UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

BIOSTATISTICA BAYESIANA CON "MATCHING PRIORS"

Relatore: Prof.ssa Laura Ventura
Dipartimento di Scienze Statistiche

Laureanda: Giulia Ranzato
Matricola N. 1132340

Anno Accademico 2017/2018

"Se hai paura di qualcosa un buon approccio è misurarla"
-Aledebez

Indice

Introduzione

L'utilizzo di metodi bayesiani nella teoria e nella pratica statistica ha avuto un forte incremento in questi ultimi anni. La loro implementazione richiede la specificazione di una distribuzione a priori e di una funzione di verosimiglianza. Solitamente la distribuzione a priori rappresenta l'informazione che il ricercatore ha in merito al fenomeno che sta analizzando prima di iniziare lo studio. Tuttavia, se non si è in possesso di un numero adeguato di dati storici, non è possibile ottenere una distribuzione a priori adeguata. Per questo motivo, in situazioni come quella appena descritta, possono essere utilizzate le cosiddette distribuzioni a priori oggettive, che non dipendono da osservazioni passate inerenti al caso di studio. Tra queste, negli ultimi venti anni, hanno preso sempre più importanza le *matching priors*, distribuzioni che rappresentano un ponte tra il mondo bayesiano e quello frequentista richiedendo un accordo tra l'inferenza bayesiana e quella frequentista.
In questa tesi si utilizzano tali distribuzioni per analizzare casi di studio in ambito medico, con lo scopo di proporre una distribuzione a priori alternativa a quelle già trattate in letteratura, che richiede l'elicitazione sul solo parametro di interesse.
La tesi è strutturata nel seguente modo. Nel primo capitolo viene sinteticamente descritto l'approccio bayesiano, soffermandosi sulla descrizione del Teorema di Bayes e della distribuzione a posteriori, sulla caratterizzazione delle diverse distribuzioni a priori e in particolare sulla definizione delle *matching priors*. Ci si è soffermati infine sulla misura di evidenza di Pereira-Stern, una misura di evidenza bayesiana utilizzata per problemi di verifica di ipotesi, valida per qualsiasi distribuzione a priori utilizzata. Nel secondo capitolo vengono illustrati alcuni casi di studio trattati in Biostatistica. In partico-

lare, viene analizzata l'inferenza su una proporzione, su due proporzioni, sulla media della normale, su un parametro scalare di regressione e sull'area sotto la curva ROC. In ogni caso considerato, la proposta bayesiana con *matching prior* viene confrontata con l'usuale approccio frequentista e l'eventuale proposta bayesiana presenti in letteratura. In particolare, con riferimento a quest'ultime, per il caso dell'inferenza su una proporzione, l'approccio proposto consiste nella specificazione di una distribuzione a priori Beta, stessa distribuzione a priori che propongono per l'inferenza su due proporzioni. Per la media di una normale caratterizzano una *JZS prior*, una distribuzione a priori che prevede la specificazione di una distribuzione uniforme per σ^2 e di una *Cauchy* per l'effetto, ovvero per $\delta = \mu/\sigma$. Per i parametri di un modello di regressione lineare propongono la *g-prior*, che riadattata, viene utilizzata anche come a priori per i parametri del modello di regressione logistico. Infine, per l'inferenza sull'area sotto la curva ROC, a seconda delle distribuzioni messe a confronto, viene specificata una a priori Gamma nel caso di due esponenziali e una Jeffreys nel caso di due normali. Nel terzo capitolo infine, si svolge un'applicazione di tutti questi casi di studio a dati reali. Per tale applicazione viene utilizzato il linguaggio di programmazione `R`, il quale, tramite la libreria `likelihoodAsy`, permette la facile implementazione delle *matching priors*.

Capitolo 1

L'approccio bayesiano

La probabilità è un concetto apparentemente semplice da capire, almeno dal punto di vista intuitivo. La sua formalizzazione, tuttavia, è più articolata: gli studiosi, per questo, si sono divisi in diverse scuole di pensiero. Alcuni ritengono che, per rendere in modo scientifico la probabilità, si debba depurarla di tutti gli elementi soggettivi che la riguardano e la caratterizzano; altri invece definiscono che proprio questi elementi soggettivi siano fondamentali per la sua definizione e ne fanno il punto di partenza. In particolare, l'inferenza bayesiana basa le sue idee sulla concezione soggettiva della probabilità, mentre l'inferenza frequentista su quella oggettiva (Cifarelli e Muliere, 1989; Liseo, 2008; Bolstad e Curran, 2016).
L'inferenza frequentista, per formalizzare tale concezione, si avvale di due principi fondamentali: quello della verosimiglianza e quello della ripetizione dell'esperimento. Il primo definisce che tutte le informazioni fornite da un generico campione sono contenute nella funzione di verosimiglianza, la quale misura la plausibilità delle possibili rappresentazioni del fenomeno. Il secondo ritiene il campione che si va ad analizzare uno dei possibili campioni che si sarebbero potuti ricavare ripetendo un gran numero di volte, nelle stesse condizioni, l'operazione di campionamento. L'inferenza bayesiana invece, per formalizzare la sua idea di probabilità, utilizza il Teorema di Bayes, rappresentandola quindi come il grado di fiducia che ogni individuo associa alla realizzazione di un dato evento.

Per quanto riguarda il punto di vista storico, l'approccio bayesiano sta

acquisendo un ruolo sempre più importante nella letteratura statistica (Liseo, 2008; Lee, 2012; Koch, 2007). Le ragioni di questa improvvisa accelerazione, iniziata più o meno negli anni '90 del secolo scorso, sono molteplici, ma riconducibili a tre categorie essenziali:

- *ragioni epistemologiche*: l'impostazione bayesiana dell'inferenza statistica formalizza in modo semplice e diretto il ragionamento induttivo di un individuo razionale che vuole calcolare la probabilità di eventi futuri;
- *ragioni pragmatiche*: nel corso degli anni sono via via aumentate le applicazioni statistiche in cui vi è l'esigenza di tener conto di informazioni extra-sperimentali;
- *ragioni di natura computazionale*: l'enorme sviluppo di nuove metodologie computazionali, come ad esempio il metodo Monte Carlo, consente ormai di analizzare, all'interno di questa impostazione, modelli statistici estremamente complessi.

1.1 Il Teorema di Bayes e la distribuzione a posteriori

Al fine di presentare in modo semplice e chiaro le peculiarità di un approccio bayesiano, si descrive un semplice esempio di applicazione di un test diagnostico.
Si consideri un test diagnostico dicotomico Y_1, con in particolare gli esiti

$$\begin{cases} Y_1 = 1, & \text{qualora il test risulti positivo} \\ Y_1 = 0, & \text{qualora il test risulti negativo.} \end{cases}$$

Inoltre sia

$$\begin{cases} \theta = 1, & \text{il paziente ha veramente la malattia} \\ \theta = 0, & \text{il paziente non ha la malattia.} \end{cases}$$

Si supponga, quindi, di conoscere le seguenti caratteristiche del test:

$$\begin{cases} \mathcal{P}(Y_1 = 1|\theta = 0) = 0.40, & \text{probabilità di falsi positivi} \\ \mathcal{P}(Y_1 = 0|\theta = 1) = 0.05, & \text{probabilità di falsi negativi.} \end{cases}$$

Tali probabilità sono importanti in quanto rappresentano gli elementi cruciali

per definire se un test diagnostico abbia una buona performance, rappresentata da un buon compromesso tra *specificità* e *sensibilità*[1].
Infine sia: $\mathcal{P}(\theta = 1)$=0.7 la prevalenza[2] della malattia.
Si suppone di aver osservato $Y_1 = 1$, ovvero il test è risultato positivo.
Avremo quindi che la probabilità di essere affetti dalla malattia risulta

$$\mathcal{P}(\theta = 1|Y_1 = 1) = \frac{\mathcal{P}(\theta = 1, Y_1 = 1)}{\mathcal{P}(Y_1 = 1)} = \frac{\mathcal{P}(\theta = 1)\mathcal{P}(Y_1 = 1|\theta = 1)}{\mathcal{P}(Y_1 = 1)} = 0.847,$$

mentre quella di non essere affetti è pari a

$$\mathcal{P}(\theta = 0|Y_1 = 1) = \frac{\mathcal{P}(\theta = 0, Y_1 = 1)}{\mathcal{P}(Y_1 = 1)} = \frac{\mathcal{P}(\theta = 0)\mathcal{P}(Y_1 = 1|\theta = 0)}{\mathcal{P}(Y_1 = 1)} = 0.153,$$

con $\mathcal{P}(Y = 1) = \mathcal{P}(\theta = 1)\mathcal{P}(Y_1 = 1|\theta = 1) + \mathcal{P}(\theta = 0)\mathcal{P}(Y_1 = 1|\theta = 0)$.

Come si può osservare, conoscere il risultato del test $Y_1 = 1$ ha incrementato la probabilità di essere affetti dalla malattia da 0.70 a 0.847.

Successivamente, il paziente decide di sottoporsi ad un secondo test Y_2, più accurato, ossia tale che:

$$\begin{cases} \mathcal{P}(Y_2 = 1|\theta = 0) = 0.04, & \text{probabilità di falsi positivi} \\ \mathcal{P}(Y_2 = 0|\theta = 1) = 0.01, & \text{probabilità di falsi negativi.} \end{cases}$$

Si supponga, quindi, di osservare $Y_2 = 0$, ovvero il risultato del secondo test diagnostico è risultato negativo. Definendo $Y = (Y_1, Y_2)$ si avrà:
$\mathcal{P}(\theta = 1|Y) \propto \mathcal{P}(Y_2 = 0|\theta = 1)\mathcal{P}(\theta = 1|Y_1 = 1) = 0.01 \times 0.847 = 0.0085$
$\mathcal{P}(\theta = 0|Y) \propto \mathcal{P}(Y_2 = 0|\theta = 0)\mathcal{P}(\theta = 0|Y_1 = 1) = 0.961 \times 0.153 = 0.1468.$

[1] Si definisce *specificità* di un test diagnostico la capacità di identificare correttamente i soggetti sani, ovvero non affetti dalla malattia o dalla condizione che ci si propone di individuare. Se un test ha un'ottima specificità, allora è basso il rischio di falsi positivi, cioè di soggetti che pur presentando valori anomali non sono affetti dalla patologia che si sta ricercando. Alternativamente, si definisce *sensibilità* di un test diagnostico la capacità di identificare correttamente i soggetti ammalati, ovvero affetti dalla malattia o dalla condizione che ci si propone di individuare. Se un test ha un'ottima sensibilità, allora è basso il rischio di falsi negativi, cioè di soggetti che pur presentando valori normali sono comunque affetti dalla patologia o dalla condizione che si sta ricercando.

[2] La prevalenza di una malattia rappresenta la proporzione di casi presenti a un dato istante nella popolazione di studio.

e infine
$\mathcal{P}(\theta = 1|Y) = 0.0085/(0.0085 + 0.1468) = 0.0547.$

Sulla base di quanto appena descritto, si nota che le opinioni si sono aggiornate condizionatamente a quello che si è osservato in precedenza, in questo modo:

$$\begin{cases} \mathcal{P}(\theta = 1) = 0.700, & \text{prima di svolgere i test } Y_1 \text{ e } Y_2 \\ \mathcal{P}(\theta = 1|Y_1) = 0.847, & \text{prima di svolgere il test } Y_2 \text{ ma dopo } Y_1 \\ \mathcal{P}(\theta = 1|Y) = 0.0547, & \text{dopo aver svolto i test } Y_1 \text{ e } Y_2 \ . \end{cases}$$

In sintesi, lo strumento che permette di unire le informazioni che si hanno a priori sulla quantità di interesse con quelle derivanti dall'osservazione dei dati è il Teorema di Bayes. Tale teorema rappresenta il meccanismo attraverso il quale si giunge all'acquisizione di una nuova conoscenza, ovvero l'aggiornamento delle opinioni iniziali sulla base degli eventi osservati (Cifarelli e Muliere, 1989); si veda anche Liseo (2008), Lee (2012) e Koch (2007). Formalmente, considerando un evento condizionato A e un evento condizionante B, questo teorema permette di calcolare la probabilità di A dato B, come

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A)\mathcal{P}(B|A)}{\mathcal{P}(B)},$$

dove:

- $\mathcal{P}(A|B)$, *probabilità a posteriori*, rappresenta la probabilità dell'evento A dato il verificarsi dell'evento B;
- $\mathcal{P}(A)$, *probabilità a priori*, descrive la probabilità dell'evento A;
- $\mathcal{P}(B|A)$ caratterizza la probabilità di B dato il verificarsi di A;
- $\mathcal{P}(B)$ è un fattore di normalizzazione.

In molte applicazioni, in generale, si assume di avere una variabile causale $Y \sim p(y;\theta)$, dove $\theta \in \Theta \subseteq \mathbb{R}^d$, $d \geq 1$, è il parametro incognito. Viene estratto un campione casuale semplice di n osservazioni $y = (y_1, ..., y_n)$. Data una distribuzione a priori $\pi(\theta)$ per θ, il Teorema di Bayes fornisce

$$\pi(\theta|y) = \frac{\pi(\theta)\mathrm{L}(\theta;y)}{\int_\theta \pi(\theta)\mathrm{L}(\theta;y)d\theta},$$

dove $\int_\theta \pi(\theta)\mathrm{L}(\theta; y)d\theta$ è un fattore di normalizzazione dipendente unicamente dal campione e $\pi(\theta|y)$, $\pi(\theta)$ e $\mathrm{L}(\theta; y)$ rappresentano i tre ingredienti fondamentali dell'inferenza bayesiana (Liseo, 2008). In particolare:

- $\pi(\theta)$ è la *distribuzione a priori*: riflette l'informazione che il ricercatore ha sul fenomeno oggetto di studio prima ancora di aver raccolto i dati campionari; generalmente si basa sulle evidenze di letteratura o su informazioni già note allo sperimentatore in base alla sua esperienza precedente. In questo contesto, quindi, contrariamente a quanto accade nell'approccio frequentista, il parametro non è più un valore fisso ed incognito, ma diventa una variabile casuale con una propria distribuzione di probabilità.
- $\mathrm{L}(\theta; y)$ è la *funzione di verosimiglianza*: la probabilità che il ricercatore assegnerebbe ai dati osservati se conoscesse il parametro di interesse.
- $\pi(\theta|y)$ è la *distribuzione a posteriori*: dopo aver osservato il campione y il ricercatore desidera aggiornare le proprie aspettative su θ, ovvero desidera combinare le proprie conoscenze a priori con le nuove conoscenze derivate dall'osservazione dei dati.

In molte applicazioni, il parametro θ viene partizionato come $\theta = (\psi, \lambda)$, dove ψ rappresenta il parametro di interesse scalare e λ quello di disturbo $(d-1)$-dimensionale. In tal caso, la distribuzione a posteriori marginale per il parametro di interesse è determinata come

$$\pi_m(\psi|y) = \int \pi(\psi, \lambda|y)d\lambda.$$

1.2 Scelta della distribuzione iniziale

L'aspetto cruciale della statistica bayesiana è la scelta della distribuzione a priori. Di fatto, la distribuzione iniziale è l'ingrediente attraverso il quale 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 (Liseo, 2008). È inoltre importante sottolineare le problematiche computazionali che possono nascere con la scelta della distribuzione a priori. Qualora, infatti, non si utilizzino delle

strutture analitiche particolari, le cosiddette distribuzioni coniugate, ottenere distribuzioni a posteriori analiticamente in forma esplicita risulta difficile.

Gli studiosi bayesiani, nel determinare la distribuzione iniziale, si suddividono in due grandi scuole di pensiero, soggettiva e oggettiva:

- *soggettiva*: implica la scelta della distribuzione a priori che meglio rappresenta le convinzioni iniziali dello studioso. Dunque, per rappresentare lo stesso fenomeno, soggetti diversi potranno specificare sia le stesse distribuzioni, ma con valori dei parametri che le caratterizzano differenti, sia distribuzioni di probabilità completamente diverse.
- *oggettiva*: propone la determinazione e l'adozione di distribuzioni a priori di tipo convenzionale, derivabili sulla base delle proprietà matematiche del modello statistico utilizzato. Così, da un lato si cerca di conservare l'oggettività delle procedure inferenziali che dipendono in questo modo esclusivamente dal modello statistico prescelto; dall'altro l'uso di una legge iniziale consente comunque l'utilizzo del Teorema di Bayes e quindi di produrre conclusioni inferenziali tramite il linguaggio probabilistico.

1.2.1 Assegnazione diretta della distribuzione iniziale

In questo paragrafo vengono brevemente illustrati alcuni dei principali metodi per esprimere le opinioni iniziali intorno ad uno o più parametri. Si veda ad esempio Cifarelli e Muliere (1989), Koch (2007), Liseo (2008), Lee (2012) o Bolstad e Curran (2016).

Metodo dell'istogramma

Si supponga che l'insieme dei valori ammissibili di Θ sia un intervallo limitato. Si ripartisce tale intervallo in sub-intervalli I_1, $I_2, ...$ e si vuole determinare la massa di probabilità che compete ad ognuno di questi intervalli. Per determinare ciò si utilizza il meccanismo della scommessa. La quantità p_i rappresenterà quindi quanto siamo disposti a pagare per ricevere il compenso legato alla verifica dell'evento

$$\text{E}=\{\text{il valore di } \theta \text{ appartiene all'intervallo } I_i\}.$$

In questo modo si ottiene così un istogramma, e da qui non resta che descriverlo con una funzione più o meno regolare la quale rappresenterà la funzione di densità di θ.

Specificazione di una forma funzionale

Questo metodo consiste nel definire inizialmente una forma funzionale della densità $\pi(\theta)$ ed in seguito passare alla specificazione dei parametri da cui dipende (detti iper-parametri). Se, ad esempio, si sceglie come densità per rappresentare la nostra opinione iniziale una Beta(α, β), quello che si deve fare in seguito è fissare i parametri α e β che la caratterizzano. Per far ciò si può procedere in due modi:

- si esprimono soggettivamente i primi due momenti della distribuzione e poi si ricavano i parametri;
- si fissano due quantili, anzichè fissare i primi due momenti, e attraverso il meccanismo della scommessa si ricavano in seguito i due parametri.

1.2.2 Distribuzioni a priori coniugate al modello

Qualora gli studiosi non siano completamente all'oscuro rispetto a informazioni a priori sul fenomeno di interesse e stiano ricercando una distribuzione a posteriori facilmente determinabile in modo esplicito, uno strumento utile è dato dalle distribuzioni coniugate. Tali distribuzioni godono di un'importante proprietà: se la distribuzione iniziale appartiene ad una tale classe di distribuzioni anche la distribuzione finale vi appartiene, cioè ha la stessa forma funzionale e differisce solo per il valore di alcuni parametri.

Sia data la funzione di verosimiglianza $\mathrm{L}(\theta; y)$ e sia $\mathbf{D}$ una classe di densità a priori per θ. La famiglia di distribuzioni iniziali $\mathbf{D}$ è coniugata al modello se avviene che

$$\pi(\theta|y) \in \mathbf{D}.$$

Pertanto ciò che muta nella distribuzione a posteriori non è la forma della distribuzione ma solo i parametri. Per uno specchietto riassuntivo delle più note coniugazioni tra modelli statistici e distribuzioni iniziali si veda la Tabella 1.1 (Liseo, 2008).

Tabella 1.1: Principali distribuzioni coniugate

Modello	Distr. iniziale	Distr. finale	Notazione
$\text{Ber}(\theta)$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + k, \beta + n - k)$	k=numero di successi
$\text{N}(\mu, \sigma_0^2)$	$\text{N}(\mu_0, \tau^2)$	$\text{N}(\frac{\mu_0\sigma^2 + \bar{y}n\tau^2}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2})$	σ_0^2 noto
$\text{Po}(\theta)$	$\text{Gamma}(\lambda, \delta)$	$\text{Gamma}(\lambda + n\bar{y}, \delta + n)$	
$\text{Exp}(\theta)$	$\text{Gamma}(\lambda, \delta)$	$\text{Gamma}(\lambda + n, \delta + n\bar{y})$	
$\text{U}(0, \theta)$	$\text{Pa}(\alpha, \varepsilon)$	$\text{Pa}(\alpha + n, w)$	$w = \max(y_{(n)}, \varepsilon)$

1.2.3 Distribuzioni a priori non informative o improprie

L'idea delle distribuzioni iniziali non informative nasce dalla concezione di voler ottenere conclusioni oggettive pur seguendo un approccio bayesiano. Da questo deriva che tali distribuzioni debbano essere ricavate dal modello statistico scelto per rappresentare il fenomeno in esame e non sulla base delle informazioni che lo sperimentatore ha a priori sull'evento in analisi. Tali distribuzioni possono essere delle distribuzioni improprie. Con il termine di distribuzione di probabilità impropria si designa la distribuzione con "funzione di densità" il cui integrale, esteso a tutto il supporto, diverge (Cifarelli e Muliere, 1989). Questa caratteristica dovrebbe essere sufficiente a sconsigliare l'utilizzo di simili densità, ma vi sono tre motivazioni sostanziali che portano uno statistico ad adottarle per la specificazione delle sue densità iniziali:

- fornire quella o quelle distribuzioni iniziali che lasciano parlare i dati aggiungendo "scarsa informazione" a quella già fornita dal campione;
- caratterizzare quelle distribuzioni iniziali cui si può attribuire il ruolo di espressione dello "stato di ignoranza" in cui un ricercatore potrebbe dichiarare di trovarsi;
- approssimare una effettiva distribuzione di probabilità iniziale dopo aver appurato che, nella situazione concreta in esame, la funzione di verosimiglianza domina, ai fini inferenziali, l'opinione iniziale.

Jeffreys (1961) giustifica l'introduzione delle distribuzioni improprie nel caso che queste vogliano riprodurre la situazione di completa ignoranza, distinguendo il caso in cui il parametro di interesse possa assumere tutti i

valori dello spazio parametrico, da quello dove può assumere solo valori positivi, assumendo per entrambi i casi una dimensione scalare per il parametro θ.

θ può assumere valori da $-\infty$ a $+\infty$

In questo contesto la "situazione di ignoranza" può essere rappresentata come segue. Se si considerano due intervalli qualsiasi del tipo $(-\infty, \alpha)$ e $(\beta, +\infty)$, il soggetto non è in grado di stabilire quale dei due eventi $\theta \in (-\infty, \alpha)$ e $\theta \in (\beta, +\infty)$ sia il più probabile, ed è quindi più appropriato attribuire la stessa probabilità a questi due eventi. Una distribuzione a priori impropria suggerita, che soddisfi questa condizione, è $\pi(\theta) = C$, ovvero una distribuzione uniforme, con C rappresentante una costante.

θ assume solo valori positivi

Jeffreys (1961) ha suggerito di assumere per $\eta = \log \theta$ una distribuzione uniforme, come, ad esempio, quella definita nel caso precedente. Per il parametro di interesse θ avremo quindi

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

Anche in questo caso non risulta possibile fare i confronti tra $\mathcal{P}(0 < \theta < \alpha)$ e $\mathcal{P}(\beta < \theta < +\infty)$.

Se si considera ora il caso in cui $\theta \in (a, b)$, la distribuzione per il parametro di interesse risulta essere

$$\pi(\theta) = \frac{1}{(\theta - a)(b - \theta)}.$$

A priori di Jeffreys

In base al problema in analisi, sono stati proposti diversi metodi ad hoc per determinare le distribuzioni iniziali non informative. Il più utilizzato è quello proposto da Jeffreys (1961) che ha suggerito di scegliere come distribuzione non informativa per un parametro scalare la funzione:

$$\pi(\theta) \propto i(\theta)^{1/2}, \tag{1.1}$$

dove $i(\theta)$ rappresenta la matrice dell'informazione attesa di Fisher rispetto al parametro di interesse per una osservazione.
Tale distribuzione in questo caso rappresenta anche una *reference prior* in quanto massimizza la distanza tra le distribuzioni a priori e posteriori calcolata tramite la divergenza di Kullback-Leibler (Bernardo, 1979).
Qualora il parametro fosse multi-dimensionale, usualmente, si assume vi sia indipendenza tra le componenti di θ e quindi la distribuzione a priori congiunta sarà rappresentata dal prodotto delle distribuzioni a priori delle singole componenti. In alternativa si può considerare il determinante della matrice di informazione attesa. La distribuzione a priori sarà quindi definita come

$$\pi(\theta) \propto \sqrt{\det(i(\theta))}.$$

1.3 *Matching priors*

Le distribuzioni *matching priors*, sviluppate da Peers (1965) e Tibshirani (1989), appartengono alla classe delle distribuzioni a priori oggettive. Esse costituiscono un ponte tra l'inferenza bayesiana e quella frequentista in quanto vanno a determinare regioni di credibilità bayesiane con validità frequentista (Datta e Mukerjee, 2004).

