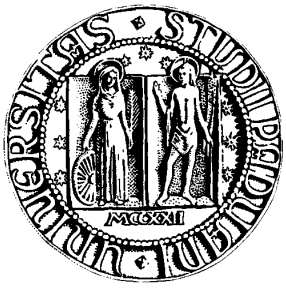
UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E TECNOLOGIE INFORMATICHE



TESI DI LAUREA

INFERENZA SULL'AREA SOTTO LA CURVA ROC BASATA SULLA VEROSIMIGLIANZA PROFILO MODIFICATA

Relatore: Ch.mo Prof. Laura Ventura

Laureando: Luigi Patuzzi

ANNO ACCADEMICO 2009-2010

Für die Von Patuzzis,
mit Liebe und Dankbarkeit.

INTRODUZIONE

Fin dall'antichità, la guerra è stata uno tra un più importanti motori per l'evoluzione di molti settori della scienza, quali per esempio la tecnologia, l'edilizia, la logistica o la medicina. Anche le scienze statistiche non sfuggono a questo fenomeno.
Nel corso della seconda guerra mondiale i sistemi radar consentivano di caratterizzare un oggetto secondo i parametri di latitudine, longitudine, altitudine e velocità, che tuttavia non erano sufficienti per classificarlo in termini di amico, nemico o rumore. L'identificazione seguiva una procedura come la seguente: innanzitutto si scartavano tutti gli oggetti definiti da parametri non compatibili (come velocità e quota per esempio di stormi di uccelli, palloni aerostatici, oggetti paracadutati...), successivamente si passava ad analizzare gli oggetti ammissibili.
Nel contesto aeronautico[1], l'operatore radar controllava se tra i piani di volo referenziati in suo possesso ce n'era uno compatibile coi parametri osservati. Se corrispondevano, allora veniva identificato e di conseguenza segnalato come amico. In caso di dubbio, la torre di controllo tentava di stabilire un contatto radio cifrato. In casi estremi era previsto l'uso di intercettori in quota per stabilire un contatto diretto. E' naturale che tanto più tempestivo era il riconoscimento dei velivoli, tanto più si poteva evitare l'intervento degli intercettori, con un risparmio in termini economici, oppure in caso di attacco effettivo si poteva ottenere un vantaggio strategico in termini di tempo. Così fu proprio durante il secondo conflitto mondiale che per la prima volta gli ingegneri dell'esercito statunitense, soprattutto dopo l'attacco a *Pearl Harbor*, adottarono un metodo grafico per valutare la bontà del criterio usato dagli operatori radar per stabilire se un oggetto era da considerare rumore dovuto a uccelli,

[1] Testimonianza diretta del dott. Mario Patuzzi, operatore radar presso la base NATO Primo R.O.C. sul Monte Venda negli anni ’60, periodo in cui si usavano tecniche e strumenti pressoché invariati rispetto a quelli del secondo conflitto mondiale.

fattori meteorologici o altro oppure un oggetto amico o nemico. Lo schema, che prese il nome di curva ROC (*Receiver Operating Characteristic*) si può applicare facilmente a molti casi di classificazione binaria e con qualche accorgimento anche a quelli con più di due classi (si veda, ad esempio, Azzalini e Scarpa, 2004, capitolo 5).
Già dagli anni '50 l'uso della curva ROC aveva superato i confini del mondo bellico ottenendo successo nel campo della psicofisica, e da lì quindi in medicina, radiologia, epidemiologia fino ad ambiti più recenti come il *data mining*.
L'uso estensivo della curva ROC in campi così diversi della scienza si può spiegare non solo con la sua genericità, che la rende adatta per molti scopi, ma anche alla sua relativa semplicità di costruzione. Considerando due gruppi di unità statistiche e un test per la loro classificazione, la curva ROC prevede infatti di rappresentare la sensibilità (vero positivo) in ordinata e il complemento a uno della specificità (vero negativo) in ascissa.
Una delle sintesi più utilizzate della curva ROC è l'area sottesa ad essa (AUC, *Area Under the ROC Curve*), che, come si vedrà in questa tesi, consente di stimare la probabilità di assegnare un'unità statistica al suo reale gruppo di appartenenza e quindi di valutare la bontà del criterio usato per la classificazione.
Un metodo parametrico classico per stimare l'AUC è avvalersi della teoria della verosimiglianza applicata al modello sollecitazione-resistenza (*stress-strength model*); si veda, per una trattazione completa di questo modello, Kotz et al. (2000). Dalla metà del secolo scorso, questo approccio si è diffuso per valutare l'affidabilità di un componente considerando un test fisico in cui una variabile X rappresenta la sollecitazione e una variabile Y la resistenza che il componente dimostra. Se la sollecitazione supera la resistenza, cioè se $X>Y$, il componente si rompe: l'affidabilità si può quindi esprimere come la probabilità che non si rompa, cioè $P(X<Y)$. Questo modello si può applicare anche al caso di una classificazione binaria e quindi al caso della curva ROC: conoscendo la distribuzione delle due variabili, si può pensare di ottenere una riparametrizzazione con la quale riuscire a esprimere la probabilità che $X>Y$ e quindi che la classificazione abbia un reale potere discriminatorio non casuale.

Lo scopo di questa tesi è discutere l'utilizzo della verosimiglianza profilo modificata (si veda, ad esempio, Pace e Salvan, 1996, capitolo 11), un miglioramento della verosimiglianza profilo attraverso l'introduzione di un fattore di modificazione, che permette un'inferenza più accurata, in particolare per dimensioni campionarie piccole o moderate.

Nei capitoli seguenti si forniranno dei richiami di base alla teoria della verosimiglianza profilo e profilo modificata (Capitolo 1), si mostreranno poi i passaggi per ottenere $P(X<Y)$ (Capitolo 2) e il metodo grafico ROC. Si procederà quindi con l'analisi di un caso reale (Capitolo 3), applicando tali strumenti, e di un caso simulato per valutarne l'efficacia.

CAPITOLO 1
Teoria della Verosimiglianza

Prima di entrare nel merito del problema di classificazione legato alla curva ROC e al modello sollecitazione-resistenza, si richiamano brevemente in questo capitolo alcuni concetti basilari della teoria della verosimiglianza, che torneranno utili per l'inferenza sulla quantità P(X<Y). I principali riferimenti bibliografici per gli argomenti richiamati in questo capitolo sono Pace e Salvan (2001) e Azzalini (2001).

1.1 VEROSIMIGLIANZA

Sia $y=(y_1, \ldots, y_n)$ un campione casuale semplice di numerosità n, realizzazione di una variabile casuale Y, con legge di probabilità F(y; θ), $\theta \in \Theta \subseteq IR^p$, $p>1$. Sia $p(y;\theta)$ la funzione di densità corrispondente. La funzione

$$L(\theta) = \prod_{i=1}^{n} p(y_i;\theta)$$

è detta **funzione di verosimiglianza** di θ basata sui dati y. Due funzioni di verosimiglianza che differiscono per una costante moltiplicativa si dicono equivalenti. Spesso le procedure di inferenza basate su $L(\theta)$ sono espresse tramite la **funzione di log-verosimiglianza**, una trasformazione monotona crescente della verosimiglianza, data da

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log p(y_i;\theta).$$

Due funzioni di log-verosimiglianza che differiscono per una costante additiva si dicono equivalenti.

Un valore $\hat{\theta} \in \Theta$ tale che $L(\hat{\theta}) \geq L(\theta)$, o analogamente $l(\hat{\theta}) \geq l(\theta)$, per ogni $\theta \in \Theta$ è detto **stima di massima verosimiglianza** di θ. Nei modelli con verosimiglianza regolare, $\hat{\theta}$ si individua come la soluzione dell'**equazione di verosimiglianza**

$$l_*(\theta) = 0,$$

dove

$$l_*(\theta) = \frac{\partial l(\theta)}{\partial \theta}$$

è detta **funzione di punteggio** o funzione score.

La matrice $(p \times p)$ delle derivate parziali seconde della log-verosimiglianza cambiate di segno è la **matrice di informazione osservata di Fisher**, ed è indicata con

$$j(\theta) = -l_{**}(\theta) = -\frac{\partial l_*(\theta)}{\partial \theta^T}.$$

Nei modelli statistici parametrici regolari, lo stimatore di massima verosimiglianza è consistente. Inoltre, la distribuzione stimata approssimata di $\hat{\theta}$ è

$$\hat{\theta} \dot{\sim} N_p\left(\theta, j(\hat{\theta})^{-1}\right). \tag{1}$$

La (1) permette di ottenere la statistica test di Wald

$$W_e = (\hat{\theta} - \theta)^T j(\hat{\theta})(\hat{\theta} - \theta) \tag{2}$$

con distribuzione asintotica χ^2_p. Ciò permette di costruire regioni di confidenza con livello nominale 1-α, di forma

$$\{\theta \in \Theta : W_e(\theta) < \chi^2_{p;1-\alpha}\},$$

dove $\chi^2_{p;1-\alpha}$ è il quantile $(1-\alpha)$ della distribuzione χ^2_p.

