti di sufficienza, consistenza, efficienza e correttezza. Questo però è vero solo in base a quello che si può valutare prima di aver osservato il campione, mentre comunque nell'approccio bayesiano tutte quelle che sono le procedure inferenziali sono condizionate dall'aver effettivamente osservato un campione ed averlo disponibile. In aggiunta nell'approccio bayesiano il peso del campione si realizza attraverso l'utilizzo della funzione di veromiglianza quindi è fondamentale il principio di verosimiglianza, cioè la fedeltà a quanto osservato, al contrario di quanto avviene nell'approccio classico dove invece è ritenuto importante il principio dell'esperimento ripetuto e nelle stesse condizioni. Il valore infatti dell'approccio bayesiano vale qualunque sia la dimensione campionaria in quanto tutte le procedure inferenziali sono basate sulla distribuzione *a posteriori* mentre nell'approccio frequentista si fa molto spesso riferimento alla teoria asintotica e quindi una dimensione campionaria adeguata e rappresentativa è fondamentale. Una differenza fondamentale poi tra due approcci è che quello frequentista si basa essenzialmente sulla massimizzazione di funzioni mentre quello bayesiano è improntato sull'integrazione.

Infine una grande spinta all'approccio bayesiano è stato l'utilizzo di nuove tecniche di integrazione quali i metodi MCMC, che hanno migliorato anche i costi di computazionali della distribuzione *a posteriori*. L'approccio bayesiano poi a differenza

di quello classico, permette di fare inferenza direttamente sul parametro con notevoli vantaggi anche dal punto di vista interpretativo in quanto anche l'interpretazione dei risultati risente di questa immediatezza.

Bibliografia

- D. Piccolo (2000). *Statistica*, Il Mulino.
- A. Azzalini (2001). *Inferenza Statistica*, Springer.
- A. Baccheri, G. Della Cioppa (2004). *Fondamenti di ricerca clinica*, Springer.
- E. Lesaffre, A.B. Lawson (2012). *Bayesian Biostatistics*, Wiley.
- J. Wakefield (2013). *Bayesian and Frequentist Regression Methods*. Springer.
- A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin (2014). *Bayesian Data Analysis*, CRC Press.
- S. Brooks, A.Gelman, G.L.Jones, X.L. Meng (2011). *Handbook of Markov Chain Monte Carlo*, CRC Press.
- G.Casella (2008). *Statistical Design*, Springer.
- *Bayesian Analysis using SAS*, Course note.
- R.C. Littell, J. Pendergast, R. Natarajan (2000). *Modelling covariance structure in the analysis of repeated measures data*. Statist. Med.; 19:1793-1819.
- B. Cantell (1997). *Using linear regression Analysis and the Gibbs Sampler to estimate the probability of a part being within specification*. Proceedings of the 22nd Annual SAS user group international conference; paper 264.
- J. Mandrekar, D.J. Sargent, P.J. Novotny, J.A. Sloan. (1999). *A general Gibbs sampling algorithm for analyzing linear models using the SAS System*, Proceedings of the 24th Annual SAS user group international conference; paper 266.
- A.Gelman, D. B. Rubin (1992). *Inference from iterative simulation using multiple sequences*. Statistical Science; 7: 457-511.

- J. Geweke (1992). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. In Bayesian Statistics 4, eds. J. M. Bernardo, J. Berger, A. P. Dawid, and A.F.M. Smith, Oxford, U.K.: Oxford University Press, pp. 169-193.
- N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller (1953). *Equations of State Calculations by Fast Computing Machines*. Journal of Chemical Physics; 21: 1087-1091.
- N. Metropolis, S. Ulam (1949). *The Monte Carlo Method*. Journal of the American Statistical Association; 44: 335-341.
- W.K Hastings (1970). *Monte Carlo sampling methods using Markov chains and their applications*. Biometrikam; 57: 97-109.
- M. K. Cowles and B. P. Carlin (1996). *Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review*. Journal of the American Statistical Association; 91: 883-904.
- H. Jeffreys (1939). *Theory of Probability*, 1st ed. The Clarendon Press, Oxford.
- H. Jeffreys (1948). *Theory of Probability*, 2nd ed. The Clarendon Press, Oxford.
- H. Jeffreys (1961). *Theory of Probability*, 3rd ed. Oxford Classic Texts in the Physical Sciences. Oxford Univ. Press, Oxford
- J. Haldane (1932). *A note on inverse probability*. Proc. Cambridge Philos. Soc. 28: 55–61.
- R.E. Kass and A.E. Raftery (1995). *Bayes Factor*. of the American Statistical Association; 90: 773-795.

- W R. Gilks and G. Roberts (1996). *Improving MCMC mixing*. Markov Chain Monte Carlo in Practice (eds W R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall.
- S. Geman and D. Geman (1984). *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*. IEEE Trans. Pattn Anal. Mach. Intell.; 6: 721-741.
- A. E. Gelfand and A. E M. Smith (1990). *Sampling based approaches to calculating marginal densities*. J Am. Statist. Ass.; 85: 398-409.
- A. E. Gelfand, S. E. Hills, A. Racine-Poon and A. E M. Smith (1990). *Illustration of Bayesian inference in normal data models using Gibbs sampling*. J. Am. Statist. Ass.; 85: 972-985.
- C. P. Robert and G. Casella (1999). *Monte Carlo Statistical Methods*. Springer Verlag New York.
- G. Casella and E. I. George (1992). *Explaining the Gibbs sampler*. The American Statistician; 46: 167-174.
- S. Chib and E. Greenberg (1995). *Understanding the Metropolis – Hastings algorithm*. American Statistical Association; 49: 327 – 335.