Definizione 1 *Si definisce matching prior una distribuzione a priori per la quale, sotto un certo ordine di approssimazione, degli intervalli di credibilità a posteriori hanno copertura frequentista (Datta e Mukerjee, 2004).*

Detto in altri termini, l'intervallo di credibilità a posteriori derivante da una *matching prior* equivale, asintoticamente, all'intervallo di confidenza unilaterale ottenuto seguendo un approccio frequentista. Analiticamente, le *matching priors*, sono tali per cui

$$\mathcal{P}(\theta \leq \theta^{1-\alpha}|y) = \mathcal{P}_\theta(\theta^{1-\alpha} \geq \theta) + O(n^{-1}), \tag{1.2}$$

con $\mathcal{P}(\theta \leq \theta^{1-\alpha}|y) = 1-\alpha$, dove $\theta^{1-\alpha}$ rappresenta il quantile $(1-\alpha)$ della distribuzione a posteriori $\pi(\theta|y)$, $\mathcal{P}_\theta(\theta^{1-\alpha} \geq \theta) = 1-\alpha$ la probabilità per $p(y;\theta)$ e $O(n^{-1})$ definisce l'ordine di errore (Staicu e Reid, 2008).

Se si considera il caso in cui il parametro di interesse θ sia scalare, la *matching prior* è definita come la distribuzione a priori non informativa di Jeffreys (1.1) (Staicu e Reid, 2008).
In presenza di parametri di disturbo, per definire una distribuzione che soddisfi la (1.2), bisogna innanzitutto suddividere il caso in cui vi sia una parametrizzazione ortogonale o meno.

1.3.1 Parametrizzazione ortogonale

Definendo con ψ il parametro di interesse scalare e con λ quello di disturbo, Tibshirani (1989) e Nicolaou (1993) hanno dimostrato che, in presenza di parametrizzazione ortogonale[3], la distribuzione che soddisfa la (1.2) è la seguente

$$\pi_{mp}(\psi, \lambda) \propto g(\lambda) i_{\psi\psi}(\psi, \lambda)^{1/2},$$

dove $g(\lambda)$ rappresenta una funzione arbitraria positiva, mentre $i_{\psi\psi}(\psi, \lambda)$ definisce l'elemento di posizione (ψ, ψ) dell'informazione di Fisher $i(\psi, \lambda)$. Ricordando inoltre che in presenza di parametrizzazione ortogonale si ha che $\hat{\lambda}_\psi = \hat{\lambda} + O_p(n^{-1})$, dove $\hat{\lambda}_\psi$ rappresenta la stima di massima verosimiglianza di λ vincolata al valore di ψ, da cui deriva che $g(\hat{\lambda}_\psi) = g(\hat{\lambda}) + O_p(n^{-1})$, è possibile ottenere la distribuzione a priori in funzione solamente del parametro di interesse inserendo la stima di λ nella specificazione della distribuzione a priori. La *matching prior* per il parametro di interesse risulta quindi essere (Ventura *et al.*, 2009)

$$\pi_{mp}(\psi) \propto i_{\psi\psi}(\psi, \hat{\lambda}_\psi)^{1/2}. \tag{1.3}$$

La distribuzione a posteriori marginale, associata a questa distribuzione a priori, può essere espressa come

$$\pi_m(\psi|y) \propto i_{\psi\psi}(\psi, \hat{\lambda}_\psi)^{1/2} \mathrm{L}_{CA}(\psi), \tag{1.4}$$

[3]Partendo dalla partizione del parametro $\theta = (\psi, \lambda)$ si dice di essere in presenza di parametrizzazione ortogonale qualora i vettori *score* ℓ_ψ e ℓ_λ risultino incorrelati, o equivalentemente se $i_{\psi\lambda} = 0$, dove $i_{\psi\lambda}$ rappresenta la matrice di informazione di Fisher. La conseguenza principale dell'ortogonalità è che le stime di massima verosimiglianza $\hat{\psi}$ e $\hat{\lambda}$ risultano asintoticamente indipendenti.

dove $\mathrm{L}_{CA}(\psi)$ rappresenta la funzione di verosimiglianza condizionata approssimata, definita come $\mathrm{L}_{CA}(\psi) = \mathrm{L}_P(\psi)|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{-1/2}$. La funzione $\mathrm{L}_P(\psi)$=$\mathrm{L}(\psi, \hat{\lambda}_\psi)$ è la verosimiglianza profilo per il parametro di interesse e $j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$ raffigura l'elemento (λ, λ) della matrice di informazione osservata calcolata in $(\psi, \hat{\lambda}_\psi)$.

1.3.2 Parametrizzazione non ortogonale

La funzione di verosimiglianza condizionata approssimata $\mathrm{L}_{CA}(\psi)$ rappresenta uno strumento utile per l'eliminazione dei parametri di disturbo; tuttavia questa non è invariante rispetto a riparametrizzazioni che non alterano il parametro d'interesse e inoltre richiede una parametrizzazione ortogonale. La distribuzione *matching prior* per ψ, quando non si è in presenza di parametrizzazione ortogonale, è

$$\pi_{mp}(\psi) \propto i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi)^{1/2}, \tag{1.5}$$

dove $i_{\psi\psi.\lambda}(\psi, \lambda) = i_{\psi\psi}(\psi, \lambda) - i_{\psi\lambda}(\psi, \lambda) i_{\lambda\lambda}(\psi, \lambda)^{-1} i_{\lambda\psi}(\psi, \lambda)$ rappresenta l'informazione parziale per ψ e $i_{\psi\lambda}(\psi, \lambda)$, $i_{\lambda\lambda}(\psi, \lambda)$,$i_{\lambda\psi}(\psi, \lambda)$ rappresentano gli elementi della matrice dell'informazione attesa di Fisher. Si vedano Ventura *et al.* (2009) e Ventura e Reid (2014).
La distribuzione a posteriori marginale risulta

$$\pi_m(\psi|y) \propto i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi)^{1/2} \mathrm{L}_{MP}(\psi), \tag{1.6}$$

dove $\mathrm{L}_{MP}(\psi) = \mathrm{L}_P(\psi) M(\psi)$ è la verosimiglianza profilo modificata, con $M(\psi)$ opportuno termine di aggiustamento (Ventura *et al.*, 2009; Ventura *et al.*, 2013).

1.3.3 Approssimazione della distribuzione a posteriori con *matching prior*

Ventura *et al.* (2013) propongono un'approssimazione del secondo ordine per la distribuzione a posteriori marginale (1.6). Questa è data da

$$\pi_m(\psi|y) \dot{\propto} \exp\left(-\frac{1}{2} r_p^*(\psi)^2\right) \left|\frac{s_p(\psi)}{r_p(\psi)}\right|, \tag{1.7}$$

dove $\dot{\propto}$ indica l'equivalenza del secondo ordine, $s_p(\psi) = \frac{\ell'_p(\psi)}{j_p(\hat{\psi})^{\frac{1}{2}}}$ rappresenta il test score, con $\ell_p(\psi)$ log-verosimiglianza profilo per il parametro di interesse, $\ell'_p(\psi) = \frac{\partial \ell_p(\psi)}{\partial \psi}$ e $j_p(\psi) = -\frac{\partial \ell'_p(\psi)}{\partial \psi}$, $r_p(\psi) = sign(\hat{\psi} - \psi)2(\ell_p(\hat{\psi}) - \ell_p(\psi))^{1/2}$ e

$$r_p^* = r_p(\psi) + \frac{1}{r_p(\psi)} \log \frac{q(\psi)}{r_p(\psi)},$$

con

$$q(\psi) = \frac{\ell'_p(\psi)}{j_p(\hat{\psi})^{1/2}} \frac{i_{\psi\psi.\lambda}(\hat{\psi}, \hat{\lambda})^{1/2}}{i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi)^{1/2}} \frac{1}{M(\psi)}.$$

La statistica $r_p^*(\psi)$ rappresenta una modifica della statistica $r_p(\psi)$ derivante da un'approssimazione di ordine superiore della densità di probabilità dello stimatore di massima verosimiglianza. La sua approssimazione alla normale standard è più accurata; l'ordine dell'errore di approssimazione è infatti $O(n^{-3/2})$ (Brazzale *et al.*, 2007).

Il vantaggio di utilizzare l'espressione (1.7) sta nel fatto che questa include automaticamente la *matching prior* e non sono quindi necessarie operazioni analitiche o computazionali per la sua definizione esplicita.

Inoltre, poiché $r_p(\psi) = s_p(\psi) + o_p(1)$, la (1.7) al primo ordine di approssimazione diventa semplicemente

$$\pi_m(\psi|y) \dot{\propto} \exp\left(-\frac{1}{2} r_p^*(\psi)^2\right),$$

dove $\dot{\propto}$ indica l'equivalenza del secondo ordine.

Dall'espressione (1.7) si può derivare facilmente un'approssimazione dell'area sotto la coda. Questa è definita integrando direttamente $\pi_m(\psi|y)$:

$$\int_{-\infty}^{\psi_0} \pi_m(\psi|y) d\psi \equiv \Phi(r_p^*(\psi_0)), \tag{1.8}$$

dove " $\equiv$ " rappresenta l'equivalenza asintotica del terzo ordine e $\Phi(\cdot)$ è la funzione di ripartizione della normale standard (Ventura *et al.*, 2013).

Utilizzando tale approssimazione, si può derivare un intervallo di credibilità asintotico *equi-tailed* definito come

$$\text{CI} = \{\psi : |r_p^*(\psi)| \leq z_{1-\alpha/2}\},$$

dove $z_{1-\alpha/2}$ rappresenta il quantile $(1 - \alpha/2)$ della distribuzione normale standard. Tale intervallo di credibilità coincide con l'intervallo di confidenza

per ψ basato su $r_p^*(\psi)$ di livello approssimato $(1-\alpha)$ (Ventura *et al.*, 2013). Dall'approssimazione (1.8) è possibile ricavare facilmente anche la mediana a posteriori, definita come la soluzione $\hat{\psi}^*$ in ψ dell'equazione $r_p^*(\psi) = 0$ (Ventura *et al.*, 2013).

1.4 Selezione del modello

Per verificare

$$\mathrm{H}_0 : \theta \in \Theta_0 \qquad \text{contro} \qquad \mathrm{H}_1 : \theta \in \Theta_1,$$

con $\Theta_0 \cup \Theta_1 = \Theta$ e $\Theta_0 \cap \Theta_1 = \emptyset$, il confronto tra le due ipotesi avviene calcolando $\mathcal{P}(\mathrm{H}_i|y)$, con $i = 0, 1$, ovvero calcolando le probabilità a posteriori

$$\mathcal{P}(\mathrm{H}_i|y) = \int_{\Theta_i} \pi_i(\theta|y) d\theta,$$

dove $\pi_i(\theta|y)$ è la distribuzione a posteriori sotto H_i, $i = 0, 1$ (Liseo, 2008). Ad esempio, considerando θ scalare, per verificare le ipotesi

$$\mathrm{H}_0 : \theta = \theta_0 \qquad \text{contro} \qquad \mathrm{H}_1 : \theta = \theta_1,$$

si può utilizzare il rapporto

$$\frac{\mathcal{P}(\mathrm{H}_0|y)}{\mathcal{P}(\mathrm{H}_1|y)} = \frac{\pi(\theta_0|y)}{\pi(\theta_1|y)} = \frac{\pi_0}{1-\pi_0} \frac{\mathrm{L}(\theta_0; y)}{\mathrm{L}(\theta_1; y)}, \tag{1.9}$$

dove $\pi_0 = \mathcal{P}(\mathrm{H}_0)$ e $\pi_1 = 1 - \pi_0 = \mathcal{P}(\mathrm{H}_1)$ rappresentano le probabilità a priori rispettivamente sotto H_0 e H_1.
Tale rapporto, come si può notare, è composto da due quantità: $\frac{\pi_0}{1-\pi_0}$ che rappresenta il peso relativo delle due ipotesi prima di osservare i dati (opinione iniziale); $\frac{\mathrm{L}(\theta_0;y)}{\mathrm{L}(\theta_1;y)}$ che costituisce il rapporto di verosimiglianza, ovvero l'evidenza empirica. Quest'ultima quantità viene chiamata *fattore di Bayes* (BF) e rappresenta il fattore moltiplicativo che trasforma l'odds a priori in quello a posteriori; in altre parole, definisce una misura di quanto l'evidenza empirica aggiorni l'opinione iniziale. In sintesi, il *fattore di Bayes* nel caso di ipotesi nulla e alternativa semplici è definito come

$$\mathrm{B}_{01} = \frac{\mathrm{L}(\theta_0; y)}{\mathrm{L}(\theta_1; y)}.$$

Valori di B_{01} maggiori di 1 stanno ad indicare che i dati supportano θ_0, al contrario valori inferiori di 1 supportano θ_1. Kass e Raftery (1995) propongono una trasformazione logaritmica del *fattore di Bayes*: $2\log(B_{01})$. Tale trasformazione permette di leggere il *fattore di Bayes* sulla stessa base del rapporto di verosimiglianza dell'approccio frequentista. Si veda la Tabella 1.2 per avere un'idea di come possono essere interpretati i valori.
Quando l'ipotesi alternativa e nulla sono composite, la (1.9) diventa

$$\frac{\mathcal{P}(H_0|y)}{\mathcal{P}(H_1|y)} = \frac{\pi_0}{1-\pi_0}\frac{\int_{\Theta_0} L(\theta;y)g_0(\theta)d\theta}{\int_{\Theta_1} L(\theta;y)g_1(\theta)d\theta},$$

con $\pi_0 = \mathcal{P}(H_0)$, $\pi_1 = 1-\pi_0 = \mathcal{P}(H_1)$ e considerando la specificazione della

Tabella 1.2: Interpretazione del *fattore di Bayes*

B_{01}	$2\log B_{01}$	Evidenza contro H_1
$1-3$	$0-2$	Debole
$3-20$	$2-6$	Sostanziale
$20-150$	$6-10$	Forte
>150	>10	Molto forte

distribuzione a priori per θ nel seguente modo: $\pi(\theta) = \pi_i g_i(\theta)$, dove $g_i(\theta)$ corrisponde alla densità a priori per θ sotto le due ipotesi considerate H_i, $i = 0, 1$.
Il *fattore di Bayes* in questo caso è rappresentato dal rapporto tra le due densità marginali ottenute rispettivamente sotto l'ipotesi nulla e quella alternativa. In sintesi, il *fattore di Bayes* nel caso di ipotesi alternativa e nulla composite è definito come

$$B_{01} = \frac{\int_{\Theta_0} L(\theta;y)g_0(\theta)d\theta}{\int_{\Theta_1} L(\theta;y)g_1(\theta)d\theta}.$$

Risulta importante sottolineare che, per come è definito, il *fattore di Bayes* non è determinabile in presenza di distribuzioni a priori improprie. Si ricorda infatti che le distribuzioni improprie integrano ad infinito. Per questo motivo, la specificazione della distribuzione a priori gioca un ruolo molto rilevante nella determinazione del *fattore di Bayes*.

1.4.1 Misura di evidenza di Pereira-Stern

Pereira e Stern (1999) hanno introdotto una misura di evidenza bayesiana in favore di un'ipotesi H_0 per superare il problema del paradosso Jeffreys-Lindley[4]. La misura di evidenza di Pereira-Stern infatti è valida qualsiasi sia la distribuzione a priori utilizzata. Questa misura, supponendo di voler verificare

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0, \end{cases}$$

comprende tutti i punti dello spazio parametrico la cui densità a posteriori è minore dell'estremo superiore calcolato in Θ_0 (Madruga *et al.*, 2001). Essa è definita come (si veda anche la Figura 1.1 come esempio considerando θ scalare)

$$\mathrm{EV} = 1 - \mathcal{P}(\theta \in \mathrm{T}(y)|y), \tag{1.10}$$

dove $\mathrm{T}(y) = \{\theta : \pi(\theta|y) > \sup_{\Theta_0}\pi(\theta|y)\}$.
In particolare, nel caso in cui la moda a posteriori risulti maggiore di θ_0, si ha

$$\mathrm{EV} = \int_{-\infty}^{\theta_0} \pi(\theta|y)d\theta + 1 - \int_{-\infty}^{\theta_1} \pi(\theta|y)d\theta. \tag{1.11}$$

In caso di distribuzione a posteriori simmetrica invece, si nota che le due aree nelle code sono uguali. Conseguentemente la (1.11) diventa

$$\mathrm{EV} = 2\int_{-\infty}^{\theta_0} \pi(\theta|y)d\theta. \tag{1.12}$$

Un valore elevato di questa misura implica che θ_0 si trova in una regione di Θ ad alta probabilità, dunque sta a significare che i dati supportano l'ipotesi nulla; d'altro canto valori piccoli portano evidenza a favore di quella alternativa.
Con la Figura 1.1 viene fornita una rappresentazione grafica della (1.10). Tale grafico pone in evidenza tutte le parti che caratterizzano la misura di evidenza di Pereira-Stern; in particolare $\mathrm{T}(y)$ è rappresentato dalla porzione di ascissa evidenziata, compresa tra θ_0 e θ_1, si ricordi infatti che rappresenta

[4]Il paradosso di Jeffreys-Lindley definisce che utilizzando distribuzioni a priori non informative per il parametro di interesse, il Fattore di Bayes porta ad accettare quasi sempre il modello specificato sotto l'ipotesi nulla.

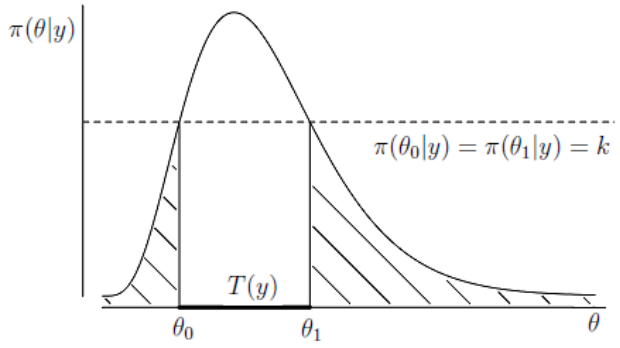

Figura 1.1: Misura di evidenza EV di Pereira e Stern: parte tratteggiata

l'insieme dei θ tale per cui $\pi(\theta|y) > \sup_{\Theta_0}\pi(\theta|y)$.

Nel caso in cui il parametro di interesse ψ sia scalare, se si utilizza come distribuzione a priori una *matching prior*, utilizzando la (1.8), si può ottenere l'approssimazione (Cabras *et al.*, 2016; Ventura e Reid, 2014)

$$\text{EV} \equiv 1 - \Phi(r_p^*(\psi_1)) + \Phi(r_p^*(\psi_0)), \tag{1.13}$$

assumendo la moda a posteriori maggiore di ψ_0. In caso contrario si avrà $\text{EV} \equiv 1 - \Phi(r_p^*(\psi_0)) + \Phi(r_p^*(\psi_1))$. Quest'approssimazione risulta essere utile in quanto non richiede la specificazione della distribuzione a priori per il parametro di disturbo.
L'approssimazione in caso di distribuzione a posteriori simmetrica è

$$\text{EV} \equiv 2\Phi(r_p^*(\psi_0)), \tag{1.14}$$

che coincide quindi con il $p - value$ basato su $r_p^*(\psi)$.
Si ritiene infine interessante sottolineare una proprietà di questa misura, ossia risulta essere una misura calibrata. Sotto l'ipotesi nulla dunque si distribuisce come un'uniforme (Ventura *et al.*, 2013).

1.5 Conclusioni

In questo capitolo è stato sinteticamente descritto l'approccio bayesiano, soffermandosi sulla descrizione del Teorema di Bayes e della distribuzione a

posteriori, sulla caratterizzazione delle diverse distribuzioni a priori e in particolare sulla definizione delle *matching priors*. Ci si è soffermati infine sulla descrizione dell'inferenza bayesiana per la verifica di ipotesi. Ciò è stato svolto con l'obiettivo di fornire gli strumenti di base che verranno poi utilizzati nel seguito della tesi.
Nel Capitolo 2 vedremo alcuni esempi di casi di studio in Biostatistica confrontando l'approccio frequentista, che usualmente viene utilizzato per risolverli, con un possibile approccio bayesiano proposto in letteratura e quello bayesiano con *matching prior* che si propone con questa tesi.

Capitolo 2

Casi di studio in Biostatistica

Con il termine studio clinico si intende, in generale, un esperimento rigorosamente controllato ed eticamente progettato nel quale i soggetti sottoposti ad analisi, sono assegnati alle diverse modalità di intervento in modo casuale e nello stesso intervallo di tempo programmato per lo studio. Nella pratica, comunque, non sempre uno studio clinico ha carattere sperimentale in senso stretto (Ventura e Racugno, 2017).
In particolare, uno studio clinico può essere:

- *Descrittivo*, qualora lo scopo primario dell'analisi sia rappresentato nel descrivere come si distribuiscono certe variabili senza valutare l'associazione tra queste, formulando quindi ipotesi di interesse in merito al caso di studio;
- *Analitico*, qualora lo scopo sia quello di rispondere ad ipotesi formulate a priori sul fenomeno di interesse.

A sua volta, lo studio analitico si suddivide in sperimentale e osservazionale. Con il termine sperimentale si intende un esperimento programmato che riguarda soggetti caratterizzati da una particolarità specifica e per la quale si vuole determinare il trattamento più appropriato. Nello specifico, in questi tipi di studio, si crea una situazione tale per cui solo uno dei fattori (il trattamento) che influenza l'oggetto di interesse può variare; e dunque, tranne in casi di variazioni casuali, i soggetti in analisi differiscono volutamente solo per il trattamento e per nessun altro fattore (Santos, 1999). I principali tipi

di studio sperimentale sono: *Randomized Clinical trial*, *Field trial*, *cluster randomized trials* e *trial di comunità*.
Per quanto riguarda lo studio analitico osservazionale invece, esso rappresenta un'indagine nella quale non avviene alcuna manipolazione nè dei soggetti nè delle variabili di interesse; in questo quindi, il ricercatore non determina l'assegnazione dei soggetti a ciascun gruppo in osservazione, ma si limita ad osservare ciò che accade in diversi tempi della sperimentazione. Questo tipo di studio è generalmente più diffuso soprattutto per considerazioni di natura etica. Qualora l'analisi fosse concentrata su un fattore di rischio di una malattia anzichè su un trattamento, è evidente infatti che non si possono esporre al rischio dei soggetti, ma si devono considerare soggetti già esposti e quindi la randomizzazione non è dunque possibile.
Rispetto al momento in cui vi è l'identificazione della malattia e l'esposizione al rischio di riscontrarla, gli studi osservazionali si suddividono in prospettici e retrospettivi:

- *Prospettico*: studio nel quale si considerano due campioni di individui rappresentati gli esposti e i non esposti, sui quali viene rilevato il numero dei malati e dei non malati dopo un certo intervallo di tempo prefissato. Lo scopo di questo studio sarà quindi valutare la differenza tra la frequenza dei malati negli esposti e non. In questo tipo di studio dunque avviene prima l'esposizione al rischio e poi l'identificazione della malattia.
- *Retrospettivo*: studio nel quale si considerano due campioni di individui rappresentanti i malati e i sani, sui quali si andrà a calcolare il numero degli esposti e dei non esposti al fattore di rischio che si sta analizzando. Lo scopo di questo studio sarà quindi valutare la differenza tra la frequenza degli esposti nei malati e non. In questo tipo di studio quindi avviene prima l'identificazione della malattia e in seguito l'esposizione al rischio di riscontrarla.

Rispetto invece ad alcune caratteristiche dello studio stesso, lo studio osservazionale si suddivide in trasversale, di coorte, caso-controllo ed ecologico.

- *Trasversale*: studio nel quale i soggetti da analizzare vengono selezionati in base alla loro patologia riscontrata. Fissato un istante temporale, si

rileva la frequenza dei fenomeni di salute-malattia e dei fattori a essi correlati, permettendo quindi il calcolo della prevalenza della malattia ma non dell'incidenza[1] di questa.

- *Di coorte*: in questo studio i soggetti da analizzare vengono selezionati in base al valore della variabile di esposizione, individuando in tale modo le cosiddette coorti. Differenti gruppi di esposti sono quindi seguiti per studiare l'occorrenza della malattia. Sulla base di quanto appena delineato, si può comprendere che uno studio di coorte è uno studio prospettico.
- *Caso-controllo*: questo studio prevede la suddivisione dei soggetti sulla base della presenza (casi) o assenza (controlli) di una malattia. L'operazione di campionamento dei soggetti porta ad una maggiore efficienza ma rappresenta anche una fonte di distorsione. Inoltre il fatto di campionare implica che in questo studio si possono ottenere solamente misure relative dell'effetto e non assolute, consentendo di calcolare solamente il rischio relativo e il rapporto tra quote. Per questi motivi, uno studio di questo tipo è tipicamente retrospettivo.
- *Ecologico*: in questo studio le unità di osservazione sono rappresentate da gruppi di persone anzichè da individui, fornendo dati in forma aggregata. Analizzare dati aggregati può portare alla cosiddetta *Fallacia ecologica*, ovvero alla possibilità di rilevare relazioni non veritiere dal momento che i dati costruiti in questo modo non permettono di identificare eventuali fattori confondenti.

Di seguito verranno presentati alcuni casi di studio classici, andando a confrontare l'usuale approccio frequentista legato ad essi, alcune proposte bayesiane note in letteratura e l'approccio bayesiano basato su *matching priors*, che rappresenta il contributo di questa tesi.

[1] L'incidenza di una malattia è identificata come il rapporto tra il numero di casi insorti nel periodo in analisi e la "massa" a rischio, dove con "massa" si intende la somma dei tempi di osservazione di ciascun soggetto fra la sua entrata nello studio e la sua uscita dal gruppo di candidati alla malattia.

