Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per l'Economia e l'Impresa

Intervalli di confidenza per la sensibilità di un test diagnostico allo stadio iniziale di una malattia

Relatore: prof. Gianfranco Adimari
Dipartimento di Scienze Statistiche

Laureando: Chiara Pisani
Matricola n. 1222467

Anno Accademico 2021/2022

Indice

Introduzione

In diversi ambiti di studio quali il machine learning, il data mining e la psicometria, vengono affrontati problemi riguardanti la classificazione. Quest'ultima consiste nell'assegnare ciascun oggetto di un campione a un gruppo differente, in base alle sue caratteristiche.

In ambito medico e sanitario, tale aspetto è evidente nei test diagnostici, strumenti di supporto per la diagnosi di una malattia. Un test diagnostico vuole, infatti, classificare un paziente come sano o malato o individuare lo stadio della malattia in cui egli si trova, qualora vi siano diversi livelli di gravità. Un test diagnostico è tanto più affidabile quanto più è in grado di classificare correttamente i soggetti.

La letteratura è spesso incentrata su test diagnostici che hanno lo scopo di discriminare la popolazione dei sani da quella dei malati. In particolare, quando il test offre risultati su scala continua, si individua un valore soglia (*cut off*), ossia quel valore al di sotto del quale il paziente è classificato dal test come sano e al di sopra del quale è valutato malato. Si utilizza in tale contesto la curva ROC, strumento grafico per la valutazione di una classificazione binaria.

Nella realtà, però, ci si può trovare di fronte a problemi che implicano la classificazione in più livelli. In ambito medico esistono malattie il cui decorso si articola in più stadi di gravità. Può essere utile, ad esempio, utilizzare test diagnostici in grado di discriminare i soggetti in tre livelli che in questa tesi verranno indicati genericamente come "assenza di malattia", "stadio iniziale" e "malattia conclamata". Un particolare interesse viene riservato allo stadio iniziale: essere in grado di identificare la malattia a uno stadio iniziale, infatti, accresce la possibilità di rendere il trattamento efficace.

La struttura della tesi è la seguente.

Nel Capitolo 1 vengono presentati i test diagnostici per la classificazione binaria (sano/malato), soffermandosi sui principali indici di valutazione del test (§1.1). Sono introdotti i concetti di curva ROC(§1.2) e di AUC(§1.2.1), ossia l'area sottesa alla curva come indice di sintesi della ROC. Si fa poi un cenno al problema della scelta della soglia ottimale, concentrandosi in particolar modo sull'indice di Youden(§1.3).

Nel Capitolo 2 si estendono i concetti presentati precedentemente al caso di tre stadi della malattia: sensibilità e specificità vengono generalizzate dalle *True Class Fraction*(§2.1), la curva ROC diventa superficie ROC(§2.2) e si introduce il VUS(§2.2.1), vale a dire il volume sotto la superficie ROC. La letteratura presenta anche una generalizzazione dell'indice di Youden(§2.3) per la selezione dei due cut off che specificano i tre livelli.

Nel Capitolo 3 viene discusso il tema principale della tesi, ossia l'inferenza sulla sensibilità di un test diagnostico nello stadio iniziale della malattia. Vengono esposti diversi metodi per la costruzione di intervalli di confidenza della sensibilità allo stadio iniziale, quando la specificità e la sensibilità nel gruppo dei malati è fissata.

Capitolo 1

I test diagnostici

1.1 Test diagnostici e indici principali

Un test diagnostico è una procedura in grado di classificare un soggetto come sano o malato.
Alcuni esempi sono i test di screening (Pap test per il tumore del collo dell'utero, mammografia per il tumore al seno, ecc.), gli esami di laboratorio (misurazione della glicemia, dosaggio TSH, dosaggio PSA, ecc.) o gli esami diagnostici per immagini (Risonanza magnetica, TAC, Radiografia, ecc.).
I test diagnostici non sono infallibili; tuttavia, costituiscono uno strumento fondamentale per la diagnosi poiché la conoscenza del vero stato di salute di un paziente (tramite un *gold standard*[1]) può richiedere procedure costose e invasive e, talvolta, è possibile solo tramite autopsia. Per tale ragione, si ricorre a test diagnostici (meno costosi e più sicuri per il paziente), seppur soggetti a errori di classificazione.
Un test può offrire una risposta:

- **dicotomica**: di natura binaria, come la presenza o assenza di un sintomo;
- **quantitativa**: discreta o continua, fornisce una misura numerica, come la misurazione della glicemia nel sangue per la diagnosi del diabete.