Una statistica test asintoticamente equivalente è data dalla statistica **log-rapporto di verosimiglianza**, nella forma

$$W(\theta) = 2(l(\hat{\theta}) - l(\theta)). \tag{3}$$

Poiché la distribuzione asintotica di $W(\theta)$ è χ^2_p, una regione di confidenza di livello nominale 1-α è

$$\hat{\Theta}(y) = \{\theta \in \Theta : W(\theta) \leq \chi^2_{p;1-\alpha}\}.$$

Quando il parametro è scalare si può considerare **la radice con segno** di $W(\theta)$, definita da

$$r(\theta) = \text{sgn}(\hat{\theta} - \theta)\sqrt{W(\theta)},$$

con sgn(·) funzione segno. Sotto condizioni di regolarità, la sua distribuzione asintotica nulla è semplicemente $N(0,1)$. Un intervallo di confidenza con livello nominale 1-α è dato allora da

$$\{\theta \in \Theta : -\mathrm{Z}_{1-\alpha/2} < r(\theta) < \mathrm{Z}_{1-\alpha/2}\},$$

con $\mathrm{Z}_{1-\alpha/2}$ quantile (1-α) di $N(0,1)$. Inoltre per $p = 1$ la (2) diventa

$$Z_e = (\hat{\theta} - \theta) j(\hat{\theta})^{1/2}$$

e un intervallo di confidenza alla Wald con livello nominale 1-α è semplicemente dato da

$$\{\theta \in \Theta : -\mathrm{Z}_{1-\alpha/2} < Z_e(\theta) < \mathrm{Z}_{1-\alpha/2}\}.$$

1.2 VEROSIMIGLIANZA PROFILO

In molte applicazioni il parametro θ può essere scomponibile in due componenti, ossia $\theta = (\psi, \lambda)$, dove il parametro scalare ψ indica il **parametro d'interesse**, mentre λ, di dimensione p-1, denota il **parametro di disturbo**. Il parametro di disturbo rende il modello più realistico, ma appunto non è di diretto interesse.

Nell'inferenza basata sulla verosimiglianza, una soluzione di ampia applicabilità per l'inferenza su ψ considera la sostituzione del parametro di disturbo λ con una sua opportuna stima. In particolare, la **verosimiglianza profilo** prevede di sostituire il parametro di disturbo λ con la sua stima di massima verosimiglianza vincolata al

parametro di interesse ψ fissato. Formalmente, la verosimiglianza profilo è definita come

$$L_P(\psi) = L\left(\psi, \hat{\lambda}_\psi\right),$$

dove $\hat{\lambda}_\psi$ è soluzione in λ di

$$\frac{\partial l(\psi,\lambda)}{\partial \lambda} = l_\lambda(\psi,\lambda) = 0\,.$$

La funzione di verosimiglianza profilo non è una verosimiglianza in senso proprio. Tuttavia essa gode di interessanti proprietà che la assimilano a una verosimiglianza propria:

1) La stima di massima verosimiglianza profilo di ψ basata su $L_P(\psi)$ coincide con la stima di massima verosimiglianza di ψ, $\hat{\psi}$, basata su $L(\psi,\lambda)$.
2) la statistica test log-rapporto di verosimiglianza profilo

   $$W_P(\psi) = 2\{l_P(\hat{\psi}) - l_P(\psi)\} = 2\left[l\left(\hat{\psi},\hat{\lambda}\right) - l\left(\psi,\hat{\lambda}_\psi\right)\right],$$

   con $l_P(\psi) = \log L_P(\psi)$, ha distribuzione asintotica ψ^2_1, sotto opportune assunzioni di regolarità. Pertanto, un intervallo di confidenza di livello nominale 1-α per ψ ha forma

   $$\left\{\psi \in \Psi : W_P(\psi) < \chi^2_{1;1-\alpha}\right\}$$

   La versione unilaterale di $W_P(\psi)$ è data dalla radice con segno di $W_P(\psi)$, ossia

   $$r_P(\psi) = \operatorname{sgn}(\hat{\psi} - \psi)\sqrt{W_P(\psi)}\,,$$

   la cui distribuzione asintotica è $N(0,1)$.
3) l'informazione osservata profilo è

   $$j_P(\psi) = -\left(\frac{\partial^2 l_P(\psi)}{\partial \psi^2}\right) = -\left(\frac{\partial^2 l\left(\psi,\hat{\lambda}_\psi\right)}{\partial \psi^2}\right) = -l_{\psi\psi}(\psi,\lambda) + l_{\psi\lambda}(\psi,\lambda)(l_{\lambda\lambda}(\psi,\lambda))^{-1} l_{\lambda\psi}(\psi,\lambda),$$

   dove $l_{\psi\psi}(\cdot)$, $l_{\psi\lambda}(\cdot)$, $l_{\lambda\psi}(\cdot)$ e $l_{\lambda\lambda}(\cdot)$ indicano gli elementi della partizione in ψ e λ di $j(\theta) = j(\psi,\lambda)$.

   Il test alla Wald per ψ può essere espresso come

   $$Z_{e_P}(\psi) = (\hat{\psi} - \psi) j_P(\hat{\psi})^{1/2}\,,$$

con distribuzione asintotica $N(0,1)$.

Le proprietà 1) – 3) rendono la verosimiglianza profilo uno strumento inferenziale interessante. Tuttavia, come già accennato, la $L_P(\psi)$ non è una verosimiglianza in senso proprio. Quando la dimensione di λ è elevata e/o la numerosità campionaria piccola o moderata, l'inferenza basata su $L_P(\psi)$ può risultare inaccurata (si veda, ad esempio, Pace e Salvan, 1996, capitoli 4 e 11).

1.3 VEROSIMIGLIANZA PROFILO MODIFICATA

La verosimiglianza profilo $L_P(\psi)$ costituisce uno strumento assai semplice e generale. Tuttavia, essa presenta alcuni limiti. Ricorrere alla verosimiglianza profilo equivale a comportarsi come se λ fosse noto e pari a $\hat{\lambda}_\psi$. Ciò può essere poco appropriato in presenza di campioni con numerosità piccola o in presenza di parametri di disturbo con dimensione elevata. Per questo motivo, i risultati dell'inferenza su ψ possono essere poco accurati. In anni recenti (vedi Pace e Salvan, 1996, capitolo 11), considerazioni di carattere asintotico hanno suggerito miglioramenti di $L_P(\psi)$ attraverso l'introduzione di fattori di modificazione.

Varie proposte di modificazione di $L_P(\psi)$ sono state studiate. In questa tesi viene considerata la **verosimiglianza profilo modificata** di Barndorff-Nielsen (1983, 1988), definita come

$$L_{MP}(\psi) = L_P(\psi) M(\psi),$$

dove $M(\psi)$ è il fattore di aggiustamento definito come

$$M(\psi) = \left| \frac{\partial \hat{\lambda}_\psi}{\partial \hat{\lambda}} \right|^{-1} \left| j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) \right|^{-1/2}$$

È possibile mostrare che $L_P(\psi)$ e $L_{MP}(\psi)$ sono asintoticamente equivalenti. Inoltre, è importante sottolineare che le statistiche del tipo (2) o (3) definite a partire dalla verosimiglianza profilo modificata, hanno la medesima distribuzione asintotica di

$W_P(\psi)$ e di $Z_{e_P}(\psi)$. È vero, tuttavia, che ci si possono attendere in pratica valutazioni più accurate, in particolare se (p-1) è elevato.

CAPITOLO 2
La curva ROC e l'AUC

Un modello di classificazione è una regola che consente di assegnare le unità statistiche a una certa classe, o gruppo, in riferimento a una o più variabili rilevate su quelle unità. Gli esempi che si possono fare sono molti: sulla base dei dati raccolti dalle carte fedeltà di una catena di supermercati si vuole dividere in categorie la clientela per prendere decisioni di marketing; un'equipe di medici vuole stabilire se un paziente è sano o malato basandosi sul valore di alcuni indicatori; un gruppo di archeologi vuole attribuire un oggetto a un sito o a un'epoca sulla base di alcune misure relativamente facili da eseguire e meno costose rispetto a strumenti più precisi ma indubbiamente meno economici; e così via. La curva ROC (si veda, ad esempio, Azzalini e Scarpa, 2004, capitolo 5) trova diretta applicazione nei casi di classificazione binaria, cioè tutte quelle volte in cui ci si trova in presenza di problemi in cui si considerano due possibili classi, e valuta l'accuratezza del modello basandosi sugli errori di classificazione. Un'estensione di questo metodo ad una situazione con più di due classi è possibile se si procede a costruire diverse curve ROC in ciascuna delle quali si confronta una classe con l'unione delle altre (Azzalini e Scarpa, 2004, paragrafo 5.2.4).

2.1 CLASSIFICAZIONE BINARIA

Per classificazione binaria si intende un problema in cui bisogna stabilire se un'unità appartiene alle classi, o gruppi, G_1 o G_2. Ad esempio, in uno studio clinico, attraverso un test diagnostico i pazienti sono classificati in sani o malati.

