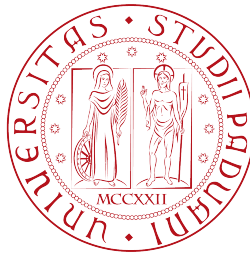


Università degli studi di Padova  
Dipartimento di Scienze Statistiche  
Corso di laurea magistrale in  
Scienze Statistiche



TESI DI LAUREA

**Inferenza bayesiana sul volume sotteso  
alla superficie ROC**

Relatore Chi.ma Prof.ssa Laura Ventura

Dipartimento di Scienze Statistiche

Correlatore Dott. Erlis Ruli

Dipartimento di Scienze Statistiche

Laureando: Antonio Zigrino

Matricola N 1039954

Anno Accademico 2014/2015



# Indice

<b>INTRODUZIONE</b>	<b>i</b>
<b>1 La curva ROC</b>	<b>1</b>
1.1 Matrice di confusione e indici sintetici . . . . .	2
1.2 La curva ROC . . . . .	5
1.2.1 AUC - Area Under the Roc Curve . . . . .	7
1.3 Inferenza su AUC . . . . .	8
1.3.1 Approccio frequentista . . . . .	8
1.3.2 Esempio: Modello Normale . . . . .	12
1.3.3 Approccio bayesiano . . . . .	15
1.3.4 Esempio Normale . . . . .	18
1.4 Indice $J$ di Youden . . . . .	19
1.4.1 Esempio Normale . . . . .	21
1.5 Conclusioni . . . . .	21
<b>2 Volume sotteso alla curva ROC</b>	<b>23</b>
2.1 Superficie ROC . . . . .	24
2.1.1 Esempio normale . . . . .	27
2.1.2 Inferenza su VUS . . . . .	27
2.2 Generalizzazione dell'Indice $J$ di Youden . . . . .	30
2.2.1 Esempio Normale . . . . .	31
2.2.2 Esempio di applicazione di tipo non parametrico . . . . .	33
2.3 Conclusioni . . . . .	34
<b>3 Inferenza parametrica bayesiana sul VUS</b>	<b>37</b>
3.1 Il modello parametrico . . . . .	37
3.2 Calcolo di distribuzioni marginali a posteriori . . . . .	39
3.2.1 Metodi MCMC . . . . .	40
3.3 MCMC applicato al VUS . . . . .	42
3.3.1 Studi di simulazione . . . . .	42

3.4	Applicazione del metodo MCMC a dati reali . . . . .	56
3.5	Conclusioni . . . . .	59
	<b>CONCLUSIONI</b>	<b>61</b>
	<b>APPENDICE</b>	<b>63</b>
	<b>Bibliografia</b>	<b>67</b>

# INTRODUZIONE

La curva ROC (*Receiver Operating Characteristic curve*) nasce a metà del secolo scorso durante la Seconda Guerra Mondiale. Essa è uno strumento che, dalla sua nascita fino ad oggi, viene utilizzato in diversi ambiti scientifici in un contesto di analisi di classificazione. Nel corso degli anni tale strumento è stato utilizzato soprattutto in campi come la biologia e la medicina, trovando anche un'importante sviluppo applicativo in contesti come il *Machine Learning* e *Data Mining*.

Nell'ambito della medicina l'analisi della curva ROC è utilizzata nelle analisi di classificazione, e in particolare per la valutazione di un test diagnostico. Esso è una procedura che permette di classificare i pazienti sulla base dell'informazione raccolta, al fine di valutare la presenza o meno di una particolare patologia.

In letteratura sono presenti differenti approcci sull'inferenza della curva ROC (si veda, ad esempio, Kotzet *al.*, 2003). Gli approcci frequentisti parametrici e non parametrici per l'inferenza sulla curva ROC hanno dato un forte contributo allo sviluppo di questo strumento, mentre i primi sviluppi bayesiani sono arrivati alla fine degli anni '90.

In generale, le procedure inferenziali connesse all'analisi della curva ROC riguardano specifiche misure e indici sintetici, dedotti da una matrice a doppia entrata, detta *Matrice di confusione*, che rappresenta la sintesi della classificazione. In ambito medico, dato un test diagnostico quantitativo, si vuole discriminare nel modo migliore possibile una popolazione sana da una popolazione malata, in funzione di un determinato valore soglia (o *cut-off*). In ambito statistico le due popolazioni possono essere rappresentate da opportune variabili casuali che, in funzione della separazione dei due gruppi di pazienti, saranno più o meno sovrapposte tra loro. Quindi, in caso di perfetta separazione dei gruppi, le due variabili risultano assolutamente distinte tra loro ma, tale situazione è alquanto impossibile nella realtà.

In relazione ad un prefissato valore soglia, nella matrice di confusione si presentano quattro possibili casi: TP (veri positivi), FP (falsi positivi), TN (veri negativi) e FN (falsi negativi). In questo particolare contesto, quello che si vuole ottenere è il miglior *trade-off* delle due misure di errata classificazione (FP e FN). Dalla matrice

di confusione sono deducibili due importanti indici, la *Sensibilità* ( $S_e$ ) e *Specificità* ( $S_p$ ), che sono inversamente correlate tra loro.

La curva ROC è costruita in funzione della sensibilità e specificità. In particolare essa considera tutti i possibili valori del test e per ognuno di essi calcola il valore di  $S_e$  e  $(1 - S_p)$  al variare della soglia  $c$ . Un importante indice deducibile dalla curva ROC è l'AUC (*Area under the ROC curve*), definita anche come il *modello sollecitazione-resistenza*, ossia come  $P(X < Y)$ . L'inferenza frequentista sull'AUC in ambito parametrico si basa sulla verosimiglianza ed in particolare sfrutta la proprietà di equivarianza degli stimatori di massima verosimiglianza. La stima dell'AUC si ottiene utilizzando il test alla *Wald* o la statistica *test radice con segno* del log-rapporto di verosimiglianza.

Per l'inferenza sull'AUC si può considerare anche un approccio bayesiano. In questo contesto, il parametro di interesse è una variabile casuale con una certa distribuzione *a priori*. Una stima bayesiana dell'AUC è data, ad esempio, dalla media a posteriori, ottenuta come valore atteso della distribuzione a posteriori marginale del parametro. In ambito bayesiano, un problema di stima intervallare per il parametro di interesse consiste nella costruzione di un intervallo di credibilità. Aspetto cruciale della metodologia bayesiana è la scelta della distribuzione a priori. In letteratura sono presenti diverse alternative come, ad esempio, le distribuzioni a priori coniugate, la distribuzione a priori non informativa di Jeffreys, le *reference priors*, le *conditional reference priors* e le *matching priors* (si vedano, ad esempio, Kotz *et al.*, 2003; Datta e Mukerjee, 2004; Liseo, 2008).

In ambito medico-sanitario potrebbe essere di interesse una classificazione su tre gruppi di pazienti. Tale contesto è affrontato nell'analisi della superficie ROC, metodologia introdotta da Scurfield (1996) e poi sviluppata negli anni da diversi autori (ad esempio, Mossman, 1999; Heckerling, 2001). Lo scopo dell'inferenza sulla superficie ROC è ottenere una stima del VUS (*Volume Under the ROC surface*). Tale quantità rappresenta l'equivalente dell'AUC nella curva ROC, e può essere rappresentato dalla probabilità  $P(X < Y < Z)$ , con  $X$ ,  $Y$  e  $Z$  variabili casuali indipendenti. Allo stato attuale della letteratura i metodi sviluppati per l'inferenza sul VUS riguardano approcci frequentisti parametrici e non parametrici (si veda, ad esempio, Yu, 2012; Ratnasamy *et al.*, 2008; Yiannoutsos *et al.*, 2008; Nakas, 2014). Sono invece assenti degli sviluppi bayesiani parametrici per l'inferenza sul VUS.

Obiettivo di questa tesi è sviluppare uno strumento bayesiano parametrico al fine di ottenere un'approssimazione della distribuzione marginale a posteriori per il VUS. A tale scopo si sono considerati i metodi Markov Chain Monte Carlo. In particolare, è stato utilizzato l'algoritmo Metropolis-Hastings con passeggiata casuale in una

sua versione adattiva (Scheidegger, 2012), che consiste nel fissare a priori il tasso di accettazione desiderato. Inoltre, per la distribuzione iniziale è stata considerata la distribuzione a priori di Jeffreys.

Per valutare il metodo proposto sono stati considerati sia degli studi di simulazione sia un'applicazione su un dataset reale su uno studio che ha riguardato la classificazione di tre diversi gradi di meningioma (Cassoni *et al.*, 2015).





# Capitolo 1

## La curva ROC

Una delle ragioni principali alla base delle misurazioni cliniche è fornire uno strumento di supporto alle diagnosi. Questo si traduce nell'individuazione di una diagnosi tra le tante possibili o nell'individuazione di persone con particolari malattie all'interno di una popolazione apparentemente sana (detto anche *screening*). In entrambi i casi, l'utilizzo di un test diagnostico permette di classificare i soggetti in due gruppi (Bland, 2000). Un'analisi di classificazione ha l'obiettivo di individuare per ogni unità statistica la sua opportuna classe (o gruppo), in funzione delle proprie caratteristiche. Ad esempio, in ambito medico si vuole classificare i risultati osservati di un test diagnostico sulla base di un *gold standard*<sup>1</sup> o dello stato del paziente..

In questo capitolo verranno definite alcune metodologie per analisi di classificazione medico-sanitarie. In particolare, verranno prima illustrati gli indici più importanti dedotti dalla matrice di confusione, i quali riferiscono sulla qualità della classificazione, quindi sull'efficacia del test diagnostico (§1.1). Successivamente verrà presentata la curva ROC (*Receiver Operating Characteristic curve*) e le procedure inferenziali connesse alla misura dell'AUC (*Area Under the ROC Curve*), nota anche come modello sollecitazione-resistenza (Kotz *et al.*, 2003). Le procedure inferenziali riguardano due approcci inferenziali, ossia l'approccio frequentista e l'approccio Bayesiano (§1.3). Infine, verrà illustrato l'indice  $J$  di Youden, uno strumento utile per la valutazione della classificazione, quindi dell'efficacia diagnostica generale (§1.4).

In letteratura sono presenti diversi approcci per l'inferenza sulla curva ROC (si veda Kotz *et al.*, 2003). In ambito frequentista, Faraggi e Reiser (2002) hanno esaminato due approcci parametrici e due approcci non-parametrici per l'inferenza sull'AUC. Recentemente, Cortese e Ventura (2012) hanno affrontato il problema dell'inferenza sul modello sollecitazione-resistenza secondo un approccio basato sui metodi asin-

---

<sup>1</sup>In medicina un *gold standard* è l'esame diagnostico più accurato, sulla base del quale ogni nuovo test deve rapportarsi per avere validità diagnostica. Esso rappresenta la diagnosi corretta per una malattia e potrebbe consistere ad esempio in un referto istologico.

totici di ordine superiore. Altri studi si sono focalizzati su un approccio semi-parametrico e non-parametrico per l'inferenza sul modello sollecitazione-resistenza (si veda, Adimari e Chiogna, 2006).

L'abilità di un test diagnostico nel discriminare tra una popolazione sana e una malata può essere influenzata da diversi fattori. Quindi, valutare l'influenza di covariate, riguardo differenti popolazioni e condizioni, potrebbe portare informazioni utili ai fini della valutazione dell'accuratezza di un test diagnostico (de Carvalho *et al.*, 2013). Al contrario, ignorare l'effetto delle covariate potrebbe portare ad una distorsione dell'analisi inferenziale sulle performance del test diagnostico. L'effetto delle covariate sull'analisi della curva ROC è stato affrontato in diversi studi (si veda ad esempio Peng e Hall, 1996; Branscum *et al.*, 2008; de Carvalho, 2013). In un contesto bayesiano, uno dei primi contributi all'analisi della curva ROC è stato fornito da Peng e Hall (1996) e Hellmich *et al.* (1998). Sulla base di questi studi, altri autori hanno affrontato l'inferenza bayesiana per la stima della curva ROC considerando approcci semi-parametrici e non-parametrici (Erkanli *et al.*, 2007; Gu *et al.*, 2008; Inácio, 2012; Branscum *et al.*, 2013). In particolare, Hanson *et al.* (2008) e de Carvalho *et al.* (2013) hanno fornito due metodi non parametrici per l'inferenza sulla curva ROC. Un contributo recente all'inferenza parametrica in ambito bayesiano è stato fornito da Ventura e Racugno (2011).

L'Indice  $J$  di Youden e i valori di *cut-off*, per un'analisi di classificazione binaria, sono stati ampiamente discussi in letteratura, fornendo anche degli studi di simulazione e delle applicazioni in ambito medico-sanitario (si veda Hilden e Glasziou, 1996; Fluss *et al.*, 2005; Perkins e Schisterman, 2005, 2006; Schisterman e Perkins, 2007).

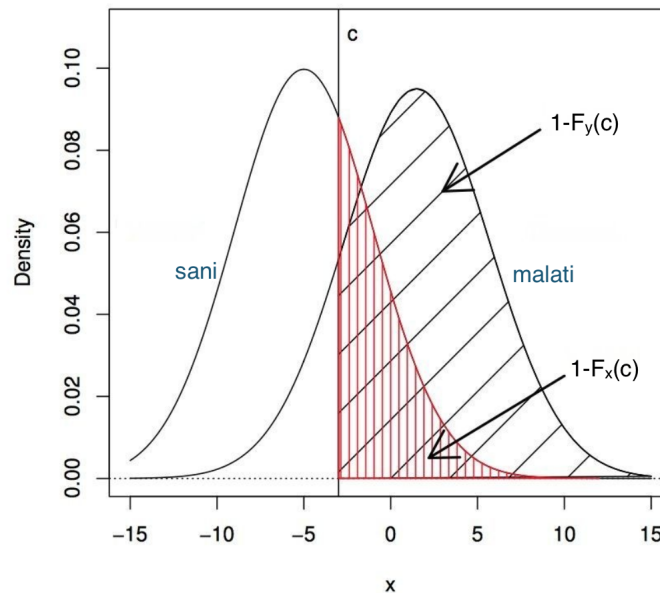
## 1.1 Matrice di confusione e indici sintetici

Dato un test diagnostico misurato su due diverse popolazioni, rappresentanti pazienti sani e malati, la migliore classificazione possibile dei pazienti si traduce in una perfetta separazione tra i due gruppi, cioè una perfetta separazione delle due distribuzioni che descrivono le due popolazioni. Questa situazione è quanto mai utopica: infatti, solitamente le distribuzioni tendono a sovrapporsi.

Dalla Figura 1.1 si nota che la soglia  $c$  (*criterion value* o *cut-off*) individua la classificazione tra i pazienti sani e malati. Fissato un valore della soglia  $c$  si possono individuare quattro casi:

- **TP (True Positive)**: pazienti malati correttamente classificati (veri positivi);
- **TN (True Negative)**: pazienti sani correttamente classificati (veri negativi);

- **FP (False Positive)**: pazienti sani classificati come malati (falsi positivi);
- **FN (False Negative)**: pazienti malati classificati come sani (falsi negativi).



**Figura 1.1:** Distribuzioni del test diagnostico nei pazienti sani e malati.

Obiettivo è ottenere la migliore classificazione possibile. Questo corrisponde a individuare il valore soglia  $c$  per il quale si ottiene il miglior *trade-off* tra i due errori di errata classificazione (FN e FP). Si può dimostrare che, quando le distribuzioni dei valori del test sono normali, allora il valore soglia  $c$  è individuato dall'ascissa del punto di intersezione delle due distribuzioni, ottenendo quindi una minimizzazione degli errori di classificazione (Bottarelli e Parodi, 2003).

Per una data soglia  $c$ , i valori del test possono essere rappresentati in una tabella a doppia entrata, detta **Matrice di confusione** o **Tabella di errata classificazione** (Tabella 1.1).

	Malato	Sano	Totale
Test Positivo	TP (Veri positivi)	FP (Falsi Positivi)	TP+FP
Test Negativo	FN (Falsi Negativi)	TN (Veri Negativi)	FN+TN
Totale	TP+FN	FP+TN	

**Tabella 1.1:** Matrice di confusione.

Ovviamente, minori sono i valori di FP e FN, migliore sarà l'accuratezza del test diagnostico.

Per poter valutare il potere diagnostico di un test possiamo servirci di alcuni indici sintetici deducibili dalla matrice di confusione. Essi sono:

- **Sensibilità:** è la proporzione di individui malati che risultano positivi al test. Essa è data dal rapporto tra veri positivi e il totale di individui realmente malati, ossia

$$S_e = \frac{TP}{TP + FN}.$$

- **Specificità:** è la proporzione di individui sani che risultano negativi al test. Essa è data dal rapporto tra veri negativi e il totale di individui realmente sani, ossia

$$S_p = \frac{TN}{TN + FP}.$$

- **Accuratezza:** è la proporzione di individui classificati correttamente dal test ed è quindi una misura generale della bontà della classificazione, data da

$$A_c = \frac{TP + TN}{TP + FN + TN + FP}.$$

Per alti valori di FN e FP si avrà un valore basso di  $A_c$ , quindi si può cercare di individuare quel valore del *cut-off* per il quale si ottiene una misura dell'accuratezza più alta.

- **Odds Ratio Diagnostico:** è il rapporto tra l'odds dei soggetti malati e l'odds dei soggetti sani. Può essere espresso in funzione della sensibilità e della specificità, come

$$DOR = \frac{TP/FN}{FP/TN} = \frac{S_e/(1 - S_e)}{(1 - S_p)/S_p}.$$

DOR rappresenta una misura generale per l'accuratezza diagnostica e può essere usato per confrontare due o più test. DOR assume valori in  $(0, \infty)$ , dove alti valori esprimono una buona discriminazione tra pazienti sani e malati. Quando esso assume valore pari a uno, allora il test non riesce a discriminare i soggetti.

- **Potere predittivo positivo:** è la probabilità che un soggetto, che risulti positivo al test, risulti essere un vero positivo (cioè che la classificazione del soggetto come "malato" sia corretta). Essa è dato da

$$PPP = \frac{TP}{TP + FP}.$$

- **Potere predittivo negativo:** è la probabilità che un soggetto, che risulti negativo al test, risulti essere un vero negativo (cioè che la classificazione del soggetto come "sano" sia corretta). Essa è dato da

$$PPN = \frac{TN}{FN + TN}.$$

In sintesi, se il fine è quello di individuare il maggior numero di malati, il test migliore è quello con sensibilità maggiore, in quanto riesce meglio nella previsione dell'esito negativo. Altrimenti, se il fine è quello di individuare il maggior numero di pazienti sani, il test migliore è quello con specificità maggiore in quanto meglio riesce nella previsione dell'esito positivo del test.

Il valore soglia  $c$  ha dunque grande importanza ai fini dell'analisi e dei risultati ottenuti. Sensibilità e specificità sono tra loro inversamente correlate, quindi la soglia scelta per l'analisi porterà ad un aumento dell'una con conseguente diminuzione dell'altra. Risulta opportuno precisare che, la scelta del *cut-off* non può prescindere solo da teorie probabilistiche, ma che è opportuno anche considerare quello che è l'obiettivo dell'analisi e il tipo di dati che si stanno analizzando.

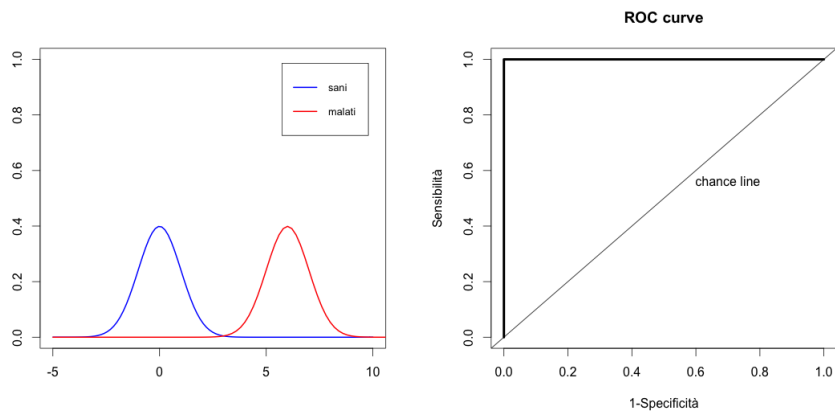
## 1.2 La curva ROC

La curva ROC è un importante strumento grafico per la valutazione di un test diagnostico e si basa sugli indici di sensibilità e specificità al variare della soglia  $c$ . Come mostrato nelle Figure 1.2, 1.3 e 1.4, la curva ROC è un grafico cartesiano in cui sono riportati i punti di coordinate (1-specificità, sensibilità). La costruzione della curva ROC considera tutti i possibili valori del test e per ognuno di essi calcola  $S_e$  e  $(1 - S_p)$  al variare della soglia  $c$ .

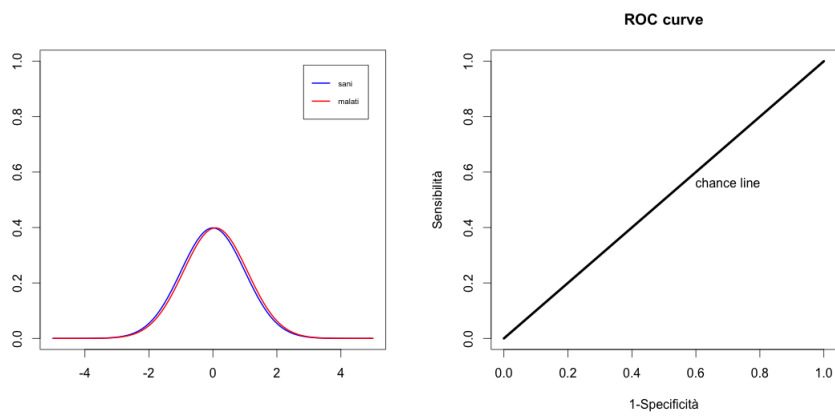
In un sistema di assi cartesiani, la relazione tra i due indici viene riportata mettendo in ascissa la proporzione di falsi positivi e in ordinata la proporzione dei veri positivi.

Per interpretare la curva ROC dobbiamo considerare che la bisettrice del grafico (*chance line*) rappresenta la classificazione casuale dei soggetti. Per ottenere la migliore classificazione possibile dovremmo ottenere una curva che stia quanto più possibile al di sopra di tale retta (si veda ad esempio Azzalini e Scarpa, 2009, Cap 5).