[1]Un gold standard è un test privo di errore, che discrimina perfettamente i pazienti in sani e malati. Il gold standard è il test diagnostico più accurato a cui qualsiasi altro nuovo test deve rapportarsi per avere validità diagnostica.

Quando il risultato del test assume valori su scala continua, si sceglie una soglia (detta *cut off*) che dicotomizza il suo esito: solitamente si assume che il test sia positivo (e il paziente malato) per valori superiori al cut off e che il test sia negativo (e il paziente sano) altrimenti.
Conoscendo il vero stato di salute del paziente tramite l'utilizzo di un gold standard, per un valore soglia c prefissato, si individuano:

- **veri positivi** (TP, *True Positive*): pazienti malati che il test classifica come positivi;
- **veri negativi** (TN, *True Negative*): pazienti sani che il test classifica come negativi;
- **falsi positivi** (FP, *False Positive*): pazienti sani che il test classifica come malati;
- **falsi negativi** (FN, *False Negative*): pazienti malati che il test classifica come sani.

Le quattro tipologie di individui TP, TN, FP, FN possono essere rappresentati in una tabella a doppia entrata (*Tabella 1.1*) che riporta il numero di casi classificati in ciascuna modalità. Tale tabella è detta **matrice di confusione.**

	Paziente malato(M+)	**Paziente sano(M-)**	
Test Positivo (T+)	TP(Veri positivi)	FP(Falsi positivi)	TP+FP
Test Negativo(T-)	FN(Falsi negativi)	TN(Veri negativi)	FN+TN
Totale	TP+FN	FP+TN	

Tabella 1.1: Matrice di confusione

A partire dalla classificazione in *Tabella 1.1*, si possono definire degli indici utili per valutare complessivamente la bontà del test, per un valore fissato del cut off. Essi sono:

- **sensibilità**: la proporzione di pazienti malati che risultano positivi al test, data dal rapporto tra i veri positivi e il totale dei soggetti malati

$$Se = \frac{TP}{TP + FN}$$

- **specificità**: la proporzione di sani che risultano negativi, data dal rapporto tra i veri negativi e il totale di soggetti sani

$$Sp = \frac{TN}{FP + TN}$$

- **accuratezza**: la proporzione di soggetti classificati correttamente (malati come positivi, sani come negativi) sul totale dei pazienti considerati, ossia

$$Ac = \frac{TP + TN}{TP + FN + FP + TN}$$

- **potere predittivo positivo:** la proporzione di malati effettivi sul totale di pazienti risultati positivi al test, ossia

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

- **potere predittivo negativo:** la proporzione di veri sani sul totale di pazienti risultati negativi al test, ossia

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

Una rappresentazione alternativa del risultato del test (quando si esprime su scala continua) si ha utilizzando la distribuzione del test nel gruppo dei sani e in quello dei malati.

L'errore nella classificazione è dovuto a una sovrapposizione delle due distribuzioni, per cui la scelta del valore soglia individua in ogni caso falsi positivi o falsi negativi. Il caso limite è quello di una sovrapposizione completa, per cui il test assegna casualmente il soggetto alla popolazione dei sani o dei malati, come si vede in *Figura 1.1.* Nella realtà, però, si verifica spesso una sovrapposizione parziale (*Figura 1.2*), che porta a un margine d'errore.
Un test perfetto si ha, invece, quando non si hanno sovrapposizioni delle distribuzione, ed è il caso di un gold standard (*Figura 1.3*): esiste almeno un valore del cut off che consente di discriminare senza errore i sani e i malati.