2.1 Test su una proporzione

Sia $y = (y_1, ..., y_n)$ un campione casuale semplice estratto da una variabile casuale Y con distribuzione di Bernoulli di parametro θ, ovvero $Y \sim \text{Ber}(\theta)$, con $Y = 1$ qualora il paziente abbia riscontrato la malattia, $Y = 0$ in caso contrario. La verosimiglianza per il parametro θ è

$$\begin{aligned} \text{L}(\theta; y) &\propto \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{(1-y_i)} \propto \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{\sum_{i=1}^{n} 1-y_i} \\ &\propto \theta^k(1-\theta)^{n-k}, \end{aligned}$$

con $k = \sum_{i=1}^{n} y_i$ pari al numero dei successi.
Si supponga di essere interessati a valutare il seguente sistema di ipotesi:

$$\begin{cases} \text{H}_0 : \theta = \theta_0 \\ \text{H}_1 : \theta \neq \theta_0, \end{cases}$$

dove θ rappresenta la proporzione di pazienti che hanno riscontrato una certa malattia. Tale test, utilizzato spesso per definire la prevalenza di una malattia, riguarda la verifica di ipotesi su una proporzione.
Per verificare tale ipotesi seguendo un approccio frequentista, l'usuale statistica test, nel caso in cui n sia sufficientemente elevato, è

$$z = \frac{\bar{Y} - \theta_0}{\sqrt{(\theta_0(1-\theta_0)/n)}},$$

che ha distribuzione approssimata $\text{N}(0, 1)$ sotto H_0, con $\bar{Y} = \sum_{i=1}^{n} Y_i/n$. Tale approssimazione viene considerata buona qualora le quantità $n\theta_0$ e $n(1-\theta_0)$ risultino maggiori o uguali a 5. Nel caso non lo fossero, si può utilizzare la statistica test

$$\text{T} = \sum_{i=1}^{n} Y_i = n\bar{Y},$$

dove sotto l'ipotesi nulla si ha che $\text{T} \sim \text{Bin}(n, \theta_0)$.

Considerando un approccio bayesiano al problema, si propone come distribuzione a priori per il parametro di interesse θ una *matching prior*. In

presenza di parametro scalare e in assenza di parametri di disturbo, la *matching prior* coincide con la Jeffreys, ossia $\pi(\theta) \propto i(\theta)^{1/2}$. Nel caso in analisi si ha

$$\pi(\theta) \propto \left(\frac{1}{\theta(1-\theta)}\right)^{1/2},$$

che corrisponde ad una distribuzione Beta$(1/2, 1/2)$, detta distribuzione arcoseno, una particolare coniugata della Bernoulli.
La posteriori risulta

$$\begin{aligned}\pi(\theta|y) \propto \pi(\theta)\mathrm{L}(\theta; y) &= \theta^{\sum_{i=1}^n y_i - 1/2}(1-\theta)^{\sum_{i=1}^n (1-y_i)-1/2} \\ &= \theta^{k-1/2}(1-\theta)^{n-k-1/2},\end{aligned}$$

con $k = \sum_{i=1}^n y_i$ pari al numero dei successi, che è una Beta$(k + 1/2, n - k + 1/2)$.
Con l'obiettivo di rispondere al sistema di ipotesi, si può calcolare la misura di evidenza di Pereira-Stern. Con la *matching prior*, considerando l'approssimazione della distribuzione a posteriori descritta nella Sezione 1.3.3, si ha $\mathrm{EV} \equiv 1 - \Phi(r^*(\theta_1)) + \Phi(r^*(\theta_0))$ se la moda a posteriori risulta maggiore di θ_0.
La statistica $r^*(\theta)$ è definita come (Brazzale *et al.*, 2007)

$$r^*(\theta) = r(\theta) + \frac{1}{r(\theta)} \log\left(\frac{q(\theta)}{r(\theta)}\right),$$

con

$$\begin{aligned} r(\theta) &= sign(\hat{\theta} - \theta)\left[2(\ell(\hat{\theta}) - \ell(\theta))^{1/2}\right] \\ &= sign(\hat{\theta} - \theta)\left[2\left(\sum_{i=1}^n y_i \log\left(\frac{\hat{\theta}}{\theta}\right) + \sum_{i=1}^n (1-y_i)\log\left(\frac{1-\hat{\theta}}{1-\theta}\right)\right)^{1/2}\right], \end{aligned}$$

$\hat{\theta} = \bar{y}$ e con

$$q(\theta) = \frac{\ell'(\theta)}{(j(\hat{\theta}))^{1/2}} = \frac{\frac{\sum_{i=1}^n y_i}{\theta} - \frac{\sum_{i=1}^n (1-y_i)}{1-\theta}}{\left(\frac{\sum_{i=1}^n y_i}{\hat{\theta}^2} + \frac{\sum_{i=1}^n (1-y_i)}{(1-\hat{\theta})^2}\right)^{1/2}}.$$

Inoltre, dal momento che in questo caso la distribuzione a posteriori ottenuta è una Beta, la misura di evidenza può essere calcolata esattamente dalla Beta

tramite la (1.11).

In letteratura sono state proposte altre distribuzioni a priori per il parametro di interesse. In particolare, caratterizzano come distribuzione del parametro una Beta(α, β) dal momento che questa risulta essere una distribuzione coniugata alla Binomiale, e quindi la posteriori sulla quale si basa l'inferenza è facilmente determinabile. I parametri di tale distribuzione, α e β, vengono definiti iper-parametri e devono essere fissati. Considerando quindi come distribuzione a priori una Beta(α, β) e come modello una Ber(θ), la distribuzione a posteriori risulta essere una Beta$(\alpha + k, \beta + n - k)$, dove $k = \sum_{i=1}^{n} y_i$ rappresenta il numero dei successi, ovvero il numero delle volte in cui si è riscontrata la presenza della malattia.

2.2 Test su due proporzioni

Si supponga ora di non essere interessati unicamente alla prevalenza della malattia, ma di voler verificare anche come questa sia distribuita all'interno di due popolazioni diverse e indipendenti tra di loro. Si tratta dunque il problema di inferenza su due proporzioni. Si considerino due variabili casuali indipendenti distribuite rispettivamente come due binomiali: $X_1 \sim \text{Bin}(n_1, \theta_1)$ e $X_2 \sim \text{Bin}(n_2, \theta_2)$. La verosimiglianza per (θ_1, θ_2) risulta essere

$$\text{L}(\theta_1, \theta_2; x_1, x_2) \propto \binom{n_1}{x_1} \theta_1^{x_1} (1-\theta_1)^{n_1 - x_1} \binom{n_2}{x_2} \theta_2^{x_2} (1-\theta_2)^{n_2 - x_2}.$$

Si supponga quindi di essere interessati al seguente sistema di ipotesi:

$$\begin{cases} \text{H}_0 : \theta_1 = \theta_2 \\ \text{H}_1 : \theta_1 \neq \theta_2. \end{cases}$$

Seguendo un approccio frequentista al problema, la statistica test per verificare l'omogeneità delle popolazioni è

$$z = \frac{\hat{\theta_1} - \hat{\theta_2}}{\sqrt{\hat{\theta}(1-\hat{\theta})(\frac{1}{n_2} + \frac{1}{n_2})}},$$

dove $\hat{\theta_1}$, $\hat{\theta_2}$ e $\hat{\theta}$ rappresentano le stime rispettivamente di θ_1, θ_2 e θ, ovvero $\hat{\theta_1} = \frac{x_1}{n_1}$, $\hat{\theta_2} = \frac{x_2}{n_2}$ e $\hat{\theta} = \frac{x_1+x_2}{n_1+n_2}$. La distribuzione sotto l'ipotesi nulla di questa statistica, per n_1 e n_2 sufficientemente grandi, è approssimativamente una normale standard.

Come nel caso trattato nel paragrafo precedente, per risolvere questo tipo di problema seguendo un approccio bayesiano, si propone come distribuzione a priori per il parametro di interesse, che in questo caso di studio è rappresentato da $\psi = (\theta_1 - \theta_2)$, una *matching prior*. Definendo come parametro di disturbo $\lambda = \theta_2$, la log-verosimiglianza per (ψ, λ) risulta essere

$$\begin{aligned}\ell(\psi, \lambda) =& x_1 \log(\psi + \lambda) + (n_1 - x_1) \log(1 - \psi - \lambda) + x_2 \log(\lambda) + \\ & (n_2 - x_2) \log(1 - \lambda),\end{aligned}$$

e la log-verosimiglianza profilo per ψ è pari a

$$\begin{aligned}\ell_p(\psi) =& x_1 \log(\psi + \hat{\lambda}_\psi) + (n_1 - x_1) \log(1 - \psi - \hat{\lambda}_\psi) + x_2 \log(\hat{\lambda}_\psi) + \\ & (n_2 - x_2) \log(1 - \hat{\lambda}_\psi),\end{aligned}$$

dove $\hat{\lambda}_\psi$ rappresenta la stima del parametro di disturbo vincolata al valore di ψ.
Infine, la matrice di informazione osservata calcolata in $(\psi, \hat{\lambda}_\psi)$ è:

$$j(\psi, \hat{\lambda}_\psi) = \begin{bmatrix} \frac{x_1}{(\psi+\hat{\lambda}_\psi)^2} + \frac{n_1-x_1}{(1-\psi-\hat{\lambda}_\psi)^2} & \frac{x_1}{(\psi+\hat{\lambda}_\psi)^2} + \frac{n_1-x_1}{(1-\psi-\hat{\lambda}_\psi)^2} \\ \frac{x_1}{(\psi+\hat{\lambda}_\psi)^2} + \frac{n_1-x_1}{(1-\psi-\hat{\lambda}_\psi)^2} & \frac{x_1}{(\psi+\hat{\lambda}_\psi)^2} + \frac{n_1-x_1}{(1-\psi-\hat{\lambda}_\psi)^2} + \frac{x_2}{\hat{\lambda}_\psi^2} + \frac{n_2-x_2}{(1-\hat{\lambda}_\psi)^2} \end{bmatrix}.$$

Definiti questi elementi, tramite la (1.5) si può ricavare la *matching prior* per ψ, mentre è possibile avere la specificazione della distribuzione a posteriori marginale considerando la (1.6). Infine grazie alla (1.7) si può ottenere un'approssimazione del secondo ordine per la distribuzione a posteriori. Facendo riferimento alla (1.13) si ottiene la misura di evidenza di Pereira-Stern, per rispondere al sistema di ipotesi trattato.

È inoltre possibile considerare un approccio bayesiano per il caso in analisi utilizzando una distribuzione a priori Beta, coniugata alla distribuzione bernoulliana. In particolare, dal momento che i parametri ignoti sono due (θ_1, θ_2),

la distribuzione a priori sarà data dalla congiunta delle due distribuzioni Beta specificate rispettivamente per ciascun parametro ignoto:

$$\pi(\theta_1, \theta_2) = \frac{\theta_1^{\alpha-1}(1-\theta_1)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \frac{\theta_2^{\gamma-1}(1-\theta_2)^{\delta-1}}{\mathrm{B}(\gamma,\delta)},$$

con $\mathrm{B}(\alpha,\beta) = \int_0^1 \theta_1^{\alpha-1}(1-\theta_1)^{\beta-1}d\theta_1$ e $\mathrm{B}(\gamma,\delta) = \int_0^1 \theta_2^{\gamma-1}(1-\theta_2)^{\delta-1}d\theta_2$, rispettivamente.
La distribuzione a posteriori risulta essere la congiunta di due Beta: la prima $\mathrm{Beta}(\alpha+k, \beta+n_1-k)$, la seconda $\mathrm{Beta}(\gamma+w, \delta+n_2-w)$, dove k e w rappresentano il numero dei successi per le due distribuzioni Beta rispettivamente. Considerando che il parametro di interesse per questo caso di studio è rappresentato da $\psi = \theta_1 - \theta_2$, è possibile ottenere la distribuzione a posteriori per quest'ultimo attraverso la trasformazione $\psi = \theta_1 - \theta_2$, $\lambda = \theta_2$ e l'inversa $\theta_1 = \psi + \lambda$, $\theta_2 = \lambda$. Lo jacobiano della trasformazione risulta

$$|J_Q(\psi,\lambda)| = \det \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = 1.$$

Conseguentemente la distribuzione a posteriori per (ψ, λ) è

$$\pi(\psi,\lambda|x_1,x_2) = (\psi-\lambda)^{\alpha+x_1}(1-\psi+\lambda)^{\beta+n_1-x_1}\lambda^{\gamma+x_2}(1-\lambda)^{\gamma+n_2-x_2}.$$

Integrando questa distribuzione rispetto al parametro di disturbo λ è possibile ottenere la distribuzione a posteriori marginale per il parametro di interesse.
Anche in questo caso si può ottenere la misura di evidenza di Pereira-Stern in risposta al sistema di ipotesi considerato inizialmente, considerando la (1.11).

2.3 T-test a un campione

Si consideri $y = (y_1, ..., y_n)$ un campione casuale semplice estratto da una variabile casuale Y con distribuzione normale, di media μ e varianza σ^2, ovvero $Y \sim \mathrm{N}(\mu, \sigma^2)$. La funzione di verosimiglianza per (μ, σ^2) risulta essere

$$\mathrm{L}(\mu,\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}.$$

Si supponga di essere interessati al seguente sistema di ipotesi:

$$\begin{cases} \mathrm{H}_0 : \mu = \mu_0 \\ \mathrm{H}_1 : \mu \neq \mu_0. \end{cases}$$

Iniziando con un approccio frequentista, per rispondere al problema in analisi, tale sistema di ipotesi può essere ricondotto ad un *t-test*. Il *t-test* è un test statistico che va a valutare se il valore medio di una distribuzione si discosta significativamente da un certo valore di riferimento. Tale test è dato da

$$t = \frac{|\bar{Y} - \mu_0|}{\sqrt{\frac{S^2}{n}}},$$

dove S^2 rappresenta la varianza campionaria corretta, definita come $S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$. La statistica test t, sotto l'ipotesi nulla, segue una distribuzione t_{n-1}.

Volendo rispondere al sistema di ipotesi appena definito seguendo un approccio bayesiano, si propone come distribuzione a priori per la media μ una *matching prior*. Definendo σ^2 quale parametro di disturbo ($\lambda = \sigma^2$), si può ricavare la log-verosimiglianza profilo per $\psi = \mu$:

$$\ell_p(\psi) = -\frac{n}{2}\log(\hat{\lambda}_\psi) - \frac{1}{2\hat{\lambda}_\psi}\sum_{i=1}^{n}(y_i - \psi)^2,$$

con $\hat{\lambda}_\psi = \frac{\sum_{i=1}^{n}(y_i-\psi)^2}{n}$.
La matrice di informazione osservata calcolata in $(\psi, \hat{\lambda}_\psi)$ è

$$j(\psi, \hat{\lambda}_\psi) = \begin{bmatrix} \frac{n^2}{\sum_{i=1}^{n}(y_i-\psi)^2} & \frac{n^3(\bar{y}-\psi)}{\left(\sum_{i=1}^{n}(y_i-\psi)^2\right)^2} \\ \frac{n^3(\bar{y}-\psi)}{\left(\sum_{i=1}^{n}(y_i-\psi)^2\right)^2} & \frac{n^3}{2(\sum_{i=1}^{n}(y_i-\psi)^2)^2} \end{bmatrix}.$$

Per la (1.3) la *matching prior* per ψ risulta pari a

$$\pi_{mp}(\psi) \propto i_{\psi\psi}(\psi, \hat{\lambda}_\psi)^{1/2} \propto \frac{1}{\hat{\lambda}_\psi}.$$

La distribuzione a posteriori marginale associata risulta essere

$$\pi_m(\psi|y) \propto \frac{1}{\hat{\lambda}_\psi}\mathrm{L}_{CA}(\psi),$$

con

$$\mathrm{L}_{CA}(\psi) = \mathrm{L}_P(\psi)|j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)|^{-1/2} \propto \left(\sum_{i=1}^{n}(y_i-\psi)^2\right)^{-n/2} \left|\frac{n^3}{2(\sum_{i=1}^{n}(y_i-\psi)^2)^2}\right|^{-\frac{1}{2}}$$
$$\propto \left(\sum_{i=1}^{n}(y_i-\psi)^2\right)^{-(n/2)+1}.$$

Infine, grazie alla (1.7) si può ottenere un'approssimazione del secondo ordine per la distribuzione a posteriori. Sfruttando questa approssimazione, si può calcolare facilmente la misura di evidenza di Pereira-Stern come risposta al sistema di ipotesi specificato grazie alla (1.13).

Rouder *et al.* (2009) propongono una distribuzione a priori alternativa per la media. Nel caso in analisi, il modello sotto l'ipotesi nulla ha solamente un parametro ignoto: la varianza. Sotto l'ipotesi alternativa invece, ad essere ignoti sono sia la media che la varianza. Dunque, risulta necessario specificare una distribuzione a priori per σ^2 sotto l'ipotesi nulla, mentre una distribuzione a priori congiunta per μ e σ^2 sotto l'ipotesi alternativa.
Nel valutare quale distribuzione a priori utilizzare, si deve ricordare il paradosso di Jeffreys-Lindley: utilizzando distribuzioni a priori non informative si è portati ad accettare quasi sempre il modello specificato sotto l'ipotesi nulla.
Rouder *et al.* (2009) specificano allora come distribuzione a priori per μ una normale

$$\mu \sim \mathrm{N}(0, \sigma_\mu^2),$$

dove l'iper-parametro σ_μ^2 rappresenta la varianza della distribuzione a priori. In alternativa, si potrebbe specificare una distribuzione per l'effetto standardizzato rappresentato da $\delta = \frac{\mu}{\sigma}$ anzichè per la media, assumendo come distribuzione per il parametro δ una $\mathrm{N}(0, \sigma_\delta^2)$, dove l'iper-parametro σ_δ^2 rappresenta la varianza della distribuzione a priori.
Tuttavia, utilizzando queste due distribuzioni, il *fattore di Bayes*, è strettamente legato al valore della varianza della distribuzione a priori. Per ovviare a questo problema, Zellner e Siow (1980) suggeriscono di specificare una

distribuzione per l'iper-parametro σ_δ^2 della forma

$$\frac{1}{\sigma_\delta^2} \sim \chi^2(1).$$

In seguito, Liang *et al.* (2008) mostrano che integrando la distribuzione a priori rispetto a δ, si può ottenere la corrispondente distribuzione a priori sull'effetto, ovvero, $\delta \sim Cauchy(1)$, distribuzione che corrisponde ad una t di Student con un grado di libertà.
Infine, una distribuzione non informativa utilizzata spesso in letteratura come distribuzione a priori per σ^2 è l'uniforme: $\pi(\sigma^2) \propto 1/\sigma^2$ (Jeffreys, 1961). Tale distribuzione viene trattata spesso in quanto gode di un forte vantaggio: fornisce la stessa informazione anche quando si usa una riparametrizzazione del parametro. Tuttavia, si deve ricordare che questa distribuzione non è propria.
Rouder *et al.* (2009) propongono quindi di combinare queste due distribuzioni (la distribuzione di *Cauchy* per l'effetto e la distribuzione uniforme per σ^2) definendo in questo modo la *JZS prior*, chiamata così per sottolineare il contributo di Jeffrey, Zellner e Siow. Distribuzione che risulta essere propria (Rouder *et al.*, 2009).
Analizzando quindi la verifica di ipotesi sulla media di una distribuzione Normale, si può caratterizzare il *fattore di Bayes JZS* nel seguente modo

$$\mathrm{B}_{01} = \frac{(1+\frac{t^2}{v})^{-(v+1)/2}}{\int_0^\infty (1+\mathrm{N}g)^{-1/2}(1+\frac{t^2}{(1+\mathrm{N}g)v})^{-(v+1)/2}(2\pi)^{-1/2}g^{-3/2}e^{-1/(2g)}dg},$$

dove t rappresenta la convenzionale statistica t, $v = n-1$ i gradi di libertà e per g specificano le seguente distribuzione:

$$\pi(g) = \frac{(1/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-1/2g}.$$

Per rispondere al sistema di ipotesi considerato è possibile calcolare inoltre la misura di evidenza di Pereira-Stern considerando la (1.11).

2.4 Modello di regressione lineare

Il modello di regressione lineare assume la seguente espressione analitica

$$Y = \beta_1 x_1 + ... + \beta_p x_p + \varepsilon,$$

dove i parametri β spiegano la relazione tra la variabile risposta e le esplicative e ε rappresenta il termine di errore, ovvero rappresenta la parte di variabilità della variabile risposta che non è spiegata dalla relazione che questa ha con le variabili esplicative. Per tale modello si assume che $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ e $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$, $i, j = 1, ..., n$.
Il parametro β viene stimato tramite il *metodo dei minimi quadrati* che fornisce $\hat{\beta} = (X^T X)^{-1} X^T Y$, dove X rappresenta la matrice del disegno. Sotto l'ipotesi di normalità si ha che $\hat{\beta} \sim \mathrm{N}_p(\beta, \sigma^2 (X^T X)^{-1})$.
In particolare, si consideri un modello di regressione normale con la seguente funzione di verosimiglianza

$$\mathrm{L}(\beta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{(Y - X\beta)^T (Y - X\beta)}{2\sigma^2} \right\}.$$

Seguendo un approccio frequentista, per verificare il seguente sistema di ipotesi:

$$\begin{cases} \mathrm{H}_0 : \beta_j = 0 \\ \mathrm{H}_1 : \beta_j \neq 0, \end{cases}$$

si può analizzare la statistica

$$t_j = \frac{\hat{\beta}_j}{\sqrt{S^2 (X^T X)^{-1}_{jj}}},$$

dove $(X^T X)^{-1}_{jj}$ indica l'elemento di posizione (j, j) della matrice $(X^T X)^{-1}$ e $S^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n-p}$, con $S^2 \sigma^2 \sim \chi^2(1)$.

Consideriamo la *matching prior* per il parametro di interesse dato da β_j. Ponendo quindi $\psi = \beta_j$ e considerando come parametro di disturbo p-dimensionale il vettore $\lambda = (\beta_1, ..., \beta_{j-1}, \beta_{j+1}, ..., \beta_p, \sigma^2)$, la log-verosimiglianza

per (ψ, λ) risulta

$$\ell(\psi, \lambda) = -\frac{n}{2}\log\lambda_p - \frac{1}{2\lambda_p}\sum_{i=1}^{n}\left[\left(\sum_{k\neq j}(y_i - x_{ik}\lambda_k)^2\right) + (y_i - x_{ij}\psi)^2\right].$$

In particolare, la verosimiglianza profilo per ψ risulta pari a

$$\ell_p(\psi) = -\frac{n}{2}\log\hat{\lambda}_{p\psi} - \frac{1}{2\hat{\lambda}_{p\psi}}\sum_{i=1}^{n}\left[\left(\sum_{k\neq j}(y_i - x_{ik}\hat{\lambda}_{k\psi})^2\right) + (y_i - x_{ij}\psi)^2\right],$$

dove $\hat{\lambda}_\psi$ rappresenta la stima di λ vincolata al valore di ψ. Ricordando, infine, che l'informazione attesa di Fisher per il modello così specificato è partizionabile come

$$i_{\psi\lambda}(\psi, \lambda) = \begin{bmatrix} \lambda_p(X^TX)^{-1} & 0 \\ 0 & \frac{n}{2(\lambda_p)^2} \end{bmatrix},$$

tramite la (1.5) è possibile ricavare la *matching prior* per ψ. La (1.6) permette invece la specificazione della distribuzione a posteriori marginale. Inoltre, grazie alla (1.7), si può ottenere un'approssimazione del secondo ordine per la distribuzione a posteriori, approssimazione che permette di ricavare in modo semplice la misura di evidenza di Pereira-Stern per rispondere al sistema di ipotesi considerato tramite la (1.13).

In letteratura sono state proposte distribuzioni a priori alternative per il parametro di interesse. In particolare si è caratterizzata la distribuzione a priori per β sia utilizzando una distribuzione uniforme, sia una distribuzione *g-prior* (Bayarri e García-Donato, 2007). Iniziando dalla prima proposta, con la distribuzione uniforme, non si hanno informazioni a priori sul fenomeno di interesse. La distribuzione a priori così caratterizzata risulta essere

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Seppure la distribuzione uniforme sia una possibile distribuzione per il parametro di interesse, si deve ricordare che questa non risulta essere una distribuzione propria.