Fissata un'opportuna soglia *c*, è possibile costruire una **tabella di contingenza** a partire da un criterio di classificazione. Questa tabella si costruisce a partire dai risultati della classificazione, che nel caso binario potranno essere quattro: un'unità realmente appartenente al gruppo G_1 è stata correttamente classificata come G_1; oppure, un'unità è stata classificata come G_1, mentre in realtà era G_2; o ancora, un'unità è stata correttamente classificata come G_2 oppure erroneamente classificata G_2 quando in realtà era G_1 (si veda la Figura 2.1). La tabella è rappresentabile come:

		classe prevista		
		G_1	**G_2**	
classe reale	**G_1**	n_{11}	n_{12}	$n_{1.}$
	G_2	n_{21}	n_{22}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	n

Dalla tabella si possono calcolare diverse statistiche, che hanno preso nome dal mondo della medicina (con G_2 paziente malato, quindi esito positivo, e G_1 paziente sano, quindi esito negativo).

La proporzione di falsi positivi (FP), cioè la proporzione di quelle unità che sono state assegnate a G_2 mentre in realtà appartengono a G_1, è data da

$$FP = n_{12} / n_{1.} .$$

La proporzione di falsi negativi (FN), cioè quelle unità che sono state assegnate a G_1 quando in realtà appartengono a G_2, è data da

$$FN = n_{21} / n_{n2.} .$$

Se aumenta una, diminuisce l'altra e viceversa.

La proporzione di veri positivi (VP), detta anche **sensibilità**, è il complemento a uno di quella di falsi negativi, ossia

$$VP = n_{22} / n_{2.} = 1 - FN .$$

La proporzione di veri negativi (VN), detta anche **specificità**, è il complemento a uno di quella di falso positivo, ossia

$$VN = n_{11} / n_{1.} = 1 - FP \, .$$

In conclusione, fissato c, si ha una tabella di contingenza con una propria sensibilità e una propria specificità.

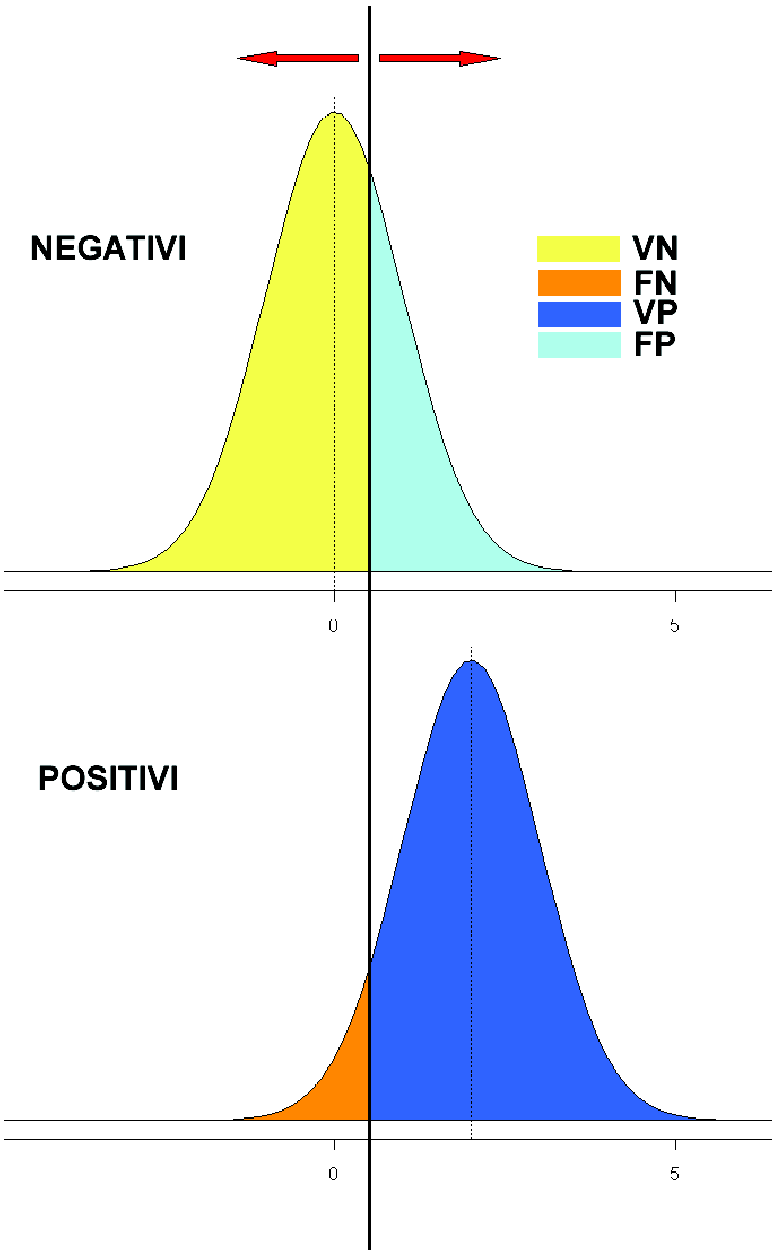


Figura 2.1: *Le due curve rappresentano le distribuzioni di probabilità dei due gruppi reali, negativi e positivi. La linea verticale rappresenta la soglia c. Al di sotto della soglia c le unità sono classificate come negative, al di sopra come positive. Al variare di c cambiano anche le proporzioni VN, FN, FP e VP.*

2.2 CURVA ROC

Nella Figura 2.2 si vede come per diversi valori della soglia c si ottengono proporzioni di veri positivi e falsi positivi diverse. La curva ROC è definita come il grafico delle coppie (1-specificità, sensibilità) al variare di c ossia

$$(FP, 1 - FN)$$

La curva ROC permette di vedere la relazione tra la probabilità di assegnare correttamente un'unità al gruppo G_2 e quella di falso positivo.

Si noti che la curva ROC è contenuta nel quadrato $[0,1]\times[0,1]$ e che, al crescere di c, si ha che sia la sensibilità che il complemento a uno della specificità crescono. Ovviamente si vorrebbe che quest'ultima quantità sia quanto più piccola possibile, mentre si auspica una sensibilità più vicina a uno: i due casi limite che si presentano e che permettono di avere un'idea per interpretare la curva ROC sono quello di una classificazione casuale e quello di una classificazione perfetta.

Il primo caso si manifesta con una curva ROC rappresentante la diagonale del quadrato, che viene chiamata linea di non discriminazione, dal punto (0,0) al punto (1,1), cioè tutti quei punti dove sensibilità = 1-specificità, cioè non esiste un criterio per assegnare un'unità a uno dei due gruppi, ma questo avviene in modo completamente casuale.

Il secondo caso, invece, si presenta quando la probabilità di vero positivo è uno, mentre quella di falso positivo è zero, quindi una spezzata passante per il punto (0,1), detto anche classificazione perfetta. Intuitivamente, il test che stiamo valutando sarà tanto migliore quanto più la curva ROC si allontana dalla diagonale e si avvicina al punto (0,1).

Nella Figura 2.3 viene riportato un esempio di curva ROC. I punti sono stati ottenuti calcolando, a intervalli regolari di c, la sensibilità e 1-specificità, a partire da due campioni simulati con distribuzione normale, il primo di media 0 e il secondo di

media 2, e uguale varianza pari a 1. Si nota che la curva è contenuta nel quadrato [0,1]×[0,1] e che si scosta dalla linea di non discriminazione avvicinandosi al punto (0,1). A seconda dei casi, bisognerà stabilire se la variabile scelta offre una discriminazione sufficientemente apprezzabile.

Un indicatore molto utilizzato della bontà della regola di classificazione è costituito dall'area sottesa alla curva ROC (o AUC, *Area Under the ROC Curve*). Esistono diversi approcci per calcolare questa misura. Una prima possibilità prevede di interpolare i punti calcolati per ottenere la curva ROC e calcolarne approssimativamente l'integrale. Oppure, nel caso di assunzioni parametriche, si può ricorrere alla teoria della verosimiglianza.