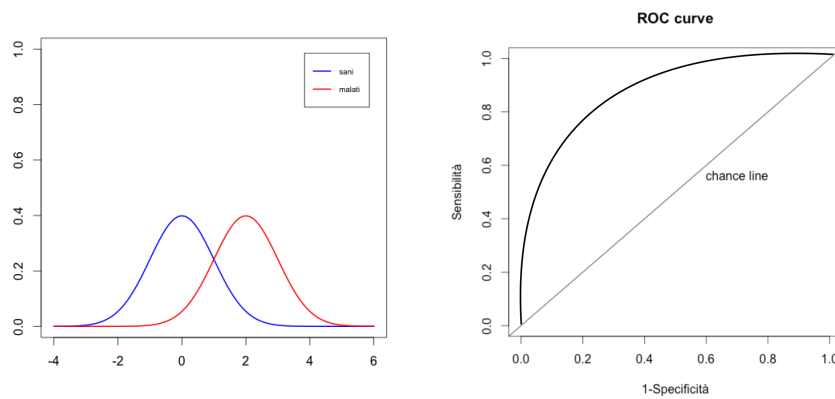
La curva ROC risulta quindi uno strumento grafico di immediata interpretazione anche nel caso di un confronto tra diversi test diagnostici riferiti ad una stessa



**Figura 1.2:** Esempio di discriminazione perfetta tra pazienti sani e malati e curva ROC relativa.



**Figura 1.3:** Esempio di discriminazione casuale dei pazienti e curva ROC relativa.



**Figura 1.4:** Esempio di discriminazione dei pazienti e curva ROC relativa.

patologia. Inoltre, dalla curva ROC è possibile derivare un indice sintetico dell'accuratezza di un test. Esso consiste nell'area sottesa alla curva ed è indicato con **AUC** (*Area Under the ROC Curve*).

### 1.2.1 AUC - Area Under the Roc Curve

Sia  $X$  la variabile che rappresenta la misura del test nei pazienti sani e  $Y$  la variabile che rappresenta la misura del test nei pazienti malati. L'AUC può essere espressa come

$$AUC = P(X < Y). \quad (1.1)$$

Tale formula è nota anche come *modello sollecitazione-resistenza* (o *stress-strength model*) (si vedano ad esempio, Bamber (1975) e Kotz *et al.*, 2003).

Inoltre, l'AUC è strettamente legata al test non-parametrico  $T_{MW}$  di Mann-Whitney (Bamber, 1975), in quanto

$$AUC = 1 - \frac{T_{MV}}{n_x n_y},$$

dove  $n_x$  e  $n_y$  rappresentano le numerosità dei campioni casuali semplici  $(x_1, \dots, x_{n_x})$  e  $(y_1, \dots, y_{n_y})$ , indipendenti, tratti rispettivamente da  $X$  e  $Y$ . L'area sottesa alla curva ROC empirica coincide con la statistica  $T$  di M-W, la quale conduce ad uno stimatore non-parametrico non distorto per l'AUC (Faraggi e Reiser, 2002).

Senza perdita di generalità e per un fissato punto  $c$  di *cut-off*, si può affermare che il risultato del test sarà positivo se risulta maggiore di  $c$ , sarà negativo altrimenti. Siano  $F_x$  e  $F_y$  le funzioni di ripartizione delle variabili casuali  $X$  e  $Y$ , rispettivamente. Data una soglia  $c$ , la sensibilità del test è definita come  $Se(c) = 1 - F_y(c)$  e la specificità come  $Sp(c) = F_x(c)$ . La curva ROC è definita come il grafico di  $Se(c)$  in funzione di  $1 - Sp(c)$ , dato  $-\infty \leq c \leq \infty$ , oppure come

$$ROC(t) = 1 - F_y(F_x^{-1}(1 - t)), \quad (1.2)$$

dove  $t \in [0, 1]$ ,  $F_x^{-1}(1 - t) = \inf\{x \in \mathbb{R} : F_x(x) \geq 1 - t\}$ .

Possiamo quindi definire l'AUC come

$$AUC = \int_0^1 ROC(u) du. \quad (1.3)$$

## 1.3 Inferenza su AUC

Come già affermato in precedenza, il modello sollecitazione resistenza (1.1) corrisponde all'AUC (Bamber, 1975). La stima dell'AUC è un campo di ricerca molto attivo e può essere affrontato secondo differenti approcci.

### 1.3.1 Approccio frequentista

Sotto ipotesi parametriche, in genere si assume che  $X$  e  $Y$  siano due variabili casuali con funzioni di densità, rispettivamente,  $f_x(x; \theta_x)$  e  $f_y(y; \theta_y)$ , con  $\theta_x \in \Theta_x \subseteq \mathbb{R}^{p_x}$  e  $\theta_y \in \Theta_y \subseteq \mathbb{R}^{p_y}$ . Sotto queste ipotesi, l'AUC può essere espressa come una funzione del parametro complessivo  $\theta = (\theta_x, \theta_y)$ , attraverso la relazione

$$AUC(\theta) = R(\theta) = P(X < Y) = \int_{-\infty}^{\infty} F_x(t; \theta_x) dF_y(t; \theta_y), \quad (1.4)$$

dove  $F_x(x; \theta_x)$  e  $F_y(y; \theta_y)$  sono, rispettivamente, le funzioni di ripartizione di  $X$  e  $Y$ . Espressioni teoriche per l'AUC sono disponibili con riferimento a diverse assunzioni distributive sulle variabili  $X$  e  $Y$  (Kotz *et al.*, 2003).

Una stima parametrica dell'AUC può essere ottenuta utilizzando la proprietà di equivarianza degli stimatori di massima verosimiglianza (cfr. Pace e Salvani, 2001, Cap.3). La funzione di verosimiglianza per  $\theta$  è

$$L(\theta) = L(\theta_x, \theta_y) = \prod_{i=1}^{n_x} f_x(x_i; \theta_x) \prod_{j=1}^{n_y} f_y(y_j; \theta_y), \quad (1.5)$$

e la funzione di log-verosimiglianza per  $\theta$  è

$$l(\theta) = \sum_{i=1}^{n_x} \log f_x(x_i; \theta_x) + \sum_{j=1}^{n_y} \log f_y(y_j; \theta_y). \quad (1.6)$$

La stima di massima verosimiglianza (SMV)  $\hat{\theta} = (\hat{\theta}_x, \hat{\theta}_y)$  per  $\theta = (\theta_x, \theta_y)$  può essere ottenuta come soluzione dell'equazione di verosimiglianza

$$l_{\theta}(\theta) = \frac{\partial l(\theta)}{\partial \theta} = 0. \quad (1.7)$$

La varianza asintotica approssimata di  $\hat{\theta}$  è data dall'inversa della matrice dell'informazione osservata

$$j(\theta) = -l_{\theta\theta}(\theta) = \frac{-\partial^2 l(\theta)}{\partial \theta \partial \theta^T}. \quad (1.8)$$



Inoltre, l'informazione attesa di Fisher è

$$i(\theta) = -E_\theta[j(\theta)] = -E_\theta \left[ \frac{-\partial^2 l(\theta)}{\partial \theta \partial \theta^T} \right]. \quad (1.9)$$

Sotto opportune condizioni di regolarità, lo stimatore di massima verosimiglianza  $\hat{\theta}$  ha distribuzione approssimata (Pace e Salvan, 2001, Cap. 6)

$$\hat{\theta} \sim N_p(\theta, j(\hat{\theta})^{-1}),$$

o equivalentemente

$$\hat{\theta} \sim N_p(\theta, i(\hat{\theta})^{-1}),$$

dove  $p$  è la dimensione del parametro  $\theta$ .

Nell'inferenza di verosimiglianza, quando il parametro  $\theta$  è scalare, una statistica test basata sulla verosimiglianza è la **radice con segno** del log-rapporto di verosimiglianza, ossia

$$r(\theta) = \text{sign}(\hat{\theta} - \theta) \sqrt{2(l(\hat{\theta}) - l(\theta))}, \quad (1.10)$$

che, sotto condizioni di regolarità, ha distribuzione asintotica nulla  $N(0, 1)$ . Un'intervallo di confidenza per  $\theta$ , con livello nominale  $(1 - \alpha)$ , è dato da

$$\{\theta \in \Theta : -z_{1-\alpha/2} < r(\theta) < z_{1-\alpha/2}\}, \quad (1.11)$$

dove  $z_{1-\alpha/2}$  è il quantile  $(1 - \alpha/2)$  della normale standard.

Se si dispone di un modello statistico caratterizzato da un parametro di interesse  $\psi$  scalare e di un parametro da disturbo  $\lambda$  di dimensione  $(p - 1)$ , risulta opportuno concentrare l'inferenza solo sulla componente di interesse. Risulta a tale fine utile disporre di un'opportuna *pseudo-verosimiglianza*, ossia di una funzione dipendente, oltre che dai dati, dal solo parametro di interesse.

Un metodo largamente utilizzato per ottenere una pseudo-verosimiglianza per il parametro di interesse  $\psi$  è la **verosimiglianza profilo**. Partendo dalla verosimiglianza completa, il parametro di disturbo  $\lambda$  viene sostituito con la sua SMV  $\hat{\lambda}_\psi$  vincolata ad un fissato valore del parametro di interesse  $\psi$ . La funzione di verosimiglianza profilo così definita assume la forma (Severini, 2000, Cap. 4)

$$L_p(\psi) = L(\psi, \hat{\lambda}_\psi), \quad (1.12)$$

e la corrispondente funzione di log-verosimiglianza profilo è

$$l_p(\psi) = l(\psi, \hat{\lambda}_\psi). \quad (1.13)$$

Nel contesto della curva ROC, il parametro di interesse è l'AUC= $P(X < Y)$ . Risulta necessaria una riparametrizzazione  $\theta = (\theta_x, \theta_y) = (\psi, \lambda)$ , dove il parametro scalare di interesse risulta  $\psi = AUC = AUC(\theta_x, \theta_y)$ , mentre  $\lambda = \lambda(\theta_x, \theta_y)$  è il parametro di disturbo. Per la proprietà di equivarianza si ha quindi che la SMV per  $\psi$  è data da  $\hat{\psi} = AUC(\hat{\theta})$ .

Le procedure inferenziali solitamente si basano sulla statistica di Wald profilo, data da

$$w_p = w_p(\psi) = \sqrt{j_p(\hat{\psi})}(\hat{\psi} - \psi), \quad (1.14)$$

o sulla statistica segno del rapporto di log-verosimiglianza profilo

$$r_p = r_p(\psi) = \text{sign}(\hat{\psi} - \psi) \sqrt{2(l_p(\hat{\psi}) - l_p(\psi))}, \quad (1.15)$$

le quali, sotto opportune condizioni di regolarità, hanno distribuzione asintotica nulla normale standard al primo ordine. L'informazione profilo osservata  $j_p(\psi)$  può essere espressa attraverso l'informazione osservata  $j(\theta)$ , come

$$j_p(\psi) = j_{\psi\psi}(\hat{\theta}_\psi) - j_{\psi\lambda}(\hat{\theta}_\psi) j_{\lambda\lambda}^{-1}(\hat{\theta}_\psi) j_{\lambda\psi}(\hat{\theta}_\psi), \quad (1.16)$$

dove  $j_{\psi\psi}(\psi, \lambda)$ ,  $j_{\psi\lambda}(\psi, \lambda)$ ,  $j_{\lambda\psi}(\psi, \lambda)$  e  $j_{\lambda\lambda}(\psi, \lambda)$  sono le sottomatrici dell'informazione osservata  $j(\theta)$ .

Per la costruzione degli intervalli di confidenza per  $\psi$  ci si può riferire alle statistiche  $w_p(\psi)$  e  $r_p(\psi)$ . Un intervallo di confidenza approssimato di livello  $(1 - \alpha)$  basato sulla statistica di Wald profilo assume la forma

$$\left( \hat{\psi} - z_{1-\alpha/2} j_p(\hat{\psi})^{-1/2}, \hat{\psi} + z_{1-\alpha/2} j_p(\hat{\psi})^{-1/2} \right). \quad (1.17)$$

Invece, un intervallo di confidenza approssimato di livello  $(1 - \alpha)$  basato su  $r_p(\psi)$  assume la forma

$$\{\psi : |r_p(\psi)| \leq z_{1-\alpha/2}\}. \quad (1.18)$$

Solitamente l'intervallo di confidenza basato sulla statistica di Wald viene preferito per la sua facilità di calcolo. Tuttavia, è noto che tale statistica non gode della proprietà di invarianza alla riparametrizzazione ed è meno accurata della statistica log-rapporto di verosimiglianza (Severini, 2000; Brazzale *et al.*, 2007).

La verosimiglianza profilo non costituisce una verosimiglianza in senso proprio, ciò nonostante è considerata uno strumento inferenziale interessante in quanto gode di alcune proprietà che la associano ad una verosimiglianza propria. Tuttavia, se i dati non contengono informazioni sufficienti sul parametro di disturbo, l'inferenza su  $\psi$  condotta sulla base della verosimiglianza profilo, può risultare inaccurata. Inoltre, per piccoli campioni l'approssimazione al primo ordine è spesso inaccurata e restituisce risultati insoddisfacenti, specialmente quando la dimensione del parametro di disturbo è elevata. Per il modello sollecitazione-resistenza, i risultati sono insoddisfacenti in termini di accuratezza quando l'AUC è prossima ad uno (Newcombe, 1998).

Per ovviare a questo problema, sono state proposte varie versioni modificate della verosimiglianza profilo, le quali compensano la carenza di informazione sul parametro di disturbo  $\lambda$ . Un esempio è la **verosimiglianza profilo modificata** introdotta da Barndorff-Nielsen (1980, 1983). La verosimiglianza profilo modificata è definita come

$$L_M(\psi) = L_P(\psi)M(\psi), \quad (1.19)$$

dove  $M(\psi)$  è un opportuno fattore di modificazione (Severini, 2000, Cap.8, Pace e Salvan, 2006). La log-verosimiglianza profilo modificata è

$$l_M(\psi) = \log L_M(\psi) = l_P(\psi) + \log M(\psi). \quad (1.20)$$

Una proprietà importante della verosimiglianza profilo modificata è l'invarianza rispetto a riparametrazioni. Un altro vantaggio è rappresentato dal fatto che  $L_P(\psi)$  e  $L_M(\psi)$  sono asintoticamente equivalenti al primo ordine, in quanto il fattore di aggiustamento  $M(\psi)$  è di ordine  $O(1)$ . Tuttavia, la distorsione della funzione score profilo modificata si riduce da  $O(1)$  a  $O(n^{-1})$ .

E' inoltre disponibile una versione più accurata della statistica  $r_p(\psi)$ , la cui distribuzione nulla è normale standard con ordine  $O(n^{-3/2})$ . Si assume che esista una statistica ancillare  $a$ , esatta o approssimata al secondo ordine, tale che  $l(\theta) = l(\theta; y) = l(\theta; \hat{\theta}, a)$ . La **radice con segno** modificata per  $\psi$  è data da

$$r_p^* = r_p^*(\psi) = r_p(\psi) + \frac{1}{r_p(\psi)} \log \frac{q(\psi)}{r_p(\psi)}, \quad (1.21)$$

dove  $q(\psi)$  è un'appropriata quantità di verosimiglianza, per la quale sono valide diverse espressioni.

La radice con segno modificata  $r_p^*(\psi)$  è una quantità pivotale avente distribuzione nulla normale standard all'ordine  $O(n^{-3/2})$ . Inoltre risulta soddisfatta la proprietà

di equivarianza alla parametrizzazione.

Un intervallo di confidenza per  $\psi$  con un livello approssimato  $(1 - \alpha)$  basato su  $r_p^*(\psi)$  è dato da  $(\psi_1^*, \psi_2^*)$ , soluzioni in  $\psi$  delle equazioni

$$r_p^*(\psi) = z_{1-\alpha/2} \quad r_p^*(\psi) = z_{\alpha/2}.$$

Equivalentemente, un intervallo di confidenza di livello approssimato  $(1 - \alpha)$  per  $\psi$  basato su  $r_p^*(\psi)$  ha la seguente forma

$$\{\psi : |r_p^*(\psi)| \leq z_{1-\alpha/2}\}. \quad (1.22)$$

La radice con segno modificata  $r_p^*(\psi)$  può essere utilizzata anche per definire un'equazione di stima per  $\psi$ , della forma

$$r_p^*(\psi) = 0, \quad (1.23)$$

che richiede una procedura numerica per la risoluzione. Lo stimatore  $\hat{\psi}^*$ , soluzione dell'equazione (1.23), è un miglioramento dello stimatore  $\hat{\psi}$ . In funzione delle proprietà di  $r_p^*(\psi)$ , lo stimatore  $\hat{\psi}^*$  è non distorto in media, così come in mediana al terzo ordine. Inoltre, dato che  $r_p^*(\psi)$  è invariante rispetto a riparametrazioni del parametro di interesse, allora  $\hat{\psi}^*$  è uno stimatore equivariante di  $\psi$ . Numerosi studi di simulazione dimostrano che gli stimatori basati su  $r_p^*(\psi)$  migliorano lo SMV (Pace e Salvan, 1999; Giummolé e Ventura, 2002; Cortese e Ventura, 2012).

### 1.3.2 Esempio: Modello Normale

Siano  $X$  e  $Y$  due variabili indipendenti aventi distribuzione normale, ossia  $X \sim N(\mu_x, \sigma_x^2)$  e  $Y \sim N(\mu_y, \sigma_y^2)$ , con  $\mu_x$ ,  $\mu_y$ ,  $\sigma_x^2$  e  $\sigma_y^2$  ignoti. Il parametro  $\theta$  è dato da  $\theta = (\mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$  e l'AUC può essere espressa come (Kotz *et al.*, 2003)

$$\psi = \psi(\theta) = \Phi \left( \frac{\mu_y - \mu_x}{\sqrt{\sigma_x^2 + \sigma_y^2}} \right). \quad (1.24)$$

Per la proprietà di equivarianza, la SMV per  $\psi$  è

$$\hat{\psi} = \psi(\hat{\theta}) = \Phi \left( \frac{\hat{\mu}_y - \hat{\mu}_x}{\sqrt{\hat{\sigma}_x^2 + \hat{\sigma}_y^2}} \right), \quad (1.25)$$

dove  $\hat{\mu}_x = (1/n_x) \sum x_i$ ,  $\hat{\mu}_y = (1/n_y) \sum y_i$ ,  $\hat{\sigma}_x^2 = (1/n_x) \sum (x_i - \hat{\mu}_x)^2$  e  $\hat{\sigma}_y^2 = (1/n_y) \sum (y_i - \hat{\mu}_y)^2$ .

In Mukherjee and Sharan (1985) è disponibile la varianza asintotica dello stimatore  $\hat{\psi}$ . Essa è data da

$$\text{Var}(\hat{\psi}) = \frac{1}{2\pi\hat{\sigma}^2} \exp \left\{ \frac{-(\hat{\mu}_x - \hat{\mu}_y)^2}{\hat{\sigma}^2} \right\} \left[ \frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y} + \frac{1}{2} \frac{(\hat{\mu}_x - \hat{\mu}_y)^2}{\hat{\sigma}^4} \left( \frac{\hat{\sigma}_x^4}{n_x} + \frac{\hat{\sigma}_y^4}{n_y} \right) \right],$$

dove  $\sigma^2 = \sigma_x^2 + \sigma_y^2$ .

Sia data la riparametrizzazione  $\theta = (\psi, \lambda)$ , con  $\psi$  parametro di interesse e  $\lambda = (\lambda_1, \lambda_2, \lambda_3) = (\mu_x, \sigma_x^2, \sigma_y^2)$  parametro di disturbo. La log-verosimiglianza può essere espressa come

$$\begin{aligned} l(\psi, \lambda_1, \lambda_2, \lambda_3) = & -\frac{1}{2}(n_x \log(\lambda_2) + n_y \log(\lambda_3)) - \frac{n_x}{2\lambda_2} \left[ \hat{\lambda}_2 + (\lambda_1 + \hat{\lambda}_1)^2 \right] \\ & - \frac{n_y}{2\lambda_3} \left[ (D_1(\hat{\theta}) - D_1(\theta))^2 + \hat{\lambda}_3 \right], \end{aligned} \quad (1.26)$$

dove  $D_1(\theta) = \sqrt{\lambda_2 + \lambda_3} \Phi^{-1}(\psi) + \lambda_1$ . Da questo risultato, in accordo con le formule (1.14), (1.15) e (1.16), si ottengono, rispettivamente, la statistica di Wald profilo  $w_p(\psi)$ , la radice con segno profilo  $r_p(\psi)$  e l'informazione profilo osservata  $j_p(\psi)$  per l'esempio normale. Quindi, risulta immediata la costruzione di un intervallo di confidenza approssimato di livello  $(1 - \alpha)$  basato sulla statistica di Wald profilo, come rappresentato nella (1.17), oppure un intervallo di confidenza basato sulla statistica radice con segno profilo, come mostrato nella (1.18). Inoltre, dalla (1.21) è possibile ricavare la versione modificata della statistica radice con segno profilo, e quindi il relativo intervallo di confidenza come nella formula (1.22).

In Cortese e Ventura (2012) sono stati considerati degli studi di simulazione per verificare la performance di  $r_p^*(\psi)$  per la costruzione di intervalli di confidenza per l'AUC e per la sua stima .

L'accuratezza di  $r_p^*(\psi)$  sotto ipotesi di normalità è stata illustrata attraverso una simulazione Monte Carlo basata su 5000 prove. E' risultato che, quando la numerosità campionaria è piccola,  $r_p^*(\psi)$  è più accurato di  $r_p(\psi)$  sia in termini di copertura degli intervalli di confidenza sia in termini di misure di dispersione (distorsione in media e in mediana, errore quadratico medio, deviazione totale dalla mediana). Le coperture empiriche per  $r_p^*(\psi)$  sono più vicine ai livelli nominali rispetto a quelle per  $r_p(\psi)$ , specialmente per  $n_x = n_y = 5, 10$ ; le differenze tendono ad essere meno evidenti con l'aumentare della numerosità campionaria. I risultati delle simulazioni sono mostrati in Tabella 1.2.

Per quanto riguarda la simmetria di copertura si vede che per  $r_p(\psi)$  è asimmetrica e lo è ancora di più per  $w_p(\psi)$ ; tali risultati sono ancora più evidenti in presenza di piccoli campioni e alti valori di  $\psi$ . Per quanto riguarda  $r_p^*(\psi)$  invece si ottengono degli scostamenti abbastanza simmetrici per tutte le dimensioni del campione considerate e per qualsiasi valore di  $\psi$ .