Zellner (1983) introduce quindi una distribuzione a priori oggettiva, detta $g-prior$. Tale distribuzione è definita da

$$\beta|\sigma^2 \sim \mathrm{N}_p\left(\beta_0, \frac{g}{\phi}(X^TX)^{-1}\right),$$

con g parametro scalare positivo, β_0 iper-parametro rappresentante la media della distribuzione e $\phi = 1/\sigma^2$ parametro di precisione al quale si associa una distribuzione a priori di Jeffreys. Queste distribuzioni a priori risultano essere coniugate alle distribuzioni Gamma-Normali[2] ed è dunque possibile calcolare analiticamente la verosimiglianza marginale ad esse associata.
La scelta del valore di g risulta essere molto importante: se si sceglie $g = 1$ si associa alla distribuzione a priori lo stesso peso del campione e con $g = 2$ tale distribuzione è 1/2 tanto importante quanto il campione. Generalmente si pone $g = n$, ad indicare che la distribuzione a priori contiene la stessa informazione contenuta in una singola osservazione; un'a priori così definita prende il nome di *unit information prior*. Foster e George (1994) suggeriscono di porre $g = p^2$, ovvero di porlo pari al quadrato del numero dei predittori del modello; mentre Fernandez *et al.* (2001) suggeriscono di porre $g = \max\{n, p^2\}$. Una distribuzione a priori così definita prende il nome di *benchmark prior*.
Anche in questo caso, per rispondere al problema in analisi, si può calcolare la misura di evidenza di Pereira-Stern grazie alla (1.11). Per la sua determinazione, si consideri la distribuzione marginale per il parametro di interesse $\psi = \beta_j$, ottenuta integrando la distribuzione a posteriori rispetto a tutti i parametri di disturbo $(\beta_1, ..., \beta_{j-1}, \beta_{j+1}, ..., \beta_p, \sigma^2)$.

2.5 Modello di regressione logistica

In un modello di regressione logistica la variabile risposta è dicotomica e quello che si va a modellare è la media di tale variabile, rappresentante la probabilità di successo. In particolare avremo:

[2]La distribuzione Gamma-Normale $(\mu, \lambda, \alpha, \beta)$ è una famiglia bivariata di distribuzioni di probabilità continue a quattro parametri. La funzione di densità di probabilità è: $\mathrm{f}(x, \tau|\mu, \lambda, \alpha, \beta) = \frac{\beta^\alpha\sqrt{\lambda}}{\Gamma(\alpha)\sqrt{2\pi}}\tau^{\alpha-1/2}e^{-\beta\tau}e^{-\frac{\lambda\tau(x-\mu)^2}{2}}$.

$$\begin{cases} \mathcal{P}(Y=y)=\theta, & \text{qualora} \quad y=1, \\ \mathcal{P}(Y=y)=1-\theta, & \text{qualora} \quad y=0, \end{cases}$$

ovvero $Y_i \sim \text{Ber}(\theta)$, $i=1,..,n$ indipendenti.
La dipendenza della media della variabile risposta rispetto alle variabili esplicative è espressa come

$$g(\theta)=x^T\beta=\beta_1 x_1+...+\beta_p x_p,$$

dove $g(\cdot)$ rappresenta la funzione legame. Per la specificazione di tale funzione sono possibili diverse alternative. La più usata è la funzione *logit* data da

$$\text{logit}(\theta)=\log\left(\frac{\theta}{1-\theta}\right)=\beta_1 x_1+...+\beta_p x_p=x^T\beta,$$

da cui si ricava

$$\theta=\frac{e^{x^T\beta}}{1+e^{x^T\beta}}.$$

Un modello di regressione logistica ha la seguente funzione di verosimiglianza

$$\text{L}(\beta;y)\propto\prod_{i=1}^{n}\theta_i^{y_i}(1-\theta_i)^{(1-y_i)},$$

con $\theta_i=\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$.
In contesti medico-sanitari è interessante interpretare i coefficienti del modello logistico introducendo il concetto di quota e rapporto di quote. Dato un evento A con probabilità θ si dice quota di un evento il rapporto tra la probabilità che l'evento A si verifichi e la probabilità che esso non si verifichi:

$$\text{o}_A=\frac{\theta}{1-\theta}=\frac{\mathcal{P}(A)}{1-\mathcal{P}(A)}.$$

Il rapporto delle quote, invece, rappresenta una misura di quanto maggiore siano le quote del manifestarsi di un certo evento in quei soggetti esposti a un determinato fattore di rischio. Dato l'evento A, considerando $\mathcal{P}(A|N)$ e $\mathcal{P}(A|S)$ le probabilità che l'evento si verifichi rispetto ai due scenari differenti N e S, il rapporto tra quote è definito come:

$$\text{OR}=\frac{\text{o}_{A|N}}{\text{o}_{A|S}}=\frac{\frac{\mathcal{P}(A|N)}{1-\mathcal{P}(A|N)}}{\frac{\mathcal{P}(A|S)}{1-\mathcal{P}(A|S)}}.$$

È di interesse proporre un possibile approccio bayesiano al caso in analisi. A tale scopo si caratterizza una distribuzione *matching prior* per il parametro di interesse. Fissando come parametro di interesse $\psi = \beta_j$ e come parametro di disturbo $(p-1)$ dimensionale il vettore $\lambda = (\beta_1, ..., \beta_{j-1}, \beta_{j+1}, ..., \beta_p)$, la log-verosimiglianza per (ψ, λ) risulta

$$\ell(\psi, \lambda) = \sum_{i=1}^{n} \left[\sum_{k \neq j} \left(y_i x_{ik} \lambda_k - \log(1 + e^{x_{ik}\lambda_k}) \right) + y_i x_{ij} \psi - \log(1 + e^{x_{ij}\psi}) \right].$$

In particolare la verosimiglianza profilo per ψ è

$$\ell_p(\psi) = \sum_{i=1}^{n} \left[\sum_{k \neq j} \left(y_i x_{ik} \hat{\lambda}_{k\psi} - \log(1 + e^{x_{ik}\hat{\lambda}_{k\psi}}) \right) + y_i x_{ij} \psi - \log(1 + e^{x_{ij}\psi}) \right],$$

dove $\hat{\lambda}_\psi$ è la stima di massima verosimiglianza del parametro di disturbo vincolata al valore di ψ.
Dalla (1.5) è possibile ricavare la distribuzione a priori *matching prior* per ψ. Considerando invece la (1.6) si può caratterizzare la distribuzione a posteriori marginale. Infine, grazie alla (1.7), si può ottenere un'approssimazione del secondo ordine per la distribuzione a posteriori. La misura di evidenza di Pereira-Stern, utilizzabile come risposta al sistema di ipotesi considerato, grazie a tale approssimazione è facilmente determinabile (si veda la (1.13)).

In letteratura si propone un approccio bayesiano al caso in analisi riadattando la *g-prior*, definita la distribuzione a priori di default nel caso di regressione lineare normale, in questo contesto (Hanson *et al.*, 2014). La *g-prior* per β risulta essere

$$\beta \sim \mathrm{N}_p(be_1, gn(X^T X)^{-1}),$$

dove X rappresenta la matrice del disegno, e_1 è un vettore p-dimensionale il cui primo elemento è pari a 1 mentre i restanti sono 0 e b raffigura la media a priori per l'intercetta. Per quanto riguarda il termine scalare g, questo può essere modellato tramite una distribuzione Gamma-Inversa. Hanson *et al.*

(2014) lo fissano ad un valore costante e positivo.
Anche per questo caso, si può calcolare la misura di evidenza di Pereira-Stern in risposta al sistema di ipotesi considerato tramite la (1.11). Per la sua determinazione, si deve considerare la distribuzione a posteriori marginale per il parametro di interesse $\psi = \beta_j$, ottenuta integrando la distribuzione a posteriori rispetto i parametri di disturbo $(\beta_1, ..., \beta_{j-1}, \beta_{j+1}, ..., \beta_p)$.

2.6 Curva ROC e AUC

In ambito medico la curva ROC risulta utile per valutare la performance diagnostica di un test, ovvero l'accuratezza del test nel discriminare tra pazienti malati e pazienti sani, o anche per confrontare tra loro le performance di differenti test diagnostici (Ventura e Racugno, 2017).
Analizzando il risultato di questi test si deve tenere in considerazione che è difficile osservare una separazione netta tra malati e sani, quindi, le distribuzioni del test nei due gruppi in genere si sovrappongono parzialmente (si veda la Figura 2.1).
La curva ROC rappresenta uno strumento statistico ottenuto a partire dalla sensibilità e dalla specificità. Considerando che queste due misure dipendono dalla soglia discriminante k, per la costruzione della curva si fa variare tale soglia e per ogni k si calcolano queste due misure. Graficamente, si veda la Figura 2.2, risulterà un diagramma cartesiano in cui in ascissa viene rappresentato il complemento a uno della specificità (1-specificità), mentre in ordinata la sensibilità.

È da preferire un test diagnostico che ha una curva il più possibile vicina all'angolo superiore sinistro del diagramma cartesiano, punto che sta ad indicare che il test ha massima specificità e massima sensibilità. Qualora la curva risultasse molto vicina alla bisettrice invece, si è in presenza di un test che classifica i pazienti in modo casuale, quindi di un test poco accurato dal momento che un paziente risulta sano o malato con la stessa probabilità.
Un indicatore di sintesi di quest'area è rappresentato dall'area sottostante a questa, indicatore che prende il nome di AUC. Per la sua determinazione si può seguire sia un approccio parametrico che uno non parametrico.
Per quanto riguarda l'approccio parametrico, si consideri X la variabile che

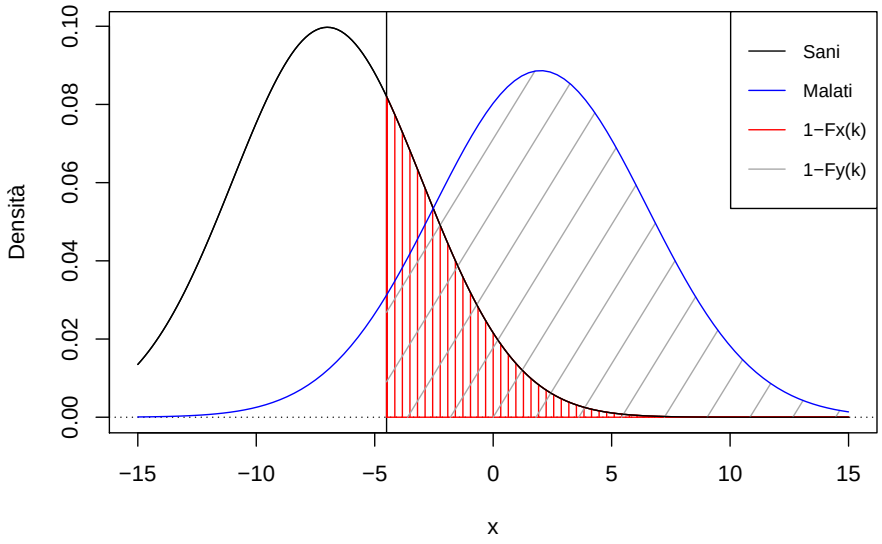

Figura 2.1: Distribuzioni del test diagnostico nei sani e nei malati

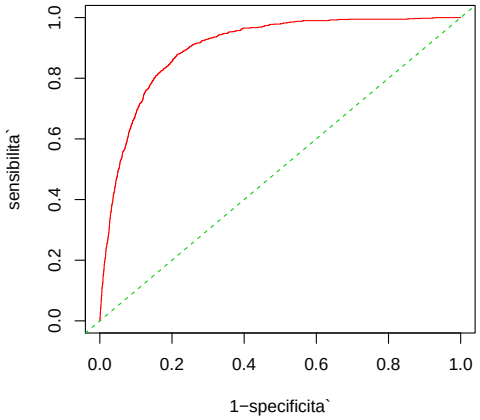

Figura 2.2: Esempio del grafico della curva ROC

descrive la misura nel gruppo dei pazienti sani e Y quella dei pazienti malati. Si assume che queste due variabili abbiano rispettivamente funzione di ripartizione $\mathrm{F}_x(x;\theta_x)$ e $\mathrm{F}_y(y;\theta_y)$, con $\theta_x \in \Theta_x \subseteq \mathbb{R}^{\mathrm{P}_x}$ e $\theta_y \in \Theta_y \subseteq \mathbb{R}^{\mathrm{P}_y}$.
La funzione di verosimiglianza per $\theta = (\theta_x, \theta_y)$ risulta essere

$$\mathrm{L}(\theta) = \mathrm{L}(\theta_x, \theta_y) = \prod_{i=1}^{n_x} f_x(x_i;\theta_x) \prod_{j=1}^{n_y} f_y(y_j;\theta_y),$$

dove $f_x(x;\theta_x)$ e $f_y(y;\theta_y)$ rappresentano le funzioni di densità o densità di probabilità di X e Y.
L'AUC può essere espressa come

$$\mathrm{AUC} = \mathrm{AUC}(\theta) = \mathcal{P}(X < Y) = \int_{-\infty}^{+\infty} \mathrm{F}_x(t;\theta_x) d\mathrm{F}_y(t;\theta_y),$$

misura che viene rappresentata come funzione del parametro complessivo $\theta = (\theta_x, \theta_y)$, nota anche come modello *sollecitazione-resistenza* (o *stress-strength model*) (Kotz *et al.*, 2003).
Inoltre, dal momento che la curva ROC può essere definita come

$$\mathrm{ROC}(t) = 1 - \mathrm{F}_y(\mathrm{F}_x^{-1}(1-t)),$$

con $t \in [0,1]$ e $\mathrm{F}_x^{-1}(1-t) = \inf\{x \in \mathbb{R} : f_x(x) \geq 1-t\}$, l'AUC è esprimibile come

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(t) dt.$$

In base a quanto appena definito quindi, se si assume, ad esempio, che $X \sim \mathrm{N}(\mu_x, \sigma_x^2)$ e $Y \sim \mathrm{N}(\mu_y, \sigma_y^2)$, si ha

$$\mathrm{AUC} = \mathrm{AUC}(\theta) = \mathcal{P}(X < Y) = \Phi\left(-\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}\right).$$

Nel caso in cui invece $X \sim \mathrm{Exp}(\alpha)$ e $Y \sim \mathrm{Exp}(\beta)$, l'AUC è definito come

$$\mathrm{AUC} = \mathrm{AUC}(\theta) = \mathcal{P}(X < Y) = \frac{\alpha}{\alpha + \beta}.$$

Per quanto riguarda l'interpretazione del valore dell'AUC vale (Swets, 1988)

- $\mathrm{AUC} = 0.5$: il test non è informativo;

- $0.5 < \text{AUC} \leq 0.7$: il test è poco accurato;
- $0.7 < \text{AUC} \leq 0.9$: il test è moderatamente accurato;
- $0.9 < \text{AUC} < 1.0$: il test è altamente accurato;
- $\text{AUC} = 1.0$: il test è perfetto.

Anche per questo caso di studio si propone un possibile approccio bayesiano al problema specificando una distribuzione *matching prior* per il parametro di interesse. Parametro che nel contesto della curva ROC risulta essere $\psi = \text{AUC} = \mathcal{P}(X < Y)$. Dalla (1.5) è possibile ricavare la distribuzione a priori *matching prior* per ψ. Considerando invece la (1.6) si può caratterizzare la distribuzione a posteriori marginale. Infine grazie alla (1.7) si può ottenere un'approssimazione del secondo ordine per la distribuzione a posteriori. La misura di Pereira-Stern, utilizzabile come risposta al sistema di ipotesi:

$$\begin{cases} \text{H}_0 : \psi = 0 \\ \text{H}_1 : \psi \neq 0, \end{cases}$$

grazie a tale approssimazione è facilmente determinabile (si veda la (1.13)).

Kotz *et al.* (2003) discutono un approccio bayesiano classico. Si consideri $X \sim f(x; \theta_x)$ e $Y \sim f(y; \theta_y)$, e si definisca con $\pi(\theta)$ la distribuzione a priori congiunta per (θ_x, θ_y) e con $\pi(\theta|x, y)$ la relativa distribuzione a posteriori. La distribuzione a posteriori per il parametro di interesse $\psi = \text{AUC}$ può essere ottenuta considerando una trasformazione. In particolare, si consideri la trasformazione $\text{F}: \theta \longrightarrow (\psi, \lambda)$ con la relativa inversa $\text{Q} = \text{F}^{-1}$. La distribuzione a posteriori per (ψ, λ) è data da $\pi(\text{Q}(\psi, \lambda)|x, y)|J_{\text{Q}(\psi,\lambda)}|$, dove $|J_{\text{Q}(\psi,\lambda)}|$ è lo jacobiano della trasformazione Q. È ora possibile ottenere la distribuzione a posteriori marginale per ψ integrando la distribuzione a posteriori ottenuta rispetto a λ, ovvero

$$\pi(\psi|x, y) = \int \pi(\text{Q}(\psi, \lambda)|x, y)|J_{\text{Q}(\psi,\lambda)}|d\lambda.$$

Per rispondere al sistema di ipotesi definito precedentemente, a partire dalla distribuzione a posteriori marginale per il parametro di interesse, si può calcolare la misura di evidenza di Pereira-Stern tramite la (1.11).

2.7 Conclusioni

In questo capitolo sono stati delineati alcuni casi di studio classici della Biostatistica. In particolare dopo una piccola rassegna sul concetto di studio clinico, si è passati alla descrizione dell'inferenza su una proporzione, su due proporzioni, sulla media della normale, su un parametro scalare di regressione e sull'area sotto la curva ROC. Per ciascun test si è descritto l'usuale approccio frequentista e le proposte bayesiane presenti in letteratura. Inoltre si è definito un approccio bayesiano che utilizza come a priori una distribuzione *matching prior*. Nel capitolo seguente vedremo un'applicazione di tutti questi casi di studio.

Capitolo 3

Applicazione a dati reali in ambito medico

In questo capitolo si presentano delle applicazioni a dati reali. Le analisi basate sulle *matching prior*, in particolare, saranno anche confrontate con altre già trattate in letteratura.

3.1 Aspetti computazionali

Prima di procedere con l'analisi dei vari casi di studio, si ritiene utile spiegare brevemente la libreria di `R` utilizzata per la caratterizzazione delle *matching priors*. Questo pacchetto, sviluppato da Bellio e Pierce (2015) per analisi frequentiste, contiene funzioni utili per la determinazione delle statistiche r_p e r_p^*, nonché i relativi intervalli di confidenza per il parametro di interesse, basate sul metodo d'integrazione Monte Carlo. Per utilizzare questo pacchetto si devono specificare tre funzioni:

- la funzione di log-verosimiglianza $\ell(\theta)$;
- la funzione che genera i dati a partire dal modello parametrico assunto;
- il parametro di interesse ψ su cui si vuole fare inferenza.

Tra le funzioni di questo pacchetto risultano di particolare interesse `rstar` e `rstar.ci`. La prima è utile soprattutto per il calcolo di r_p^* e r_p, la seconda

per la determinazione degli intervalli di confidenza.
Seppur questo pacchetto risulti un pacchetto frequentista a tutti gli effetti, per le analisi che seguono viene utilizzato da un punto di vista bayesiano. Ricordando infatti l'approssimazione della distribuzione a posteriori con *matching prior* descritta nella Sezione 1.3.3, si deduce che tramite le funzioni contenute in questo pacchetto si possono ottenere anche tutti gli elementi che caratterizzano la distribuzione a posteriori con *matching prior* e tutte le statistiche di sintesi ad essa associate.
Per un esempio di applicazione in `R` si veda l'Appendice B.

3.2 Caso di studio: Misure della creatinchinasi nei pazienti con angina instabile

I dati utilizzati provengono da uno studio riguardante la diagnosi precoce dell'infarto (Bland, 2009). In particolare, viene misurata la quantità dell'enzima creatinchinasi (CK) situata nel tessuto muscolare. Tale enzima risulta presente soprattutto nelle fibre cardiache a seguito di stress importanti o danni muscolari gravi come ad esempio un evento infartuale al miocardio. Si analizzano le rilevazioni, che avvengono mediante una semplice analisi del sangue, di n=15 pazienti con angina instabile (AI). Considerando che si ritiene malato un paziente con valore di creatinchinasi inferiori a 75 U/L, lo scopo dello studio è valutare se la prevalenza di tale malattia riscontrata nella popolazione in analisi risulta essere uguale a quella della popolazione italiana, pari a 0.3.

Si tratta dunque di una verifica di ipotesi su una proporzione.
Seguendo l'usuale approccio frequentista, per rispondere a tale quesito, ci si può ricondurre alla statistica z (si veda la Sezione 2.1). Quello che si andrà a presentare in seguito è invece un approccio bayesiano al problema.

Trattandosi di un caso di verifica di ipotesi su una proporzione, si considera $Y \sim \text{Ber}(\theta)$.

3.2.1 Modello bayesiano con *matching prior*

Con l'obiettivo di rispondere al problema in analisi si specifica per il parametro di interesse, ovvero la prevalenza della malattia, una *matching prior*. Per la sua caratterizzazione, in particolare, si considera l'approssimazione del secondo ordine della distribuzione a posteriori (si veda la Sezione 1.3.3). Tramite la libreria di R `LikelihoodAsy` e in particolare tramite le sue funzioni `rstar` e `rstar.ci` è possibile ricavare tutti gli elementi che compongono l'approssimazione della distribuzione a posteriori con *matching prior* e tutte le statistiche di sintesi ad esse associate. Si ritiene importante sottolineare quanto dimostrato nella Sezione 2.1: in questo caso specifico la distribuzione a posteriori con *matching prior* è una Beta $(\sum_{i=1}^{n} y_i + 1/2, \sum_{i=1}^{n}(1-y_i) + 1/2)$. La mediana di tale distribuzione risulta essere 0.6, mentre l'intervallo di credibilità *equi-tailed* al 95% è pari a $(0.3503, 0.8148)$, intervallo che coincide con quello di confidenza basato su r_p^*.
Tramite la funzione `rstar.ci` è possibile confrontare le curve relative alla statistica r_p e alla statistica r_p^*. Come si può notare dal grafico riportato in (Figura 3.1) le due curve praticamente coincidono ad indicare che l'intervallo basato su r_p, pari a $(0.3511, 0.8144)$, risulta molto simile a quello basato su r_p^*.

Questo grafico mette in ascissa il parametro di interesse e in ordinata i valori delle statistiche r_p e r_p^*. La curva nera rappresenta i valori della statistica r_p al variare di ψ, mentre quella rossa quella della statistica r_p^*. Per un α fissato, tale per cui $r_p(\psi) = \alpha$, si tracci una parallela all'asse orizzontale corrispondente, e la sua simmetrica rispetto all'asse di destra; l'intervallo di confidenza basato su r_p avrà come estremi i valori dell'ascissa derivanti dai punti di intersezione delle parallele con la curva nera, mentre quello basato su r_p^*, dai punti di intersezione delle parallele con la curva rossa.
In generale, si può affermare che maggiore è la distanza tra le curve, migliore risulterà essere l'approssimazione alla normale standard della statistica r_p^* rispetto alla statistica r_p.
Calcolando infine la misura di Evidenza di Pereira-Stern per

$$\begin{cases} \mathrm{H}_0: & \theta = 0.3 \\ \mathrm{H}_1: & \theta \neq 0.3, \end{cases}$$

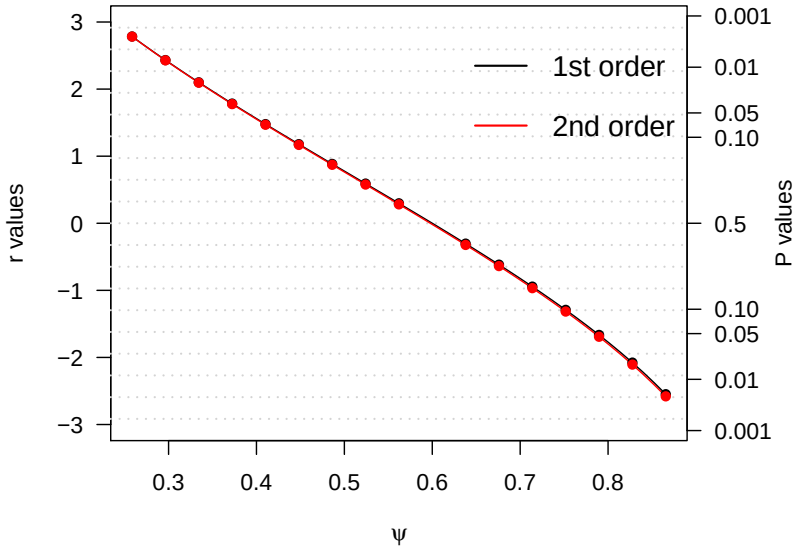

Figura 3.1: Grafico della statistica r_p (linea nera) e statistica r_p^* (linea rossa).

si è ottenuto EV = 0.0018. Si può quindi definire che vi è evidenza contro l'ipotesi nulla a favore di quella alternativa, ovvero H_1 è maggiormente supportata dai dati campionari rispetto a H_0.