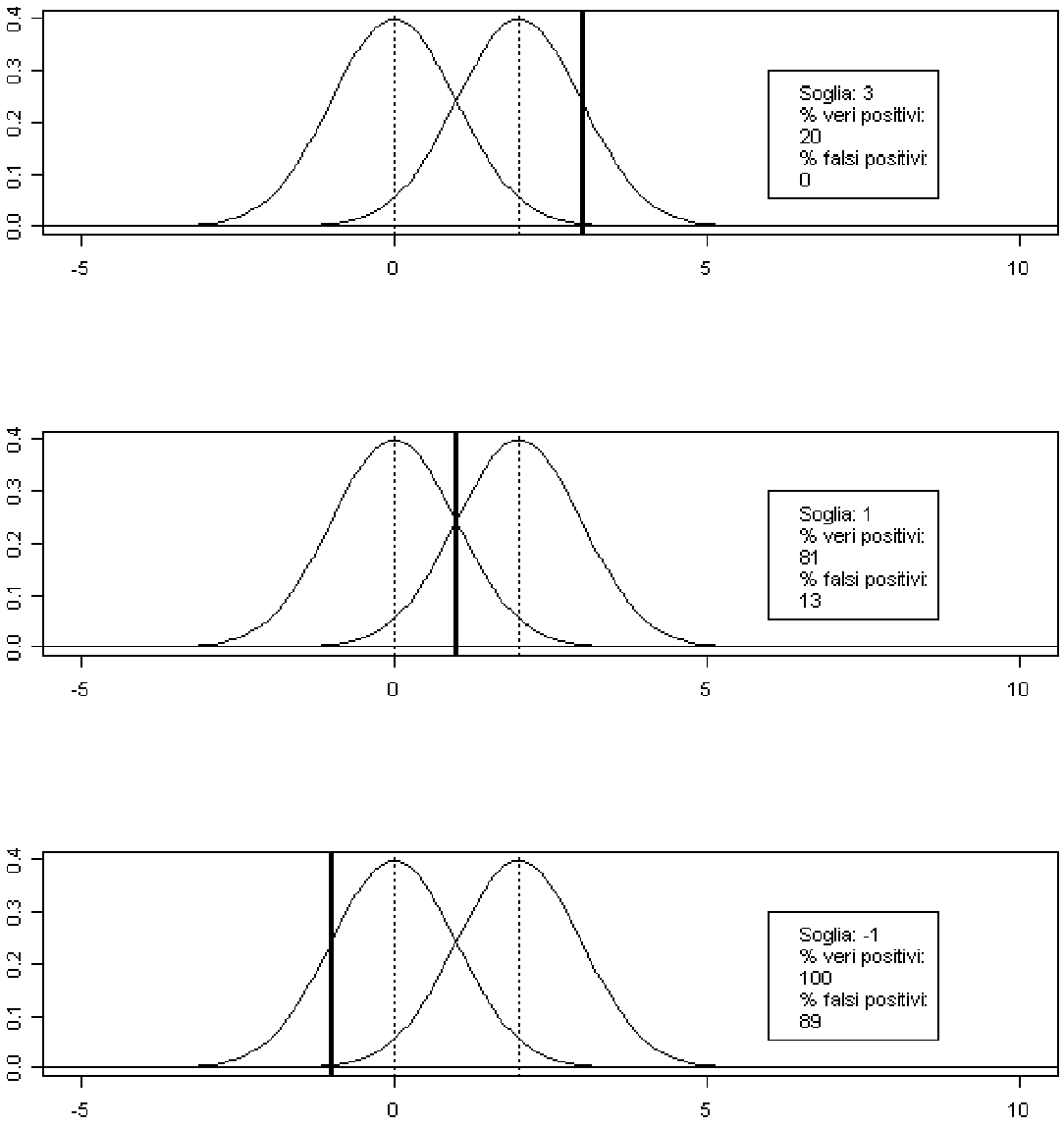


Figura 2.2: *Le due curve rappresentano le distribuzioni di probabilità di due campioni, la linea verticale più spessa rappresenta il livello della soglia. Nella legenda compaiono le percentuali di vero positivo e falso positivo calcolate per quella soglia. Per questo esempio sono stati simulati due campioni con distribuzione normale, rispettivamente, di media 0 e 2 e varianza 1.*

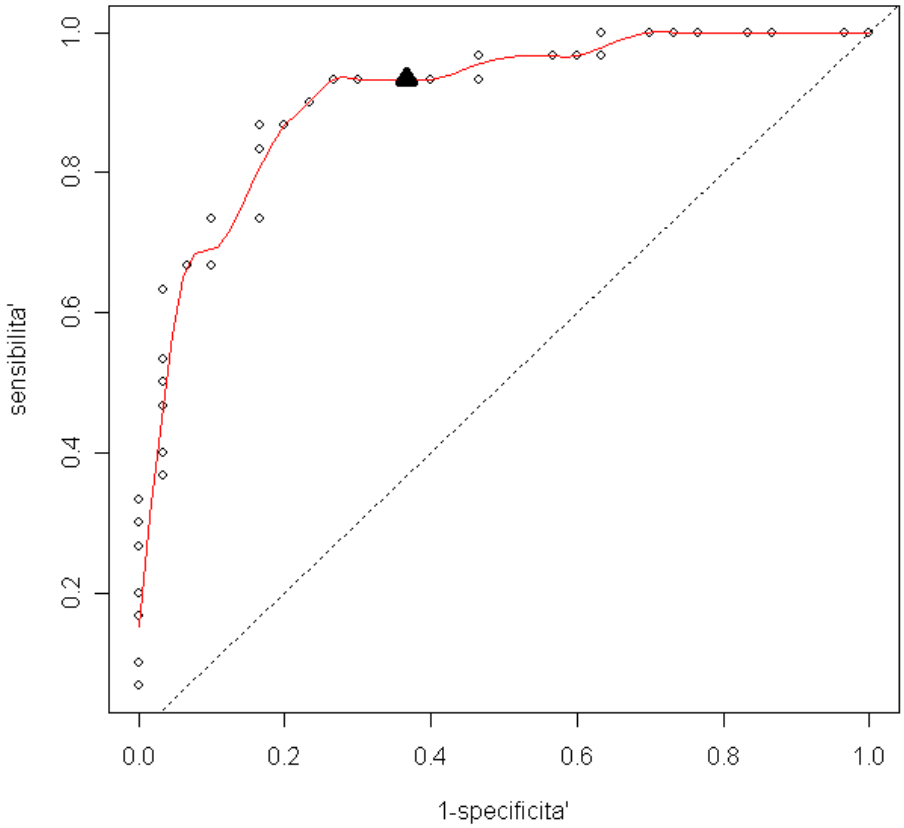


Figura 2.3: *Un esempio di curva ROC per due campioni simulati con distribuzione normale di media, rispettivamente, 0 e 2 e varianza 1. Il triangolo nero rappresenta la soglia c = 0.5. La linea tratteggiata è quella indicante la classificazione casuale. La linea rossa rappresenta un'approssimazione della curva ROC.*

2.3 L'AREA SOTTO LA CURVA ROC

Sia X una variabile che rappresenta la misura nel gruppo G_1 e Y quella nel gruppo G_2. La quantità P(X<Y) può essere interpretata come l'AUC e in letteratura è nota anche come modello sollecitazione-resistenza (*stress-strength model*); per una trattazione su questo modello si veda Kotz et al. (2003).

Diverse assunzioni parametriche e non parametriche sulle variabili X e Y sono possibili. In questa tesi si assume che X e Y sono variabili casuali indipendenti e appartenenti alla stessa famiglia di distribuzioni. In particolare, nel Capitolo 3 si discuterà il caso di due esponenziali.

Formalmente, siano X e Y due variabili casuali con funzioni di densità, rispettivamente, $p_x(x;\theta_x)$ e $p_y(y;\theta_y)$, con $\theta_x \in \Theta_x \subseteq IR^{p_x}$ e $\theta_y \in \Theta_y \subseteq IR^{p_y}$.

L'AUC può essere espressa come una funzione dell'intero parametro $\theta = (\theta_x, \theta_y)$, attraverso la relazione

$$\psi = \psi(\theta) = P(X < Y) = \int_{-\infty}^{+\infty} F_x(t;\theta_x) dF_y(t;\theta_y), \tag{4}$$

con $F_x(\cdot)$ e $F_y(\cdot)$ funzioni di ripartizione di X e Y, rispettivamente. Espressioni teoriche per l'AUC sono disponibili con riferimento a diverse assunzioni distributive sia su X che su Y (si veda, ad esempio, Kotz et al, 2003).

Per l'inferenza sulla (4) per molte distribuzioni di probabilità usualmente si procede per via parametrica utilizzando la verosimiglianza profilo.

2.4 TEORIA DELLA VEROSIMIGLIANZA

Siano $x = (x_1,...,x_{n_x})$ e $y = (y_1,...,y_{n_y})$ due campioni casuali semplici di numerosità n_x e n_y tratti, rispettivamente, da X e Y. Poiché la funzione di verosimiglianza è invariante rispetto a riparametrizzazioni del modello, si ha che la stima di massima verosimiglianza dell'AUC è legata a quella di θ dalla relazione

$$\hat{\psi} = \psi(\hat{\theta}),$$

con $\hat{\theta} = (\hat{\theta}_x, \hat{\theta}_y)$ stima di massima verosimiglianza di θ ottenuta massimizzando

$$L(\theta) = L(\theta_x, \theta_y) = \prod_{i=1}^{n_x} p_x(x_i;\theta_x) \prod_{j=1}^{n_y} p_y(y_j;\theta_y) = L(\theta_x;x) L(\theta_y;y).$$

Ponendo $\theta = (\psi, \lambda)$, con $\psi = \psi(\theta) = P(X < Y)$ e λ opportuno parametro di disturbo, intervalli di confidenza per ψ possono essere ottenuti a partire dalla verosimiglianza profilo, come illustrato nel Paragrafo 1.2. Pertanto, un intervallo di confidenza approssimato di livello nominale 1-α per l'AUC può essere costruito come

$$\{\psi \in (0,1) : -\mathrm{Z}_{1-\alpha/2} < r_P(\psi) < \mathrm{Z}_{1-\alpha/2}\}$$

o, utilizzando il test alla Wald, come

$$\{\psi \in (0,1) : -\mathrm{Z}_{1-\alpha/2} < Z_{e_P}(\psi) < \mathrm{Z}_{1-\alpha/2}\}.$$

Risultati più accurati si possono ottenere con la verosimiglianza profilo modificata

$$L_{MP}(\psi) = L_P(\psi) M(\psi),$$

illustrata nel Paragrafo 1.3. In particolare, con $L_{MP}(\psi)$ si può ottenere $r_{MP}(\psi)$ per costruire intervalli di confidenza approssimati, in modo analogo alla verosimiglianza profilo.