Gli studi di simulazione sono stati effettuati anche per confrontare i due stimatori di  $\psi$ , quello basato su  $r_p^*(\psi)$  ( $\hat{\psi}^*$ ) e la SMV  $\hat{\psi}$ . I risultati sono riportati in Tabella 1.3. I due stimatori sono stati confrontati in termini di distorsione in media e in mediana. Specialmente in presenza di piccoli campioni e alti valori di  $\psi$ ,  $\hat{\psi}^*$  risulta preferibile a  $\hat{\psi}$ , soprattutto per la distorsione in mediana. Per quanto riguarda la distorsione in media, i due stimatori risultano avere differenze trascurabili nella maggior parte dei casi.

**Tabella 1.2:** Risultati delle simulazione per gli intervalli di confidenza per  $\psi$ , sotto assunzioni di normalità e nel caso di omoschedasticità. Fonte: Cortese e Ventura (2012).

$(n_x, n_y)$	Method	$\psi = 0.5$ ( $\mu_x = \mu_y = 5$ )		$\psi = 0.8$ ( $\mu_x = 5, \mu_y = 6.33$ )		$\psi = 0.95$ ( $\mu_x = 5, \mu_y = 7.60$ )	
		90%	95%	90%	95%	90%	95%
(5, 5)	$w_p(\psi)$	0.760 (0.122, 0.118)	0.776 (0.010, 0.214)	0.722 (0.034, 0.244)	0.767 (0.016, 0.217)	0.638 (0.003, 0.359)	0.656 (0.002, 0.343)
	$r_p(\psi)$	0.832 (0.083, 0.085)	0.915 (0.017, 0.068)	0.848 (0.047, 0.105)	0.915 (0.021, 0.064)	0.838 (0.025, 0.137)	0.908 (0.011, 0.081)
	$r_p^*(\psi)$	0.891 (0.050, 0.059)	0.962 (0.017, 0.021)	0.900 (0.051, 0.049)	0.958 (0.022, 0.021)	0.904 (0.048, 0.048)	0.957 (0.022, 0.021)
(10, 10)	$w_p(\psi)$	0.850 (0.076, 0.074)	0.906 (0.047, 0.047)	0.812 (0.027, 0.161)	0.859 (0.012, 0.129)	0.733 (0.004, 0.263)	0.776 (0.001, 0.223)
	$r_p(\psi)$	0.885 (0.058, 0.057)	0.940 (0.029, 0.031)	0.878 (0.045, 0.077)	0.936 (0.022, 0.042)	0.875 (0.034, 0.091)	0.929 (0.015, 0.056)
	$r_p^*(\psi)$	0.909 (0.045, 0.045)	0.954 (0.022, 0.024)	0.903 (0.053, 0.044)	0.952 (0.025, 0.023)	0.897 (0.054, 0.049)	0.945 (0.025, 0.030)
(20, 20)	$w_p(\psi)$	0.874 (0.066, 0.060)	0.930 (0.036, 0.034)	0.860 (0.029, 0.112)	0.907 (0.011, 0.082)	0.815 (0.006, 0.180)	0.845 (0.002, 0.154)
	$r_p(\psi)$	0.893 (0.055, 0.052)	0.946 (0.029, 0.025)	0.892 (0.043, 0.065)	0.943 (0.023, 0.034)	0.895 (0.034, 0.071)	0.940 (0.016, 0.044)
	$r_p^*(\psi)$	0.903 (0.049, 0.048)	0.952 (0.025, 0.023)	0.903 (0.051, 0.046)	0.950 (0.028, 0.023)	0.907 (0.048, 0.046)	0.947 (0.024, 0.029)
(30, 30)	$w_p(\psi)$	0.884 (0.059, 0.057)	0.937 (0.031, 0.031)	0.868 (0.031, 0.101)	0.918 (0.010, 0.072)	0.839 (0.008, 0.152)	0.875 (0.001, 0.124)
	$r_p(\psi)$	0.895 (0.053, 0.052)	0.946 (0.028, 0.026)	0.889 (0.043, 0.068)	0.947 (0.020, 0.033)	0.895 (0.035, 0.070)	0.943 (0.019, 0.038)
	$r_p^*(\psi)$	0.902 (0.048, 0.049)	0.949 (0.026, 0.024)	0.899 (0.050, 0.051)	0.951 (0.024, 0.025)	0.903 (0.047, 0.050)	0.951 (0.026, 0.024)

**Tabella 1.3:** Risultati delle simulazione per la distorsione in media (BI), l'errore quadratico medio (MSE), distosione in mediana (meBI) e per la deviazione totale dalla mediana (MAD), sotto assunzioni di normalità e nel caso di omoschedasticità e non omoschedasticità. Fonte: *Cortese e Ventura (2012)*.

$(n_x, n_y)$	Estimator	Equal variance				Different variance			
		BI	MSE	meBI	MAD	BI	MSE	meBI	MAD
$\psi = 0.5$									
(5, 5)	$\hat{\psi}$	-0.0003	0.038	0.0017	0.205	0.0064	0.043	0.0177	0.230
	$\hat{\psi}^*$	-0.0052	0.035	0.0017	0.182	0.0002	0.040	0.0160	0.199
(10, 10)	$\hat{\psi}$	0.0005	0.017	-0.0019	0.135	0.0005	0.018	0.0018	0.135
	$\hat{\psi}^*$	-0.0012	0.016	0.0017	0.128	-0.0014	0.017	0.0019	0.126
(20, 20)	$\hat{\psi}$	-0.0001	0.008	0.0003	0.092	0.0003	0.008	0.0012	0.092
	$\hat{\psi}^*$	-0.0007	0.008	0.0001	0.089	-0.0003	0.008	0.0013	0.089
(30, 30)	$\hat{\psi}$	0.0002	0.005	-0.0001	0.075	0.0001	0.005	-0.0003	0.074
	$\hat{\psi}^*$	-0.0005	0.005	0.00004	0.073	-0.0003	0.005	-0.0004	0.072
$\psi = 0.8$									
(5, 5)	$\hat{\psi}$	0.0096	0.019	0.0336	0.134	0.0157	0.019	0.0387	0.146
	$\hat{\psi}^*$	-0.0147	0.018	0.0013	0.133	-0.0120	0.019	0.0001	0.144
(10, 10)	$\hat{\psi}$	0.0075	0.009	0.0179	0.097	0.0109	0.010	0.0196	0.102
	$\hat{\psi}^*$	-0.0061	0.009	0.0023	0.095	-0.0045	0.009	0.0019	0.101
(20, 20)	$\hat{\psi}$	0.0028	0.004	0.0068	0.068	0.0048	0.005	0.0089	0.071
	$\hat{\psi}^*$	-0.0043	0.004	0.0008	0.067	-0.0034	0.005	0.00004	0.070
(30, 30)	$\hat{\psi}$	0.0017	0.003	0.0043	0.057	0.0033	0.003	0.0055	0.058
	$\hat{\psi}^*$	-0.0031	0.003	-0.0007	0.056	-0.0022	0.003	-0.0003	0.058
$\psi = 0.95$									
(5, 5)	$\hat{\psi}$	0.0018	0.003	0.0237	0.034	0.0012	0.004	0.0235	0.037
	$\hat{\psi}^*$	-0.0146	0.005	0.0043	0.051	-0.0168	0.006	-0.0005	0.060
(10, 10)	$\hat{\psi}$	0.0012	0.002	0.0113	0.035	0.0005	0.002	0.0135	0.038
	$\hat{\psi}^*$	-0.0086	0.002	0.0008	0.040	-0.0106	0.003	0.0012	0.045
(20, 20)	$\hat{\psi}$	0.0006	0.001	0.0061	0.027	0.0003	0.001	0.0063	0.030
	$\hat{\psi}^*$	-0.0045	0.001	0.0007	0.029	-0.0056	0.001	-0.00002	0.033
(30, 30)	$\hat{\psi}$	-0.0003	0.001	0.0032	0.024	0.000002	0.001	0.0039	0.026
	$\hat{\psi}^*$	-0.0038	0.001	-0.0004	0.024	-0.0040	0.001	-0.0004	0.027

### 1.3.3 Approccio bayesiano

In un'ottica bayesiana, il parametro  $\theta$  non viene più considerato come una costante ignota ma viene visto come una variabile casuale con funzione di densità  $\pi(\theta)$ , chiamata distribuzione a priori (Kotz *et al.*, 2003).

Data la distribuzione a priori  $\pi(\theta)$  per il parametro  $\theta$  e data la funzione di verosimiglianza per  $\theta = (\theta_x, \theta_y)$ , dal Teorema di Bayes si ottiene la distribuzione a posteriori per  $\theta$ , che assume la forma

$$\pi(\theta | x, y) = \frac{L(\theta; x, y)\pi(\theta)}{\int_{\Theta} L(\theta; x, y)\pi(\theta)d\theta}. \quad (1.27)$$

La (1.27) rappresenta la distribuzione del parametro  $\theta$  condizionata all'esperimento, nonché tutta l'informazione disponibile sullo stesso parametro. L'espressione al denominatore  $\int_{\Theta} L(x, y | \theta)\pi(\theta)d\theta$ , la costante di normalizzazione, è una quantità che non dipende da  $\theta$  (Liseo, 2008). L'espressione della distribuzione a posteriori può essere riscritta come

$$\pi(\theta | x, y) \propto L(\theta; x, y)\pi(\theta). \quad (1.28)$$

Dalla distribuzione a posteriori è possibile ottenere alcuni indicatori utili all'analisi inferenziale, quali ad esempio, il valore atteso a posteriori, la mediana a posteriori e

la moda a posteriori. Ad esempio, la media a posteriori per  $\theta$  è data da

$$\tilde{\theta} = \int \theta \pi(\theta | x, y) d\theta. \quad (1.29)$$

Sia  $\theta = (\psi, \lambda)$ , dove  $\psi$  è il parametro di interesse e  $\lambda$  il parametro di disturbo. L'inferenza bayesiana su  $\psi$  si basa sulla distribuzione marginale a posteriori

$$\pi(\psi | x, y) = \int \pi(\psi, \lambda | x, y) d\lambda. \quad (1.30)$$

Uno stimatore bayesiano per  $\psi = P(X < Y)$  può essere ottenuto come media a posteriori della (1.30) (Kotz *et al.*, 2003), ossia

$$\tilde{\psi} = \int \psi \pi(\psi | x, y) d\psi. \quad (1.31)$$

Come ben noto, in ambito bayesiano, un problema di stima intervallare per il parametro di interesse si traduce nella costruzione di un intervallo di credibilità.

Sia  $\pi(\psi | x, y)$  la distribuzione a posteriori marginale di  $\psi$ . Quindi, per ogni insieme  $W \in [0, 1]$ , si ha

$$P_\pi(\psi \in W | x, y) = \int_W \pi(\psi | x, y) d\psi. \quad (1.32)$$

Come si può notare, esistono infiniti insiemi di credibilità  $W$  che soddisfano la condizione  $P_\pi(\psi \in W | x, y) = 1 - \alpha$ . Dato che è preferibile considerare l'insieme più corto si sceglierà l'insieme  $W$  che comprende la più alta densità a posteriori (HPD - *Highest Posterior Density*) (Kotz *et al.*, 2003, Cap. 3), che assume la forma

$$W(x, y) = \{\psi : \pi(\psi | x, y) \geq c_\alpha\}, \quad (1.33)$$

dove  $c_\alpha$  è scelto in modo che

$$\int_{W(x, y)} \pi(\psi | x, y) d\psi = 1 - \alpha. \quad (1.34)$$

Nell'inferenza bayesiana risulta cruciale la scelta della distribuzione a priori sul parametro. In letteratura sono presenti diverse alternative come, ad esempio, le distribuzioni a priori coniugate, la distribuzione a priori non informativa di Jeffreys, le *reference priors*, le *conditional reference priors* e le *matching priors* (si veda ad esempio Kotz *et al.*, 2003; Datta e Mukerjee, 2004; Liseo, 2008).

Una distribuzione di probabilità iniziale  $\pi(\theta)$  si dice coniugata al modello se la forma funzionale della distribuzione iniziale e della distribuzione finale sono uguali (Liseo,



2008). Si consideri ad esempio il caso di un modello normale con media ignota e varianza  $\sigma^2$  nota, ossia

$$f(x; \mu, \sigma^2) \sim N(\mu, \sigma^2),$$

e una distribuzione a priori normale

$$\pi(\mu) \sim N(\mu_0, \tau^2).$$

con  $\mu_0$  e  $\tau$  noti. Tale distribuzione a priori è coniugata poichè la distribuzione a posteriori ha la forma

$$\pi(\mu | x) \sim N\left(\frac{\mu_0\sigma^2 + \bar{x}n\tau^2}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right).$$

La distribuzione a priori di Jeffreys appartiene alla classe delle distribuzioni non informative. Essa è definita come (Liseo, 2008)

$$\pi^J(\theta) \propto \sqrt{\det i(\theta)}, \quad (1.35)$$

dove  $i(\theta)$  è la matrice di informazione attesa di Fisher. Affinchè sia possibile utilizzare la distribuzione a priori di Jeffreys è necessario che la matrice  $i(\theta)$  esista e sia definita positiva, condizione verificata nella quasi totalità dei modelli statistici di uso comune. Il vantaggio nell'uso della distribuzione (1.35) sta nel fatto che essa gode della proprietà di invarianza alla riparametrizzazione (Liseo, 2008). Le *matching priors* appartengono alla classe delle distribuzioni a priori non informative. Esse sono distribuzioni a priori per la quale gli intervalli di credibilità a posteriori hanno una copertura frequentista ad un certo ordine di approssimazione (Kotz *et al.*, 2003; Datta e Mukerjee, 2004).

Quindi le *matching priors* sono tali per cui (si veda ad esempio Staicu e Reid, 2008)

$$Pr_\pi\{\theta \leq \theta^{(1-\alpha)}(x) | x\} = Pr_\theta\{\theta^\alpha(X) \geq \theta\} + O(n^{-1}), \quad (1.36)$$

dove  $\theta^{(1-\alpha)}(x)$  è il quantile  $(1 - \alpha)$  della distribuzione a posteriori  $\pi(\theta | x)$ . Secondo Datta e Mukerjee (2004) tali distribuzioni a priori sono *matching priors* del primo ordine. In un modello con parametro scalare, Welch e Peers (1963) hanno mostrato che la *matching prior* è  $\pi(\theta) \propto i(\theta)^{-1/2}$  e coincide con la a priori di Jeffreys. In modelli con parametri di disturbo, con parametri  $\psi$  e  $\lambda$  ortogonali, con riferimento all'informazione attesa di Fisher ( $i_{\psi\lambda}(\psi, \lambda) = 0$ ), Tibshirani (1989) e Nicolau (1993)

hanno mostrato che la *matching prior* assume la forma

$$\pi(\psi, \lambda) \propto i_{\psi\psi}(\psi\lambda)^{1/2}g(\lambda), \quad (1.37)$$

dove  $g(\lambda)$  è una funzione arbitraria di  $\lambda$ . Staicu e Reid (2008) mostrano che la scelta più semplice è  $g(\lambda) = 1$ . La distribuzione a posteriori è esprimibile come

$$\pi_{ps}(\psi | x, y) \propto \pi_{mp}(\psi)L_{MP}(\psi), \quad (1.38)$$

che risulta una distribuzione a posteriori propria (Ventura *et al.*, 2009). Nella (1.38) la *matching prors* associata alla verosimiglianza profilo modificata è (Ventura *et al.*, 2009)

$$\pi_{mp}(\psi) \propto i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi)^{1/2}, \quad (1.39)$$

dove

$$i_{\psi\psi.\lambda}(\psi, \lambda) = i_{\psi\psi}(\psi, \lambda) - i_{\psi\lambda}(\psi, \lambda)i_{\lambda\lambda}(\psi, \lambda)^{-1}i_{\lambda\psi}(\psi, \lambda)$$

è l'informazione parziale per  $\psi$ .

### 1.3.4 Esempio Normale

Siano  $X$  e  $Y$  due variabili indipendenti aventi distribuzione Normale, ossia  $X \sim N(\mu_x, \sigma^2)$  e  $Y \sim N(\mu_y, \sigma^2)$ , con  $\mu_x, \mu_y$  ignoti e uguale varianza  $\sigma^2$  ignota. Il parametro  $\theta$  è dato da  $\theta = (\mu_x, \mu_y, \sigma^2)$  e l'AUC può essere espressa come (Ventura e Racugno, 2011)

$$\psi = \psi(\theta) = \Phi\left(\frac{\mu_y - \mu_x}{\sqrt{2\sigma^2}}\right).$$

Si consideri la riparametrizzazione  $\theta = (\psi, \lambda)$ , con  $\psi = (\mu_y - \mu_x)/\sigma$  parametro di interesse e  $\lambda = (\lambda_1, \lambda_2) = (\mu_y, \sigma^2)$  parametro di disturbo. Per ottenere la distribuzione a posteriori per  $\psi$  si può considerare la *reference prior* (si veda Ghosh e Sun, 1998), data da

$$\pi(\psi, \lambda) \propto \frac{1}{\lambda_2} \frac{1}{(2(1 + m^{-1})(1 + m) + \psi^2)^{1/2}}, \quad (1.40)$$

con  $m = n_x/n_y$ . Così come riferito in Ghosh e Sun (1998), la distribuzione a posteriori marginale  $\pi(\psi | x, y)$  è analiticamente complessa da calcolare. Al fine di superare questo problema si può considerare un'approssimazione del secondo ordine

di  $\pi(\psi | x, y)$ , basato sulla log-verosimiglianza profilo

$$l_p(\psi) = -n \left( \log \tilde{\lambda}_2 + \frac{\hat{\lambda}_2^2}{2\tilde{\lambda}_2^2} \right) - \frac{1}{2\tilde{\lambda}_2^2} \left( n_y(\hat{\lambda}_1 - \tilde{\lambda}_1)^2 + n_x(\hat{\lambda}_2\psi + \hat{\lambda}_1 - \tilde{\lambda}_2\psi - \tilde{\lambda}_1)^2 \right), \quad (1.41)$$

dove  $n = n_x + n_y$ , e le stime vincolate di massima verosimiglianza  $\hat{\lambda}_\psi = (\hat{\lambda}_1, \hat{\lambda}_2)$  sono ottenute numericamente. La distribuzione a posteriori marginale per  $\psi$  è

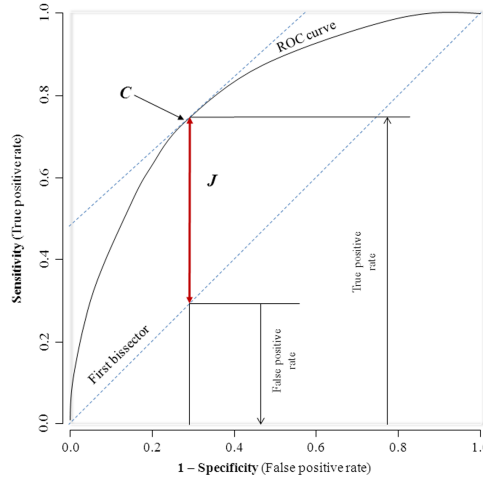
$$\pi^H(\psi | x, y) \propto \frac{\exp \{l_p(\psi)\} |j_{\lambda,\lambda}(\psi, \hat{\lambda}_\psi)|^{-1/2}}{\tilde{\lambda}_2(2(1+m^{-1})(1+m) + \psi^2)^{1/2}}, \quad (1.42)$$

dove  $j_{\lambda,\lambda}(\psi, \hat{\lambda}_\psi)$  è l'informazione osservata per  $\psi$ , ossia

$$|j_{\lambda,\lambda}(\psi, \hat{\lambda}_\psi)| = \frac{2n^2}{\tilde{\lambda}_2^4} - \frac{2n_x n_y \psi (\hat{\lambda}_2 \hat{\psi} + \hat{\lambda}_1 - \tilde{\lambda}_2 - \tilde{\lambda}_1)}{\tilde{\lambda}_2^5} + \frac{n_x n_y \psi^2}{\tilde{\lambda}_2^4}. \quad (1.43)$$

## 1.4 Indice $J$ di Youden

Un altro strumento utile a valutare l'accuratezza di un test diagnostico è l'indice  $J$  di Youden, introdotto da Youden (1950). L'indice  $J$  rappresenta la distanza massima tra un punto  $(x, y)$ , che giace sulla bisettrice, e un punto che giace sulla curva ROC.



**Figura 1.5:** Indice  $J$  di Youden.

La statistica  $J$  rappresenta la massimizzazione di un'opportuna funzione della Sensibilità e Specificità per un valore soglia  $c^*$ , che rappresenta quel valore di *cut-off* che meglio riesce a catturare la distinzione tra popolazione di sani e di malati. L'indice è così definito

$$J = \max_c \{Se(c) + Sp(c) - 1\}, \quad J \in [0, 1]. \quad (1.44)$$

L'indice  $J$  assume il suo valore massimo, pari a uno, quando un test diagnostico separa perfettamente la popolazione dei sani dalla popolazione dei malati, mentre assume valore minimo, pari a zero, quando tale separazione non risulta migliore di una scelta casuale (Luo e Xiong, 2012). Inoltre, è possibile fornire dei criteri per la scelta del valore di *cut-off* "ottimale"  $c^*$ , in funzione del quale viene massimizzata l'espressione dell'indice  $J$  (Greiner *et al.*, 2000). Una proprietà importante dell'indice  $J$  è l'invarianza rispetto a trasformazioni monotone del test diagnostico.

Per la stima di  $J$  e  $c^*$ , ci si riferisce ai risultati di Faraggi e Reiser (2002). La sensibilità e la specificità, per un qualsiasi valore soglia  $c$ , possono essere espressi come  $Se(c) = 1 - F_x(c)$  e  $Sp(c) = F_y(c)$ . L'espressione (1.44) dell'Indice  $J$  di Youden può essere riscritta come

$$J = \max_c \{F_y(c) - F_x(c)\}. \quad (1.45)$$

Il valore soglia ottimale  $c^*$  sarà tale da massimizzare questa espressione. La ricerca di tale *cut-off*, e quindi la stima di  $J$  ( $\hat{J}$ ), può essere ottenuta mediante un approccio non parametrico considerando le stime empiriche delle due funzioni  $F_x$  e  $F_y$  (dette  $\hat{F}_x$  e  $\hat{F}_y$ ) sostituite nell'espressione (1.45), ossia

$$\hat{J} = \max_c \{\hat{F}_y(c) - \hat{F}_x(c)\}. \quad (1.46)$$

Un secondo approccio parametrico parte dal presupposto che la Curva ROC è invariante rispetto a qualsiasi trasformazione dei dati. Questo metodo consiste nel considerare una trasformazione monotona  $t(\cdot)$ , per la quale si ha che  $t(X)$  e  $t(Y)$  seguono una distribuzione normale. Dopo aver applicato la trasformazione dei dati possiamo usare le stime  $\hat{J}$  e  $c^*$  per i dati trasformati  $(c_t^*, J_t)$ , ottenendo quindi le relative stime  $\hat{J} = \hat{J}_t$  e  $\hat{c}^* = t^{-1}(\hat{c}_t^*)$ .