3.2.2 Modello bayesiano con a priori Beta

Una possibile distribuzione a priori alternativa per il parametro di interesse corrisponde alla distribuzione Beta(α, β), distribuzione coniugata alla bernoulliana (si veda la Sezione 2.2). In particolare, in assenza di informazione a priori sulla prevalenza della malattia nei pazienti studiati, si fissano i valori degli iper-parametri in modo tale che la varianza della Beta sia elevata, dando quindi maggior peso ai dati campionari. Si pone a tale fine $\alpha = \beta = 1$. Dalla distribuzione a posteriori così ottenuta è possibile determinare tutte le statistiche di sintesi ad essa associate. In particolare, la mediana a posteriori risulta essere 0.62 e l'intervallo di credibilità a posteriori *equi-tailed* al 95% è $(0.3528, 0.8124)$. L'intervallo di credibilità dunque presenta un'ampiezza

molto simile a quella ottenuta con la *matching prior*. Volendo rispondere al sistema di ipotesi specificato nel paragrafo precedente, è stata anche calcolata la misura di Evidenza di Pereira-Stern che, risultando pari a 0.015, indica che i dati supportano l'ipotesi alternativa rispetto a quella nulla. Si giunge quindi alle medesime conclusioni alle quali si era giunti con la *matching prior*. Data la ridotta numerosità campionaria, si ritiene, inoltre, interessante valutare quanto incidano i valori degli iper-parametri della Beta sulla distribuzione a posteriori. Si provano quindi diversi valori per ciascun iper-parametro calcolandone rispettivamente il valore della mediana a posteriori, l'intervallo di credibilità *equi-tailed* al 95% e la misura di evidenza di Pereira-Stern. I risultati sono riportati nella Tabella 3.1. Come si può osservare, le mediane

Tabella 3.1: Statitiche di sintesi delle varie distribuzioni Beta ed EV

Iper-parametri	Mediana	Intervallo	EV
$\alpha = \beta = 1$	0.62	(0.3528,0.8124)	0.015
$\alpha = \beta = 2$	0.58	(0.3574,0.7847)	0.012
$\alpha = 1, \beta = 3$	0.53	(0.3076,0.7388)	0.042
$\alpha = 5, \beta = 1$	0.68	(0.4572, 0.8460)	0.0005

ponendo $\alpha = \beta = 2$ e $\alpha = \beta = 1$ risultano simili, quella relativa alla coppia $\alpha = 1, \beta = 3$ ha un valore leggermente inferiore, mentre quella riferita a $\alpha = 5, \beta = 1$ è traslata verso destra. Per quanto riguarda gli intervalli di credibilità, appare meno ampio quello relativo alla distribuzione che ha come iper-parametri $\alpha = 5, \beta = 1$, tuttavia, le loro ampiezze non differiscono di molto.
Considerando, infine, la misura di evidenza di Pereira-Stern, si può affermare che tutte le distribuzioni, seppur con evidenza differente, portano allo stesso risultato: i dati campionari supportano l'ipotesi alternativa rispetto a quella nulla.
La Figura 3.2 riporta le varie distribuzioni a posteriori e conferma quanto detto. Le distribuzioni Beta ottenute fissando gli iper-parametri pari a $\alpha = 1, \beta = 3$ o $\alpha = 5, \beta = 1$ si discostano dalle altre distribuzioni, in particolare la distribuzione Beta(5,1) risulta con varianza minore. Va comunque sottolineato che la Beta(5,1) è una distribuzione soggettiva, mentre la

matching prior è oggettiva. Risulta quindi di rilevante importanza la scelta degli iper-parametri della distribuzione a priori Beta. Iper-parametri che la *matching prior* non richiede di fissare.

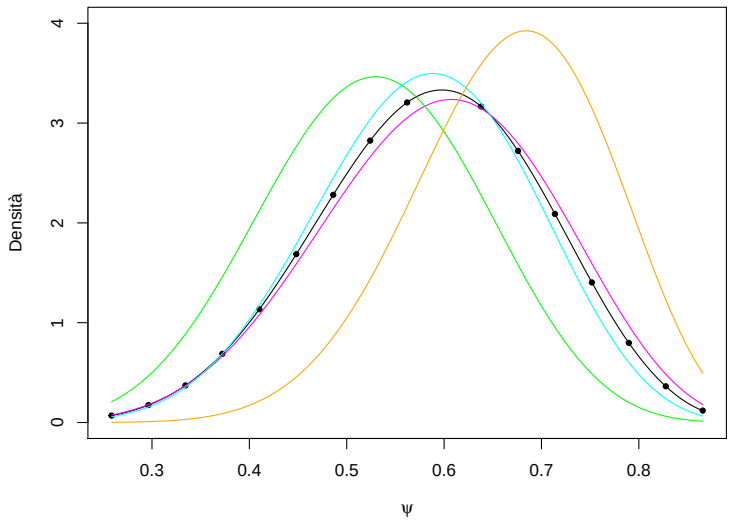

Figura 3.2: Distribuzioni a posteriori con *matching prior* (linea nera punteggiata), con a priori Beta(1,1) (linea magenta), Beta(2,2) (linea azzurra), Beta(1,3) (linea verde), Beta(5,1) (linea arancione).

3.3 Caso di studio: Malattia ostruttiva dell'arteria coronarica

I dati utilizzati in questa Sezione derivano da uno studio condotto su 20 individui, 10 maschi e 10 femmine, nei quali viene valutata la presenza/assenza della malattia ostruttiva dell'arteria coronarica (CAD) (Bland, 2009). La malattia coronarica fa parte delle malattie più frequenti della medicina cardiovascolare. Con questo termine, si intende, una qualsiasi alterazione, anatomica o funzionale, delle arterie coronarie, cioè dei vasi sanguigni che irrorano il miocardio. L'angina pectoris, l'infarto cardiaco o addirittura l'ar-

resto cardiocircolatorio, risultano essere i principali effetti portati da questa alterazione.
Scopo dell'analisi è valutare qualora la proporzione di CAD nei maschi possa essere ritenuta significativamente uguale a quella delle femmine.

Si tratta dunque di un test di verifica di ipotesi su due proporzioni.
Seguendo l'usuale approccio frequentista si può considerare la statistica test definita nella Sezione 2.2. Quello che si vedrà di seguito invece è un possibile approccio bayesiano al problema.

Si definisca con θ_1 la prevalenza della malattia dei maschi e con θ_2 quella delle femmine. Si avrà quindi che la popolazione maschile è descritta da $Y \sim \text{Ber}(\theta_1)$, mentre quella femminile da $X \sim \text{Ber}(\theta_2)$.

3.3.1 Modello bayesiano con *matching prior*

Volendo verificare se le prevalenze siano uguali, il parametro di interesse dello studio sarà $\psi = \theta_1 - \theta_2$. Per tale parametro si specifica una *matching prior*. In particolare, come per il caso precedente, si considera l'approssimazione del secondo ordine della distribuzione a posteriori con *matching prior* (si veda la Sezione 1.3.3). Tramite le funzioni presenti nella libreria `LikelihoodAsy` si possono ottenere tutti gli elementi per la caratterizzazione della distribuzione a posteriori e le misure di sintesi ad essa associate. La mediana a posteriori risulta essere 0.39 mentre l'intervallo di credibilità *equi-tailed* ad un livello del 95% ha come estremi $(0.2331, 0.5801)$, intervallo che coincide con quello di confidenza basato su r_p^*.
Nel grafico riportato in Figura 3.3 sono messi a confronto gli intervalli di confidenza basati su r_p e su r_p^*. Come si può vedere le due curve si discostano solamente in corrispondenza di valori piccoli di ψ: l'intervallo di confidenza basato su r_p, pari a $(0.2432, 0.5802)$, risulta leggermente più ampio di quello basato su r_p^*.
Infine è stata calcolata la misura di evidenza di Pererira-Stern per

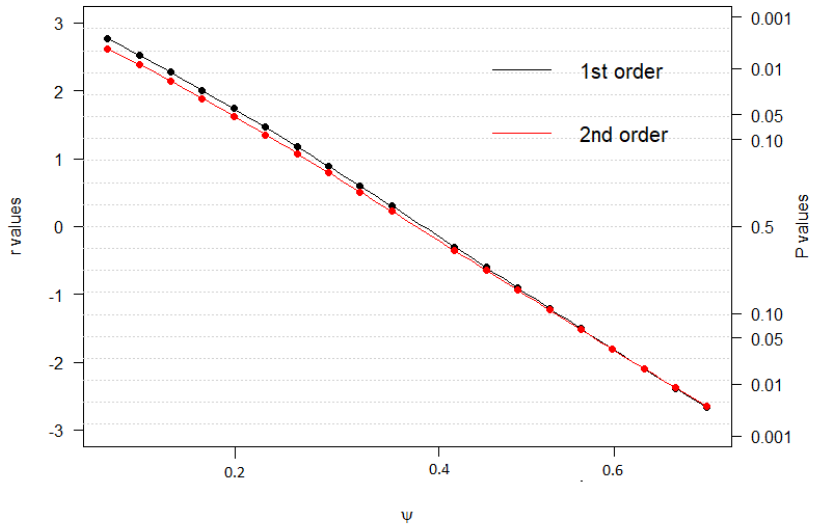

Figura 3.3: Grafico della statistica r_p (linea nera) e statistica r_p^* (linea rossa).

$$\begin{cases} \mathrm{H}_0: & \psi = 0 \\ \mathrm{H}_1: & \psi \neq 0, \end{cases}$$

che risulta essere 0.0003. Un valore così piccolo di questa misura porta evidenza contro l'ipotesi nulla, ovvero i dati supportano H_1 rispetto a H_0.

3.3.2 Modello bayesiano con a priori Beta

Una possibile distribuzione a priori alternativa per il problema in analisi è rappresentata dalla Beta, distribuzione coniugata alla Bernoulli. In particolare, viene specificata una distribuzione a priori Beta sia per θ_1 ($\mathrm{Beta}(\alpha, \beta)$) che per θ_2 ($\mathrm{Beta}(\gamma, \lambda)$), fissando tutti gli iper-parametri pari a 0.5. Per ottenere la distribuzione a priori per il parametro di interesse $\psi = \theta_1 - \theta_2$ e la relativa distribuzione a posteriori marginale, si considera la trasformazione di variabili aleatorie descritta nella Sezione 2.2.
La mediana a posteriori risulta essere 0.4 mentre l'intervallo di credibilità *equi-tailed* a livello 95% è pari a $(0.2334, 0.5801)$. La distribuzione appena ottenuta in pratica coincide con quella a posteriori con *matching prior*. Con

lo scopo di verificare il sistema di ipotesi descritto inizialmente si calcola la misura di evidenza di Pereira-Stern che, risultando pari a 0.0006, conferma quanto affermato nel paragrafo precedente: vi è evidenza contro l'ipotesi nulla.
Si vuole valutare inoltre quanto incida la scelta degli iper-parametri della distribuzione a posteriori ottenuta come trasformata della congiunta di due Beta. Si provano quindi diversi valori per ciascun iper-parametro e nella Tabella 3.2 vengono riportate le rispettive statistiche di sintesi e le misure di evidenza di Pereira-Stern calcolate in risposta al sistema di ipotesi specificato precedentemente. Come si può notare, la distribuzione a posteriori con

Tabella 3.2: Statitiche di sintesi delle varie distribuzioni a posteriori ed EV

Iper-parametri	Mediana	Intervallo	EV
$\alpha = \beta = \gamma = \lambda = 0.5$	0.39	(0.233,0.580)	0.0006
$\alpha = \beta = \gamma = \lambda = 2$	0.378	(0.245,0.478)	0.0008
$\alpha = \gamma = 1, \beta = \lambda = 3$	0.420	(0.241,0.581)	0.0005
$\alpha = \gamma = 5, \beta = \lambda = 1$	0.328	(0.161, 0.492)	0.004

$\alpha = \gamma = 5, \beta = \lambda = 1$ è quella che più si discosta da tutte le altre, ma in ogni caso, anche se con evidenza minore, porta allo stesso risultato: i dati supportano H_1 rispetto a H_0.
Nel grafico riportato in Figura 3.4 vengono riportate tutte le distribuzioni a posteriori marginali specificate. Come si può vedere, la distribuzione ottenuta con *matching prior* coincide con quella nella quale gli iper-parametri sono $\alpha = \beta = \gamma = \lambda = 0.5$ e non si discosta molto da quelle in cui $\alpha = \gamma = 1, \beta = \lambda = 3$ e $\alpha = \beta = \gamma = \lambda = 2$. La distribuzione in cui $\alpha = \gamma = 5, \beta = \lambda = 1$ è invece traslata rispetto alle altre. Si nota quindi come, anche per questo caso di studio, possa essere importante la scelta degli iper-parametri per la specificazione della distribuzione a posteriori trattata in letteratura, caratterizzando la *matching-prior* invece non serve fissarli.

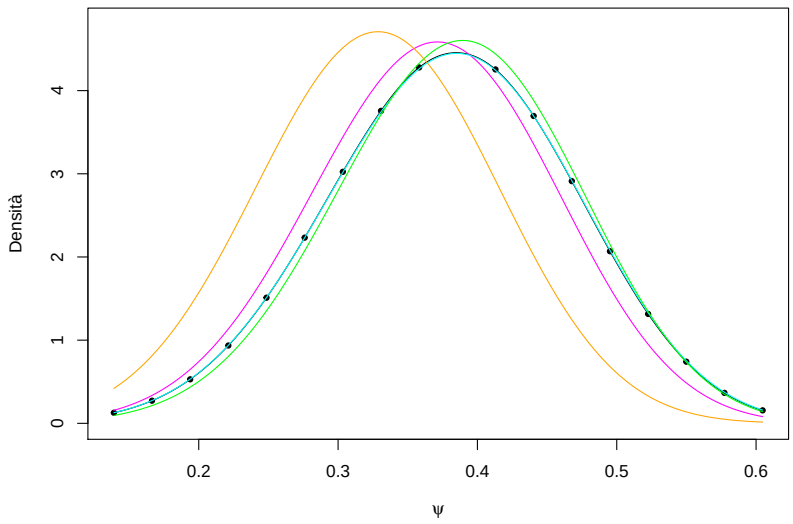

Figura 3.4: Distribuzioni a posteriori marginali con *matching prior* (linea nera punteggiata), con a priori Beta congiunta con iper-parametri $\alpha = \beta = \gamma = \lambda = 0.5$ (linea azzurra), $\alpha = \beta = \gamma = \lambda = 2$ (linea magenta), $\alpha = \gamma = 1, \beta = \lambda = 3$ (linea verde), $\alpha = \gamma = 5, \beta = \lambda = 1$ (linea arancione).

3.4 Caso di studio: Spessore delle piaghe cutanee al bicipite

I dati analizzati provengono da uno studio riguardante pazienti con malattie intestinali (Maudgal *et al.*, 1985; Bland, 2009). In questo studio sono state rilevate le misure dello spessore delle piaghe cutanee (in mm) al bicipite di 20 pazienti affetti da malattia di Crohn (primo gruppo) e di 9 pazienti con celiachia (secondo gruppo) (Figura 3.5).
Lo scopo dell'analisi è valutare qualora lo spessore medio delle piaghe nei pazienti con diagnosi differenti possa essere considerato uguale, ovvero se la differenza tra le medie delle due popolazioni sia nulla.
Seguendo l'approccio frequentista, per rispondere al problema in analisi, si può svolgere un $t - test$, dopo aver verificato le ipotesi di normalità e omo-

schedasticità (si veda la Sezione 2.3). Da una breve analisi esplorativa risulta

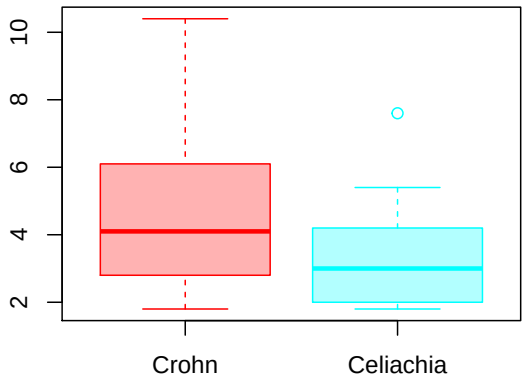

Figura 3.5: Boxplot dello spessore delle piaghe nei due gruppi di pazienti.

che lo spessore medio delle piaghe nei pazienti affetti dalla malattia di Crohn è pari a 4.72 mm (sd=2.42 mm), mentre in quelli celiaci risulta essere 3.53 mm (sd=1.96 mm). L'ipotesi di omoschedasticità è verificata, mentre vi è una lieve evidenza contro l'assunzione della normalità per quanto riguarda il primo gruppo. L'ipotesi nulla del test Shapiro-Wilk, infatti, non viene rifiutata solamente ad un livello di significatività pari all'1%. Per i celiaci invece, non rifiutiamo ad un livello del 5%.
Tenendo in considerazione il fatto di avere basse numerosità campionarie, si assume la normalità delle misure per entrambi i gruppi. Si considera quindi per la popolazione di pazienti affetti dalla malattia di Crohn $Y \sim \mathrm{N}(\mu_1, \sigma^2)$, per i celiaci invece $X \sim \mathrm{N}(\mu_2, \sigma^2)$.

3.4.1 Modello bayesiano con *matching prior*

Per rispondere al problema in analisi si procede con la specificazione di una *matching prior* come distribuzione a priori per il parametro di interesse,

parametro che in questo caso rappresenta la differenza tra lo spessore medio delle piaghe tra i pazienti affetti dalla malattia di Crohn e quelli celiaci. Si pone dunque $\psi = \mu_1 - \mu_2$.
In particolare, per svolgere le analisi in `R`, si fa riferimento all'approssimazione del secondo ordine della distribuzione a posteriori marginale.
Tramite le funzioni presenti nella libreria `LikelihoodAsy`, è possibile ricavare la statistica r_p^*, elemento cruciale per la determinazione della distribuzione a posteriori marginale, la mediana a posteriori e l'intervallo di credibilità. La mediana a posteriori risulta pari a 1.18, mentre l'intervallo di credibilità *equi-tailed* al 95% è $(-0.5965, 2.9598)$. La funzione `rstar` fornisce tutti gli elementi necessari per la determinazione della misura di evidenza di Pereira-Stern che, per

$$\begin{cases} \mathrm{H}_0 : & \psi = 0 \\ \mathrm{H}_1 : & \psi \neq 0, \end{cases}$$

risulta essere 0.15. Dal momento che il valore di questa misura è elevato, ci suggerisce che i dati supportano l'ipotesi nulla.
Un'altra funzione utile è rappresentata da `rstar.ci`. Grazie a questa infatti è possibile ottenere un grafico (Figura 3.6) che mette a confronto la statistica r_p con la statistica r_p^*. Come si può notare dalla Figura 3.6, le due curve sono molto simili, salvo per valori molto piccoli di α. L'intervallo di confidenza a livello 95% basato su r_p è pari a $(-0.488, 2.851)$. Intervallo che risulta con ampiezza leggermente inferiore rispetto a quello basato su r_p^* che coincide con l'intervallo di credibilità *equi-tailed* della distribuzione a posteriori marginale con *matching prior*.

3.4.2 Modello bayesiano con *JZS prior*

Si procede ora rispondendo al problema in analisi specificando un modello bayesiano presente in letteratura. In particolare viene caratterizzata la *JZS prior*, la distribuzione a priori che prevede la specificazione di una distribuzione uniforme per σ^2 e di una *Cauchy* per l'effetto, ovvero per $\delta = (\mu_1 - \mu_2)/\sigma$. Grazie alla funzione `ttestBF` contenuta nel pacchetto `BayesFactor` è stato possibile ottenere la distribuzione a posteriori marginale del caso trattato e

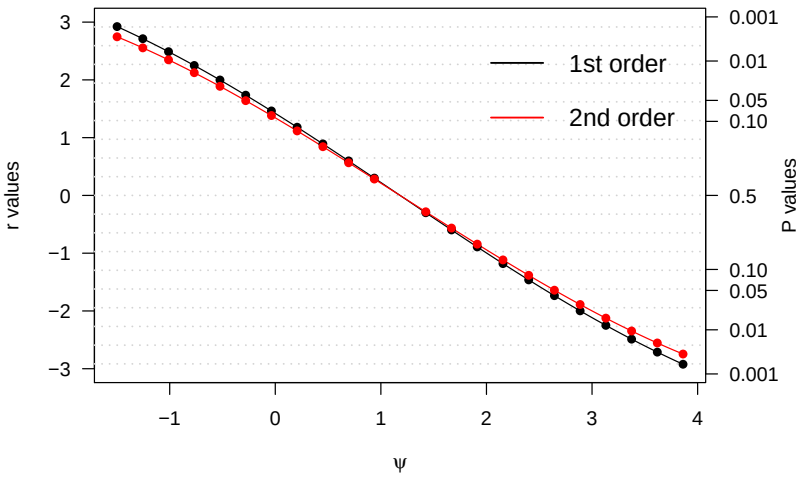

Figura 3.6: Grafico della statistica r_p (linea nera) e statistica r_p^* (linea rossa).

tutte le statistiche di sintesi ad essa associate. La mediana a posteriori risulta pari a 0.95 e l'intervallo di credibilità *equi-tailed* al 95% ha come estremi $(-0.67, 2.69)$. Il valore della mediana a posteriori risulta leggermente inferiore a quello ottenuto con la *matching prior* e l'intervallo di credibilità risulta molto simile.

Tramite la funzione `ttestBF` è stato possibile inoltre ricavare il *fattore di Bayes*. Il *fattore di Bayes* fornito da questa funzione va a testare l'ipotesi nulla in cui la media del primo gruppo sia pari alla media del secondo. In particolare va a valutare qualora il valore dell'effetto standardizzato sia pari a zero o meno. Per l'effetto usa come distribuzione a priori una Cauchy, mentre per σ^2 un'uniforme; dunque quello che si ottiene sarà un *fattore di Bayes JZS*. Il valore restituito è $B_{01} = 1.48$, valore che suggerisce che i dati supportano l'ipotesi nulla di omogeneità delle medie. Considerando infine la trasformazione di Kass e Raftery del *fattore di Bayes* si trova 1.17, e riferendosi alla Tabella 1.2 possiamo affermare vi sia evidenza debole contro H_1. Infine si è calcolata la misura di evidenza di Pereira-Stern che, risultando pari a 0.26, seppur con evidenza differente, conferma quanto affermato nel caso della *matching prior*: i dati supportano l'ipotesi nulla.

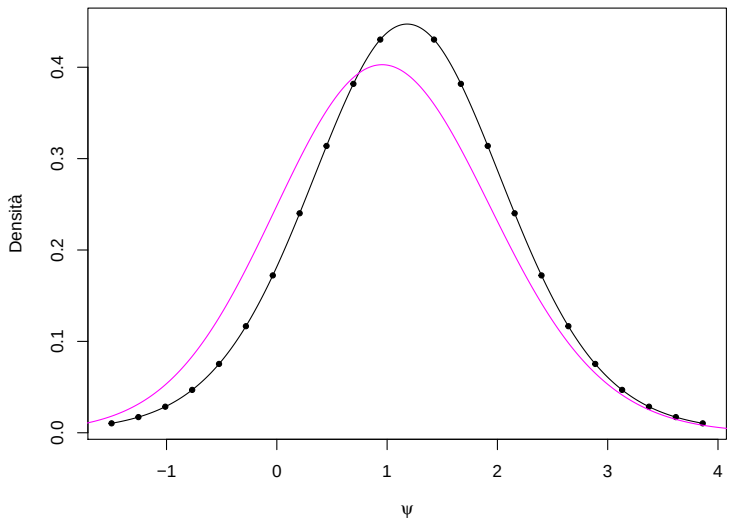

Figura 3.7: Distribuzioni a posteriori marginali con *matching prior* (linea nera punteggiata) e con *ZJS prior* (linea magenta).

Nel grafico riportato in Figura 3.7 sono state messe a confronto le distribuzioni a posteriori con *matching prior* e quella con *JZS prior*. Come si può notare, seppure gli intervalli *equi-tailed* avessero risultati molto simili, la distribuzione a posteriori con *matching prior* risulta con varianza inferiore. È da preferire la distribuzione con *matching prior* anche perché per la sua caratterizzazione basta specificare la distribuzione a priori per il parametro di interesse, nel caso di *JZS prior* invece occorre specificare la distribuzione a priori anche per quello di disturbo.

3.5 Caso di studio: Fibrosi cistica

La fibrosi cistica è un malattia genetica autosomica recessiva causata da una mutazione del gene CFTCR (*Cystic Fibrosis Trasmembrane Conductance Regulator*). Tale gene, codifica una proteina situata nella membrana cellulare dell'epitelio la cui funzione principale consiste nel trasportare il cloro attraverso le membrane cellulari. Questa malattia si manifesta in seguito

ad un'anomalia nel trasporto del cloro nella membrana delle cellule delle ghiandole a secrezione esterna, ghiandole che conseguentemente secernono un muco denso, vischioso e quindi poco fluido. Tutto ciò porta ad un'ostruzione dei dotti principali degli organi interessati, provocando l'insorgenza di gran parte delle manifestazioni cliniche tipiche della malattia, come la comparsa di infezioni polmonari ricorrenti, l'insufficienza pancreatica, stati di malnutrizione, cirrosi epatica, ostruzione intestinale e infertilità maschile.
I dati utilizzati sono contenuti nel dataset `cystfibr`, dataset messo a disposizione nella libreria `ISwR` di `R` (Dalgaard, 2008). Questo riporta i risultati di uno studio durante il quale si sono valutate le funzionalità del polmone in 25 pazienti affetti da questa malattia. In particolare vengono riportate le seguenti variabili: `age`, l'età dei pazienti in anni; `sex`, variabile dicotomica rappresentante il sesso; `height`,l'altezza in cm; `weight`, il peso in kg; `bmp`, l'indice di massa corporea; `fev1`, il volume respiratorio nel tempo di un secondo; `rv`, il volume residuo; `frc`, la capacità funzionale residua; `tlc`, la capacità totale del polmone e `pemax`, variabile risposta rappresentante l'indice della forza respiratoria massima (Altman, 1991).
La variabile risposta è interpretabile anche come misura della malnutrizione dei pazienti. Scopo dell'analisi è valutare qualora le variabili esplicative descritte in precedenza influenzino tale malnutrizione.