Il miglioramento nell'inferenza su ψ, che si ottiene con la verosimiglianza profilo modificata, sarà oggetto di studio nelle simulazioni effettuate nel Capitolo 3.

CAPITOLO 3
Il caso esponenziale

In questo capitolo si presenta un esempio su dati reali e si discutono degli studi di simulazione per valutare il miglioramento nell'inferenza sull'AUC introdotto dalla verosimiglianza profilo modificata. In questa tesi, si considera la situazione in cui X e Y seguono una distribuzione esponenziale. Una volta mostrati sinteticamente i passaggi algebrici per ottenere le quantità necessarie per l'inferenza sul parametro che rappresenta l'area sotto la curva ROC, si procederà con il calcolo vero e proprio con il software statistico R, i cui codici sono disponibili in Appendice.

3.1 DISTRIBUZIONE ESPONENZIALE

Siano $y = (y_1,...,y_{n_y})$ e $x = (x_1,...,x_{n_x})$ due campioni casuali semplici con distribuzioni di probabilità esponenziali, rispettivamente, $p(y;\alpha) = \alpha e^{-\alpha y}$ e $p(x;\beta) = \beta e^{-\beta y}$, con α>0 e β>0.

La verosimiglianza congiunta è

$$L(\theta) = L(\alpha,\beta) = \alpha^n e^{-\alpha\sum_{i=1}^{n_y} y_i} \beta^m e^{-\beta\sum_{j=1}^{n_x} x_j}$$
$$= \alpha^{n_y} e^{-\alpha n_y \bar{y}} \beta^{n_x} e^{-\beta n_x \bar{x}}$$

con $\bar{y} = \frac{1}{n_y}\sum_{i=1}^{n_y} y_i$ e $\bar{x} = \frac{1}{n_x}\sum_{j=1}^{n_x} x_j$.

La log-verosimiglianza è

$$l(\theta) = l(\alpha,\beta) = n_y \log\alpha - \alpha n_y \bar{y} + n_x \log\beta - \beta n_x \bar{x}$$

La funzione punteggio è

$$l_*(\alpha,\beta)=\begin{cases}\dfrac{\partial l(\alpha,\beta)}{\partial\alpha}=\dfrac{n_y}{\alpha}-\bar{y}n_y\\ \dfrac{\partial l(\alpha,\beta)}{\partial\beta}=\dfrac{n_x}{\beta}-\bar{x}n_x\end{cases}$$

Dall'equazione di verosimiglianza $l_*(\alpha,\beta)=0$, si ottengono la stima di massima verosimiglianza $\hat{\theta}=(\hat{\alpha},\hat{\beta})$ di θ con

$$\begin{cases}\hat{\alpha}=\dfrac{1}{\bar{y}}\\ \hat{\beta}=\dfrac{1}{\bar{x}}\end{cases}. \tag{5}$$

Bisogna esprimere i nuovi parametri ψ e λ in funzione di α e β. Il parametro ψ, che rappresenta l'AUC, e λ, parametro di disturbo, si possono esprimere come

$$\begin{cases}\psi=\dfrac{\beta}{\alpha+\beta}\\ \lambda=\alpha+\beta\end{cases}$$

La relazione tra i vecchi e i nuovi parametri è dunque

$$\begin{cases}\alpha=\lambda-\beta=\lambda-\psi\lambda=\lambda(1-\psi)\\ \beta=\psi\lambda\end{cases}$$

La log-verosimiglianza diventa

$$\begin{aligned}&l(\psi,\lambda)=n_y\log(\lambda(1-\psi))-n_y\lambda(1-\psi)\bar{y}+n_x\log(\lambda\psi)-n_x\psi\lambda\bar{x}=\\&=(n_y+n_x)\log\lambda+n_y\log(1-\psi)+n_x\log\psi-n_y\lambda(1-\psi)\bar{y}-n_x\psi\lambda\bar{x}\end{aligned}$$

Si può esprimere $\bar{x}$ e $\bar{y}$ in funzione delle stime di massima verosimiglianza (5), come

$$\begin{cases}\bar{y}=\dfrac{1}{\hat{\alpha}}=\dfrac{1}{\hat{\lambda}(1-\hat{\psi})}\\ \bar{x}=\dfrac{1}{\hat{\beta}}=\dfrac{1}{\hat{\psi}\hat{\lambda}}\end{cases}$$

e per l'invarianza della stima di massima verosimiglianza rispetto a riparametrizzazioni si ha che

$$\begin{cases}\hat{\psi}=\dfrac{\hat{\beta}}{\hat{\alpha}+\hat{\beta}}=\dfrac{\frac{1}{\bar{x}}}{\frac{1}{\bar{y}}+\frac{1}{\bar{x}}}=\dfrac{\bar{y}}{(\bar{x}+\bar{y})}\\ \hat{\lambda}=\hat{\alpha}+\hat{\beta}=\dfrac{1}{\bar{y}}+\dfrac{1}{\bar{x}}=\dfrac{\bar{x}+\bar{y}}{\bar{x}\bar{y}}\end{cases}$$

Sfruttando quanto appena visto, la log-verosimiglianza diventa

$$l(\psi,\lambda)=\left(n_x+n_y\right)\log\lambda+n_y\log(1-\psi)+n_x\log\psi-\frac{\lambda(1-\psi)n_y}{\hat{\lambda}(1-\hat{\psi})}-\frac{\lambda\psi n_x}{\hat{\lambda}\hat{\psi}}$$

La derivata parziale rispetto a λ della log-verosimiglianza è

$$\frac{\partial l(\psi,\lambda)}{\partial\lambda}=\frac{n_x+n_y}{\lambda}-\frac{n_y(1-\psi)}{(1-\hat{\psi})\hat{\lambda}}-\frac{n_x\psi}{\hat{\psi}\hat{\lambda}}=\frac{n_x+n_y}{\lambda}-\frac{\left(n_x\psi(1-\hat{\psi})+n_y\hat{\psi}(1-\psi)\right)}{\hat{\psi}(1-\hat{\psi})\hat{\lambda}}$$

Da $\frac{\partial l(\psi,\lambda)}{\partial\lambda}=0$, si ottiene $\hat{\lambda}_\psi$, dato da

$$\hat{\lambda}_\psi=\frac{\left(n_x+n_y\right)\hat{\psi}(1-\hat{\psi})\hat{\lambda}}{\left(n_x\psi(1-\hat{\psi})+n_y\hat{\psi}(1-\psi)\right)}$$

Derivando la funzione punteggio si ottengono le quantità

$$l_{\psi\psi}=\frac{\partial l(\psi,\lambda)}{\partial\psi\partial\psi}=-\frac{n_x}{\psi^2}-\frac{n_y}{(1-\psi)^2}$$

$$l_{\psi\lambda}=l_{\lambda\psi}\,\frac{\partial l(\psi,\lambda)}{\partial\psi\partial\lambda}=\frac{\partial l(\psi,\lambda)}{\partial\lambda\partial\psi}=-\frac{n_x}{\hat{\psi}\hat{\lambda}}+\frac{n_y}{(1-\hat{\psi})\hat{\lambda}}=\frac{n_y\hat{\psi}-n_x(1-\hat{\psi})}{\hat{\psi}\hat{\lambda}(1-\hat{\psi})}$$

$$l_{\lambda\lambda}=\frac{\partial l(\psi,\lambda)}{\partial\lambda\partial\lambda}=-\frac{\left(n_x+n_y\right)}{\lambda^2},$$

con le quali si ottiene $j_P(\psi)$ e quindi Z_{e_P}. In particolare, il test alla Wald profilo risulta dato da

$$\mathrm{Z}_{e_P}(\psi)=(\hat{\psi}-\psi)\sqrt{\frac{n_x n_y}{\left(n_x+n_y\right)\hat{\psi}^2(1-\psi)^2}}$$

La verosimiglianza profilo è

$$L_p(\psi)=\psi^{n_x}(1-\psi)^{n_y}\,\hat{\lambda}_\psi^{\,n_x+n_y}$$

Si ha inoltre

$$j_{\lambda\lambda}\left(\psi,\hat{\lambda}_\psi\right)=\frac{\left(n_x+n_y\right)}{\lambda^2}$$

e

$$\left|\frac{\partial\hat{\lambda}_\psi}{\partial\hat{\lambda}}\right|=\frac{\left(n_x+n_y\right)\hat{\psi}(1-\hat{\psi})}{\left(n_x\psi(1-\hat{\psi})+n_y\hat{\psi}(1-\psi)\right)}=\frac{\hat{\lambda}_\psi}{\hat{\lambda}}$$

Quindi la verosimiglianza profilo modificata risulta data da

$$L_{MP}(\psi) = L_P(\psi)\left|\frac{\partial\hat{\lambda}_\psi}{\partial\hat{\lambda}}\right|^{-1}\left|j_{\lambda\lambda}\left(\psi,\hat{\lambda}_\psi\right)\right|^{-1/2} = L_P(\psi)\frac{\hat{\lambda}}{\hat{\lambda}_\psi}\frac{\hat{\lambda}_\psi}{\sqrt{n+m}} = L_P(\psi), \qquad (6)$$

ossia risulta equivalente alla verosimiglianza profilo. Quindi nello studio di simulazione e nell'esempio con dati reali si userà solamente (6).