Per poter utilizzare questo metodo è necessario definire la trasformata  $t(\cdot)$  da applicare ai dati. Diversi autori (Zou and Hall, 2000; Zou and Hall, 2002; Faraggi and Reiser, 2002; O'Malley and Zou, 2002) hanno raccomandato l'uso della trasformata di tipo Box-Cox (Fluss *et al.*, 2005), data da

$$t(y) = y^{(\lambda)} = \begin{cases} (y^\lambda - 1)\lambda & \text{se } \lambda \neq 0 \\ \log(y) & \text{se } \lambda = 0 \end{cases}. \quad (1.47)$$

Assumendo che  $Y^{(\lambda)}$  e  $X^{(\lambda)}$  seguano una distribuzione normale,  $\lambda$  può essere stimato utilizzando le procedure di massima verosimiglianza.

### 1.4.1 Esempio Normale

Consideriamo un approccio parametrico per la stima delle due funzioni di ripartizione. Siano  $X$  e  $Y$  due variabili casuali con distribuzioni normali e indipendenti, ossia  $X \sim N(\mu_x, \sigma_x^2)$  e  $Y \sim N(\mu_y, \sigma_y^2)$ , rispettivamente. Senza perdita di generalità si assume che  $\mu_x < \mu_y$ . Si ottiene che

$$Se(c) = 1 - \Phi\left(\frac{c - \mu_y}{\sigma_y}\right) \quad \text{e} \quad Sp(c) = \Phi\left(\frac{c - \mu_x}{\sigma_x}\right), \quad (1.48)$$

$$S(c) = \Phi\left(\frac{c - \mu_x}{\sigma_x}\right) - \Phi\left(\frac{c - \mu_y}{\sigma_y}\right), \quad J \equiv \max_c \{S(c)\}.$$

Risolvendo l'equazione  $S'(c) = 0$  otteniamo

$$c^* = \frac{(\mu_y \sigma_x^2 - \mu_x \sigma_y^2) - \sigma_y \sigma_x \sqrt{(\mu_x - \mu_y)^2 + (\sigma_x^2 - \sigma_y^2) \log(\sigma_x^2 / \sigma_y^2)}}{(\sigma_x^2 - \sigma_y^2)}, \quad (1.49)$$

che risulta il punto di massimo. Sostituendo  $c^*$  a  $c$  nelle formule in (1.45) si ottiene

$$J^* = \Phi\left(\frac{c^* - \mu_x}{\sigma_x}\right) - \Phi\left(\frac{c^* - \mu_y}{\sigma_y}\right). \quad (1.50)$$

Per la stima di  $J^*$  e  $c^*$ , i parametri ignoti  $(\mu_x, \mu_y, \sigma_x, \sigma_y)$  possono essere sostituiti dalle medie campionarie e deviazione standard.

## 1.5 Conclusioni

Da quanto enfatizzato in questo primo capitolo, l'inferenza sulla curva ROC, quindi sull'AUC, si colloca nel contesto di analisi di dati medico-sanitari. In particolare, il modello sollecitazione-resistenza (si veda ad esempio Kotz *et al.*, 2003) è uno strumento molto utilizzato in un contesto di analisi di classificazione su due gruppi in vari ambiti applicativi, in particolare in quello medico. Come dimostrato dai diversi studi in letteratura, l'approccio frequentista, considerando in particolare i metodi inferenziali di ordine superiore, ha portato un efficace riscontro anche in un contesto applicativo. L'approccio bayesiano si è sviluppato anche in contesti in cui non si dispone di un *gold standard* (si veda Gonçalves *et al.*, 2012).

La rassegna fatta in questo primo capitolo è volta soprattutto ad introdurre il concetto di superficie ROC, visto come una generalizzazione del concetto di curva ROC. Nel capitolo successivo viene contestualizzata l'inferenza connessa a tale concetto,

rivolta in particolare ad un approccio parametrico di tipo bayesiano delle procedure di classificazione.

## Capitolo 2

# Volume sotteso alla curva ROC

In molti campi scientifici, soprattutto in quello medico-sanitario, potrebbe essere utile disporre di uno strumento che consenta una classificazione dei dati osservati in più di due classi. Ad esempio, si consideri il caso di una classificazione su tre modalità dei pazienti oggetto di studio. In questo capitolo si presenta una generalizzazione della teoria illustrata nel capitolo precedente. In particolare, si introduce il concetto di superficie ROC (§2.1), il cui studio in letteratura è ancora molto vivace. Successivamente viene illustrata la quantità VUS (§2.2) (*Volume Under the ROC Surface*) e l'estensione per tre variabili dell'Indice  $J$  di Youden (§2.3).

Il concetto di superficie ROC, introdotta da Scurfield (1996), è stato oggetto di numerosi studi in anni recenti. Mossman (1999) ha proposto una versione teorica della superficie ROC, poi implementata in *Mathematica* da Heckerling (2001). Partendo da Mossman (1999), Dreiseitl *et al.* (2000) hanno affrontato l'inferenza sul VUS, considerando un approccio non-parametrico, mentre Nakas e Yiannoutsos (2004), sempre in un contesto non-parametrico, hanno affrontato il problema della struttura della superficie ROC e di una generalizzazione del problema di classificazione multipla. Ferri *et al.* (2003) e Li e Zhou (2009) hanno considerato la superficie ROC come un'estensione della curva ROC per un test diagnostico su tre gruppi, mentre hanno considerato il VUS come una misura generale indicativa della potenza discriminante del test diagnostico. Recentemente, unendo gli approcci forniti da Scurfield (1996) e Mossman (1999), sono state derivate le basi teoriche per la generalizzazione del concetto di analisi della curva ROC (si veda Xiong *et al.*, 2006; Li e Fine, 2008). In particolare, Xiong *et al.* (2006) hanno considerato un approccio parametrico per l'inferenza sulla superficie ROC, estendendo il lavoro fatto da Nakas e Yiannoutsos (2004); si veda anche Kranowski e Hand (2009). Recentemente, lo studio della superficie ROC è stato oggetto di studio di numerosi autori. Questo è dimostrato dai vari studi che si sono recentemente trovati in letteratura (si veda, ad esempio, Yu,

2012; Ratnasamy *et al.*, 2008; Yiannoutsos *et al.*, 2008, Nakas, 2014).

In un contesto di analisi inferenziale, il test diagnostico per tre modalità di classificazione è spesso visto come un'estensione del caso binario, specialmente per la ricerca dei punti di *cut-off*, utili ad individuare la migliore classificazione possibile (Mossman, 1999; Landgrebe e Duin, 2007). La rassegna presentata in questo capitolo, allo stato attuale della letteratura, si riferisce soprattutto a trattazioni non parametriche in ambito frequentista e bayesiano. La metodologia bayesiana risulta ancora oggi carente di uno approccio parametrico all'inferenza sul VUS.

Nel capitolo precedente si è visto come la curva ROC e, in particolare, l'AUC sono uno strumento efficace per la valutazione dei test diagnostici in un contesto di classificazione binaria. Lo scopo di questo capitolo è quello di illustrare la generalizzazione di tale metodo, quando è di interesse la classificazione delle osservazioni in tre possibili gruppi.

## 2.1 Superficie ROC

Le nozioni sulla superficie ROC sono intese come una naturale generalizzazione della curva ROC per la valutazione di un test diagnostico in un contesto di classificazione a tre livelli.

Per definire formalmente il problema generale di classificazione, si supponga di avere a disposizione  $n_x$  misurazioni dal Gruppo I, estratte dalla variabile casuale  $X$ ,  $n_y$  misurazioni dal Gruppo II, estratte dalla variabile casuale  $Y$ ,  $n_z$  misurazioni dal Gruppo III, estratte dalla variabile casuale  $Z$ . Inoltre, siano  $F_x$ ,  $F_y$  e  $F_z$  le rispettive funzioni di ripartizione di  $X$ ,  $Y$  e  $Z$ , variabili indipendenti.

Supponendo che  $X < Y < Z$  sia l'ordinamento oggetto di interesse, la regola decisionale per la classificazione dei soggetti in una delle tre classi avviene in funzione di due valori soglia ordinati  $c_1 < c_2$ . Obiettivo dell'analisi è valutare la qualità del test diagnostico nella classificazione dei soggetti nei tre diversi gruppi.

Si consideri di voler classificare i soggetti in tre gruppi (Gruppo I, Gruppo II, Gruppo III). La costruzione della superficie ROC si basa su una regola decisionale, in funzione della quale sono possibili complessivamente nove classificazioni, di cui tre classificazioni corrette (TCF - *True Classification Fraction*) e sei classificazioni errate (FCF - *False Classification Fraction*). Le tre classificazioni corrette  $TCF_i$  (per  $i = x, y, z$ ), possono essere individuate mediante la frazione attesa di individui correttamente classificati, ossia

- il paziente è considerato sano se l'osservazione  $x$  è minore della soglia  $c_1$ . La



frazione di pazienti sani, classificati correttamente, è data da

$$\#(X \leq c_1) = TCF_x.$$

- il paziente è considerato lievemente malato se l'osservazione  $y$  si colloca tra le soglie  $c_1$  e  $c_2$ . La frazione di pazienti lievemente malati, classificati correttamente, è data da

$$\#(c_1 < Y \leq c_2) = TCF_y.$$

- il paziente è considerato malato se l'osservazione  $z$  è maggiore della soglia  $c_2$ . La frazione di pazienti malati, correttamente classificati, è data da

$$\#(Z > c_2) = TCF_z.$$

Invece, per i sei possibili casi di errata classificazione si ha

- il paziente è classificato come sano, quando invece è un malato lieve, se

$$\#(c_1 \leq X \leq c_2) = FCF_{xy},$$

- il paziente è classificato come sano, quando in realtà è malato, se

$$\#(X \geq c_2) = FCF_{xz},$$

- il paziente è classificato come malato lieve, quando in realtà è sano, se

$$\#(Y < c_1) = FCF_{yx},$$

- il paziente è classificato come malato lieve, quando in realtà è malato, se

$$\#(Y \geq c_2) = FCF_{yz},$$

- il paziente è classificato come malato, quando in realtà è sano, se

$$\#(Z < c_1) = FCF_{zx},$$

- il paziente è classificato come malato, quando in realtà è lievemente malato, se

$$\#(c_1 \leq Z \leq c_2) = FCF_{zy}.$$

Tali risultati possono essere riassunti nella Matrice di Confusione (Tabella 2.1), simile alla Tabella 1.1.

		Classe reale		
		Gruppo I	Gruppo II	Gruppo III
Classificazione	Gruppo I	$TCF_x$	$FCF_{xy}$	$FCF_{xz}$
	Gruppo II	$FCF_{yx}$	$TCF_y$	$FCF_{yz}$
	Gruppo III	$FCF_{zx}$	$FCF_{zy}$	$TCF_z$

**Tabella 2.1:** Matrice di confusione per tre gruppi.

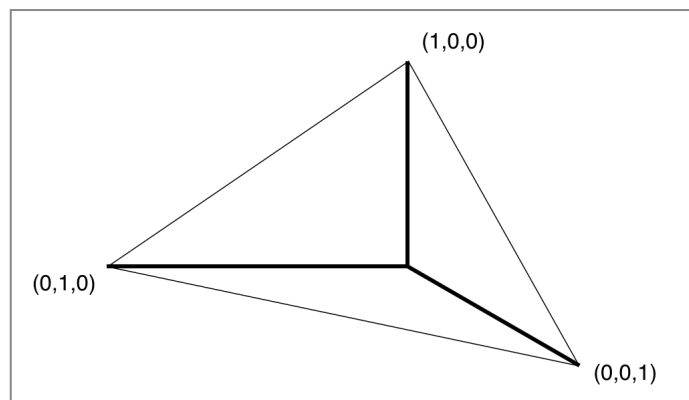
Le singole frazioni all'interno della Tabella 2.1 hanno valore in  $[0, 1]$ . Valgono quindi le seguenti relazioni

$$TCF_x + FCF_{yx} + FCF_{zx} = 1$$

$$FCF_{xy} + TCF_y + FCF_{zy} = 1$$

$$FCF_{zx} + FCF_{zy} + TCF_z = 1$$

Possiamo rappresentare la superficie ROC come un grafico tridimensionale in un'unità cubica, di coordinate  $\{F_x(c_1), F_y(c_2) - F_y(c_1), 1 - F_z(c_2)\}$ , al variare di  $(c_1, c_2)$  (con  $c_1 < c_2$ ). Gli assi del grafico tridimensionale  $(x, y, z)$ , che rappresentano i tre gruppi della classificazione, rispettivamente i pazienti sani, malati lievi e malati, hanno angoli di coordinate  $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$  (Figura 2.4).



**Figura 2.1:** Grafico tridimensionale in cui ogni asse si riferisce ad un gruppo di pazienti.

Sia  $ROC_s$  la forma funzionale della superficie ROC (Nakas e Yiannoutsos, 2004). Allora si ha

$$ROC_s(TCF_x, TCF_z) = F_y(F_z^{-1}(1 - TCF_z)) - F_y(F_x^{-1}(TCF_x)).$$

Ponendo  $TCF_x = p_x$  e  $TCF_z = p_z$ , possiamo riscrivere l'espressione come

$$ROC_s(p_x, p_z) = \begin{cases} F_y(F_z^{-1}(1 - p_z)) - F_y(F_x^{-1}(p_x)), & \text{se } F_x^{-1}(p_x) \geq F_z^{-1}(1 - p_z) \\ 0, & \text{altrimenti} \end{cases} \quad (2.1)$$

### 2.1.1 Esempio normale

Sotto l'assunzione di normalità delle variabili  $X, Y$  e  $Z$ , ossia

$$X \sim N(\mu_x, \sigma_x^2), \quad Y \sim N(\mu_y, \sigma_y^2), \quad Z \sim N(\mu_z, \sigma_z^2), \quad (2.2)$$

indipendenti con parametri ignoti, Xiong *et al.* (2006) hanno usato il modello (2.1) per descrivere la struttura generale della superficie ROC. Una forma parametrica della superficie ROC è

$$ROC_s(p_x, p_z) = \left\{ \Phi(\beta_1 + \beta_2 \Phi^{-1}(1 - p_z)) - \Phi(\beta_3 + \beta_4 \Phi^{-1}(p_x)) \right\} \\ \times I_{\{\beta_3 + \beta_4 \Phi^{-1}(p_x) \leq \beta_1 + \beta_2 \Phi^{-1}(1 - p_z)\}}(p_x, p_z) \quad (2.3)$$

dove  $I(\cdot)$  è la funzione indicatrice e  $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^T$  è il vettore dei parametri della superficie ROC dati da

$$\beta_1 = \frac{\mu_z - \mu_y}{\sigma_y}, \quad \beta_2 = \frac{\sigma_z}{\sigma_y}, \quad \beta_3 = \frac{\mu_x - \mu_y}{\sigma_y}, \quad \beta_4 = \frac{\sigma_x}{\sigma_y}.$$

### 2.1.2 Inferenza su VUS

Il volume sotteso alla superficie ROC (VUS - *Volume Under the ROC Curve*) può essere inteso come un indicatore generale della bontà di classificazione e può essere espresso come  $P(X < Y < Z)$ . Infatti, esso è l'equivalente dell'AUC nella curva ROC, discussa nel capitolo precedente. Sotto assunzioni di normalità, Xiong *et al.* (2006) definiscono il VUS come

$$VUS = \int_{-\infty}^{\infty} \Phi\left(\frac{\sigma_y^2}{\sigma_x^2}u - \frac{\mu_x - \mu_y}{\sigma_x^2}\right) \Phi\left(-\frac{\sigma_y^2}{\sigma_z^2}u + \frac{\mu_z - \mu_y}{\sigma_z^2}\right) \phi(u)du, \quad (2.4)$$

Uno stimatore non-parametrico e non distorto del VUS è dato da (Nakas, 2014)

$$\widehat{VUS} = \frac{1}{n_x n_y n_z} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \sum_{k=1}^{n_z} I(x_i, y_j, z_k), \quad (2.5)$$

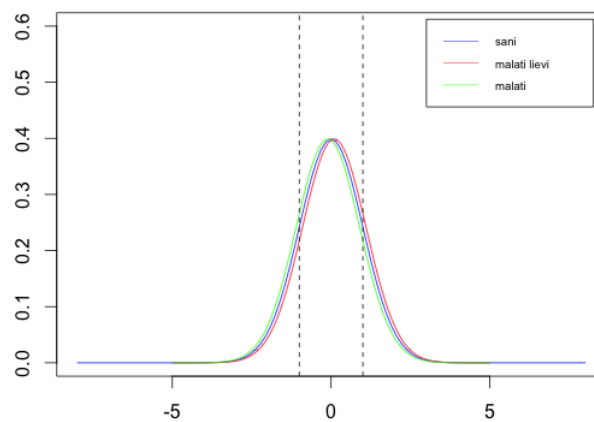
dove  $n_x, n_y$  e  $n_z$  rappresentano le numerosità campionarie dei campioni, rispettivamente. Inoltre,  $I(X, Y, Z)$  è uguale a (Dreiseitl *et al.*, 2000)

$$I(X, Y, Z) = \begin{cases} 1 & \text{se } X < Y < Z \\ \frac{1}{2} & \text{se } X = Y < Z \quad \text{o} \quad X < Y = Z \\ \frac{1}{6} & \text{se } X = Y = Z \\ 0 & \text{altrimenti} \end{cases}$$

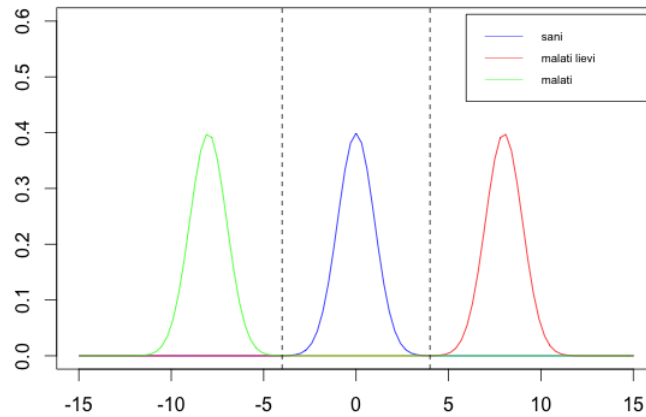
Quindi si può scrivere il valore atteso del VUS come

$$\begin{aligned} E[VUS] &= P(X < Y < Z) + \frac{1}{2}P(X < Y = Z) + \frac{1}{2}P(X = Y < Z) + \\ &+ \frac{1}{6}P(X = Y = Z). \end{aligned}$$

Il VUS assume valore  $1/3! = 1/6$  quando le tre distribuzioni sono completamente sovrapposte (Figura 2.2), mentre assume valore pari a uno quando sono completamente separate e nel giusto ordine (Figura 2.3). Una rappresentazione tridimensionale del VUS quando è pari a  $1/6$  è fornita nella Figura 2.1.

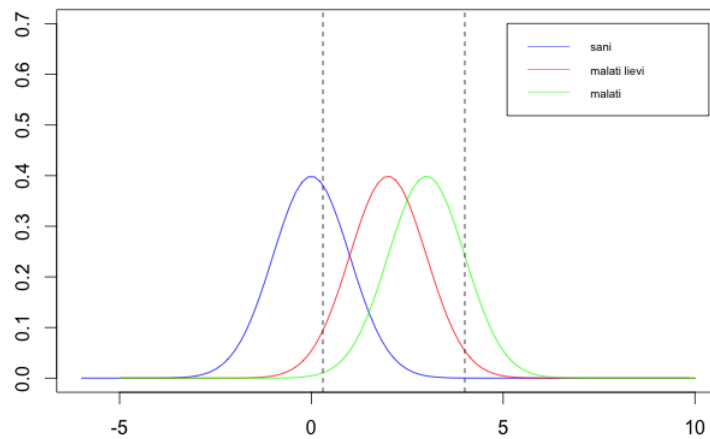


**Figura 2.2:** Caso di sovrapposizione dei tre gruppi di classificazione.



**Figura 2.3:** Caso di perfetta separazione dei tre gruppi di classificazione.

La Figura 2.2 e la Figura 2.3 mostrano due casi estremi delle distribuzioni dei gruppi, raramente riscontrabile nella realtà. In Figura 2.4 si mostra invece il caso in cui i tre gruppi di pazienti risultano parzialmente sovrapposti tra loro.



**Figura 2.4:** Caso di sovrapposizione parziale dei tre gruppi di classificazione.

Kang e Tian (2013) hanno considerato l'uso della trasformazione di Box-Cox applicata a dati non normali per la stima parametrica del VUS. Inoltre, hanno confrontato tale metodo con un approccio non-parametrico basato sulla densità kernel,

in termini di distorsione ed EQM (Errore Quadratico Medio).

## 2.2 Generalizzazione dell'Indice $J$ di Youden

Un versione dell'indice  $J$  di Youden nel caso di una classificazione su tre gruppi è stata proposta per la valutazione dell'accuratezza di un test diagnostico e per la scelta dei relativi punti di *cut-off* (Nakas *et al.*, 2010, 2013).

Si consideri il caso in cui  $X, Y$  e  $Z$  rappresentino le classi di pazienti malati, malati lievi e sani, rispettivamente. Inoltre, in funzione di quanto specificato nel §2.1, si definisce l'indice  $J$  in presenza di tre gruppi di classificazione come

$$\begin{aligned} J_3 &= \max_{c_1, c_2} \{TCF_x + TCF_y + TCF_z - 1\} \\ &= \max_{c_1, c_2} \{F_x(c_1) + F_y(c_2) - F_y(c_1) - F_z(c_2)\}. \end{aligned} \tag{2.6}$$

L'indice  $J_3$  può essere visto come la somma degli indici  $J$  nel caso di una classificazione binaria tra i gruppi sani e malati lievi e i gruppi malati lievi e malati. Infatti, l'indice  $J_3$  assume valori nell'intervallo  $[0, 2]$ . Luo e Xiong (2013) hanno definito  $J_3$  nell'intervallo  $[0, 1]$  proponendo di usare  $J_3/2$ .