Per rispondere al problema in analisi si può stimare un modello di regressione lineare. Seguendo un approccio frequentista si può procedere come spiegato nella Sezione 2.4. Tuttavia in questo caso si presenta un possibile approccio bayesiano al problema.

Da una semplice analisi esplorativa iniziale (Figura 3.8) si può definire che le variabili più correlate con l'indice della forza respiratoria massima sono l'età, l'altezza, il peso e il volume respiratorio.

Dalla Figura 3.8 si può notare inoltre che alcune variabili esplicative, come la coppia età-altezza, età-peso, altezza-peso, oppure capacità funzionale residua-volume residuo, appaiono correlate tra loro. Si decide quindi di svolgere una procedura di selezione passo passo basata sull'AIC, il criterio più utilizzato per la selezione delle variabili, per definire quali di queste siano

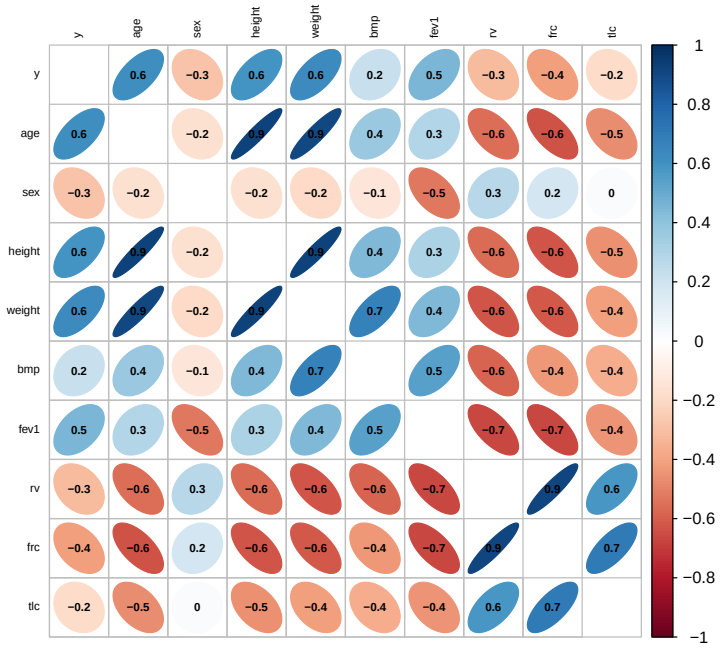

Figura 3.8: Diagramma della matrice di correlazione.

le più rilevanti penalizzando per la complessità del modello. Partendo dal modello completo, si valuta il guadagno in termini di AIC sia aggiungendo le variabili rimosse nei passi precedenti, sia togliendo variabili già presenti nel modello (procedura *stepwise both*). Le variabili selezionate da questa procedura risultano essere: il peso, l'indice di massa corporea, il volume respiratorio e il volume residuo.

Il modello del caso in analisi è definito come

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon,$$

dove x_1 rappresenta il peso, x_2 l'indice di massa corporea, x_3 il volume respiratorio e x_4 il volume residuo.

3.5.1 Modello bayesiano con *matching prior*

Dal momento che si vuole capire l'effettiva influenza del peso sull'indice della forza respiratoria massima, si procede con la specificazione di una *matching prior* come distribuzione a priori per il parametro di interesse $\psi = \beta_1$. Considerando in particolare l'approssimazione del secondo ordine della distribuzione a posteriori, tramite la libreria `LikelihoodAsy` si sono ottenuti tutti gli elementi che la caratterizzano. Da questa poi si possono ottenere tutte le statistiche di sintesi per rispondere al problema in analisi.
La mediana a posteriori risulta 1.77, mentre l'intervallo di credibilità a posteriori *equi-tailed* a livello 95% è $(0.9743, 2.5641)$. Osservando il grafico in Figura 3.9 si può valutare il grado di accuratezza dell'intervallo di confidenza basato su r_p rispetto a quello di confidenza basato su r_p^*.

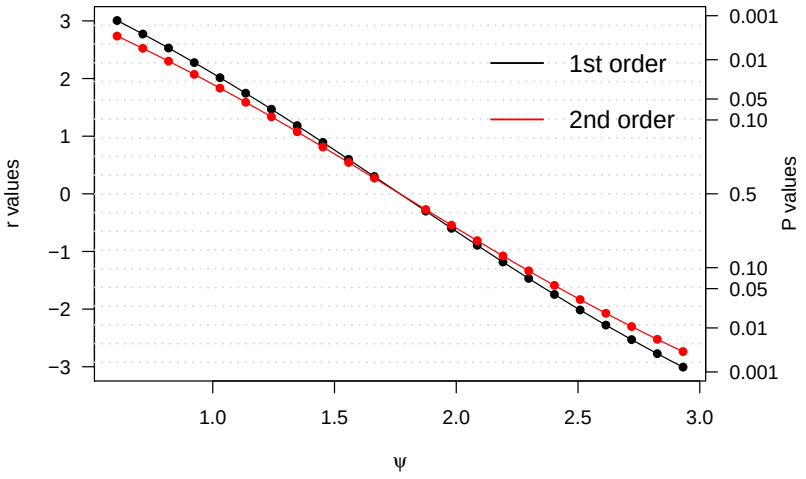

Figura 3.9: Grafico della statistica r_p (linea nera) e statistica r_p^* (linea rossa).

L'intervallo di confidenza basato su r_p al 95% è pari a $(1.052, 2.487)$ e presenta un'ampiezza leggermente inferiore rispetto a quello basato su r_p^* che coincide con quello di credibilità *equi-tailed* della distribuzione a posteriori marginale con *matching prior*. Infine, per verificare l'influenza del peso sul-

l'indice della forza respiratoria, ovvero per rispondere al seguente sistema di ipotesi:

$$\begin{cases} \mathrm{H}_0: & \beta_1 = 0 \\ \mathrm{H}_1: & \beta_1 \neq 0, \end{cases}$$

si è calcolata la misura di evidenza di Pereira-Stern. Risultando $\mathrm{EV} = 0.00005$, si può affermare che vi sia evidenza forte contro l'ipotesi nulla a favore di quella alternativa.

3.5.2 Modello bayesiano con *g-prior*

La distribuzione a priori definita di default per il modello lineare normale è la *g-prior* (Zellner, 1983). Si decide quindi di utilizzare anche questa distribuzione a priori per rispondere al caso in analisi. Per la sua specificazione si è utilizzato il metodo di campionamento di Gibbs, un algoritmo appartenente alla classe dei metodi MCMC usato spesso in ambito bayesiano in quanto crea una catena di Markov irriducibile e aperiodica, la cui distribuzione stazionaria equivale alla distribuzione a posteriori che si sta cercando. In particolare, un ciclo di simulazioni dal campionamento di Gibbs consiste nella generazione di un valore da ciascuna distribuzione condizionata (*full conditional*), ovvero dalla distribuzione di ciascun parametro di interesse condizionato a tutti gli altri. Procedendo in questo modo, viene simulata una catena di Markov che ha come distribuzione limite la distribuzione a posteriori cercata.
Nel caso in analisi, ricordando che il modello è definito come $Y = X\beta + \varepsilon$, con $\varepsilon \sim \mathrm{N}_n(0, \sigma^2 I_n)$, e che la distribuzione a priori è $\beta|\sigma^2 \sim \mathrm{N}_p(\beta_0, \frac{g}{\phi}(X^T X)^{-1})$, il cui iper-parametro $\phi = \frac{1}{\sigma^2}$ ha distribuzione a priori di Jeffreys, le distribuzioni condizionate a posteriori risultano

$$\pi(\beta|\phi, x, y) \sim \mathrm{N}_p(A^{-1}B, A^{-1}),$$

con $A = (g\sigma^2(X^T X)^{-1})^{-1} + X^T(\sigma^2 I)^{-1}X$,
con $B = (g\sigma^2(X^T X)^{-1})^{-1}\beta_0 + X^T(\sigma^2 I)^{-1}y$,
e

$$\pi(\phi|\beta, x, y) \sim \mathrm{Gamma}(C, D),$$

con $C = \frac{(n+3)}{2}$, $D = \frac{(\beta-\beta_0)^T X^T X(\beta-\beta_0)}{g} + \frac{(y-X\beta)^T(y-X\beta)}{2}$.
Per quanto riguarda il parametro scalare positivo g si decide di porlo inizialmente pari a 16 (p^2) come suggerito da Foster e George (1994) e in seguito si fissa pari a $25 = \max\{n, p^2\}$ come proposto da Fernandez *et al.* (2001), in questo caso si ottiene la cosiddetta *unit information prior*.
Per studiare la convergenza dell'algoritmo si è deciso di replicare 5 serie in parallelo ognuna composta da 10000 iterazioni[1].
Considerando inizialmente il caso in cui $g = 16$. La mediana a posteriori risulta 1.68 e l'intervallo *equi-tailed* a livello 95% pari a $(0.652, 2.68)$, valori che risultano leggermente traslati a sinistra rispetto a quelli ottenuti con la *matching prior*. Per rispondere al problema in analisi, viene calcolata la misura di evidenza di Pereira-Stern che risultando pari a 0.001 conferma quanto detto nel caso della *matching prior*, anche se con evidenza inferiore: i dati supportano H_1.
Ponendo $g = n$ si arriva alle stesse conclusioni. La mediana a posteriori risulta infatti 1.68, l'intervallo *equi-tailed* $(0.540, 2.700)$, la misura di evidenza di Pereira-Stern 0.002.
Nel grafico riportato in Figura 3.10 sono messe a confronto le distribuzioni a posteriori marginali ottenute. Come si può notare quella ottenuta con la *matching prior* risulta con una varianza inferiore rispetto a quelle ottenute con *g-prior*. Si preferisce la *matching prior* inoltre perchè quest'ultima richiede la specificazione della distribuzione a priori solamente per il parametro di interesse, a differenza di quelle ottenute con *g-prior* che necessitano la specificazione della distribuzione a priori anche per i parametri di disturbo e di fissare il valore dell'iper-parametro g. Si ritiene inoltre importante sottolineare la differenza del costo computazionale che richiedono: il campionamento di Gibbs ha un costo nettamente superiore.

3.6 Caso di studio: Cicatrici dovute al parto cesareo

Il dataset `parti` contiene informazioni riguardanti le caratteristiche del parto di 70 donne (Bland, 2009). In particolare, le variabili contenute nel

[1] In Appendice A vengono riportati i trace plot come verifica di convergenza dell'algoritmo.

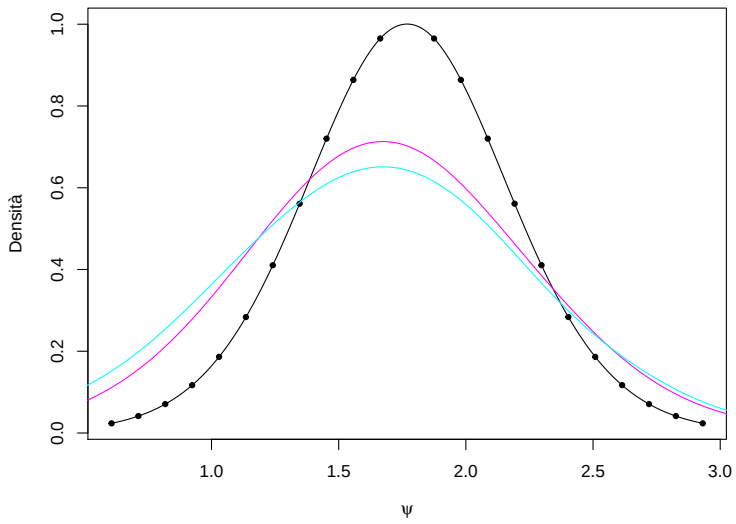

Figura 3.10: Distribuzioni a posteriori marginali con *matching prior* (linea nera punteggiata), con *g-prior* con $g = 16$ (linea azzurra) e con $g = n$ (linea magenta).

dataset sono: `ind`, presenza o assenza del parto indotto, `pvd`, aver già avuto parti naturali, `bmi`, indice di massa corporea della madre, e `y`, variabile di interesse che definisce qualora la donna abbia avuto un parto cesareo o meno. Dalla letteratura è noto che tutti questi fattori concorrono ad aumentare il rischio di parto cesareo. In particolare, aumenta il rischio per le persone obese (per i parti cesarei il BMI medio è pari a 27.27 mentre per quelli naturali è 22.02), per chi non ha già avuto parti naturali e per le donne alle quali il travaglio viene indotto.
Scopo di questo studio è quindi valutare qualora questi fattori influenzino effettivamente la probabilità di parto cesareo. Risulta di interesse capire quali siano le cause principali che portano al parto cesareo, in quanto le donne che ne hanno già affrontato uno in precedenza, cercano di avere un travaglio naturale alla gravidanza successiva per non mettere alla prova la cicatrice dovuta al parto cesareo.

Per rispondere al problema in analisi si può stimare un modello di regressione logistica. Seguendo un approccio frequentista si può procedere come descritto nella Sezione 2.5, quello che si vedrà ora invece è un possibile approccio bayesiano al problema.

Il modello per il problema trattato è definito come

$$\log\left(\frac{\theta}{1-\theta}\right) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

dove x_1 rappresenta la presenza o assenza del parto indotto, x_2 l'aver già avuto parti naturali e x_3 l'indice di massa corporea della madre.

3.6.1 Modello bayesiano con *matching prior*

Focalizzandosi sulla variabile relativa all'aver già avuto parti naturali, ovvero volendo valutare l'effetto di questa variabile sulla probabilità di parto cesareo, si considera come parametro di interesse $\psi = \beta_2$ e per questo si specifica una *matching prior*. In particolare, anche per questo caso, si considera l'approssimazione del secondo ordine della distribuzione a posteriori marginale. Una volta caratterizzata la distribuzione a posteriori grazie alle funzioni presenti nella libreria `LikelihoodAsy`, è possibile ottenere le statistiche di sintesi ad essa associate. La mediana a posteriori risulta 2.78 e l'intervallo di credibilità *equi-tailed* di livello 95% è $(1.23, 4.24)$. Tale intervallo coincide con quello di confidenza basato su r_p^*. Nella Figura 3.11 viene messo a confronto l'intervallo basato su r_p con quello basato su r_p^*. Come si può vedere la curva relativa alla statistica r_p^* si discosta da quella della statistica r_p, l'approssimazione alla normale standard dunque per r_p^* risulta migliore. L'intervallo basato su r_p^* presenta un'ampiezza inferiore rispetto a quello basato su r_p che al livello 95% è pari a $(1.323, 4.443)$. Infine, al fine di rispondere al quesito iniziale, si considera il seguente sistema di ipotesi

$$\begin{cases} \mathrm{H}_0: & \psi = 0 \\ \mathrm{H}_1: & \psi \neq 0. \end{cases}$$

Si calcola quindi la misura di evidenza di Pereira-Stern che, risultando pari a 0.0004, porta evidenza contro l'ipotesi nulla a favore di quella alternativa.

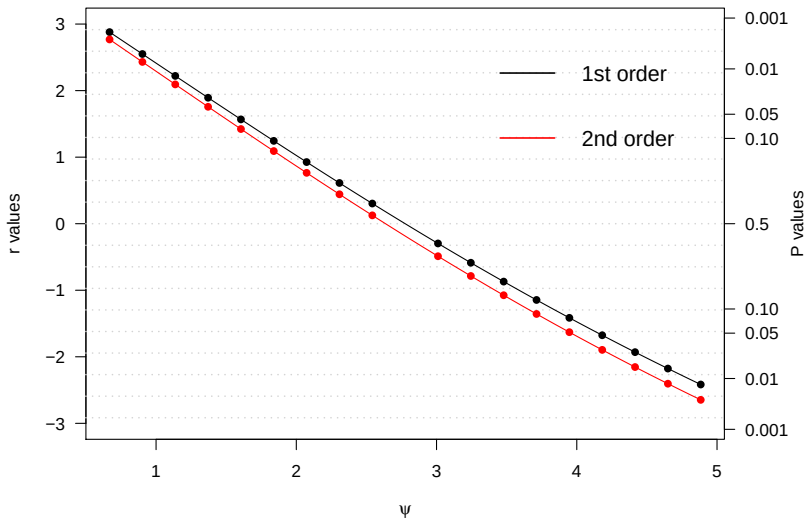

Figura 3.11: Grafico della statistica r_p (linea nera) e statistica r_p^* (linea rossa).

3.6.2 Modello bayesiano con *g-prior*

Una possibile distribuzione a priori alternativa è rappresentata dalla *g-prior*, distribuzione a priori di default per il modello di regressione lineare che Hanson *et al.* (2014) riadattano per il modello di regressione logistica. Per la caratterizzazione della distribuzione a posteriori si utilizza un algoritomo appartenente alla classe dei metodi MCMC chiamato *Metropolis-Hastings*. Tale algoritmo, comincia con un valore iniziale θ^0 e specifica una regola per simulare θ^t, il t-esimo valore in sequenza, dato θ^t. Questa procedura utilizza una *proposal density* $g(\theta^*|\theta^{t-1})$ per simulare un valore candidato θ^* e una probabilità di accettazione π, che specifica la probabilità con cui il valore candidato appena generato viene accettato come valore successivo della catena markoviana. Sotto condizioni di regolarità poste sulla *proposal density*, la sequenza di valori simulati converge ad una variabile casuale che è distribuita come la distribuzione a posteriori che si sta cercando. In particolare, dal momento che si utilizza una *proposal density* indipendente dal valore attuale della catena, l'algoritmo utilizzato prende il nome di

Metropolis-Hasting con indipendence sampler. In particolare, per studiare la convergenza dell'algoritmo si è deciso di replicare 5 serie in parallelo ognuna composta da 5000 iterazioni[2].
Come per il caso del modello lineare si decide di considerare un primo caso nel quale $g = n$ e un secondo caso nel quale si fissa $g = 9$, ovvero al quadrato del numero dei predittori.
Con $g = n$ la mediana a posteriori risulta 3.05, valore leggermente superiore a quello ottenuto con la *matching prior*, e l'intervallo di credibilità *equi-tailed* di livello 95% è $(1.17, 5.11)$. Tale intervallo risulta più ampio di quello ottenuto precedentemente.
Volendo rispondere al sistema d'ipotesi definito inizialmente si calcola anche per questo caso la misura di evidenza di Pereira-Stern che, risultando pari a 0.003 conferma quanto detto prima, seppur con evidenza inferiore: i dati supportano H_1.
Si ripetono quindi le stesse analisi ponendo $g = 9$. La mediana a posteriori è pari a 2.83, l'intervallo di credibilità *equi-tailed* di livello 95% risulta $(0.91, 4.44)$, la misura di evidenza di Pereira-Stern è 0.002. In questo caso la mediana e l'intervallo di credibilità *equi-tailed* risultano più vicini ai valori ottenuti con la *matching prior*. In ogni caso, anche questo modello conferma quanto detto precedentemente: vi è evidenza contro l'ipotesi nulla.

Nel grafico riportato in Figura 3.12 vengono messe a confronto le distribuzioni a posteriori marginali ottenute. Come si può osservare la distribuzione a posteriori con *matching prior* presenta una varianza inferiore rispetto alle altre due. Inoltre per la sua caratterizzazione non è necessario fissare il valore dell'iper-parametro scalare g e di specificare le distribuzioni a priori anche per i parametri di disturbo, cosa necessaria per la specificazione della *g-prior*. Infine, paragonando il costo computazionale richiesto, il *Metropolis-Hastings con Indipendence Sampler* presenta un costo computazionale maggiore.

[2]In appendice A vengono riportati i trace plot come verifica di convergenza dell'algoritmo.

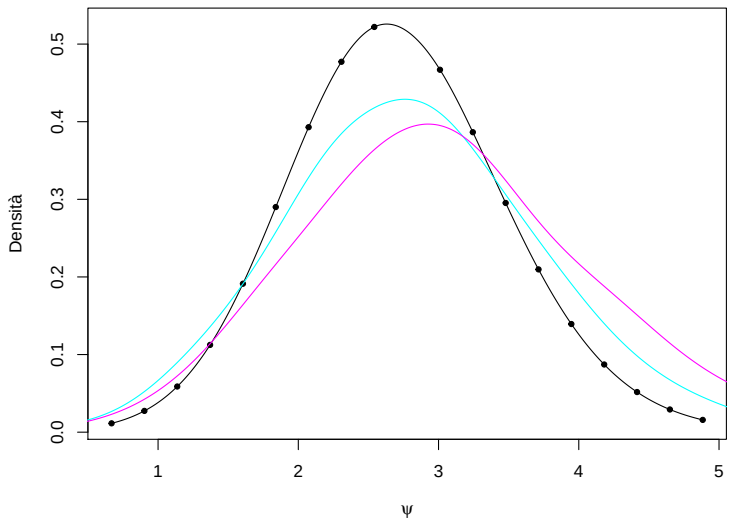

Figura 3.12: Distribuzioni a posteriori marginali con *matching prior* (linea nera punteggiata), con *g-prior* con $g = 16$ (linea azzurra) e con $g = n$ (linea magenta).

3.7 Caso di studio: Linfoma anaplastico a grandi cellule

I dati utilizzati derivano da uno studio riguardante il linfoma anaplastico a grandi cellule svolto dal centro oncologico pediatrico di Padova. In questo studio, in particolare, vengono rilevati i valori della proteina Hsp70 in pazienti con linfonodi sani (controlli) e in pazienti con linfonodi malati (casi). Nel dataset sono presenti 4 controlli e 10 casi. Scopo dell'analisi è quello di definire se tale proteina discrimini efficacemente i casi dai controlli. Ci si aspetta, infatti, che il paziente malato abbia un livello maggiore della proteina Hsp70 rispetto a quello sano.

Questo problema può essere riformulato in termini di AUC. Infatti considerando con X la variabile risposta di un gruppo di controllo e con Y quella di un trattamento, l'AUC può essere interpretato come misura dell'accura-

tezza del test diagnostico. In particolare, di seguito, si presenta una stima bayesiana di questo.

Dopo una breve analisi esplorativa (Figura 3.13) si assume che sia i casi che i controlli abbiano distribuzione esponenziale, assunzione confermata anche dal test di Kolmogorov-Smirnov.

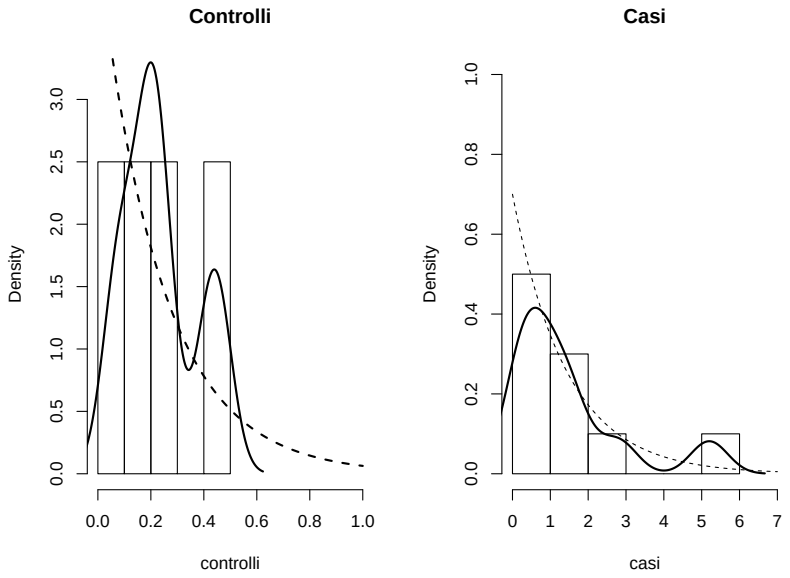

Figura 3.13: Distribuzioni empiriche e teoriche stimate dei livelli della proteina nei casi e nei pazienti.

3.7.1 Modello bayesiano con *matching prior*

Si consideri $X \sim \text{Exp}(\alpha)$ e $Y \sim \text{Exp}(\beta)$. Nel caso di distribuzione esponenziale vale $\text{AUC} = \frac{\alpha}{\alpha+\beta}$. Si pone quindi $\psi = \frac{\alpha}{\alpha+\beta}$ e per questo si specifica come distribuzione a priori una *matching prior*. In particolare, considerando l'approssimazione del secondo ordine della distribuzione a posteriori e utilizzando alcune funzioni della libreria `LikelihoodAsy` la mediana a posteriori risulta 0.859 e l'intervallo di credibilità *equi-tailed* a livello 95% è pari a $(0.6046, 0.9468)$, intervallo che coincide con quello di confidenza basato sul-

la statistica r_p^*. La funzione `rstar.ci` permette di confrontare l'intervallo di confidenza basato su r_p con quello basato su r_p^*. Come si può osservare dalla Figura 3.14, le curve delle due statistiche non coincidono e in particolare l'intervallo di confidenza basato su r_p^* risulta di ampiezza inferiore. L'intervallo di confidenza basato su r_p è pari a $(0.6265, 0.9479)$.