3.2 SIMULAZIONE VIA MONTE CARLO

Lo studio di simulazione presentato è diviso in due fasi. Nella prima si considera una stima puntuale di ψ, nella seconda si trattano le stime intervallari.

Considerando la situazione illustrata nel Paragrafo 3.1, si vuole vedere come si comporta la stima puntuale di ψ variando sia la numerosità campionaria sia il vero valore di ψ, usando la funzione `sim_exp`, che implementa la verosimiglianza profilo modificata, che, come mostrato, coincide in questo caso con la verosimiglianza profilo modificata (si veda l'Appendice). I risultati della simulazione sono riassunti nella Tabella 3.1, dove in ogni cella viene indicata la media e la deviazione standard (tra parentesi) dei vettori delle stime ottenute simulando 5000 volte due campioni di numerosità (n_x, n_y), con ψ fissato a 0.5, 0.8, 0.9 e 0.99.

(n_x, n_y)	**ψ= 0.5**	**ψ= 0.8**	**ψ= 0.9**	**ψ= 0.99**
(3,3)	0.499 (0.189)	0.771 (0.142)	0.874 (0.100)	0.985 (0.017)
(5,5)	0.498 (0.149)	0.783 (0.109)	0.884 (0.069)	0.988 (0.009)
(10,10)	0.502 (0.110)	0.792 (0.074)	0.893 (0.043)	0.989 (0.005)
(20,20)	0.501 (0.078)	0.795 (0.053)	0.897 (0.030)	0.990 (0.003)
(50,50)	0.501 (0.050)	0.798 (0.032)	0.899 (0.018)	0.990 (0.002)

Tabella 3.1 : *Studio Monte Carlo per la stima puntuale.*

Si può vedere come, per valori di ψ elevati, in particolare maggiori di 0.8 (che sono anche quelli di maggiore interesse nelle analisi in cui si usa la curva ROC), lo stimatore ottenuto dalla verosimiglianza profilo modificata è affidabile non solo in termini di media, ma anche in termini di deviazione standard (Figura 3.1 e 3.2).

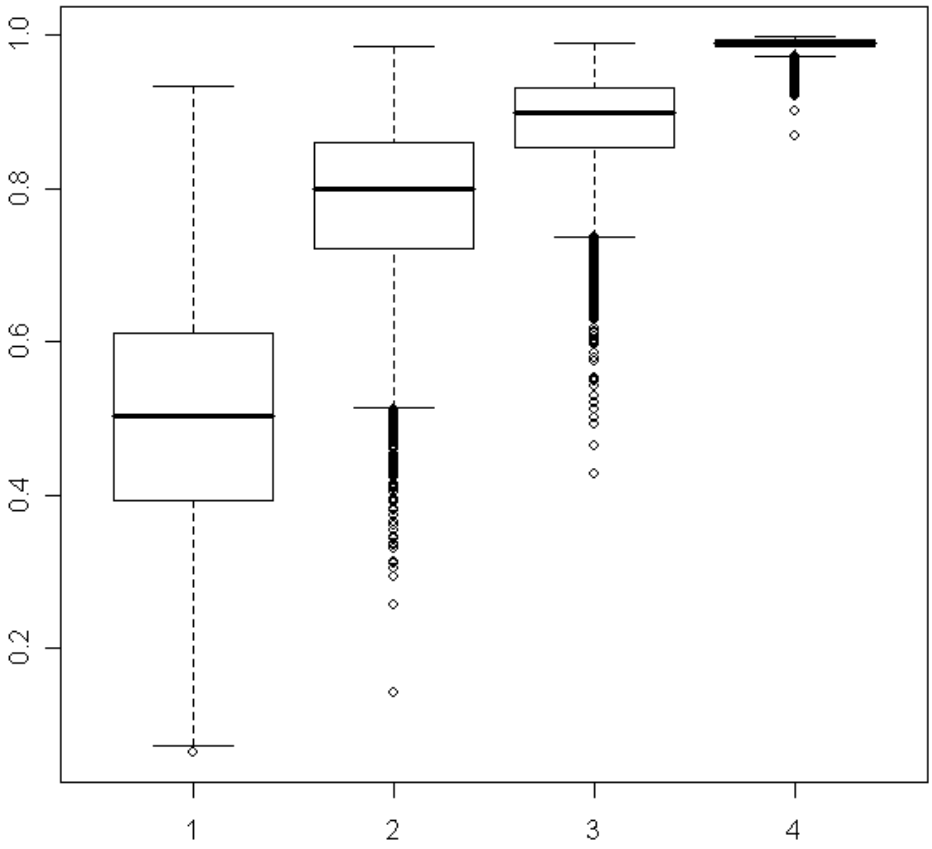


Figura 3.1:
(1) ψ=0.5, n_x=n_y=5, $\hat{\psi}$; (2) ψ=0.8, n_x=n_y=5, $\hat{\psi}$;
(3) ψ=0.9, n_x=n_y=5, $\hat{\psi}$; (4) ψ=0.99, n_x=n_y=5, $\hat{\psi}$.

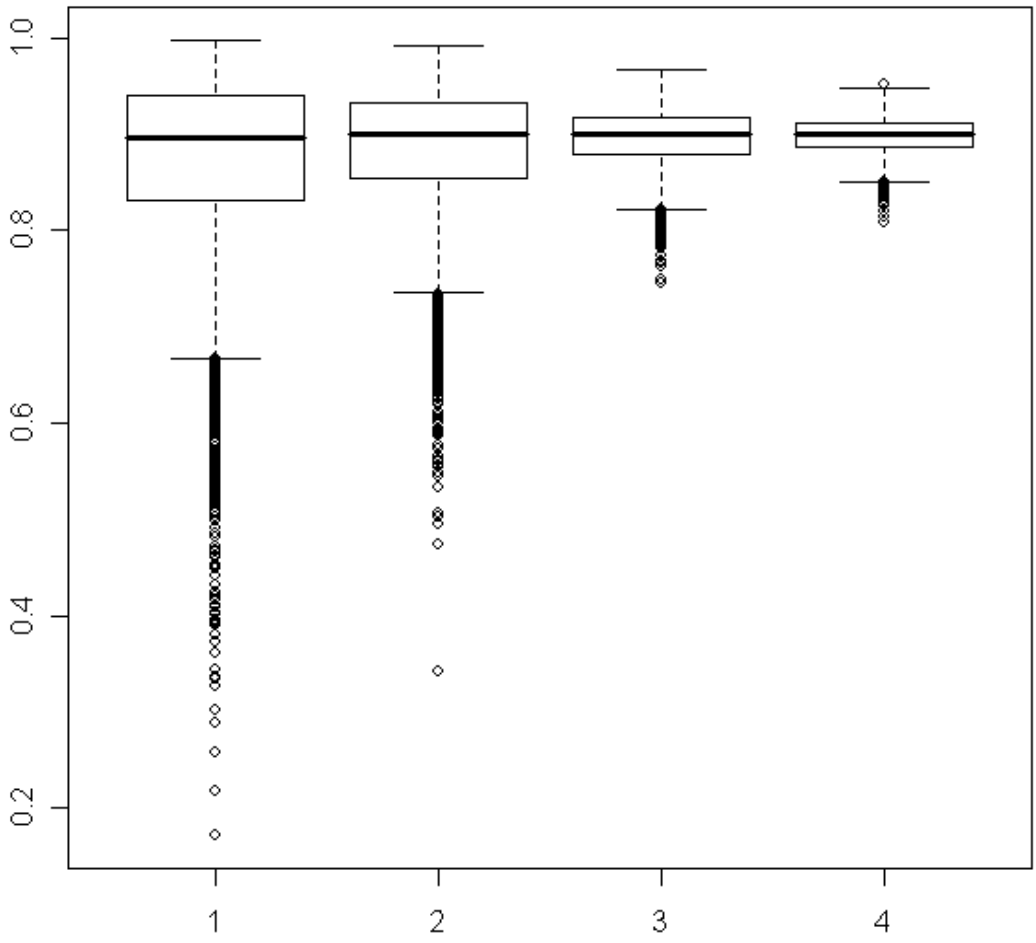


Figura 3.1:
(1) ψ=0.9, n_x=n_y=3, $\hat{\psi}$; (2) ψ=0.9, n_x=n_y=5, $\hat{\psi}$;
(3) ψ=0.9, n_x=n_y=20, $\hat{\psi}$; (4) ψ=0.9, n_x=n_y=50, $\hat{\psi}$.