Così come per l'indice  $J$  nel contesto di classificazione su due gruppi, anche in questo contesto di analisi, l'indice  $J_3$  gode dell'invarianza rispetto a trasformazioni monotone del test diagnostico.

Gli approcci alla stima dell'indice  $J_3$  presenti in letteratura riguardano metodi parametrici e non-parametrici. Nakas *et al.* (2010, 2013) hanno affrontato il problema in un approccio non-parametrico, considerando la funzione di ripartizione empirica applicata all'espressione (2.12), approfondendo precedenti studi affrontati da Silverman (1986) e Fluss *et al.* (2005). Inoltre, sempre in un contesto di inferenza non-parametrica, Coolen-Maturi *et al.* (2013) hanno fornito approfondimenti per la stima e per l'uso dell'Indice di Youden generalizzato. Luo e Xiong (2012, 2013) hanno affrontato il problema della stima dell'indice di Youden considerando un approccio parametrico. In particolare, Luo e Xiong (2012) hanno creato un pacchetto del software *R* (**DiagTest3Grp**) per la stima di VUS, per l'indice di Youden generalizzato e per la stima dei punti di *cut-off* ottimali, sotto assunzioni normali per un modello parametrico e per un modello non parametrico.

### 2.2.1 Esempio Normale

Sotto assunzioni di normalità, siano  $X, Y, Z$  le variabili casuali indipendenti che rappresentano i tre gruppi di pazienti oggetti di studio (sani, malati lievi e malati). In particolare, si assume una distribuzione  $N(\mu_i, \sigma_i^2)$  per ogni gruppo, con  $i = x, y, z$ . Senza perdita di generalità, si assume  $\mu_x < \mu_y < \mu_z$ . Le espressioni per le probabilità di corretta classificazione dei pazienti sono (Luo e Xiong, 2012)

$$S_p(c_1) = P(X \leq c_1) = \Phi\left(\frac{c_1 - \mu_x}{\sigma_x}\right), \quad S_e(c_2) = P(Z \geq c_2) = 1 - \Phi\left(\frac{c_2 - \mu_z}{\sigma_z}\right) \quad e$$

$$S_m(c_1, c_2) = P(c_1 < Y \leq c_2) = \Phi\left(\frac{c_2 - \mu_y}{\sigma_y}\right) - \Phi\left(\frac{c_1 - \mu_y}{\sigma_y}\right).$$

Sotto le assunzioni di normalità, l'indice  $J$  di Youden è dato da

$$J(c_1, c_2) = \frac{1}{2} \left\{ \left[ \Phi\left(\frac{c_1 - \mu_x}{\sigma_x}\right) - \Phi\left(\frac{c_1 - \mu_y}{\sigma_y}\right) \right] + \left[ \Phi\left(\frac{c_2 - \mu_y}{\sigma_y}\right) - \Phi\left(\frac{c_2 - \mu_z}{\sigma_z}\right) \right] \right\}. \quad (2.7)$$

La coppia di punti *cut-off* è ottenuta considerando le derivate parziali della (2.7) in  $c_1$  e  $c_2$  rispettivamente. Risolvendo in zero le derivate parziali si ottiene che

$$\begin{cases} c_1^* = \frac{(\mu_y \sigma_x^2 - \mu_x \sigma_y^2) - \sigma_x \sigma_y \sqrt{(\mu_x - \mu_y)^2 + (\sigma_x^2 - \sigma_y^2) \log\left(\frac{\sigma_x^2}{\sigma_y^2}\right)}}{\sigma_x^2 - \sigma_y^2} \\ c_2^* = \frac{(\mu_z \sigma_y^2 - \mu_y \sigma_z^2) - \sigma_y \sigma_z \sqrt{(\mu_y - \mu_z)^2 + (\sigma_y^2 - \sigma_z^2) \log\left(\frac{\sigma_y^2}{\sigma_z^2}\right)}}{\sigma_y^2 - \sigma_z^2}. \end{cases} \quad (2.8)$$

In caso di uguaglianza delle varianze ( $\sigma_x^2 = \sigma_y^2$  e/o  $\sigma_y^2 = \sigma_z^2$ ), i punti di *cut-off* ottimali diventano

$$c_1^* = \frac{\mu_x - \mu_y}{2} \quad e \quad c_2^* = \frac{\mu_y - \mu_z}{2}.$$

Assumendo una distribuzione normale per le variabili  $X, Y$  e  $Z$ , la stima dei punti di *cut-off* ottimali e dell'indice  $J$  possono ottenute sostituendo le SMV dei parametri di interesse nella (2.6) e (2.7). Inoltre, quando le assunzioni di normalità non sono verificate, si può considerare di implementare la trasformata di Box-Cox (Vexler *et al.*, 2008).

Sotto l'assunzione di normalità, l'indice  $J$  e la coppia ottimale dei punti di *cut-off* sono funzioni delle medie campionarie e delle varianze campionarie delle tre variabili. In questo contesto, sfruttando il metodo delta, è possibile ricavare la varianza asintotica delle soglie ottimali e la varianza asintotica dell'indice  $J$ . In particolare si

ha

$$\text{Var}(\hat{c}_1^*) = \text{Var}(\hat{\mu}_x) \left( \frac{\partial c_1^*}{\partial \mu_x} \right)^2 + \text{Var}(\hat{\mu}_y) \left( \frac{\partial c_1^*}{\partial \mu_y} \right)^2 + \text{Var}(\hat{\sigma}_x) \left( \frac{\partial c_1^*}{\partial \sigma_x} \right)^2 + \text{Var}(\hat{\sigma}_y) \left( \frac{\partial c_1^*}{\partial \sigma_y} \right)^2$$

$$\text{Var}(\hat{c}_2^*) = \text{Var}(\hat{\mu}_y) \left( \frac{\partial c_2^*}{\partial \mu_y} \right)^2 + \text{Var}(\hat{\mu}_z) \left( \frac{\partial c_2^*}{\partial \mu_z} \right)^2 + \text{Var}(\hat{\sigma}_y) \left( \frac{\partial c_2^*}{\partial \sigma_y} \right)^2 + \text{Var}(\hat{\sigma}_z) \left( \frac{\partial c_2^*}{\partial \sigma_z} \right)^2$$

$$\begin{aligned} \text{Var}(\hat{J}^*) &= \text{Var}(\hat{\mu}_x) \left( \frac{\partial J}{\partial \mu_x} \right)^2 + \text{Var}(\hat{\mu}_z) \left( \frac{\partial J}{\partial \mu_z} \right)^2 + \text{Var}(\hat{\mu}_y) \left( \frac{\partial J}{\partial \mu_y} \right)^2 + \\ &\quad \text{Var}(\hat{\sigma}_x) \left( \frac{\partial J}{\partial \sigma_x} \right)^2 + \text{Var}(\hat{\sigma}_z) \left( \frac{\partial J}{\partial \sigma_z} \right)^2 + \text{Var}(\hat{\sigma}_y) \left( \frac{\partial J}{\partial \sigma_y} \right)^2. \end{aligned}$$

Per le proprietà del metodo delta si ha che la stima  $\hat{J}^*$  e le stime  $(c_1^*, c_2^*)$  risultano asintoticamente non distorte con distribuzioni normali approssimate. Gli intervalli di confidenza di livello  $(1 - \alpha)$  per  $c_1^*$ ,  $c_2^*$  e  $J$  hanno rispettivamente la forma

$$\hat{c}_1^* \pm z_{\alpha/2} \sqrt{\text{Var}(\hat{c}_1^*)}, \quad (2.9)$$

$$\hat{c}_2^* \pm z_{\alpha/2} \sqrt{\text{Var}(\hat{c}_2^*)}, \quad (2.10)$$

$$\hat{J}^* \pm z_{\alpha/2} \sqrt{\text{Var}(\hat{J}^*)}. \quad (2.11)$$



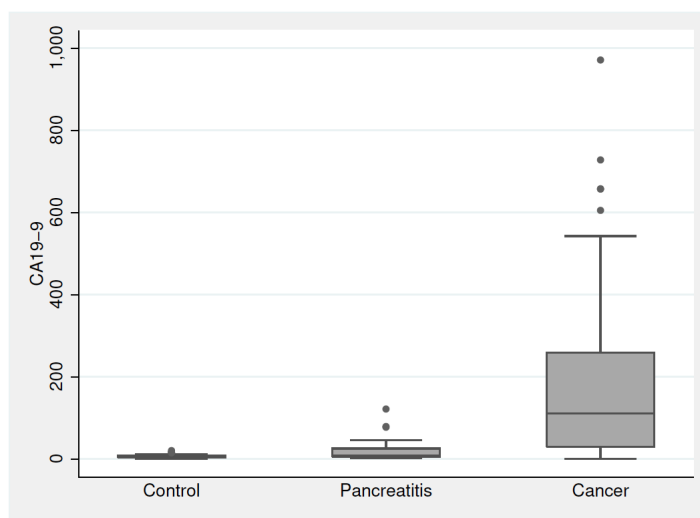
### 2.2.2 Esempio di applicazione di tipo non parametrico

Nakas (2014) ha fornito un'applicazione di tipo non parametrico per un'analisi di classificazione su tre gruppi. In particolare ha analizzato i dati relativi ad un test diagnostico sul cancro al pancreas (CA19-9). I dati si riferiscono a 40 pazienti affetti da cancro al pancreas, 23 pazienti pancreatici, e 40 pazienti sani di controllo. Lo scopo del test diagnostico CA19-9 è di differenziare i pazienti nei tre gruppi ordinali, cioè controllo < pancreatici < cancro.

La Tabella 2.1 e la Figura 2.5 riportano un'analisi descrittiva dei dati.

	Controllo	Pancreatici	Cancro
mean	6.94	22.50	200.46
sd	4.74	30.88	237.85
median	6.60	8.51	111.60
min	0.6	2.5	0.6
max	20.67	121.80	971.50
N	40	23	40

**Tabella 2.2:** Analisi descrittiva dei gruppi (Fonte: Nakas, 2014).



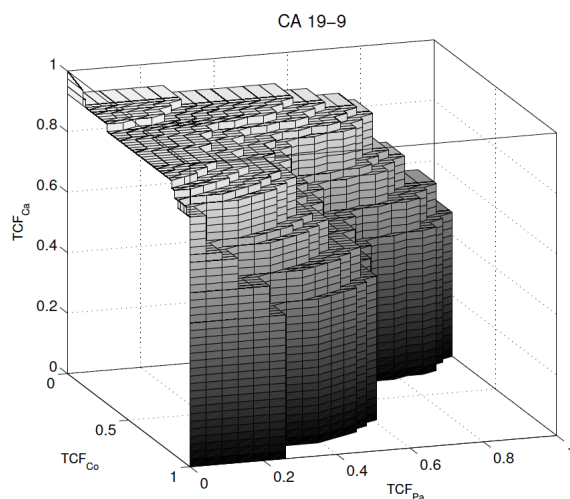
**Figura 2.5:** Boxplots per i gruppi di pazienti sottoposti al test diagnostico CA19-9 (Fonte: Nakas, 2014).

La stima non parametrica del VUS è pari a 0.528 con un intervallo di confidenza al 95% pari a (0.403, 0.654). L'indice di Youden  $J_3$  risulta 0.929 ( $0.464 \times 2$ ), ottenuto in funzione di  $c_1 = 8.40$  e  $c_2 = 25.60$ , dove  $c_1$  corrisponde al valore di *cut-off* tra pazienti affetti da pancreatite e pazienti sani, mentre  $c_2$  è il valore di *cut-off* tra i pancreatici e i pazienti affetti da cancro al pancreas. I valori di TCF sono 80% per

i pazienti sani, 30.40% per i pazienti affetti da pancreatite e l'82.50% per i pazienti affetti da cancro al pancreas. Tale risultato per il VUS evidenzia che i gruppi non sono molto discriminati tra loro a causa della basso tasso di corretta classificazione per pazienti affetti da pancreatite.

Nell'analisi parametrica condotta da Nakas (2014) sugli stessi dati, la stima del VUS è ottenuta mediante l'utilizzo della libreria di *R* **DiagTest3Grp**. I risultati delle analisi hanno fornito un  $\widehat{VUS} = 0.519(IC : 0.385, 0.653)$ ,  $\hat{J}_3 = 1.22 (0.61 \times 2)$ , ottenuto per i valori soglia  $c_1 = 16.17$  e  $c_2 = 86.62$ . La Figura 2.6 mostra una rappresentazione grafica tridimensionale della stima del VUS.

Inoltre, dal test di Shapiro-Wilk è stato verificato il rigetto delle assunzioni di normalità per i dati osservati, motivo per il quale la stima non parametrica del VUS risulta più realistica di quella ottenuta con il metodo parametrico.



**Figura 2.6:** Esempio di superficie ROC ottenuta con un approccio non parametrico.  
Fonte: Nakas, (2014).

## 2.3 Conclusioni

Le metodologie legate alla superficie ROC sono uno strumento importante nell'analisi in un problema di classificazione su tre gruppi. L'utilità di tale metodo è dimostrata dalle applicazioni fatte in diversi ambiti scientifici (si veda ad esempio Ratnasamy *et al.*, 2008; Abraham *et al.*, 2009; Wandishin e Mullen, 2009; Tremont *et al.*, 2011; Dunngalvin *et al.*, 2011; Bruña *et al.*, 2012; Cianferoni *et al.*, 2012; Yu, 2012; Coleman *et al.*, 2013; Migliaretti *et al.*, 2013; Leichtle *et al.*, 2013; Nakas, 2014). Fino ad oggi non risultano applicazioni di analisi parametriche bayesiane all'inferenza della superficie ROC.

Quindi, introdotti i concetti fondamentali di curva ROC e AUC (*Area Under the ROC curve*), superficie ROC e VUS (*Volume Under the ROC Surface*), il prossimo capitolo discute una applicazione di inferenza parametrica bayesiana per il VUS. Infatti, l'obiettivo di questa tesi è fornire un contributo all'inferenza bayesiana parametrica oggettiva per l'analisi del VUS, che risulta ancora oggi carente in letteratura. Per una stima parametrica di tipo bayesiana del VUS si può considerare un approccio Markov Chain Monte Carlo; in particolare, risultano appropriati gli algoritmi Metropolis-Hastings e Gibbs sampler.



# Capitolo 3

## Inferenza parametrica bayesiana sul VUS

Dopo aver richiamato nei primi due capitoli l'inferenza sull'AUC e sul VUS, in questo capitolo si discute l'inferenza bayesiana parametrica per il VUS che, allo stato attuale, non risulta ancora presente in letteratura. In particolare, l'analisi condotta si basa su un'impostazione soggettiva della metodologia bayesiana.

In particolare, vengono presentati i modelli parametrici utilizzati nell'analisi (§3.1); vengono illustrati gli algoritmi MCMC, utili per approssimare distribuzioni a posteriori marginali (§3.2); vengono illustrati i risultati delle simulazioni ottenute mediante l'utilizzo dell'algoritmo *Metropolis-Hastings* (§3.3); viene presentata un'applicazione dell'algoritmo su un dataset reale (§3.4).

### 3.1 Il modello parametrico

Supponiamo di voler classificare i pazienti in tre gruppi ordinali, indicati rispettivamente con *Gruppo I*, *Gruppo II* e *Gruppo III*. Le misurazioni del test diagnostico nei tre gruppi vengono rappresentate dalle variabili aleatorie  $X$ ,  $Y$  e  $Z$ , rispettivamente. Inoltre, si assume che tali variabili si distribuiscano come normali indipendenti, con medie e varianze ignote. Quindi si ha

$$X \sim N(\mu_x, \sigma_x^2) \quad Y \sim N(\mu_y, \sigma_y^2) \quad Z \sim N(\mu_z, \sigma_z^2). \quad (3.1)$$

Le rispettive funzioni di densità sono

$$f(x; \mu_x, \sigma_x^2) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp \left\{ -\frac{1}{2\sigma_x^2}(x - \mu_x)^2 \right\},$$

$$f(y; \mu_y, \sigma_y^2) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp \left\{ -\frac{1}{2\sigma_y^2} (y - \mu_y)^2 \right\} \quad (3.2)$$

e

$$f(z; \mu_z, \sigma_z^2) = \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp \left\{ -\frac{1}{2\sigma_z^2} (z - \mu_z)^2 \right\}.$$

Dati tre campioni casuali semplici di numerosità  $n_x$ ,  $n_y$  e  $n_z$ , rispettivamente tratti da  $X$ ,  $Y$  e  $Z$ , le funzioni di verosimiglianza nei tre modelli sono

$$L(\mu_x, \sigma_x^2) = \prod_{i=1}^{n_x} f_x(x_i; \mu_x, \sigma_x^2) = \left( \frac{1}{\sigma_x^2} \right)^{n_x} \exp \left\{ -\frac{1}{2\sigma_x^2} \sum_{i=1}^{n_x} (x_i - \mu_x)^2 \right\},$$

$$L(\mu_y, \sigma_y^2) = \prod_{i=1}^{n_y} f_y(y_i; \mu_y, \sigma_y^2) = \left( \frac{1}{\sigma_y^2} \right)^{n_y} \exp \left\{ -\frac{1}{2\sigma_y^2} \sum_{i=1}^{n_y} (y_i - \mu_y)^2 \right\} \quad (3.3)$$

e

$$L(\mu_z, \sigma_z^2) = \prod_{i=1}^{n_z} f_z(z_i; \mu_z, \sigma_z^2) = \left( \frac{1}{\sigma_z^2} \right)^{n_z} \exp \left\{ -\frac{1}{2\sigma_z^2} \sum_{i=1}^{n_z} (z_i - \mu_z)^2 \right\}.$$

Lo scopo dell'analisi è calcolare la probabilità  $P(X < Y < Z)$ . Si considera quindi la funzione di verosimiglianza completa per  $\theta = (\mu_x, \mu_y, \mu_z, \sigma_x^2, \sigma_y^2, \sigma_z^2)$ , che assume la forma

$$L(\theta) = \frac{1}{\sqrt{(\sigma_x^2)^{n_x} (\sigma_y^2)^{n_y} (\sigma_z^2)^{n_z}}} \exp \left\{ -\frac{1}{2\sigma_x^2} \sum_{i=1}^{n_x} (x_i - \mu_x)^2 - \frac{1}{2\sigma_y^2} \sum_{i=1}^{n_y} (y_i - \mu_y)^2 - \frac{1}{2\sigma_z^2} \sum_{i=1}^{n_z} (z_i - \mu_z)^2 \right\}. \quad (3.4)$$

La log-verosimiglianza per  $\theta$  è

$$l(\theta) = -\frac{1}{2} (n_x \log(\sigma_x^2) + n_y \log(\sigma_y^2) + n_z \log(\sigma_z^2)) - \frac{1}{2} \left[ \frac{1}{\sigma_x^2} \sum_{i=1}^{n_x} (x_i - \mu_x)^2 + \frac{1}{\sigma_y^2} \sum_{i=1}^{n_y} (y_i - \mu_y)^2 + \frac{1}{\sigma_z^2} \sum_{i=1}^{n_z} (z_i - \mu_z)^2 \right]. \quad (3.5)$$

Data la verosimiglianza per il parametro  $\theta$  e data una distribuzione a priori  $\pi(\theta)$ , dal Teorema di Bayes si ottiene la distribuzione a posteriori per  $\theta$ , data da

$$\pi(\theta | x, y, z) \propto \pi(\theta) L(\theta). \quad (3.6)$$

A partire da tale distribuzione si desidera ottenere la distribuzione a posteriori marginale per il VUS, ossia per  $P(X < Y < Z)$ . Sotto le assunzioni di normalità, il VUS può essere espresso come una funzione del parametro  $\theta$  (Xiong *et al.*, 2006). In particolare, si ha

$$\text{VUS} = \int_{-\infty}^{\infty} \Phi\left(\frac{\sigma_y^2}{\sigma_x^2}u - \frac{\mu_x - \mu_y}{\sigma_x^2}\right) \Phi\left(-\frac{\sigma_y^2}{\sigma_z^2}u + \frac{\mu_z - \mu_y}{\sigma_z^2}\right) \phi(u)du. \quad (3.7)$$

Una stima frequentista del VUS può essere ottenuta sostituendo gli ignoti parametri nella (3.7) con delle opportune stime. Ad esempio, se si utilizzassero le rispettive stime di massima verosimiglianza, si otterrebbe la SMV del VUS (Kang *et al.*, 2013). Nell'approccio bayesiano è di interesse ottenere la distribuzione marginale a posteriori per il VUS. Nel prossimo paragrafo si illustrano i metodi bayesiani per l'approssimazione delle distribuzioni marginali a posteriori.

## 3.2 Calcolo di distribuzioni marginali a posteriori

Le procedure bayesiane comportano un uso notevole di calcolo integrale per determinare distribuzioni a posteriori, marginali o predittive, e queste sono per la maggior parte delle volte irrisolvibili a livello analitico (Liseo, 2008).

Negli anni si sono sviluppate diverse tecniche per il calcolo della distribuzione a posteriori.

### METODI ASINTOTICI

1. Un metodo di calcolo basato su sviluppi asintotici è il **Metodo di Laplace** (Liseo, 2008). Permette di approssimare le distribuzioni a posteriori o a posteriori marginali, o momenti a posteriori. Dal punto di vista matematico gli integrali da calcolare sono del tipo (Liseo, 2008)

$$I_n = \int_{\Omega} g(\theta) \pi(\theta | x) d\theta,$$

per un'opportuna funzione  $g(\theta)$ . Il metodo di Laplace fornisce delle approssimazioni di secondo ordine per  $I_n$  (Liseo, 2008). Inoltre, considerando il rapporto tra due integrali, si dimostra che si ottengono delle approssimazioni del terzo ordine (Tierney e Kadane, 1986).

2. Un'alternativa al metodo di Laplace è fornita dalla tecnica del **punto di sella** (Liseo, 2008). Essa risulta particolarmente efficace per approssimare funzioni espresse mediante un integrale del tipo

$$g(\psi) = \int f(\theta, \psi) d\psi.$$

Questa tecnica ha alla base il metodo di Laplace, con la differenza che lo sviluppo in serie di Taylor avviene intorno a un valore che dipende da  $\psi$  (Liseo, 2008).