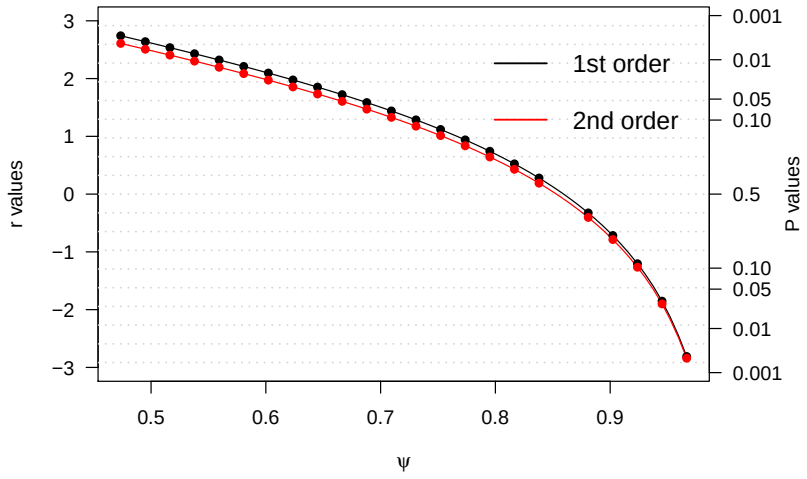

Figura 3.14: Grafico della statistica r_p (linea nera) e statistica r_p^* (linea rossa).

Per rispondere infine al problema in analisi, ovvero testando

$$\begin{cases} \mathrm{H}_0: & \psi = 0 \\ \mathrm{H}_1: & \psi \neq 0, \end{cases}$$

si calcola la misura di evidenza di Pereira-Stern. Dal momento che risulta $\mathrm{EV} = 0.004$, vi è evidenza contro l'ipotesi nulla a favore di quella alternativa.

3.7.2 Modello bayesiano con a priori Gamma

Una possibile distribuzione a priori per α e β è la distribuzione Gamma, distribuzione coniugata all'Esponenziale. La distribuzione a priori congiunta sarà quindi

$$\pi(\alpha, \beta) \propto \alpha^{\mu-1} e^{-\gamma\alpha} \beta^{\nu-1} e^{-\lambda\beta}, \tag{3.1}$$

con iper-parametri μ, γ, ν, $\lambda > 0$. La distribuzione a posteriori invece risulta

$$\pi(\alpha, \beta|x, y) \propto \alpha^{n_1+\mu-1} e^{-\alpha(\gamma+n_1\bar{X})} \beta^{n_2+\nu-1} e^{-\beta(\lambda+n_2\bar{Y})},$$

che è la congiunta di due distribuzioni gamma: Gamma($\mu^* = n_1 + \mu, \gamma^* = \gamma + n_1\bar{X}$) e Gamma($\nu^* = n_2 + \nu, \lambda^* = \lambda + n_2\bar{Y}$).
Per ottenere la distribuzione a priori per l'AUC, il parametro di interesse, definendo ψ=AUC, si considera la trasformazione F : $\psi = \frac{\alpha}{\alpha+\beta}$, $\lambda = \alpha + \beta$ e l'inversa Q : $\alpha = \psi\lambda$, $\beta = \psi(1 - \lambda)$.
Dal momento che lo jacobiano in questo caso è definito da

$$|J_Q(\psi, \lambda)| = \det \begin{bmatrix} \lambda & \psi \\ -\lambda & 1-\psi \end{bmatrix} = \lambda,$$

integrando rispetto a λ la distribuzione a posteriori marginale risulta

$$\pi(\psi|x, y) \propto \psi^{\mu^*-1}(1-\psi)^{\nu^*-1}(1-\mathrm{B}\psi)^{-(\mu^*+\nu^*)},$$

con $0 < \psi < 1$ e $\mathrm{B} = \frac{(\lambda^*-\gamma^*)}{\lambda^*} < 1$.
Definita la distribuzione a posteriori, fissando gli iper-parametri $\mu = \nu = 1$ e $\gamma = \lambda = 0.5$, è possibile ricavare tutte le statistiche di sintesi ad essa associate. La mediana a posteriori risulta 0.851 e l'intervallo di credibilità *equi-tailed* a livello 95% è $(0.5806, 0.9269)$; intervallo che presenta un'ampiezza leggermente inferiore rispetto a quello ottenuto dalla distribuzione a posteriori con *matching prior*. Anche in per questo caso viene calcolata la misura di evidenza di Pereira-Stern che risultando pari a 0.003 conferma quanto detto prima: i dati supportano H_1.
Si decide inoltre di valutare quanto incida la scelta dei valori degli iper-parametri sulla distribuzione a posteriori marginale data la numerosità campionaria ridotta. Si fissano quindi diversi valori per ciascun iper-parametro, si definisce la distribuzione a posteriori ad essa associata e per questa si calcolano le rispettive statistiche di sintesi e la misura di evidenza di Pereira-Stern in risposta al sistema di ipotesi specificato al paragrafo precedente.
Dai risultati riportati nella Tabella 3.3, si può osservare che le mediane a posteriori di tutte le distribuzioni a posteriori presentano circa lo stesso valore ad accezione di quella ottenuta fissando come iper-parametri $\alpha = \nu =$

Tabella 3.3: Statitiche di sintesi delle varie distribuzioni a posteriori ed EV

Iper-parametri	Mediana	Intervallo	EV
$\alpha = \nu = 1, \gamma = \lambda = 0.5$	0.851	(0.5806,0.9269)	0.003
$\alpha = \nu = 2, \gamma = \lambda = 0.5$	0.86	(0.6310,0.9291)	0.003
$\alpha = \nu = 3, \gamma = \lambda = 0.5$	0.867	(0.6677,0.9307)	0.002
$\alpha = \nu = 9, \gamma = \lambda = 2$	0.804	(0.6443, 0.8841)	0.004

$9, \gamma = \lambda = 2$ che risulta traslata verso sinistra, stessa distribuzione a posteriori che presenta un intervallo di credibilità *equi-tailed* al 95% di ampiezza inferiore. Tuttavia, osservando la misura di evidenza di Pereira-Stern, tutte le distribuzioni specificate, seppur con evidenza diversa, portano alla stessa conclusione: i dati supportano l'ipotesi alternativa.

3.7.3 Modello bayesiano con a priori di Jeffreys

Un'altra distribuzione alternativa per il caso in analisi è rappresentata dalla distribuzione a priori di Jeffreys. Tale distribuzione si può ottenere ponendo nella (3.1) $\gamma = \lambda = \mu = \nu = 0$. Di conseguenza la distribuzione a posteriori marginale risulta

$$\pi(\psi|x, y) \propto \psi^{n_1 - 1}(1 - \psi)^{n_2 - 1}(1 - \mathrm{B}\psi)^{-(n_1 + n_2)}.$$

La mediana a posteriori è pari a 0.887 e l'intervallo di credibilità *equi-tailed* a livello 95% è $(0.6045, 0.9468)$. Intervallo che coincide con quello ottenuto a partire dalla distribuzione a posteriori con *matching prior*, la mediana invece risulta traslata verso destra. La misura di evidenza di Pereira-Stern risulta pari a 0.003 portando le stesse conclusioni in risposta al caso in analisi ottenute con la distribuzione a posteriori con *matching prior* e con l'a priori Gamma congiunta.

Nel grafico riportato in Figura 3.15 vengono riportate tutte le distribuzioni a posteriori marginali ottenute. La distribuzione con a priori Gamma congiunta con i valori degli iper-parametri pari a $\alpha = \nu = 1, \gamma = \lambda = 0.5$ risulta preferibile a quella ottenuta con *matching prior* in quanto a parità del valore della mediana presenta varianza inferiore. Tuttavia, si nota come

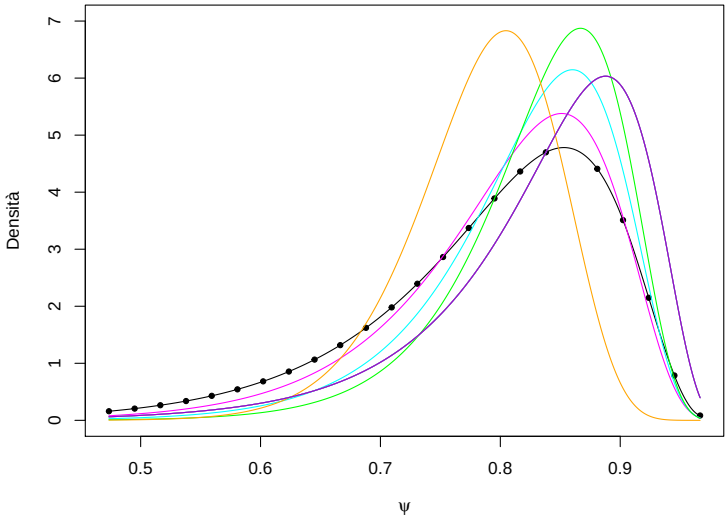

Figura 3.15: Distribuzioni a posteriori marginali con *matching prior* (linea nera punteggiata), con a priori di Jeffreys (linea viola), e con a priori Gamma congiunta con iper-parametri $\alpha = \nu = 1, \gamma = \lambda = 0.5$ (linea magenta), $\alpha = \nu = 2, \gamma = \lambda = 0.5$ (linea azzurra), $\alpha = \nu = 3, \gamma = \lambda = 0.5$ (linea verde), $\alpha = \nu = 9, \gamma = \lambda = 2$ (linea arancione).

la distribuzione con a priori Gamma congiunta cambi in base ai valori degli iper-parametri, fissando $\alpha = \nu = 9, \gamma = \lambda = 2$, ad esempio, la mediana si discosta molto da quella di tutte le altre distribuzioni a posteriori. La distribuzione a posteriori con *matching prior* non richiede, invece, di fissare gli iper-parametri per la sua specificazione.

3.8 Caso di studio: Aneurisma dell'aorta addominale

L'aneurisma dell'aorta addominale consiste in una dilatazione anomala dell'aorta addominale. Generalmente il calibro dell'aorta addominale è di 1.5 - 2 cm, qualora la misura del calibro dovesse superare di due volte questo valore si diagnostica un aneurisma.
I dati analizzati sono stati elaborati dal Dipartimento di Radiologia dell'U-

niversità degli Studi di Padova. In particolare, nel dataset `aorta` vengono riportate le misure del calibro dell'aneurisma dell'aorta di 48 pazienti suddivisi in due gruppi rispetto alla dimensione del calibro misurata, rilevate, rispettivamente, con lo strumento standard, ovvero la TAC, e un nuovo strumento che si basa sugli ultrasuoni (US).
Scopo dell'analisi è valutare la performance diagnostica del nuovo strumento di rilevazione, ovvero la sua accuratezza nel discriminare tra pazienti a rischio e pazienti non a rischio.
È di interesse avere misure accurate del calibro dell'aorta in quanto essenziali per diagnosticare e stabilire la gravità della malattia. Qualora il calibro della malattia dovesse superare i 5 *cm* la prassi clinica richiede l'intervento chirurgico, poiché è noto infatti che all'aumentare del calibro il rischio di rottura dell'aneurisma aumenta provocando la morte del paziente.

In ambito medico la performance diagnostica di un test può essere valutata mediante un'analisi statistica basata sulla curva ROC. Quello che si vedrà ora è un possibile modello bayesiano per la sua specificazione.

A seguito di una breve analisi descrittiva è risultato che la distribuzione del calibro misurato con la nuova tecnica presenta valori più elevati nel gruppo dei malati rispetto a quello dei sani. Inoltre risulta che entrambe le popolazioni sono distribuite normalmente con medie e varianze differenti (Figura 3.16). Si consideri quindi $X \sim \mathrm{N}(\mu_x, \sigma_x^2)$ e $Y \sim \mathrm{N}(\mu_y, \sigma_y^2)$.

3.8.1 Modello bayesiano con *matching prior*

Con lo scopo di valutare la performance della nuova tecnica di misurazione del calibro dell'aorta, si può considerare come parametro di interesse l'area sottesa alla curva ROC. Si ponga quindi $\psi = \Phi\left(-\frac{\mu_x-\mu_y}{\sqrt{\sigma_x^2+\sigma_y^2}}\right)$. Per il parametro di interesse si specifica una *matching prior* e in particolare per le analisi si considera l'approssimazione del secondo ordine della distribuzione a posteriori con *matching prior*. Grazie alle funzioni `rstar` e `rstar.ci` contenute nella libreria `LikelihoodAsy` è possibile ottenere tutti gli elementi che compongono la distribuzione a posteriori marginale e le statistiche di sintesi

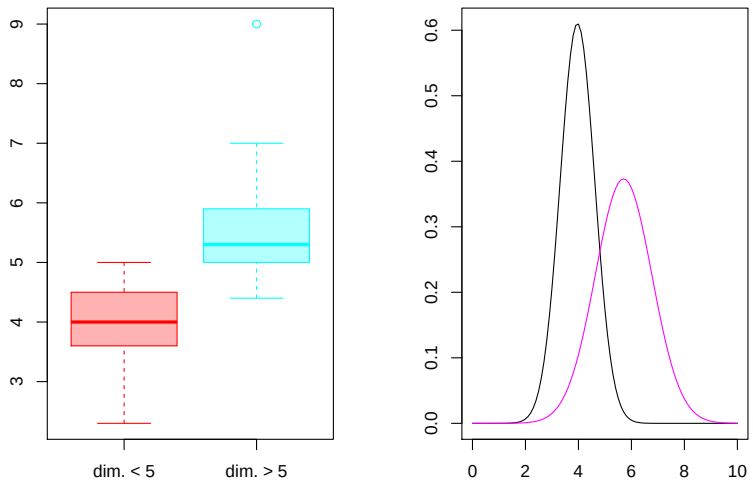

Figura 3.16: Boxplot e distribuzioni dei calibri misurati nei 2 gruppi.

ad essa associate. La mediana a posteriori risulta 0.923 e l'intervallo di credibilità *equi-tailed* al 95% è $(0.7895, 0.9733)$, intervallo che coincide con quello di confidenza basato su r_p^*. Tramite la funzione `rstar.ci` è possibile inoltre mettere a confronto gli intervalli di confidenza basati su r_p e su r_p^*. Come si può notare dal grafico riportato in Figura 3.17, l'intervallo di confidenza basato su r_p^* presenta un'ampiezza inferiore rispetto a quello basato su r_p che ha come estremi $(0.8044, 0.9761)$.

Infine per valutare la performance del test, viene calcolata la misura di evidenza di Pereira-Stern. In particolare considerando il sistema di ipotesi seguente

$$\begin{cases} \mathrm{H}_0: & \psi = 0 \\ \mathrm{H}_1: & \psi \neq 0, \end{cases}$$

è risultato $\mathrm{EV} = 0.0001$. Vi è dunque evidenza contro l'ipotesi nulla, ovvero i dati supportano H_1 rispetto a H_0.

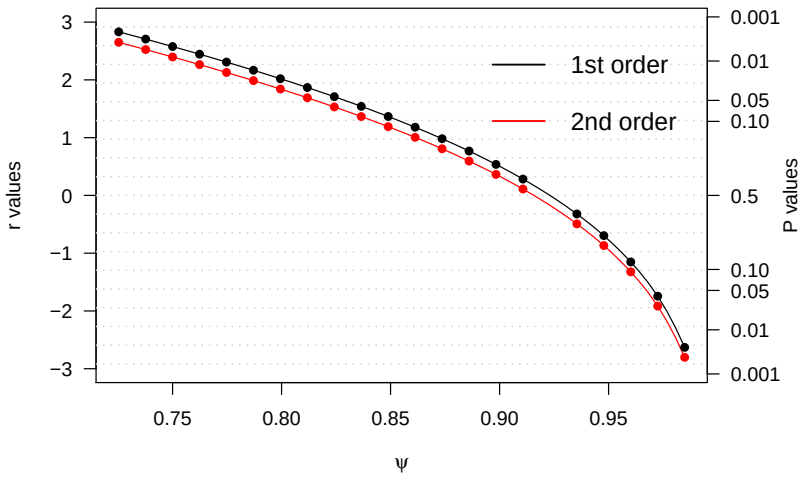

Figura 3.17: Grafico della statistica r_p (linea nera) e statistica r_p^* (linea rossa).

3.8.2 Modello bayesiano con a priori di Jeffreys

Una possibile distribuzione a priori per i parametri delle due distribuzioni normali considerate è rappresentata dalla distribuzione a priori congiunta di Jeffreys:

$$\pi(\theta) \propto \frac{1}{\sigma_x^2}\frac{1}{\sigma_y^2}.$$

La distribuzione a posteriori dunque sarà della forma

$$\pi(\theta|x,y) = (\sigma_x^2)^{-\frac{n_1}{2}-1}(\sigma_y^2)^{-\frac{n_2}{2}-1}\exp\left\{-\frac{\sum_{i=1}^{n_1}(x_i-\mu_x)^2}{2\sigma_x^2} - \frac{\sum_{i=1}^{n_2}(y_i-\mu_y)^2}{2\sigma_y^2}\right\},$$

con n_1 = numero delle osservazioni per i pazienti non a rischio, n_2 = numero delle osservazioni per i pazienti a rischio.
Con lo scopo di valutare la performance del nuovo strumento di rilevazione, il parametro di interesse dell'analisi è rappresentato dall'AUC, l'area sottesa alla curva ROC. Ponendo quindi AUC $= \psi = \Phi\left(-\frac{\mu_x-\mu_y}{\sqrt{\sigma_x^2+\sigma_y^2}}\right)$, è possibile ottenere la distribuzione a priori per ψ considerando la trasformazione F: $\psi = \Phi\left(-\frac{\mu_x-\mu_y}{\sqrt{\sigma_x^2+\sigma_y^2}}\right)$, $\mu_y = \mu_y$, $\sigma_x^2 = \sigma_x^2$, $\sigma_x^2 = \sigma_x^2$ e l'inversa Q:

$\mu_x = -\sqrt{\sigma_x^2 + \sigma_y^2}\Phi^{-1}(\psi) + \mu_y$, $\mu_y = \mu_y$, $\sigma_x^2 = \sigma_x^2$, $\sigma_x^2 = \sigma_x^2$.
Lo jacobiano della trasformata risulta dunque

$$
\begin{aligned}
|J_Q(\psi, \mu_y, \sigma_x^2, \sigma_y^2)| =& \det \begin{bmatrix} -\sqrt{\sigma_x^2 + \sigma_y^2}\frac{d(\Phi^{-1}(\psi))}{d\psi} & 1 & \frac{-\Phi^{-1}(\psi)}{2\sqrt{\sigma_x^2+\sigma_y^2}} & \frac{-\Phi^{-1}(\psi)}{2\sqrt{\sigma_x^2+\sigma_y^2}} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
=& \sqrt{\sigma_x^2 + \sigma_y^2}\frac{d(\Phi^{-1}(\psi))}{d\psi},
\end{aligned}
$$

e di conseguenza la distribuzione a posteriori della trasformata è

$$
\begin{aligned}
\pi(\psi, \mu_y, \sigma_x^2, \sigma_y^2 | x, y) \propto & (\sigma_x^2)^{-\frac{n_1}{2}-1}(\sigma_y^2)^{-\frac{n_2}{2}-1} \\
& \exp\left\{-\frac{\sum_{i=1}^{n_1}(x_i - (-\sqrt{\sigma_x^2 + \sigma_y^2}\Phi^{-1}(\psi) + \mu_y))^2}{2\sigma_x^2}\right\} \\
& \exp\left\{-\frac{\sum_{i=1}^{n_2}(y_i - \mu_y)^2}{2\sigma_y^2}\right\}\sqrt{\sigma_x^2 + \sigma_y^2}\frac{d(\Phi^{-1}(\psi))}{d\psi}.
\end{aligned}
$$

Integrando questa distribuzione rispetto ai parametri di disturbo (μ_y, σ_x^2 e σ_y^2) è possibile ottenere la distribuzione a posteriori marginale per ψ. Definita questa poi si possono ricavare tutte le statistiche di sintesi ad essa associate. La mediana a posteriori risulta pari a 0.928, valore leggermente superiore a quello ottenuto con *matching prior*, mentre l'intervallo di credibilità *equi-tailed* al 95% è $(0.8098, 0.9848)$, intervallo che presenta un'ampiezza molto simile rispetto a quello ottenuto con *matching prior*. Per rispondere al problema in analisi, anche per questo caso viene calcolata la misura di evidenza di Pereira-Stern che risultando pari a 0.0002 conferma quanto detto per il caso della *matching prior*: vi è evidenza contro l'ipotesi nulla.
Nel grafico riportato in Figura 3.18 vengono messe a confronto le curve delle distribuzioni a posteriori con *matching prior* e con distribuzione a priori di Jeffreys. In particolare, dato il costo computazionale per il calcolo della distribuzione a posteriori con a priori di Jeffreys, si decide di calcolare quest'ultima per pochi punti per poi in seguito lisciare quest'ultimi. Per rendere confrontabili le due curve sono stati scelti gli stessi punti utilizzati per il calcolo della distribuzione a posteriori con *matching prior*. Come si può notare le due curve risultano simili, quella di Jeffreys risulta leggermente traslata

verso destra; tuttavia l'onere computazionale richiesto per la specificazione della distribuzione a posteriori con Jeffreys è nettamente superiore.

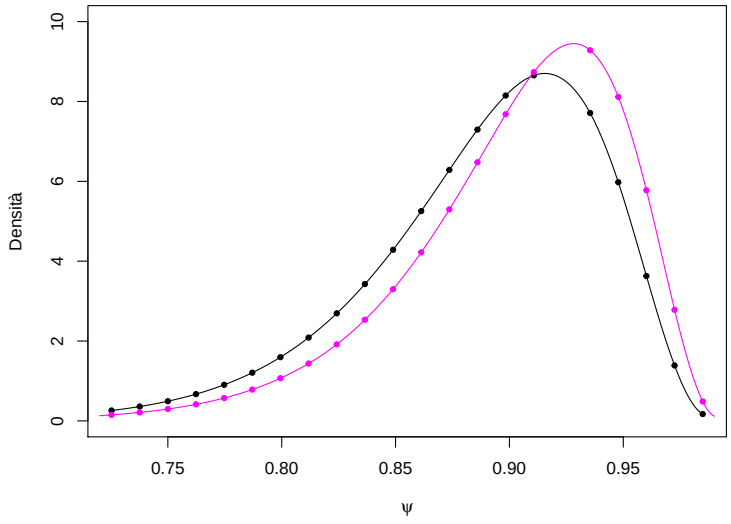

Figura 3.18: Distribuzioni a posteriori marginali con *matching prior* (linea nera punteggiata) e con a priori di Jeffreys (linea magenta punteggiata).

3.9 Conclusioni

In questo capitolo sono stati analizzati alcuni casi di studio inerenti a dati reali medici al fine di valutare le performance dell'approccio bayesiano proposto basato su *matching prior* in quest'ambito, in particolare confrontandolo con alcune proposte bayesiane note in letteratura.
In base a queste analisi si può affermare che la *matching prior* è una buona distribuzione a priori alternativa per i casi di studio trattati in quanto compete molto bene con l'approccio bayesiano trattato in letteratura, rispetto al quale inoltre non richiede la specificazione della distribuzione a priori per il parametro di disturbo e di fissare arbitrariamente il valore degli iper-parametri. In alcuni casi inoltre riduce di molto l'onere computazionale richiesto.

Conclusioni

Le distribuzioni *matching priors* appartengono alla classe delle distribuzioni a priori oggettive, ovvero alle a priori utilizzate quando non si hanno informazioni soggettive in merito al caso analizzato, e rappresentano il ponte tra il mondo bayesiano e quello frequentista andando a determinare regioni di credibilità bayesiane con validità frequentista.
In questa tesi, tramite la loro specificazione, si è proposta una distribuzione a priori alternativa a quelle già trattate in letteratura, per analizzare casi di studio in ambito medico. A tale scopo, inizialmente, ci si è soffermati su una descrizione sintetica dell'approccio bayesiano, delineando il Teorema di Bayes e la distribuzione a posteriori, caratterizzando le diverse distribuzioni a priori, con particolare attenzione alle *matching prior*, e descrivendo l'inferenza bayesiana. In seguito sono stati trattati alcuni classici casi di studio in Biostatistica. In particolare è stata analizzata l'inferenza su una proporzione, su due proporzioni, sulla media della normale, su un parametro scalare di regressione e sull'area sotto la curva ROC. Per ogni caso trattato, viene descritto l'usuale approccio frequentista, l'approccio bayesiano con *matching prior* e alcune proposte bayesiane trattate in letteratura. Infine, per valutare le performance dell'approccio bayesiano proposto, si sono analizzati casi di studio inerenti a dati reali medici confrontando i risultati ottenuti con le proposte bayesiane note in letteratura.
In base a queste analisi, si può affermare che la *matching prior* è una buona distribuzione a priori alternativa per i casi di studio trattati, infatti compete molto bene con l'approccio bayesiano trattato in letteratura, rispetto al quale inoltre non richiede la specificazione della distribuzione a priori per il parametro di disturbo e di fissare arbitrariamente il valore degli iper-parametri.

In alcuni casi inoltre riduce di molto l'onere computazionale richiesto.

Bibliografia

Altman, D. G. (1991). *Practical statistics for medical research*. Chapman e Hall.

Bayarri, M. J. e García-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika* **94**, 135–152.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B* **41**, 113–147.

Bland, M. (2009). *Statistica medica*. Apogeo Editore.

Bolstad, W. M. e Curran, J. M. (2016). *Introduction to Bayesian statistics*. John Wiley & Sons.