Per lo studio degli intervalli di confidenza per ψ è stata creata una funzione (`sim_exp2`) che calcola $Z_{e_P}(\psi)$ e $r_{MP}(\psi)$ per ciascuna delle 5000 coppie di campioni simulati, lo confronta con i quantili 0.25 e 0.975 di una normale standard e conta ogni volta che $Z_{e_P}(\psi)$ e $r_{MP}(\psi)$ cadono nell'intervallo, per avere un indicatore dell'accuratezza dei due metodi. I risultati della simulazione sono riassunti nella

Tabella 3.2, dove in ogni cella c'è il rapporto tra il numero di volte in cui il valore del test rientra nell'intervallo e il numero di ripetizioni.

(n_x, n_y)	**test**	**ψ= 0.5**	**ψ= 0.8**	**ψ= 0.9**
(3,3)	$Z_{e_P}(\psi)$	0.798	0.819	0.825
	$r_{MP}(\psi)$	0.937	0.941	0.941
(5,5)	$Z_{e_P}(\psi)$	0.831	0.841	0.858
	$r_{MP}(\psi)$	0.943	0.938	0.946
(10,10)	$Z_{e_P}(\psi)$	0.856	0.871	0.876
	$r_{MP}(\psi)$	0.945	0.944	0.948
(20,20)	$Z_{e_P}(\psi)$	0.878	0.879	0.896
	$r_{MP}(\psi)$	0.944	0.946	0.959
(50,50)	$Z_{e_P}(\psi)$	0.894	0.882	0.895
	$r_{MP}(\psi)$	0.954	0.944	0.948

Tabella 3.2

Dalla Tabella 3.2 si nota come, soprattutto con numerosità campionarie basse, $r_{MP}(\psi)$ sia più accurato di $Z_{e_P}(\psi)$e come la differenza tra $r_{MP}(\psi)$ e $Z_{e_P}(\psi)$ rimanga, anche per numerosità campionarie elevate o comunque non inferiori a 10-20 unità, sostanziale. Procedure di inferenza accurate anche per numerosità campionarie molto basse, come la verosimiglianza profilo modificata applicata al calcolo dell'AUC, sono strumenti di estremo interesse: in molte applicazioni, specialmente quelle mediche, capita che ottenere dei campioni di numerosità alta sia troppo costoso o addirittura impossibile (basti pensare a patologie particolarmente rare). Nel paragrafo seguente viene presentata un'applicazione degli strumenti usati nella simulazione a un caso reale.

3.3 I DATI ALCL

I valori contenuti nel dataset oggetto di studio sono stati raccolti dal Centro Oncoematologico Pediatrico di Padova e si dividono in due gruppi: uno di malati (variabile Y) e uno di sani (variabile X). Il Centro voleva valutare se la quantità di una certa proteina (la Hsp70) presente nel paziente potesse essere un indicatore in grado di discriminare i sani dai malati di Linfoma Anaplastico a Grandi Cellule (ALCL), un particolare tipo di cancro. Non solo sembra che i pazienti malati abbiano un livello di Hsp70 maggiore rispetto ai sani (come sembra suggerire anche il boxplot nella Figura 3.3), ma la presenza di tale proteina sembra diminuire l'efficacia della terapia.

Una delle difficoltà maggiori nel trarre delle conclusioni da questo dataset è data dalla numerosità molto bassa dei due gruppi, 10 per quello dei malati e 4 per quello dei sani. Il livello di proteina Hsp70 nel paziente è stato rilevato su scala continua e le osservazioni sono indipendenti. Inoltre, si assume che i due gruppi abbiano distribuzioni esponenziali, rispettivamente, $p(y;\alpha)=\alpha e^{-\alpha y}$ e $p(x;\beta)=\beta e^{-\beta y}$. Riprendendo la terminologia usata nei paragrafi precedenti, si ha che ψ esprime la probabilità che il livello della proteina Hsp70 nel gruppo dei malati sia minore del livello della proteina nel gruppo dei sani, cioè ψ=P(X<Y).

Applicando quindi la verosimiglianza profilo espressa nel Paragrafo 3.1, si ha che $\hat{\psi}=\frac{\bar{y}}{\bar{y}+\bar{x}}=0.8594$. Cioè si ha che la probabilità dell'evento X<Y è alta. In altre parole il livello della proteina risulta maggiore nei pazienti malati.