3. Un altro metodo asintotico per ottenere un'approssimazione della distribuzione marginale a posteriori, e le quantità ad essa connesse, per un parametro scalare è il metodo **HOTA** (*Higher-Order Tail Area approximation*) (Ruli *et al.*, 2014). Questo è un metodo basato sulla combinazione del *metodo di inversione* e di sviluppi asintotici di ordine superiore. Rispetto ai metodi MCMC, discussi nel paragrafo successivo, il metodo HOTA ha il vantaggio di restituire campioni indipendenti ad un costo computazionale trascurabile.

Un limite di questi metodi è che richiedono opportune condizioni di regolarità sulle funzioni da approssimare. In alternativa si può ricorrere a metodi computazionali.

### 3.2.1 Metodi MCMC

I metodi MCMC (Albert, 2011) sono degli strumenti di calcolo numerico che si basano sulle catene markoviane. In particolare, si basano su catene di Markov che garantiscono la *condizione di ergodicità*, ossia una catena che soddisfa contemporaneamente le condizioni di *irriucibilità*, *aperiodicità* e *ricorrenza positiva*. L'importanza della condizione di ergodicità è essenziale ai fini della validità dei diversi algoritmi che si possono utilizzare in questo contesto (Sartori, 2012).

Un esempio di catena di Markov (con passeggiata casuale) è data da

$$X^{(t+1)} = X^{(t)} + \epsilon_t,$$

dove  $\epsilon_t$  è indipendente da  $X^{(t)}$ . Se  $\epsilon_t \sim N(0, 1)$ , allora la densità di transizione Kernel  $K(x^{(t+1)}, x^{(t)})$  corrisponde ad una normale  $N(x^{(t)}, 1)$ . Inoltre, si dimostra che la condizione di ergodicità è verificata per  $K(x, \cdot) > 0$ .



In generale, la simulazione di una catena markoviana consiste nel simulare un valore  $X^{j+1}$  dato il valore al tempo precedente  $X^j$ , secondo una *matrice di transizione*  $P$  o la *densità di transizione Kernel*. Se una catena ha distribuzione limite  $\pi$ , allora l'algoritmo di simulazione restituisce una serie di valori avente distribuzione pari alla distribuzione limite  $\pi$ . Si dimostra che, per  $n$  sufficientemente grande, la distribuzione di  $X^n$  è approssimativamente pari a  $\pi$ ; quindi una serie del tipo  $(X^{n+1}, X^{n+2}, \dots)$  ha distribuzione limite  $\pi$ , la quale gode della proprietà di invarianza.

I metodi MCMC costituiscono un metodo generico di simulazione da una distribuzione  $\pi$ , dalla quale sarebbe difficile simulare mediante altri metodi. Alla base degli algoritmi MCMC c'è la necessità di costruire una catena markoviana dalla quale sia facile la simulazione e che abbia una certa distribuzione limite.

L'algoritmo Metropolis-Hastings (Sartori, 2012) costituisce un metodo generale per la simulazione da una catena markoviana. Esso rappresenta il metodo alla base delle analisi di calcolo numerico utilizzate in questo capitolo per approssimare la distribuzione marginale del VUS.

L'algoritmo Metropolis-Hastings consiste nel simulare un valore proposto da una matrice di transizione (o densità di transizione) che permetta la simulazione. Il valore proposto viene accettato o rifiutato con una certa probabilità in modo che la catena abbia sempre una distribuzione limite  $\pi$ . Inoltre, tale algoritmo presenta delle varianti a seconda del metodo utilizzato per valutare il valore proposto. Per tale scopo si può scegliere di utilizzare un algoritmo Metropolis-Hastings *random walk*, *indipendece sampler* e *Gibbs sampler*.

### 3.3 MCMC applicato al VUS

Al fine di ottenere un'approssimazione della distribuzione marginale a posteriori per il VUS, si utilizzano dei metodi MCMC, ed in particolare l'algoritmo *Metropolis-Hastings adattivo* con passeggiata casuale (Scheidegger, 2012), utilizzando il *Software-R* (<http://www.r-project.org>). L'implementazione di questa variante adattiva consiste in un algoritmo Metropolis-Hastings in cui viene fissato il tasso di accettazione desiderato.

Al fine di illustrare l'inferenza bayesiana sul VUS si considerano sia degli studi di simulazione, sia un'applicazione su dati reali (Cassoni *et al.*, 2015).

In tutti gli scenari considerati si considerano tre variabili normali indipendenti, come nella (3.1), la cui log-verosimiglianza per  $\theta$  è data nella (3.5). La distribuzione a posteriori per  $\theta$  è data nella (3.6), mentre per la distribuzione iniziale  $\pi(\theta)$  si utilizza la distribuzione a priori di Jeffreys  $\pi^J(\theta) \propto (\sigma_x \sigma_y \sigma_z)^{-1}$ .

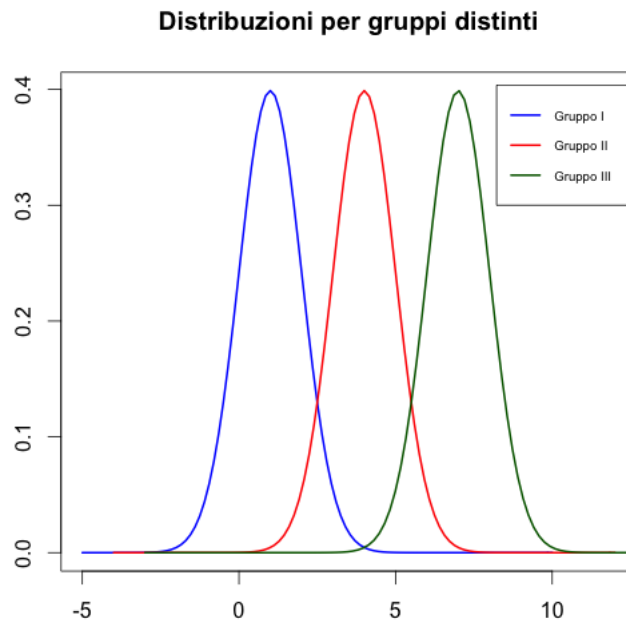
#### 3.3.1 Studi di simulazione

Gli scenari che si considerano consistono nel fissare diversi parametri per le tre variabili normali. In particolare, si considera il caso in cui i tre gruppi sono abbastanza distinti tra loro ( $VUS = 0.966$ ), il caso in cui sono in parte sovrapposti ( $VUS = 0.536$ ), ed infine il caso in cui sono quasi del tutto sovrapposti ( $VUS = 0.245$ ).

L'approccio bayesiano alla stima del VUS viene confrontato sia con il metodo frequentista parametrico sia con uno frequentista non parametrico, ottenuti utilizzando due funzioni del pacchetto R **DiagTest3Grp** (Luo e Xiong, 2012).

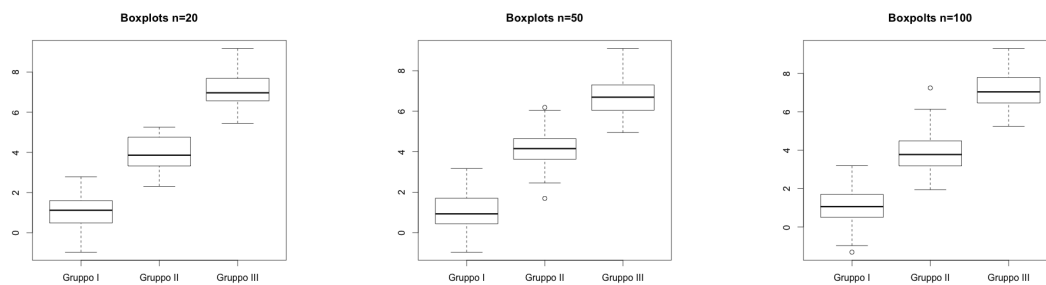
#### Simulazione 1 ( $VUS = 0.966$ )

Si consideri la simulazione da tre variabili normali:  $X \sim N(1, 1)$ ;  $Y \sim N(4, 1)$ ;  $Z \sim N(7, 1)$  ed il relativo valore reale del  $VUS=0.966$ . Inoltre, si considerano tre diversi scenari per la numerosità campionaria. In particolare, si considera il caso in cui  $n = 20, 50, 100$  (con  $n = n_x = n_y = n_z$ ). La Figura 3.1 mostra le distribuzioni di  $X$ ,  $Y$  e  $Z$ .



**Figura 3.1:** Distribuzioni di  $X$ ,  $Y$  e  $Z$  -  $VUS = 0.996$ .

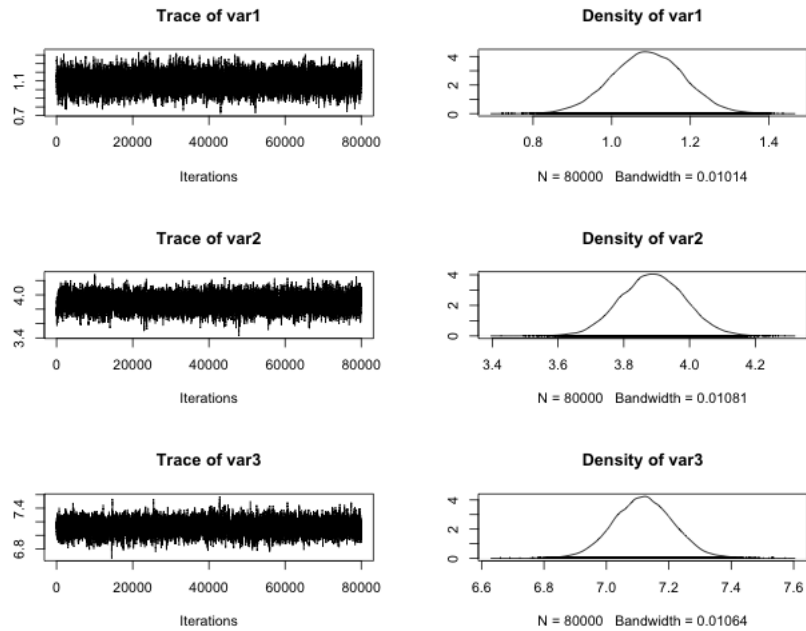
I boxplots nella Figura 3.2 mostrano le distribuzioni dei tre gruppi di pazienti, nei tre diversi scenari della numerosità campionaria.



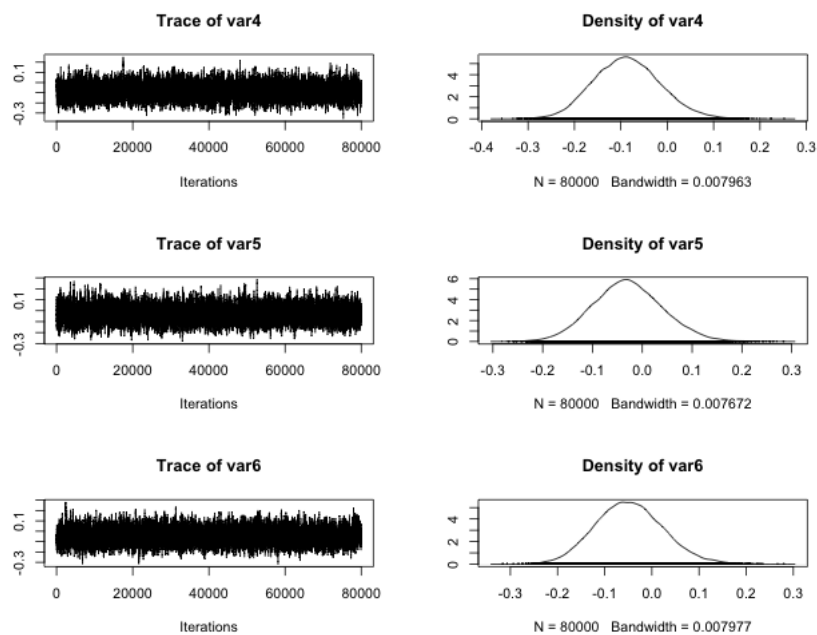
**Figura 3.2:** Boxplots per i tre scenari  $n=20,50,100$

Si considera l'utilizzo dei metodi MCMC al fine di ottenere le distribuzioni marginali a posteriori delle componenti del parametro  $\theta$ , e quindi di ricavare la distribuzione marginale a posteriori per il VUS. Si implementa a tale scopo l'algoritmo Metropolis-Hastings nella sua variante adattiva, ottenendo le distribuzioni marginali a posteriori per i singoli parametri. La Figura 3.3 mostra le distribuzioni approssimate marginali a posteriori delle medie, mentre la Figura 3.4 mostra le distribuzioni approssimate marginali a posteriori delle varianze. Inoltre, in Figura 3.5 si mostrano

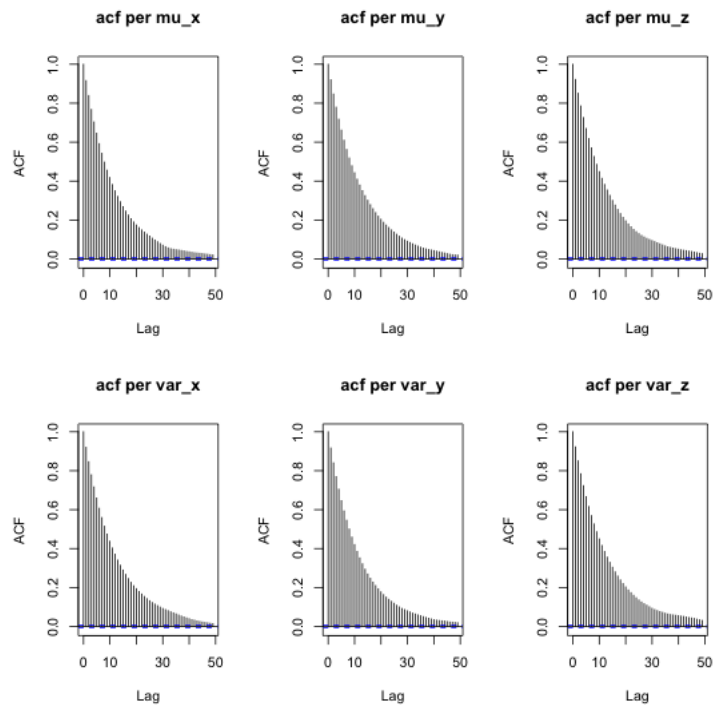
i correlogrammi per valutare il grado di autocorrelazione all'interno della serie.



**Figura 3.3:** Catena di Markov e distribuzioni a posteriori marginali approssimate delle medie per  $n=100$ .



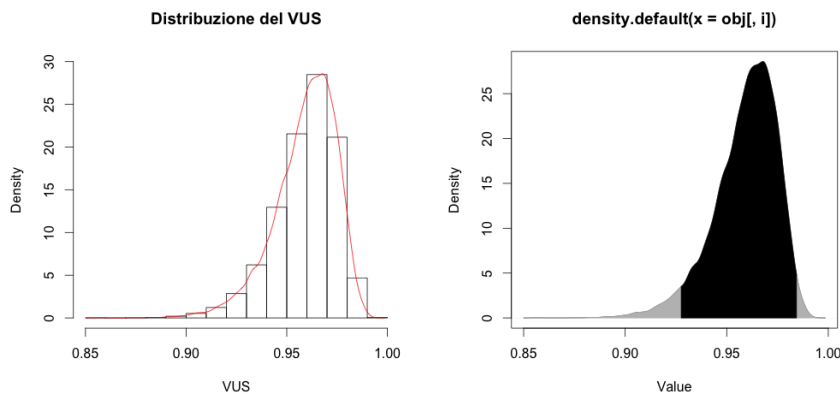
**Figura 3.4:** Catena di Markov e distribuzioni a posteriori marginali approssimate delle varianze per  $n=100$ .



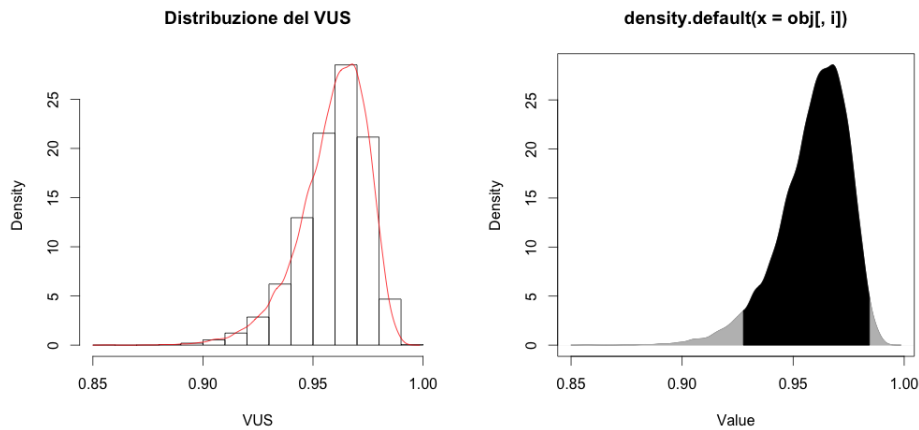
**Figura 3.5:** Autocorrelazioni nelle catene di Markov per  $n=100$ .

I risultati ottenuti per valutare l'autocorrelazione nella serie risultano molto simili tra loro nei tre scenari della numerosità campionaria. In Figura 3.5 si mostrano solo quelli per  $n=100$ .

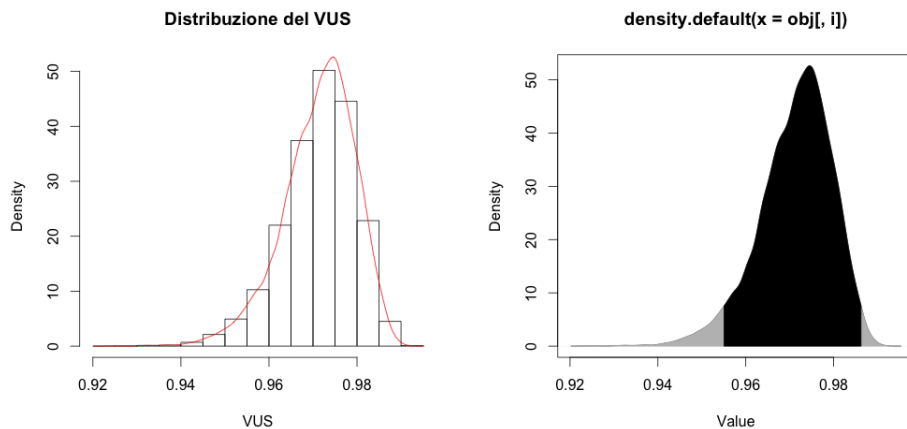
Il grafico dell'approssimazione della distribuzione marginale a posteriori per il VUS è data nella Figura 3.6 (a sinistra). A destra è data una rappresentazione grafica dell'intervallo di credibilità HPD.



**Figura 3.6:** Approssimazione della distribuzione marginale a posteriori per il VUS (a sinistra). Intervallo HPD (95%) per il VUS (a destra) ( $n=20$ ).



**Figura 3.7:** Approssimazione della distribuzione marginale a posteriori per il VUS (a sinistra). Intervallo HPD (95%) per il VUS (a destra) ( $n=50$ ).



**Figura 3.8:** Approssimazione della distribuzione marginale a posteriori per il VUS (a sinistra). Intervallo HPD (95%) per il VUS (a destra) ( $n=100$ ).

Dalla Tabella 3.1 si può notare come le stime per il VUS sono molto simili tra loro, dove le stime del VUS ottenute con l'approccio bayesiano sono molto vicine al valore reale. Inoltre, l'accuratezza delle stime bayesiane è paragonabile all'accuratezza delle stime dei due metodi frequentisti, in tutti gli scenari di questa prima simulazione.

**Tabella 3.1:** SIMULAZIONE 1: Media a posteriori del VUS ottenuta mediante i metodi MCMC, confrontata con la SMV del VUS e la stima non parametrica. Viene anche riportata la distorsione (BIAS) delle stime.

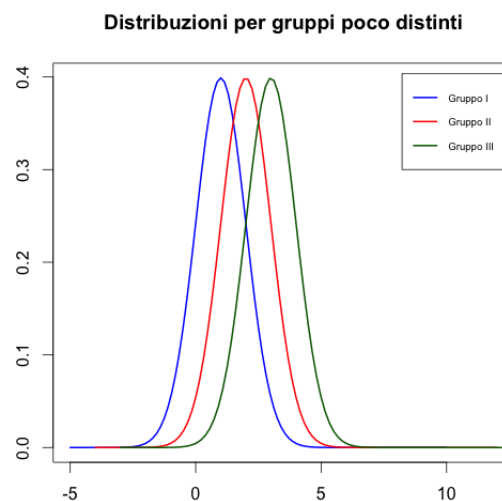
$$X \sim N(1, 1); Y \sim N(4, 1); Z \sim N(7, 1) - VUS = 0.966$$

Stime	Numerosità campionaria		
	(20,20,20)	(50,50,50)	(100,100,100)
VUS MCMC	0.966	0.959	0.971
IC HPD (95%)	(0.920 0.996)	(0.927 0.985)	(0.955 0.986)
BIAS	-0.0001	-0.007	0.005
VUS par	0.980	0.966	0.974
IC Wald (95%)	(0.953 1)	(0.940 0.991)	(0.959 0.988)
BIAS	0.014	-0.0001	0.008
VUS non par	0.993	0.971	0.975
IC (95%)	(0.961 1)	(0.942 0.989)	(0.962 0.989)
BIAS	0.027	0.005	0.009

### Simulazione 2 ( $VUS = 0.536$ )

In questa seconda simulazione si considerano le tre variabili normali:  $X \sim N(1, 1)$ ;  $Y \sim N(2, 1)$ ;  $Z \sim N(3, 1)$  ed il relativo valore reale del  $VUS=0.536$ . Anche in questo caso si considerano i tre diversi scenari per la numerosità campionaria, con  $n = 20, 50, 100$  (per  $n = n_x = n_y = n_z$ ). La Figura 3.9 mostra le distribuzioni di  $X$ ,  $Y$  e  $Z$ .