Brazzale, A. R., Davison, A. C. e Reid, N. (2007). *Applied asymptotics: case studies in small-sample statistics*. Cambridge University Press.

Cabras, S., Ventura, L. e Racugno, W. (2016). Higher-order asymptotics for Bayesian significance tests for precise null hypotheses in the presence of nuisance parameters. *Journal of Statistical Computation and Simulation* **85**, 2989–3001.

Cifarelli, D. M. e Muliere, P. (1989). *Statistica bayesiana: appunti ad uso degli studenti*. Iuculano.

Dalgaard, P. (2008). *Introductory statistics with R*. Springer.

Datta, G. S. e Mukerjee, R. (2004). *Probability matching priors higher order asymptotics*. Springer.

Fernandez, C., Ley, E. e Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100**, 381–427.

Foster, D. P. e George, E. I (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics* **22**, 1947–1975.

Hanson, T. E., Branscum, A. J. e Johnson, W. O. (2014). Informative g-Priors for Logistic Regression. *Bayesian Analysis* **9**, 597–612.

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.

Kass, R. E. e Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Koch, K.-R. (2007). *Introduction to Bayesian statistics*. Springer.

Kotz, S., Lumelskii, Y. e Pensky, M. (2003). *The stress-strength model and its generalizations: theory and applications*. World Scientific.

Lee, P. M. (2012). *Bayesian statistics: an introduction*. John Wiley & Sons.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. e Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* **103**, 410–423.

Liseo, B. (2008). *Introduzione alla statistica bayesiana*. Roma.

Madruga, M. R., Esteves, L. G. e Wechsler, S. (2001). On the Bayesianity of Pereira-Stern tests. *Test* **10**, 291–299.

Maudgal, D.P., Ang, L., Patel, S., Bland, J.M. e Maxwell, J.D. (1985). Nutritional assessment in patients with chronic gastrointestinal symptoms: comparison of functional and organic disorders. *Human nutrition. Clinical nutrition* **39**, 203–212.

Nicolaou, A. (1993). Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *Journal of the Royal Statistical Society. Series B* **55**, 377–390.

Peers, H.W. (1965). On confidence points and Bayesian probability points in the case of several parameters. *Journal of the Royal Statistical Society. Series B* **27**, 9–16.

Pereira, C. A. e Stern, J. M. (1999). Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy* **1**, 99–110.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. e Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review* **16**, 225–237.

Santos, S. I *et al.* (1999). *Cancer epidemiology, principles and methods*. IARCPress Lyon.

Staicu, A.-M. e Reid, N. M. (2008). On probability matching priors. *Canadian Journal of Statistics* **36**, 613–622.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science* **240**, 1285–1293.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604–608.

Ventura, L. e Racugno, W. (2017). *Biostatistica. Casi di studio in R*. Egea.

Ventura, L. e Reid, N. (2014). Approximate Bayesian computation with modified log-likelihood ratios. *Metron* **72**, 231–245.

Ventura, L., Cabras, S. e Racugno, W. (2009). Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *Journal of the American Statistical Association* **104**, 768–774.

Ventura, L., Sartori, N. e Racugno, W. (2013). Objective Bayesian higher-order asymptotics in models with nuisance parameters. *Computational Statistics & Data Analysis* **60**, 90–96.

Zellner, A. (1983). Applications of Bayesian analysis in econometrics. *Journal of the Royal Statistical Society* **32**, 23–34.

Zellner, A. e Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de estadística y de investigación operativa* **31**, 585–603.

Appendice A

Appendice: Materiale aggiuntivo

A.0.1 Traceplot per il modello lineare

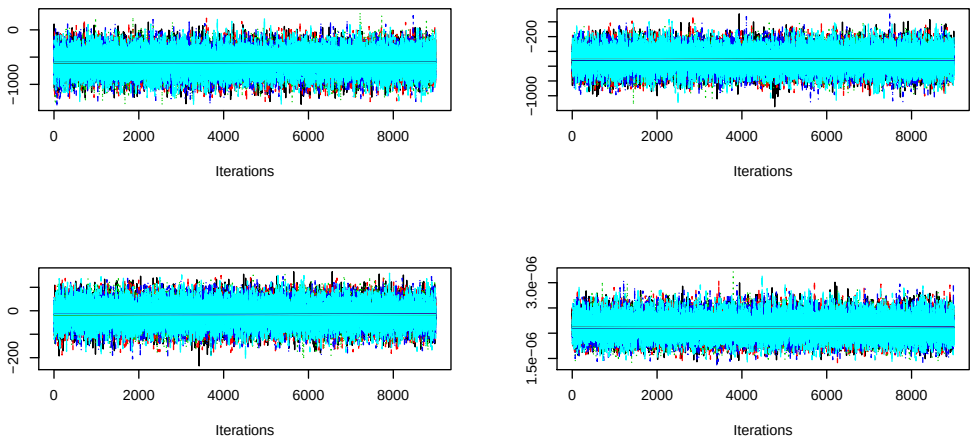

Figura A.1: traceplot.

Come si può osservare tutte le serie risultano essere stazionarie, quindi le distribuzioni che si ottengono rappresentano proprio le distribuzioni a posteriori che si stanno cercando.

A.0.2 Trace plot per la regressione logistica

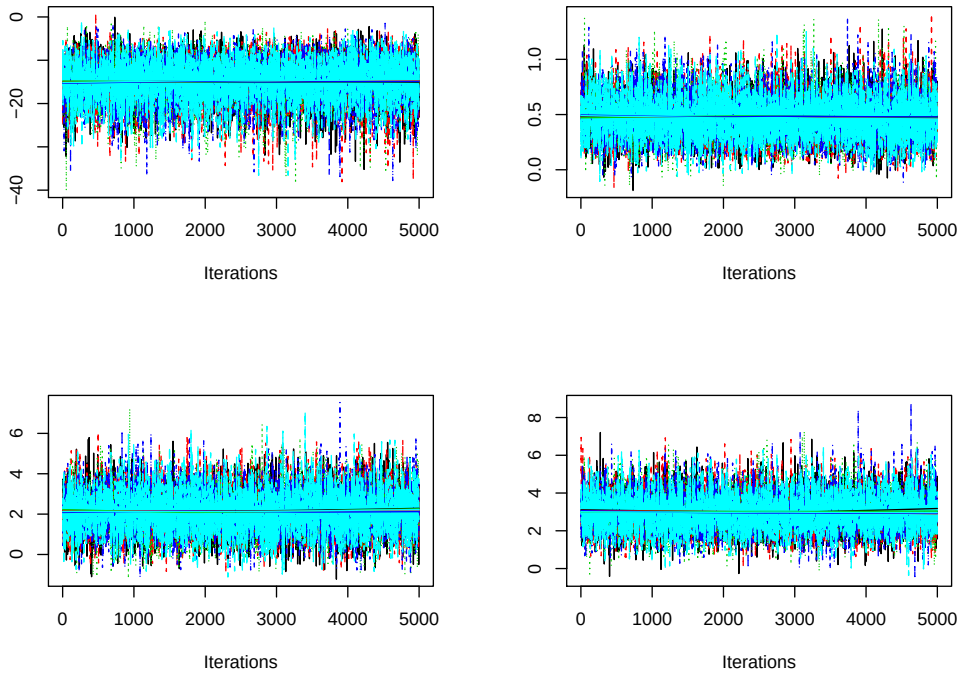

Figura A.2: traceplot.

Come per il modello lineare, anche in questo caso tutte le serie appaiono stazionarie ad indicare che le distribuzioni che si sono ottenute coincidono con le distribuzioni a posteriori che si stanno cercando.

Appendice B

Appendice: Codice R

Si riporta il codice `R` utilizzato per l'implementazione degli esempi con `likelihoodAsy`.

```
TEST SU UNA PROPORZIONE

oss<-list(y=malati)

#Log-verosimiglianza
loglik<-function(theta,data){
p1<-theta[1]
l<-sum(dbinom(data$y,1,p1, log=TRUE))
return(l)
}

#Funzione generatrice dei dati
gendat<-function(theta,data){
out<-data
w<-length(data$y)
out$y<-rbinom(w,1, theta[1])
return(out)
}

#Parametro interesse
```

```
psi<-function(theta){
ris<-theta[1]
return(ris)
}

#Calcolo rstar
rs<-rstar(data=oss, thetainit=0.5, floglik=loglik, psival=0, fpsi=psi,
    datagen=gendat)

#Calcolo intervallo di rstar
rs.ci<-rstar.ci(data=oss, thetainit=0.5, floglik=loglik, fpsi=psi,
    datagen=gendat)

#Grafico della posteriori
myint<-function(foo, npoints=10^4)
{
sm.foo<-spline(foo$x, foo$y, npoints)
area<-rep(NA, length(sm.foo$x)-1)
for(i in 1:(length(sm.foo$x)-1))
area[i]<-(sm.foo$x[i+1]-sm.foo$x[i])*mean(c(sm.foo$y[i+1],sm.foo$y[i]))
sum(area)
}

foo<-list(x=NULL, y=NULL)
foo$x<-rs.ci$psivals
foo$y<-exp(-0.5*rs.ci$rsvals^2)
foo$y<-foo$y/myint(foo)
sm.foo<-spline(foo$x, foo$y, 400)
plot(sm.foo$x, sm.foo$y, type="l", ylab="Densita'", xlab=expression(psi),
    ylim=c(0,4))
points(foo$x, foo$y, cex=1, pch=20)

#Calcolo EV
EV<-2*pnorm(0.3, rs$psi.hat,rs$se.psi.hat)
```

TEST SU DUE PROPORZIONI

```
oss<-list(y=mm, x=ff)

#Log-verosimiglianza
loglik<-function(theta,data){
p1<-theta[1]
p2<-theta[2]
l<-sum(dbinom(data$y,1,p1,log=TRUE))+
sum(dbinom(data$x,1,p2,log=TRUE))
return(l)
}

#Funzione generatrice dei dati
gendat<-function(theta,data){
out<-data
w<-length(data$y)
z<-length(data$x)
out$y<-rbinom(w,1, theta[1])
out$x<-rbinom(z,1, theta[2])
return(out)
}

#Parametro interesse
psi<-function(theta){
ris<-theta[1]-theta[2]
return(ris)
}

#calcolo rstar
rs<-rstar(data=oss, thetainit=c(0.5,0.5), floglik=loglik, psival=0, fpsi=
    psi, datagen=gendat)

#calcolo intervallo di rstar
rs.ci<-rstar.ci(data=oss, thetainit=p.init, floglik=loglik, fpsi=psi,
    datagen=gendat)
```

```
#Grafico della posteriori
myint<-function(foo, npoints=10^4)
{
sm.foo<-spline(foo$x, foo$y, npoints)
area<-rep(NA, length(sm.foo$x)-1)
for(i in 1:(length(sm.foo$x)-1))
area[i]<-(sm.foo$x[i+1]-sm.foo$x[i])*mean(c(sm.foo$y[i+1],sm.foo$y[i]))
sum(area)
}

foo<-list(x=NULL, y=NULL)
foo$x<-rs.ci$psivals
foo$y<-exp(-0.5*rs.ci$rsvals^2)
foo$y<-foo$y/myint(foo)
sm.foo<-spline(foo$x, foo$y, 400)
plot(sm.foo$x, sm.foo$y, type="l", ylab="Densita'", xlab=expression(psi),
    ylim=c(0,4.8))
points(foo$x, foo$y, cex=1, pch=20)

#Calcolo EV
EV<-2*pnorm(0, rs$psi.hat,rs$se.psi.hat)
```

```
T-TEST

oss<-list(y=rm, x=ct)

#Log-verosimiglianza
loglik<-function(theta,data){
mu1<-theta[1]
sigma1<-exp(theta[2])
mu2<-theta[3]
sigma2<-exp(theta[4])
l<-sum(dnorm(data$y,mu1,sigma1, log=TRUE))+sum(
dnorm(data$x,mu2,sigma2, log=TRUE))
return(l)
}
```

```
#Funzione generatrice dei dati
gendat<-function(theta,data){
out<-data
n<-length(data$y)
m<-length(data$x)
out$y<-rnorm(n,theta[1], exp(theta[2]))
out$x<-rnorm(m,theta[3], exp(theta[4]))
return(out)
}

#Parametro interesse
psi<-function(theta){
ris<-theta[1]-theta[3]
return(ris)
}

#Calcolo rstar
rs<-rstar(data=oss, thetainit=c(1,1,0,0), floglik=loglik, psival=0, fpsi=
    psi, datagen=gendat)

#Calcolo intervallo di rstar
rs.ci<-rstar.ci(data=oss, thetainit=p.init, floglik=loglik, fpsi=psi,
    datagen=gendat)

#Grafico della posteriori
myint<-function(foo, npoints=10^4)
{
sm.foo<-spline(foo$x, foo$y, npoints)
area<-rep(NA, length(sm.foo$x)-1)
for(i in 1:(length(sm.foo$x)-1))
area[i]<-(sm.foo$x[i+1]-sm.foo$x[i])*mean(c(sm.foo$y[i+1],sm.foo$y[i]))
sum(area)
}

foo<-list(x=NULL, y=NULL)
```

```
foo$x<-rs.ci$psivals
foo$y<-exp(-0.5*rs.ci$rsvals^2)
foo$y<-foo$y/myint(foo)
sm.foo<-spline(foo$x, foo$y, 400)
plot(sm.foo$x, sm.foo$y, type="l", ylab="Densita'", xlab=expression(psi))
points(foo$x, foo$y, cex=1, pch=20)

#Calcolo EV
EV<-2*pnorm(0, rs$psi.hat,rs$se.psi.hat)
```

```
MODELLO LINEARE

a<-dati$y
b<-dati$weigth
c<-dati$bmi
d<-dati$fev1
e<-dati$rv
oss<-list(y=a, b=b, c=c, d=d, e=e)

#Log-verosimiglianza
loglik<-function(theta,data){
beta1<-theta[1]
beta2<-theta[2]
beta3<-theta[3]
beta4<-theta[4]
sigma<-exp(theta[5])
eta<-beta1*data$b+beta2*data$c+beta3*data$d+beta4*data$e
l <- sum(dnorm(data$y, eta, sigma, log=T))
return(l)
}

#Funzione generatrice dei dati
gendat<-function(theta,data){
out<-data
beta1<-theta[1]
beta2<-theta[2]
```

```
beta3<-theta[3]
beta4<-theta[4]
sigma<-exp(theta[5])
n<-length(data$y)
eta<-beta1*data$b+beta2*data$c+beta3*data$d+beta4*data$e
out$y<-rnorm(n, eta, sigma)
return(out)
}

#Parametro interesse
psi<-function(theta){
ris<-theta[1]
return(ris)
}

#Calcolo rstar
rs<-rstar(data=oss, thetainit=c(1,1,1,1,0), floglik=loglik, psival=0, fpsi
    =psi, datagen=gendat)

#Calcolo intervallo di rstar
rs.ci<-rstar.ci(data=oss, thetainit=p.init, floglik=loglik, fpsi=psi,
    datagen=gendat)

#Grafico della posteriori
myint<-function(foo, npoints=10^4)
{
sm.foo<-spline(foo$x, foo$y, npoints)
area<-rep(NA, length(sm.foo$x)-1)
for(i in 1:(length(sm.foo$x)-1))
area[i]<-(sm.foo$x[i+1]-sm.foo$x[i])*mean(c(sm.foo$y[i+1],sm.foo$y[i]))
sum(area)
}

foo<-list(x=NULL, y=NULL)
foo$x<-rs.ci$psivals
foo$y<-exp(-0.5*rs.ci$rsvals^2)
```

```
foo$y<-foo$y/myint(foo)
sm.foo<-spline(foo$x, foo$y, 400)
plot(sm.foo$x, sm.foo$y, type="l", ylab="Densita'", xlab=expression(psi))
points(foo$x, foo$y, cex=1, pch=20)

#Calcolo EV
EV<-2*pnorm(0, rs$psi.hat,rs$se.psi.hat)
```

```
MODELLO LOGISTICO

a<-dati$y
b<-dati$bmi
c<-dati$ind
d<-dati$pvd
oss<-list(y=a, x=b, z=c, w=d)

#Log-verosimiglianza
loglik<-function(theta,data) {
lin<-theta[1]*data$x + theta[2]*data$z + theta[3]*data$w
l<-sum(data$y * lin - log( 1 + exp(lin)))
return(l)
}

#Funzione generatrice dei dati
gendat<-function(theta,data){
out<-data
n<-length(data$y)
out$y<-rlogis(n)
return(out)
}

#Parametro interesse
psi<-function(theta){
ris<-theta[3]
return(ris)
}
```

```
#Calcolo rstar
rs<-rstar(data=oss, thetainit=c(0.5,0.5,0.5), floglik=loglik, psival=0,
    fpsi=psi, datagen=gendat)

#Calcolo intervallo di rstar
rs.ci<-rstar.ci(data=oss, thetainit=p.init, floglik=loglik, fpsi=psi,
    datagen=gendat)

#Grafico della posteriori
myint<-function(foo, npoints=10^4)
{
sm.foo<-spline(foo$x, foo$y, npoints)
area<-rep(NA, length(sm.foo$x)-1)
for(i in 1:(length(sm.foo$x)-1))
area[i]<-(sm.foo$x[i+1]-sm.foo$x[i])*mean(c(sm.foo$y[i+1],sm.foo$y[i]))
sum(area)
}

foo<-list(x=NULL, y=NULL)
foo$x<-rs.ci$psivals
foo$y<-exp(-0.5*rs.ci$rsvals^2)
foo$y<-foo$y/myint(foo)
sm.foo<-spline(foo$x, foo$y, 400)
plot(sm.foo$x, sm.foo$y, type="l", ylab="Densita'", xlab=expression(psi))
points(foo$x, foo$y, cex=1, pch=20)

#Calcolo EV
EV<-2*pnorm(0, rs$psi.hat,rs$se.psi.hat)
```

```
AUC CON ESPONENZIALE

oss<-list(y=casi, x=controlli)

#log-verosimiglianza
loglik<-function(theta,data){
```

```
alpha<-exp(theta[1])
beta<-exp(theta[2])
l<-sum(dexp(data$x,alpha,log=T))+sum(dexp(data$y,beta,log=T))
return(l)
}

#Funzione generatrice dei dati
gendat<-function(theta,data){
out<-data
n<-length(data$x)
m<-length(data$y)
alpha<-exp(theta[1])
beta<-exp(theta[2])
out$x<-rexp(n,alpha)
out$y<-rexp(m,beta)
return(out)
}

#Parametro interesse
psi<-function(theta){
alpha<-exp(theta[1])
beta<-exp(theta[2])
ris<-alpha/(alpha+beta)
return(ris)
}

#Calcolo rstar
rs<-rstar(data=oss, thetainit=c(1,1), floglik=loglik, psival=0, fpsi=psi,
    datagen=gendat)

#Calcolo intervallo di rstar
rs.ci<-rstar.ci(data=oss, thetainit=p.init, floglik=loglik, fpsi=psi,
    datagen=gendat)

#Grafico della posteriori
myint<-function(foo, npoints=10^4)
```

```
{
sm.foo<-spline(foo$x, foo$y, npoints)
area<-rep(NA, length(sm.foo$x)-1)
for(i in 1:(length(sm.foo$x)-1))
area[i]<-(sm.foo$x[i+1]-sm.foo$x[i])*mean(c(sm.foo$y[i+1],sm.foo$y[i]))
sum(area)
}

foo<-list(x=NULL, y=NULL)
foo$x<-rs.ci$psivals
foo$y<-exp(-0.5*rs.ci$rsvals^2)
foo$y<-foo$y/myint(foo)
sm.foo<-spline(foo$x, foo$y, 400)
plot(sm.foo$x, sm.foo$y, type="l", ylab="Densita'", xlab=expression(psi),
    ylim=c(0,7))
points(foo$x, foo$y, cex=1, pch=20)

#Calcolo EV
posteriori<-exp(-0.5*rs.ci$rsvals^2)
med<-rs$psi.hat
per_st<-function(theta0)
{
theta1<-uniroot(function(y) posteriori(y)-posteriori(theta0),
if(theta0<med) c(med+10^-3,1-10^-8) else c(10^-8,med-10^-3))$root
x<-sort(c(theta0,theta1))
pnorm(x[1], rs$psi.hat,rs$se.psi.hat)+(1-pnorm(x[2], rs$psi.hat,rs$se.psi.
    hat))
}
per_st(0)
```

```
AUC CON NORMALE

oss<-list(y=malati, x=sani)

#Log-verosimiglianza
loglik<-function(theta,data){
```

```
mu1<-theta[1]
sigma1<-exp(theta[2])
mu2<-theta[3]
sigma2<-exp(theta[4])
l<-sum(dnorm(data$y,mu1,sigma1, log=TRUE))+sum(
dnorm(data$x,mu2,sigma2, log=TRUE))
return(l)
}

#Funzione generatrice dei dati
gendat<-function(theta,data){
out<-data
n<-length(data$y)
m<-length(data$x)
out$y<-rnorm(n,theta[1], exp(theta[2]))
out$x<-rnorm(m,theta[3], exp(theta[4]))
return(out)
}

#Parametro interesse
psi<-function(theta){
q<- -((theta[3]-theta[1])/(sqrt(exp(theta[2])^2+exp(theta[4])^2)))
ris<-pnorm(q)
return(ris)
}

#Calcolo rstar
rs<-rstar(data=oss, thetainit=c(1,0,1,0), floglik=loglik, psival=0, fpsi=
    psi, datagen=gendat)

#Calcolo intervallo di rstar
rs.ci<-rstar.ci(data=oss, thetainit=p.init, floglik=loglik, fpsi=psi,
    datagen=gendat)

#Grafico della posteriori
myint<-function(foo, npoints=10^4)
```

```
{
sm.foo<-spline(foo$x, foo$y, npoints)
area<-rep(NA, length(sm.foo$x)-1)
for(i in 1:(length(sm.foo$x)-1))
area[i]<-(sm.foo$x[i+1]-sm.foo$x[i])*mean(c(sm.foo$y[i+1],sm.foo$y[i]))
sum(area)
}

foo<-list(x=NULL, y=NULL)
foo$x<-rs.ci$psivals
foo$y<-exp(-0.5*rs.ci$rsvals^2)
foo$y<-foo$y/myint(foo)
sm.foo<-spline(foo$x, foo$y, 400)
plot(sm.foo$x, sm.foo$y, type="l", ylab="Densita'", xlab=expression(psi),
    ylim=c(0,10))
points(foo$x, foo$y, cex=1, pch=20)

#Calcolo EV
posteriori<-exp(-0.5*rs.ci$rsvals^2)
med<-rs$psi.hat
per_st <- function(theta0)
{
theta1<-uniroot(function(y) posteriori(y)-posteriori(theta0),
if(theta0<med) c(med+10^-3,1-10^-8) else c(10^-8,med-10^-3))$root
x<-sort(c(theta0,theta1))
pnorm(x[1], rs$psi.hat,rs$se.psi.hat)+(1-pnorm(x[2], rs$psi.hat,rs$se.psi.
    hat))
}
per_st(0)
```

Ringraziamenti

Le sensazioni che si provano quando si acquisisce la consapevolezza che la carriera universitaria è giunta al termine sono contrastanti: da un lato vi è la gioia e la soddisfazione di esserci riusciti, nonché di poter iniziare ad interfacciarsi con il mondo del lavoro; da un lato la nostalgia e un po' di malinconia nel ricordare tutte le esperienze di vita vissute in questa facoltà e le innumerevoli amicizie instaurate. Amicizie che in tutto questo percorso sono state fondamentali.
Un ringraziamento va quindi a Sara, la mia amica di sempre. Anche se negli ultimi anni ci siamo viste poco, nei momenti di bisogno lei c'è sempre stata.
Un grazie a mio fratello e ai miei cugini, coloro che hanno sempre creduto in me e mi hanno sopportata in tutti questi anni. Sono convinta che il nostro legame sia così forte per via degli insegnamenti dateci dal nonno Bruno, persona a cui devo veramente tutto.
Un ringraziamento va anche a Fabio, colui che mi ha insegnato che l'amicizia tra uomo e donna può esistere veramente. Con i suoi modi tutt'altro che carini e coccolosi mi ha accompagnata in tutti questi anni di università schierandosi in prima linea per superare ogni ostacolo con me. Oltre ad un grazie si merita anche uno scusa per le innumerevoli volte che presa da qualche ansia gli sono piombata in casa all'alba e nonostante tutto lui mi ha sempre accolta.
Un grazie ai miei coinquilini, il vivere a Padova con loro è stata una delle esperienze più belle della mia vita.
Un grazie anche a tutte le persone che ho avuto il piacere di conoscere in questi anni; in particolare un sentito grazie ai Soliti, mai avrei pensato di poter creare un gruppo così fantastico dentro le mura della facoltà. Tra loro un ringraziamento speciale va ad Anna e Lucia che si sono dovute subire

tutte le mie ansie e paure, e a Ciro, la persona più rappresentativa di questo percorso.
Infine un ringraziamento va alla mia relatrice, la professoressa Laura Ventura, che si è rivelata essere una persona super professionale e disponibile; un grazie anche a tutti i miei correttori di bozza e ai miei correlatori segreti Cristian e Dado, correlatore anche di vita.