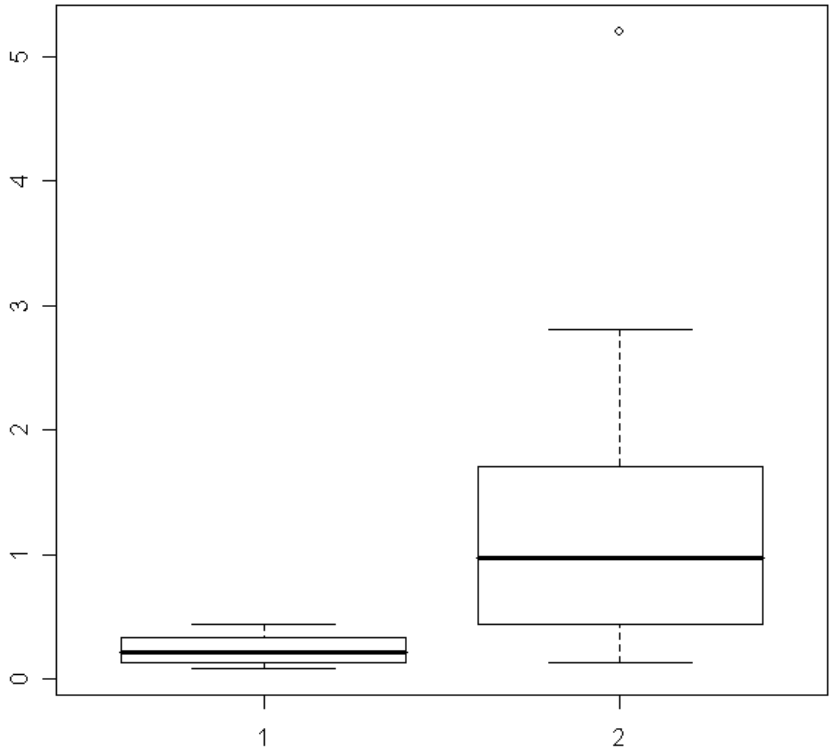


Figura 3.3: *Boxplot dei due gruppi, (1) sani e (2) malati.*

APPENDICE
Codici R

CODICI DELLE FIGURE

```
#############
# FIGURA 2.1
# GRAFICO ESEMPIO CLASSIFICAZIONE DA DUE NORMALI DI MEDIA 0 E 2
# colori e legenda aggiunti con software di grafica

# Disegniamo le due distribuzioni e la soglia esempio a 0.5
curve(1/(sqrt(2*pi)*1)*exp(-((x-0)^2/(2*1^2))), from=-5, to=10)
abline(v=0, lty="dotted")
abline(v=0.5)
abline(h=0)

curve(1/(sqrt(2*pi)*1)*exp(-((x-2)^2/(2*1^2))), from=-5, to=10)
abline(v=2, lty="dotted")
abline(v=0.5)
abline(h=0)


#############
# FIGURA 2.2
# GRAFICO PER TRE VALORI DELLA SOGLIA
# esempio: 2, 1, -1

x <- rnorm(100) # generiamo i due campioni
y <- rnorm(100,2)
dati <- c(x,y)

# Prepariamo lo spazio per i tre grafici
par(mfrow=c(3,1))

# Disegniamo il primo:
# - le distribuzioni
curve(1/(sqrt(2*pi)*1)*exp(-((x-0)^2/(2*1^2))), from=-5, to=10,
ylab="", xlab="")
abline(v=0, lty="dotted")
abline(h=0)
```

```
curve(1/(sqrt(2*pi)*1)*exp(-((x-2)^2/(2*1^2))), from=-5, to=10,
ylab="", xlab="", add=TRUE)
abline(v=2, lty="dotted")
abline(v=3, lwd=3)
abline(h=0)

# - calcoliamo le proporzioni
th <- 3
TP <- 0
FP <- 0
for (n in 1:200) {
     if ((dati[n]>th) & (n>100))
          TP = TP+1
     if ((dati[n]>th) & (n<=100))
          FP = FP+1
}

# - aggiungiamo la legenda
legend(6, 0.3, legend=c("Soglia: 3", "% veri positivi:", TP,
"% falsi positivi:", FP))

# Disegniamo il secondo:
curve(1/(sqrt(2*pi)*1)*exp(-((x-0)^2/(2*1^2))), from=-5, to=10,
ylab="", xlab="")
abline(v=0, lty="dotted")
abline(h=0)

curve(1/(sqrt(2*pi)*1)*exp(-((x-2)^2/(2*1^2))), from=-5, to=10,
ylab="", xlab="", add=TRUE)
abline(v=2, lty="dotted")
abline(v=1, lwd=3)
abline(h=0)

th <- 1
TP <- 0
FP <- 0
for (n in 1:200) {
     if ((dati[n]>th) & (n>100))
          TP = TP+1
     if ((dati[n]>th) & (n<=100))
          FP = FP+1
}

legend(6, 0.3, legend=c("Soglia: 1", "% veri positivi:", TP,
"% falsi positivi:", FP))

# e il terzo:
curve(1/(sqrt(2*pi)*1)*exp(-((x-0)^2/(2*1^2))), from=-5, to=10,
ylab="", xlab="")
abline(v=0, lty="dotted")
abline(h=0)

curve(1/(sqrt(2*pi)*1)*exp(-((x-2)^2/(2*1^2))), from=-5, to=10,
```

```
ylab="", xlab="",  add=TRUE)
abline(v=2, lty="dotted")
abline(v=-1, lwd=3)
abline(h=0)

th <- -1
TP <- 0
FP <- 0
for (n in 1:200) {
     if ((dati[n]>th) & (n>100))
          TP = TP+1
     if ((dati[n]>th) & (n<=100))
          FP = FP+1
}

legend(6, 0.3, legend=c("Soglia: -1", "% veri positivi:", TP,
"% falsi positivi:", FP))


############
# FIGURA 2.3
# DATI PER ESEMPIO DI GRAFICO ROC DA DUE NORMALI DI MEDIA 0 E 2

# rm(list=ls())

ROC <- matrix(nrow=66,ncol=2)
x <- rnorm(30) # generiamo i due campioni
y <- rnorm(30,2)
dati <- c(x,y)
i <- 4.33 # facciamo variare la soglia tra i valori tipici di una
# normale
r <- 1
while (r<=66) { # generiamo i vettori sensibilità e 1-specificità
     FN <- 0
     FP <- 0
     for (n in 1:60) {
          if ((dati[n]<=i) & (n>30))
               FN = FN+1
          if ((dati[n]>i) & (n<=30))
               FP = FP+1
     }
     ROC[r,1] <- FP/30
     ROC[r,2] <- 1-(FN/30)
     r = r+1
     i = i-0.1
}

# ROC

# disegniamo i punti ROC
plot(ROC, xlab="1-specificita'", ylab="sensibilita'")

# aggiungiamo la curva che approssima i punti ROC
```

```
lines(spline(ROC[,1], ROC[,2]), col="red")

# aggiungiamo corrispondente alla soglia 0.5
th <- 0.5
FN <- 0
FP <- 0
for (n in 1:60) {
     if ((dati[n]<=th) & (n>30))
          FN = FN+1
     if ((dati[n]>th) & (n<=30))
          FP = FP+1
}
points(FP/30, 1-(FN/30), pch=24, bg="black", lwd=5)

# aggiungiamo la linea di non discrimazione
abline(coef=c(0,1), lty="dotted")

############
# FIGURA 3.1

# Simuliamo con psi crescente e numerosità campionarie fisse.
# La funzione sim_exp è illustrata nel paragrafo successivo.
ris1 = sim_exp(5000,5,5,1,1)
ris2 = sim_exp(5000,5,5,1/4,1)
ris3 = sim_exp(5000,5,5,1/9,1)
ris4 = sim_exp(5000,5,5,1/99,1)

boxplot(ris1, ris2, ris3, ris4)

############
# FIGURA 3.2

# Simuliamo con psi fisso e numerosità campionarie crescenti.
ris1 = sim_exp(5000,3,3,1/9,1)
ris2 = sim_exp(5000,5,5,1/9,1)
ris3 = sim_exp(5000,20,20,1/9,1)
ris4 = sim_exp(5000,50,50,1/9,1)

boxplot(ris1, ris2, ris3, ris4)

############
# FIGURA 3.3

boxplot(sani, malati)
```

CODICI DELLE SIMULAZIONI

```
# Funzione per calcolare la SMV profilo puntuale
# Riceve come parametri i due campioni xdat e ydat
mle <- function(xdat,ydat){
my <- mean(ydat)
```

```
mx <- mean(xdat)
psi.hat <- my/(mx+my)
}


# Funzione per nr simulazioni di due campioni di numerosità
# nx e ny e con distribuzione esponenziale con parametri be e al
sim_exp <- function(nr,nx,ny,be,al){
p.list <- 0 # il vettore che conterrà le stime
psi.v <- be/(al+be) # psi vero
print(psi.v)
for(i in 1:nr){
xx <- rexp(nx,rate=be)
yy <- rexp(ny,rate=al)
p.list[i] <- mle(xx,yy)
}
p.list
}


# Esempio:
# ris1 = sim_exp(5000,5,5,1,1)
# mean(ris1)
# mean(ris1-0.5)
# sd(ris1)

# Funzione per calcolare il test alla Wald
# Riceve i parametri xdat e ydat per i campioni e psi per l'ipotesi
# nulla.
Zep <- function(xdat, ydat, psi){
      my <- mean(ydat)
      mx <- mean(xdat)
      n <- length(ydat)
      m <- length(xdat)
      phat <- my/(mx+my)
      j <- n*m/((n+m)*phat^2*(1-phat)^2)
      Z <- (phat-psi)*sqrt(j)
      Z
}

# Funzione per calcolare la radice con segno di W_p(psi)
# Riceve gli stessi parametri della funzione Ze
rp <- function(psi, xdat, ydat){
      my <- mean(ydat)
      mx <- mean(xdat)
      ny <- length(ydat)
      nx <- length(xdat)
      psi.hat <- my/(mx+my)
      lp <- function(psi){ # -log verosimiglianza profilo
      log(((nx*mx*psi/(1-psi))+(ny*my))^(-nx-ny)*(psi/(1-psi))^nx)
      }
      W <- 2*(lp(psi.hat)-lp(psi)) # log rapporto di veros.
      r <- sign(psi.hat-psi)*sqrt(W) # radice con segno di W
```

```
}


# Funzione per nr simulazioni di due campioni di numerosità
# nx e ny e con distribuzione esponenziale con parametri al e be

sim_exp2 <- function(nr,nx,ny,be,al){
      Zep.list = rp.list = 0
      psi.v <- be/(al+be)
      print(psi.v)
      for (i in 1:nr){
            xx <- rexp(nx, rate=be)
            yy <- rexp(ny, rate=al)
            Z <- Zep(xx, yy, psi.v)
            # conta solo se rientra
            if (abs(Z)<qnorm(0.95)) Zep.list <- Zep.list+1
            r <- rp(psi.v, xx, yy)
            # conta solo se rientra
            if(abs(r) < qnorm(0.975)) rp.list <- rp.list+1
      }
      c(Zep.list/nr, rp.list/nr)
}

ris2 <- sim_exp2(5000,5,5,1,1)
ris2

# Esempio:
# ris2 = sim_exp2(5000,5,5,1,1)
# ris2
```

CODICI PER L'ALCL

```
# dataset
sani= c(0.23,0.44,0.19,0.08)  # X
malati= c(2.8,1.4,0.13,0.2,0.8,0.56,0.44,5.2,1.7,1.14)  # Y


#  psi=P(X<Y)
my = mean(malati)
mx = mean(sani)
round(my/(mx+my), digits=4)
```

BIBLIOGRAFIA

Adelchi Azzalini (2001), *Inferenza statistica, una presentazione basata sul concetto di verosimiglianza*, Springer-Verlag, Berlino.

Adelchi Azzalini, Bruno Scarpa (2004), *Analisi dei dati e data mining*, Springer-Verlag, Berlino.

Stefano M. Iacus, Guido Masarotto (2007), *Laboratorio di Statistica con R*, McGraw-Hill, Milano.

Samuel Kotz, Yan Lumelski, Marianna Pensky (2003), *The Stress-Strength Model and its Generalizations – Theory and Applications*, World Scientific, Singapore.

Luigi Pace, Alessandra Salvan (2001), *Introduzione alla Statistica – II Inferenza, Verosimiglianza, Modelli*, CEDAM, Padova.

Luigi Pace, Alessandra Salvan (1996), *Teoria della Statistica – Metodi, modelli approssimazioni asintotiche*, CEDAM, Padova.

Materiale e appunti del corso di Tecniche Statistiche di Classificazione, Facoltà di Scienze Statistiche, a cura dei Professori Monica Chiogna e Francesco Pauli.

Ringraziamenti

Innanzitutto ringrazio la professoressa Laura Ventura per la cortesia e la tanta pazienza che ha avuto nel guidarmi nella stesura di questa tesi.

Non posso però non ricordare anche tutte quelle persone che mi hanno accompagnato in questi tre anni e che mi hanno sostenuto (e soprattutto aggiungerei anche sopportato), in particolare: i Von Patuzzi al completo; Sara e Giulia, cioè le mie alpiniste preferite, Madda, Elena, insieme a tutti gli altri amici di Padova; i Livii, Max, Balda, Vil e Lancia, che nonostante le grandi distanze da anni mi sono sempre vicini; tutta la compagnia di Vienna, che ricordo con tanta nostalgia, e con loro anche la dott.ssa Mura, senza il cui aiuto non avrei potuto vivere quei momenti meravigliosi.

INDICE