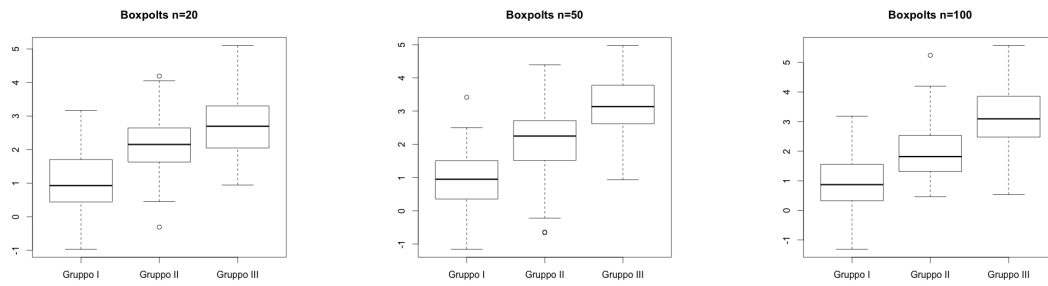
I boxplots nella Figura 3.10 mostrano le distribuzioni di tre campioni tratti dai



**Figura 3.9:** Distribuzioni di  $X$ ,  $Y$  e  $Z$  -  $VUS = 0.536$ .

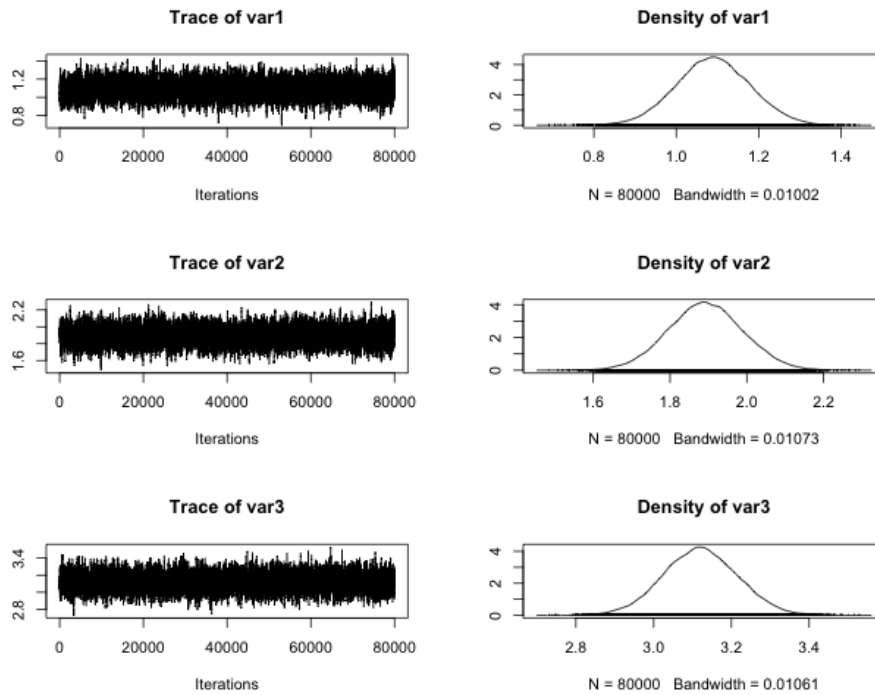
gruppi di pazienti, per i tre diversi scenari sulla numerosità campionaria.

Si implementa l'algoritmo Metropolis-Hastings, sempre nella sua variante adattiva.



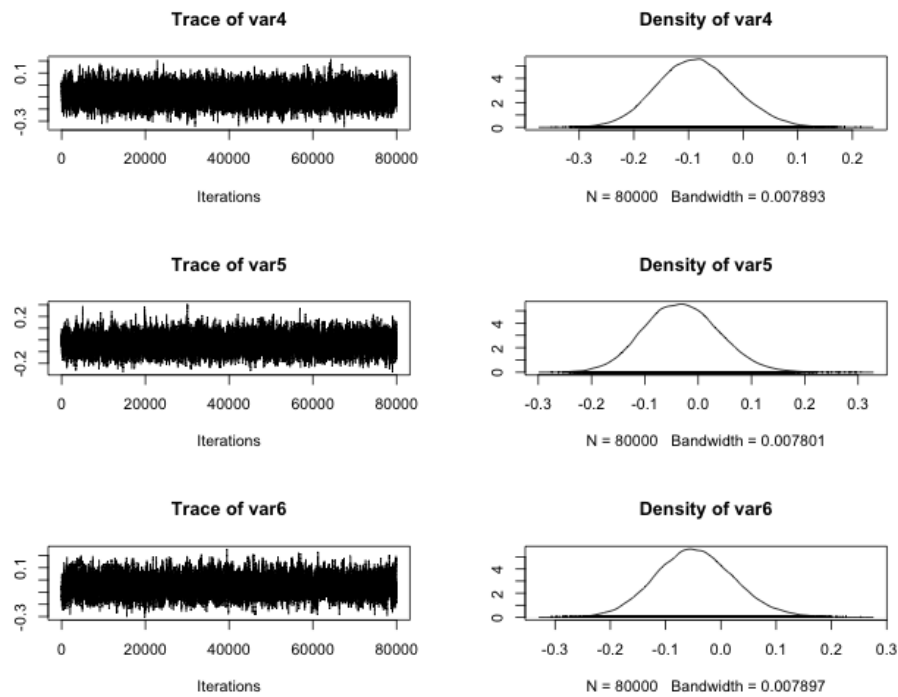
**Figura 3.10:** Boxplots per i tre scenari  $n=20,50,100$ .

La Figura 3.11 mostra le approssimazioni delle distribuzioni marginali a posteriori delle medie, mentre la Figura 3.12 mostra le approssimazioni delle distribuzioni marginali a posteriori delle varianze. Inoltre, in Figura 3.12 vengono mostrati i rispettivi correlogrammi.

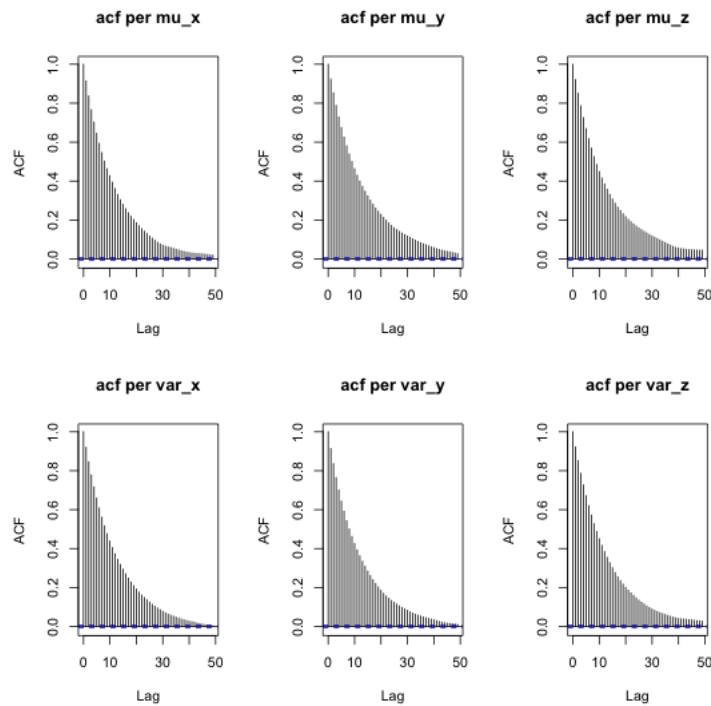


**Figura 3.11:** Catena di Markov e distribuzioni a posteriori marginali approssimate delle medie per  $n=100$ .





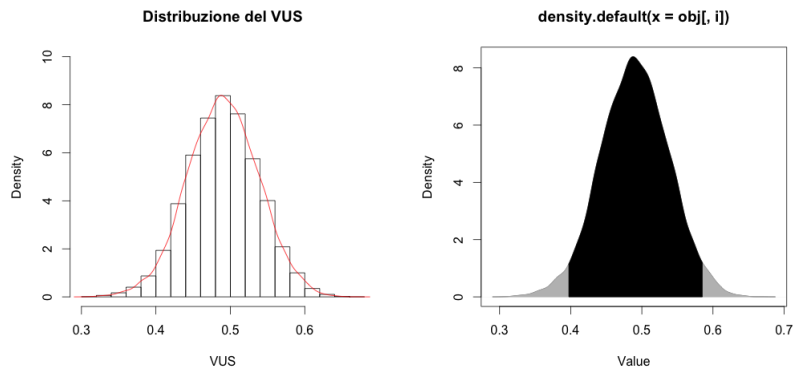
**Figura 3.12:** Catena di Markov e distribuzioni a posteriori marginali approssimate delle varianze per  $n=100$ .



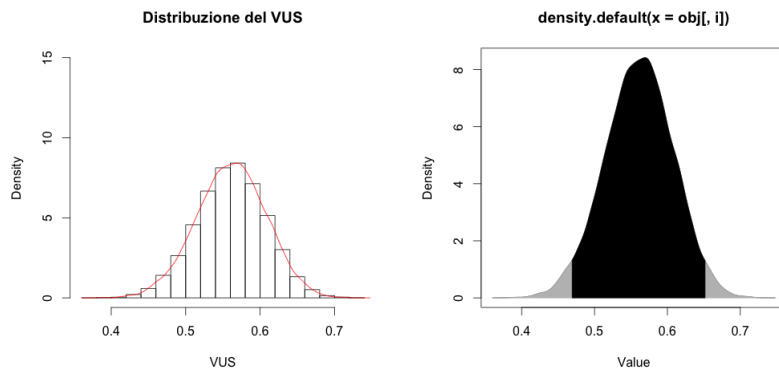
**Figura 3.13:** Autocorrelazioni nelle catene di Markov per  $n=100$ .

Il grafico dell'approssimazione della distribuzione marginale a posteriori del VUS

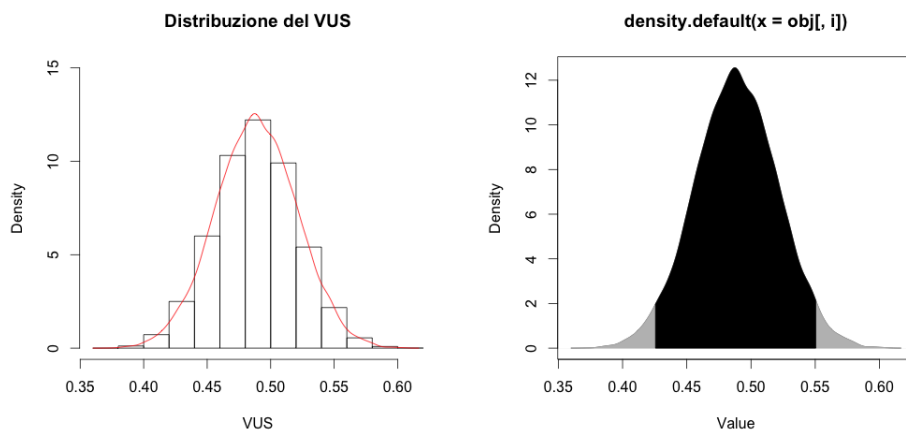
è data nelle Figura 3.14, 3.15 e 3.16, per le diverse numerosità campionarie.



**Figura 3.14:** Approssimazione della distribuzione marginale a posteriori per il VUS (a sinistra). Intervallo HPD (95%) per il VUS (a destra) ( $n=20$ ).



**Figura 3.15:** Approssimazione della distribuzione marginale a posteriori per il VUS (a sinistra). Intervallo HPD (95%) per il VUS (a destra) ( $n=50$ ).



**Figura 3.16:** Approssimazione della distribuzione marginale a posteriori per il VUS (a sinistra). Intervallo HPD (95%) per il VUS (a destra) ( $n=100$ ).

La Tabella 3.2 mostra le stime del VUS. I metodi di stima sono gli stessi dello scenario nella Simulazione 1. Si può notare come le stime del VUS sono meno accurate rispetto a quelle della simulazione precedente. La stima bayesina del VUS risulta più accurata nel caso  $n=100$ .

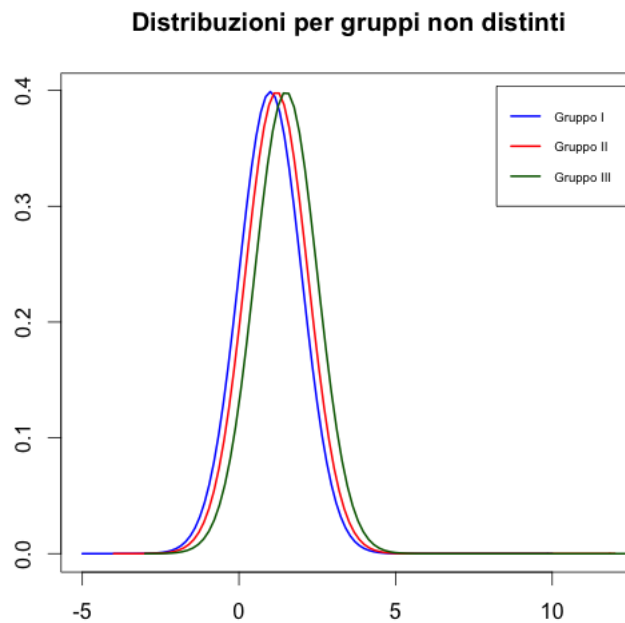
**Tabella 3.2:** SIMULAZIONE 2: Media a posteriori del VUS ottenuto mediante i metodi MCMC, SMV del VUS e stima non parametrica per il VUS. Viene anche riportata la distorsione (BIAS) delle stime.

$$X \sim N(1, 1); Y \sim N(2, 1); Z \sim N(3, 1) - VUS = 0.536$$

Stime	Numerosità campionaria		
	(20,20,20)	(50,50,50)	(100,100,100)
VUS MCMC	0.490	0.561	0.489
IC HPD (95%)	(0.397 0.585)	(0.397 0.585)	(0.425 0.551)
BIAS	-0.046	0.025	0.047
VUS parametrico	0.498	0.570	0.499
IC Wald (95%)	(0.406 0.591)	(0.479 0.662)	(0.491 0.621)
BIAS	-0.038	0.034	-0.037
VUS non parametrico	0.493	0.571	0.558
IC (95%)	(0.396 0.602)	(0.484 0.675)	(0.501 0.624)
BIAS	-0.043	0.035	0.022

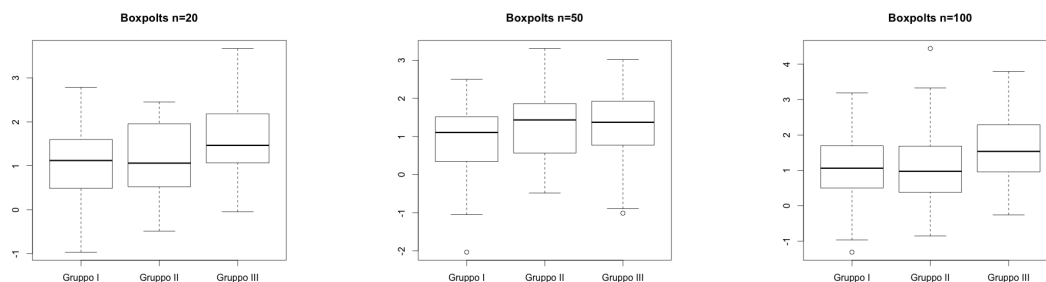
**Simulazione 3 - ( $VUS = 0.2451$ )**

In questa terzo ed ultimo esempio di simulazione si considerano tre variabili normali:  $X \sim N(1, 1)$ ;  $Y \sim N(1.5, 1)$ ;  $Z \sim N(2, 1)$  ed un relativo valore reale del  $VUS=0.245$ . Anche in questo caso si considerano i tre diversi scenari per la numerosità campionaria, dove  $n = 20, 50, 100$  (con  $n = n_x = n_y = n_z$ ). La Figura 3.17 mostra le distribuzioni di  $X$ ,  $Y$  e  $Z$ .



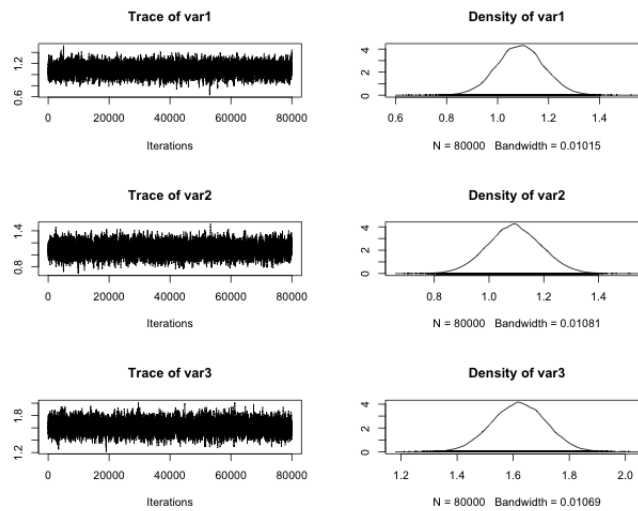
**Figura 3.17:** Distribuzioni di  $X$ ,  $Y$  e  $Z$  . $VUS=0.245$ .

La Figura 3.18 mostra i boxplots dei tre campioni indipendenti tratti dalle tre distribuzioni per i tre diversi scenari sulla numerosità campionaria.

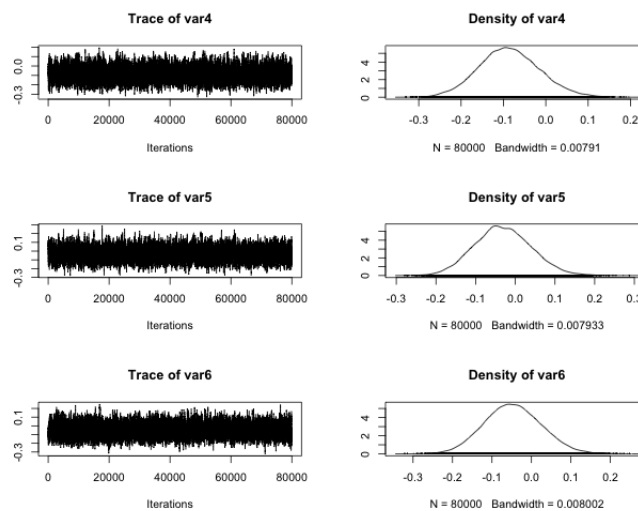


**Figura 3.18:** Boxplots per i tre scenari  $n=20, 50, 100$ .

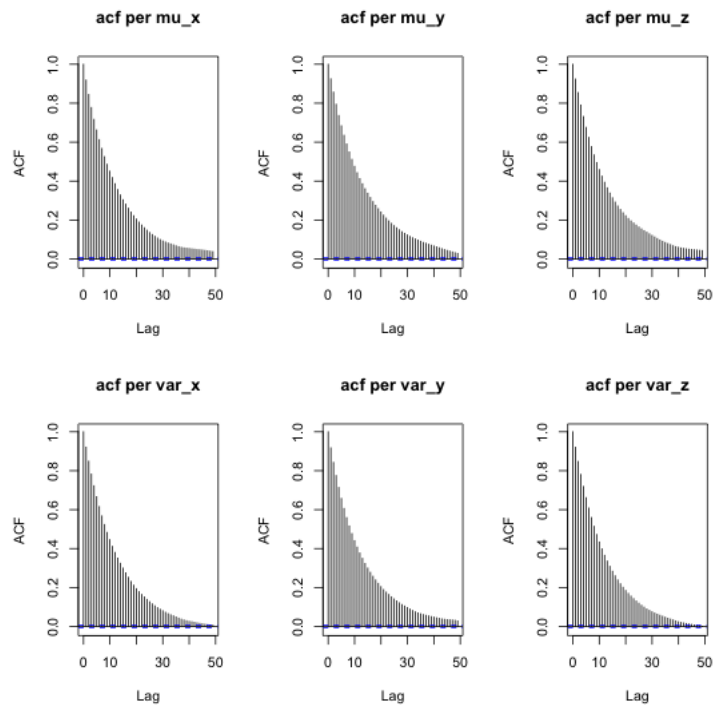
Si implementa l'algoritmo Metropolis-Hastings, sempre nella sua variante adattiva. La Figura 3.19 mostra le approssimazioni delle distribuzioni marginali a posteriori delle medie, mentre la Figura 3.20 mostra le approssimazioni delle distribuzioni marginali a posteriori delle varianze. In Figura 3.21 vengono mostrati i rispettivi correlogrammi.



**Figura 3.19:** Catena di Markov e distribuzioni a posteriori marginali approssimate delle medie per  $n=100$ .

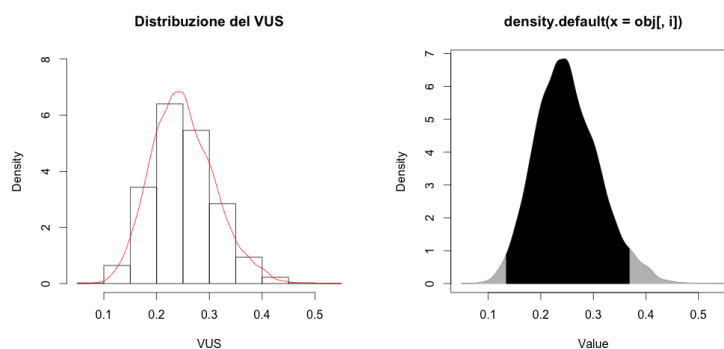


**Figura 3.20:** Catena di Markov e distribuzioni a posteriori marginali approssimate delle varianze per  $n=100$ .

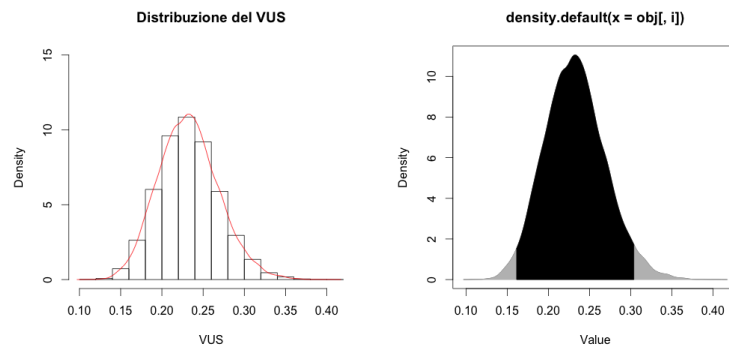


**Figura 3.21:** Autocorrelazioni nelle catene di Markov per  $n=100$ .

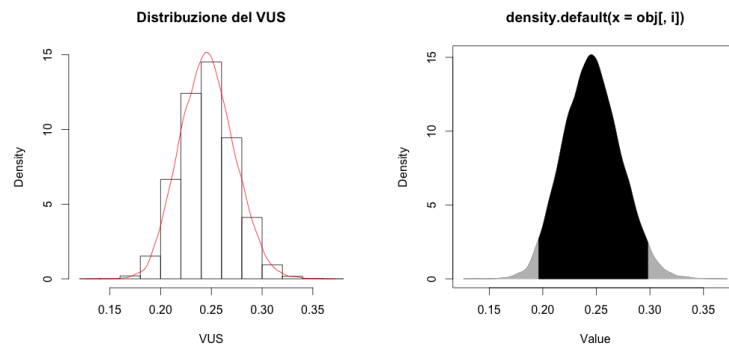
Nelle Figure 3.22, 3.23 e 3.24 si mostra la distribuzione a posteriori marginale approssimata del VUS e una rappresentazione grafica dell'intervallo di credibilità HPD nei tre scenari della numerosità campionaria.



**Figura 3.22:** Approssimazione della distribuzione marginale a posteriori per il VUS (a sinistra). Intervallo HPD (95%) per il VUS (a destra) ( $n=20$ ).



**Figura 3.23:** Approssimazione della distribuzione marginale a posteriori per il VUS (a sinistra). Intervallo HPD (95%) per il VUS (a destra) ( $n=50$ ).



**Figura 3.24:** Approssimazione della distribuzione marginale a posteriori per il VUS (a sinistra). Intervallo HPD (95%) per il VUS (a destra) ( $n=100$ ).

La Tabella 3.3 mostra i risultati per le stime del VUS. I metodi di stima sono gli stessi presentati nella Simulazione 1 e nella Simulazione 2. Dai risultati mostrati nella Tabella 3.3 si nota che, le stime frequentiste parametriche e le stime bayesiane parametriche del VUS sono molto simili. Inoltre, la stima bayesiana ha un'accuratezza solitamente migliore rispetto alle stime ottenute con i due metodi frequentisti.

**Tabella 3.3:** SIMULAZIONE 3: Media a posteriori del VUS ottenuto mediante i metodi MCMC, SMV del VUS e stima non parametrica per il VUS.

$$X \sim N(1, 1); Y \sim N(1.2, 1); Z \sim N(1.5, 1) - VUS = 0.2451$$

Stime	Numerosità campionaria		
	(20,20,20)	(50,50,50)	(100,100,100)
VUS MCMC	0.250	0.232	0.246
IC HPD (95%)	(0.134 0.369)	(0.161 0.304)	(0.138 0.270)
BIAS	0.0058	-0.0131	0.0008
VUS parametrico	0.257	0.234	0.246
IC Wald (95%)	(0.195 0.298)	(0.162 0.305)	(0.195 0.298)
BIAS	0.0118	-0.0111	0.0118
VUS non parametrico	0.235	0.235	0.241
IC (95%)	(0.130 0.370)	(0.155 0.305)	(0.191 0.294)
BIAS	-0.0101	-0.0101	-0.0041

### 3.4 Applicazione del metodo MCMC a dati reali

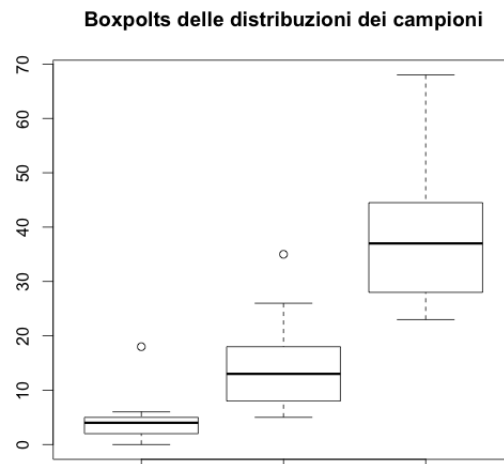
Si considera uno studio condotto su un dataset di 70 osservazioni di casi di meningioma. Lo studio aveva l'obiettivo di testare l'utilizzo dell'anticorpo **PHH3** al fine di migliorare la diagnosi del grado di differenziazione del tumore (Cassoni *et al.*, 2015). In particolare, si sono messi a confronto i risultati ottenuti con il PHH3 e quelli riferiti alle validate tecniche di laboratorio su vetro **H&E** (*Hematoxylin and Eosin*) che, a seconda della colorazione del vetrino, individuano un particolare grado del tumore (grado I, grado II, grado III). In accordo con l'*Organizzazione Mondiale della Sanità* (OMS), si è mostrato come gli indici di attività mitotica su PHH3 e H&E sono molto correlati tra loro ( $R = 0.92, p < 0.0001$ ). Inoltre, lo studio ha mostrato come PHH3, rispetto a H&E, abbia sia una migliore correlazione intraclasse e sia inter-classe.

Sempre in accordo con l'OMS, attraverso le tecniche basate sull'analisi non parametrica della superficie ROC si è stimato  $\widehat{VUS}=0.91$  ottenuto dall'utilizzo dell'anticorpo PHH3. L'ottima classificazione ottenuta mediante l'utilizzo del PHH3, ha mostrato come questo sia uno strumento diagnostico complementare al test istologico H&E.

Sulla base dello studio del VUS in un contesto non parametrico, si vuole ora condurre un'analisi sulla stima parametrica bayesiana per il VUS, prendendo in riferimento gli stessi dati dell'analisi.

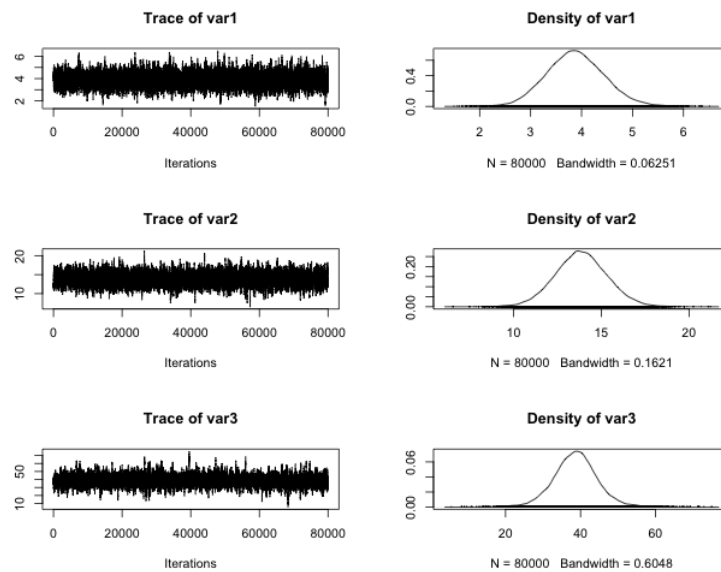
La Figura 3.25 mostra i boxplots dei tre campioni. Come nell'analisi di simulazione viste in precedenza, si considera un algoritmo Metropolis-Hastings con passeggiata casuale di tipo adattivo, nel quale il tasso di accettazione è fissato al 45%. Nella Figura 3.26 si mostra l'approssimazione delle distribuzioni marginali a posteriori delle medie, mentre nella Figura 3.27 si mostra lo stesso risultato per le varianze. Da quanto riscontrato dai grafici la catena di Markov sembra avere un andamento



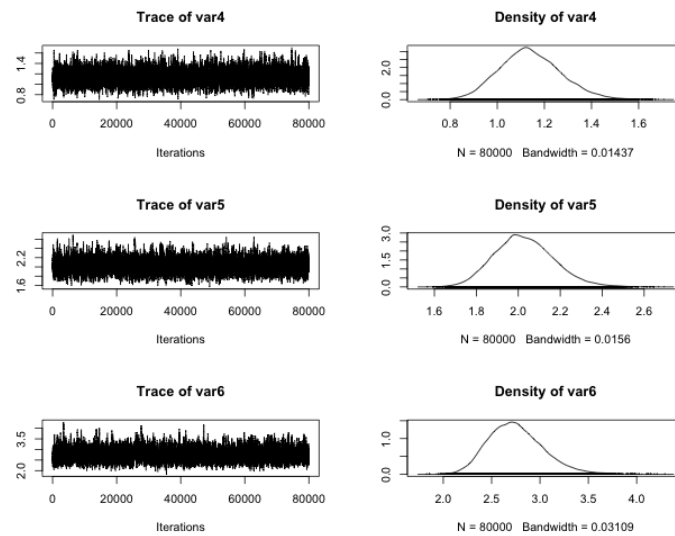


**Figura 3.25:** Boxplots dei campioni.

coerente con le assunzioni del metodo MCMC.

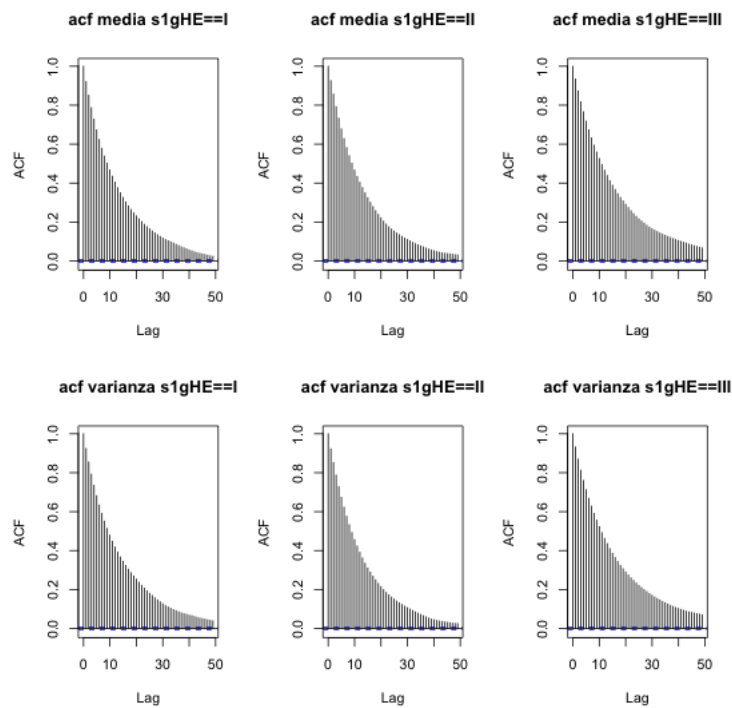


**Figura 3.26:** Catena di Markov e distribuzioni a posteriori marginali approssimate delle medie.



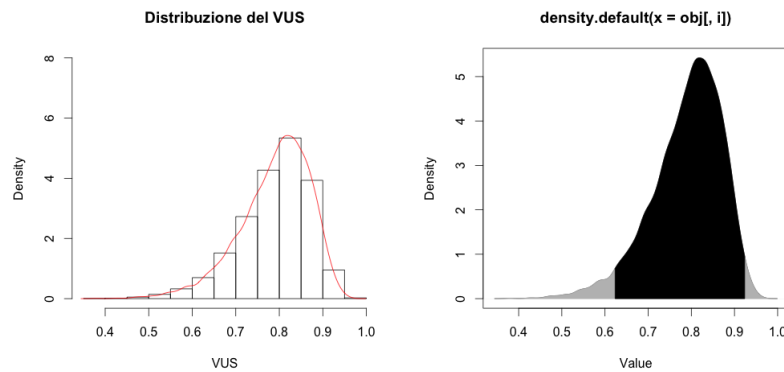
**Figura 3.27:** Catena di Markov e distribuzioni a posteriori marginali approssimate delle varianze.

La Figura 3.27 riporta i correlogrammi relativi ai singoli parametri.



**Figura 3.28:** Autocorrelazioni nelle catene di Markov.

La Figura 3.29 mostra la distribuzione a posteriori approssimata del VUS.



**Figura 3.29:** Approssimazione della distribuzione marginale a posteriori (a sinistra). A destra si mostra l'intervallo HPD (95%) per il VUS.

La Tabella 3.4 mostra i risultati ottenuti per l'approssimazione della distribuzione marginale del VUS.

**Tabella 3.4:** Implementazione su dati reali: Stima della media a posteriori del VUS ottenuto mediante i metodi MCMC, confrontato con il VUS ottenuto mediante SMV e un metodo non parametrico.

Stime		BIAS
VUS MCMC	0.790	-0.21
IC HPD (95%)	(0.623 0.924)	
VUS parametrico	0.830	-0.17
IC Wald (95%)	(0.699 0.961)	
VUS non parametrico	0.922	-0.078
IC (95%)	(0.847 0.977)	

Dalla Tabella 3.4 si nota come, nell'applicazione su dati reali, le stime per il VUS non sono del tutto concordi tra loro. In particolare, si evince che l'approccio frequentista e quello bayesiano oggettivo danno risultati simili, mentre si evidenziano maggiori differenze con la stima del VUS ottenuta con il metodo non parametrico.

## 3.5 Conclusioni

Nelle analisi di simulazione condotte in questo capitolo si riscontra che, questa prima applicazione parametrica bayesiana oggettiva ha fornito delle stime bayesiane accurate del VUS. Infatti, nei vari scenari illustrati si sono ottenute delle basse misure della distorsione in accordo sia con il metodo frequentista parametrico e sia con

quello frequentista non parametrico.

Possibili estensioni di questa applicazione consistono nell'utilizzo di altre distribuzioni *a priori*, ad esempio una distribuzione *matching prior*. Altri sviluppi del metodo proposto possono riguardare l'utilizzo di versioni modificate della verosimiglianza. Inoltre, è opportuno considerare anche l'utilizzo dell'algoritmo Gibbs Sampler per ottenere l'approssimazione della distribuzione a posteriori marginale.

# CONCLUSIONI

La metodologia della curva ROC ha avuto negli anni una continua evoluzione. I diversi contesti nei quali si pone tale metodo ha spinto molti autori a sviluppare diverse applicazioni nei vari contesti della statistica inferenziale. L'analisi della curva ROC ha avuto una indubbia rilevanza soprattutto nel campo della medicina, in un contesto di classificazione su due gruppi. Le crescenti esigenze che si sono presentate in questo particolare contesto scientifico hanno portato allo sviluppo dell'analisi della superficie ROC, la quale è vista come una generalizzazione della curva ROC. L'approccio inferenziale di cui si è discusso nell'ultimo capitolo risulta il primo contributo all'inferenza bayesiana parametrica oggettiva sul VUS. Una naturale estensione dello strumento proposto è considerare l'utilizzo di versioni modificate della verosimiglianza. Inoltre, al fine di migliorare tale strumento, si può considerare un utilizzo congiunto dei metodi MCMC e del metodo HOTA (Erlis et al., 2014).



# Appendice

Per le analisi di simulazione e per l'applicazioni sul dataset reale si è utilizzato il Software R (<http://www.r-project.org>). Di seguito viene riportato il codice principale utilizzato.

## I pacchetti R utilizzati.

```
library(LaplacesDemon)
library(MCMCpack)
library(adaptMCMC)
library(coda)
```

## Dataset

```
N <- 100
mu <- c(1,2,3)
sigma <- c(1,1,1)
set.seed(123)
dati <- data.frame(X=rnorm(N,mu[1],sigma[1]),
                  Y=rnorm(N,mu[2],sigma[2]),Z=rnorm(N,mu[3],sigma[3]))
```

## Verosimiglianza congiunta e stime

```
llog.theta <- function(theta,data)
{
  lx = sum(dnorm(dati$X,mean=theta[1],sd=exp(theta[4]),log=T))
  ly = sum(dnorm(dati$Y,mean=theta[2],sd=exp(theta[5]),log=T))
  lz = sum(dnorm(dati$Z,mean=theta[3],sd=exp(theta[6]),log=T))
  return(-lx-ly-lz)
}
mle <- nlminb(c(mu,log(sigma)),function(x) llog.theta(x,data=dati))
Jhat <- hessian(llog.theta,mle$par,data=dati1)
```

**Algoritmo Metropolis-Hastings adattivo**

```

start=c(1,1,1,0,0,0)
mcmcadaptc <- MCMC(p=function(x) -llog.theta(x, data=dati),
                  scale=solve(Jhat), n=10^5, init=start,adapt=T,
                  acc.rate=0.45, list=F)

mymcmc1 <- mcmcadaptc[-c(1:20000),]
mymcmc2 <- Thin(mymcmc1,5)

plot(mcmc(mymcmc1))

acf(mymcmc1[, ],main="acf per ")

```

**VUS**

```

vus <- function(theta.mcmc) {
  a = exp(theta.mcmc[,5]-theta.mcmc[,4])
  b = (theta.mcmc[,1]-theta.mcmc[,2])/exp(theta.mcmc[,4])
  c = exp(theta.mcmc[,5]-theta.mcmc[,6])
  d = (theta.mcmc[,3]-theta.mcmc[,2])/exp(theta.mcmc[,6])
  nsim = dim(theta.mcmc)[1]
  ans = c()
  Exval <- function(s, a, b, c, d)
    pnorm(a*s-b)*pnorm(-c*s +d)*dnorm(s)
  Exvalv <- Vectorize(Exval, vectorize.args = "s")
  for(i in 1:nsim){
    ans[i] = integrate(Exvalv, lower = -50, upper = 50,
                      a=a[i],b=b[i],c=c[i],d=d[i])$value

    if (i%%100==0) cat("\ rsimulation",i,"\ r")
  }
  return(ans)
}

```

**Media a posteriori del VUS e intervallo HPD**

```
VUS.mcmc <- vus(mymcmc2)
```



```
# media a posteriori  
mean(VUS.mcmc)
```

```
# IC HPD  
p.interval(VUS.mcmc,HPD=T,prob=0.95,plot=T)
```

### Approssimazione della distribuzione marginale del VUS

```
hist(VUS.mcmc,freq=F,xlab="VUS",main="Distribuzione del VUS"  
      ,ylim=c(0,15))
```

```
lines(density(VUS.mcmc), col = "red")
```

### Metodi frequentisti per il VUS

```
VUS.p <- VUS(dati$X,dati$Y,dati$Z, method =c("Norm"),  
             p = 0, q = 0, alpha = 0.05)
```

```
VUS.np <- VUS(dati$X,dati$Y,dati$Z, method =c("NonPar"),  
             p = 0, q = 0, alpha = 0.05)
```



# Bibliografia

- [1] Albert, J. (2009). *Bayesian Computation with R*. Springer, New York.
- [2] Azzalini, A., Scarpa, B. (2004). *Analisi dei Dati e Data Mining*. Springer, Milano.
- [3] Brazzale, A.R., Davison, A.C., Reid, N. (2007). *Applied Asymptotics. Case Studies in Small-Sample Statistics*. Cambridge University Press, Cambridge.
- [4] Cortese, G., Ventura, L. (2013). Accurate likelihood inference on the area under ROC curve for small samples. *Computational Statistics*, **28**, 1035-1059.
- [5] Cortese, G., Ventura, L. (2012). Accurate higher-order likelihood inference. *Computational Statistics*, **28**, 1035-1059.
- [6] Dreiseitl, S., Ohno-Machado, L., Binder, M. (2000). Comparing Three-class Diagnostic Test by Three-way ROC Analysis. *Medical Decision Making*, **20**, 323-331.
- [7] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861-874.
- [8] Fawcett, T. (2004). ROC Graphs: Notes and practical considerations for Researchers. *Kluwer Academic Publisher*.
- [9] Ferri, C., Hernández-Orallo, J., Salido, M.A. (2003). Volume Under the ROC Surface for Multi-class Problems. *Lecture Notes in Computer Science*, **2837**, 108-120.
- [10] Fluss, R., Faraggi, D., Reiser, B. (2005). Estimation of the Youden index and its associated cut-off point. *Biometrical Journal*, **47**, 458-72.
- [11] Gonçalves, L., Subtil, A., Oliveira, M.R., De Zea Bermudez, P. (2014). ROC curve estimation: An overview. *Statistical Journal*, **12**, 1-20.

- [12] He, X., Frey, E.C. (2008). The Meaning and Use of the Volume Under a Three-Class ROC Surface (VUS). *IEEE Transactions on medical imaging*, **27**, 577-588.
- [13] He, X., Metz, C.E., Tsui B.M.W., Links, J.M., Frey, E.C. (2006). Three-Class ROC Analysis - A Decision Theoretic Approach Under the Ideal Observer Framework. *IEEE Transactions on medical imaging*, **25**, 571-581.
- [14] Kang, L., Xiong, C., Tian L. (2013). Estimating confidence intervals for the difference in diagnostic accuracy with three ordinal diagnostic categories without a gold standard. *Computational Statistics Data Analysis*, **68**, 326-338.
- [15] Kang, L., Tian L. (2013). Estimation of the volume under the ROC surface with three ordinal diagnostic categories. *Computational Statistics Data Analysis*, **62**, 39 - 51.
- [16] Kotz, S., Lumelskii, Y., Pensky, M. (2003). *The Stress-Strength Model and its Generalizations. Theory and Applications*. Word Scientific, Singapore.
- [17] Liseo, B. (2008). *Introduzione alla statistica bayesiana*. Springer, Milano.
- [18] Luo, J., Xiong, C. (2013). Youden Index and Associated Cut-Points for Three Ordinal Diagnostic Groups. *Communications in Statistics - Simulation and Computation*, **42**, 1213-1234.
- [19] Mosmann, D. (1999). Three-Way ROCs. *Medical Decision Making*, **19**, 78-89.
- [20] Nakas, C.T., Yiannoutsos, C.T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Bayesian Analysis*, **23**, 3437-3449.
- [21] Nakas, C.T. (2014). Development in ROC surface analysis and assessment of diagnostic markers in three-class classification problems. *Statistical Journal*, **12**, 43-65.
- [22] Pace, L., Salvan, A. (2001). *Introduzione alla Statistica: Inferenza, verosimiglianza, modelli*. Cedam, Padova.
- [23] Ruli, E., Sartori, N., Ventura, L. (2014). Marginal posterior simulation via higher-order tail area approximations. *Bayesian Analysis*, **9**, 129-146.
- [24] Severini, T.A. (2000). *Likelihood Method in Statistics*. Oxford University Press.

- [25] Schsterman, E.F., Perkins, N.J., Liu, A., Bondel, H. (2005). Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology*, **16**, 73-81.
- [26] Ventura, L., Racugno, W. (2012). On interval and point estimators based on a penalization of the modified profile likelihood. *Statistics and Probability Letters*, **82**, 1285-1289.
- [27] Ventura, L., Racugno, W. (2011), Recent advances on Bayesian inference for  $P(X \min Y)$ , *Bayesian Analysis*, **6**, 411-428.
- [28] Ventura, L., Cabras, S., Racugno, W. (2009), Prior distributions from pseudo-likelihoods in the presence of nuisance parameters, *Journal of the American Statistical Association*, **104**, 68-774.
- [29] Ventura, L., Racugno, W. (2011). Recent advances on Bayesian inference for  $P(X \leq Y)$ . *Bayesian Analysis*, **6**, 411-428.