Siano, dunque, X e Y due variabili casuali indipendenti che descrivono rispettivamente l'esito del test diagnostico quando esso viene effettuato sul gruppo dei sani e

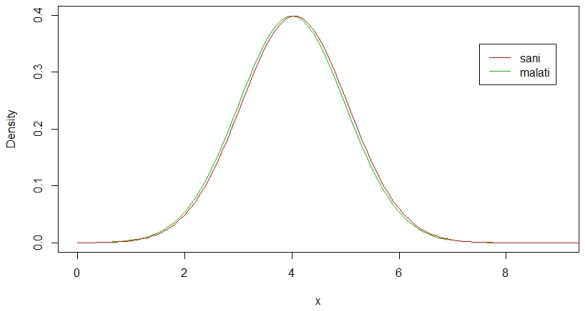

Figura 1.1: Distribuzione del test nei sani e malati quando il test non è informativo.

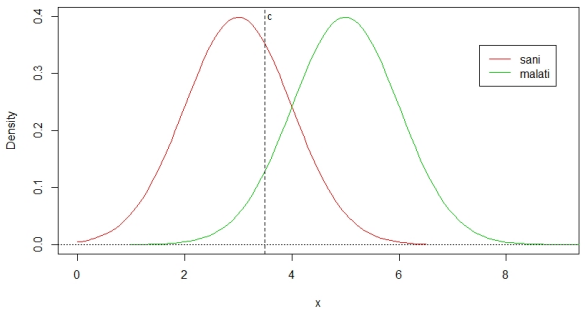

Figura 1.2: Distribuzione del test nei sani e malati. La sovrapposizione determina errori di classificazione.

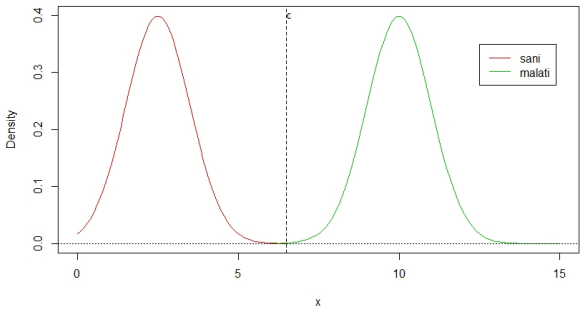

Figura 1.3: Distribuzione del test nei sani e malati nel caso di un gold standard.

dei malati (il cui vero stato di salute è determinato dal gold standard) e c il cut off. Siano F e G le funzioni di ripartizione di X e Y.
La sensibilità e la specificità sono definite rispettivamente come:

$$\mathbf{Se(c)} = \Pr\{Y > c\} = 1 - \Pr\{Y \leq c\} = 1\text{-}G(c)$$

$$\mathbf{Sp(c)} = \Pr\{X \leq c\} = F(c)$$

In questa ottica, la sensibilità è la probabilità che un malato risulti positivo al test e la specificità è la probabilità che un sano abbia esito negativo.

1.2 La curva ROC

La sensibilità e la specificità sono definite per un valore prefissato del cut off. Facendo variare la soglia su tutto il range dei possibili valori, si costruisce la curva ROC (dall'acronimo inglese *Receiver Operating Characteristics*), il principale strumento grafico per la valutazione della bontà di un test diagnostico.

La curva ROC si estende sul piano cartesiano individuato dagli assi *1-Specificità* e *Sensibilità*. La curva è formata, quindi, da tutte le possibili coppie $(1 - Sp(c), Se(c))$ ottenute al variare della soglia, dove l'ascissa del punto è il complemento a 1 della specificità e l'ordinata è la sensibilità per uno specifico valore del cut off. In altri termini, la curva ROC è la rappresentazione grafica di Se(c) in funzione di 1-Sp(c). Definendo:

$$1\text{-}Sp(c) = 1 - \Pr\{X \leq c\} = \Pr\{X > c\} = FPR(c)$$

$$Se(c) = \Pr\{Y > c\} = TPR(c)$$

dove FPR (*False Positive Rate*) è la proporzione di falsi positivi e TPR(c) (*True Positive Rate*) la proporzione di veri positivi per una soglia c, si può specificare la curva ROC come segue:

$$ROC(\cdot) = \{(FPF(c), TPF(c))\}, \text{ al variare di c}$$

Un test non è informativo quando la sua distribuzione non dipende dalla presenza o assenza della malattia: la distribuzione del test risulta identica nei due gruppi e, per qualsiasi valore del cut off, si verifica TPF(c)=FPF(c). Graficamente, le due distribuzioni sono completamente sovrapposte e la curva ROC coincide con la bisettrice nel quadrato di lato (0,1).

Al contrario, un test perfetto separa completamente i malati dai sani: le due distribuzioni non mostrano sovrapposizione, per cui per esiste un valore soglia c per il quale TPF(c)=1 e FPF(c)=0. Graficamente, le due distribuzioni sono separate e la curva ROC è formata dal segmento che collega l'origine del piano con il punto (0,1) e da quello che congiunge il punto (0,1) a (1,1).

Più comunemente, accade che le distribuzioni del test nei due gruppi siano parzialmente sovrapposte e la curva ROC è compresa tra le curve dei due casi limite.

I tre diversi casi sono riportati in *Figura 1.4*.

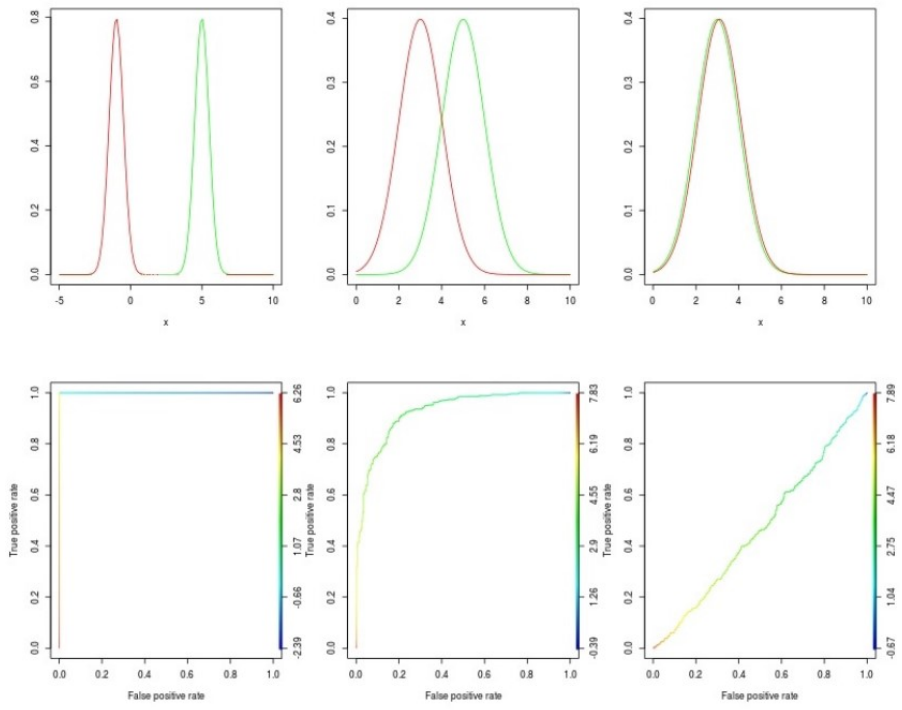

Figura 1.4: Distribuzioni del test nei sani (curva rossa) e malati(curva verde) e rispettiva curve ROC. A sinistra il caso di un gold standard, a destra il caso di un test non informativo. Al centro, il caso più comune.

Per una revisione più dettagliata dei metodi utilizzati per l'analisi della curva ROC, si rimanda a Shapiro (1999), Pepe (2003), Zhou e Qin (2005).

1.2.1 L'AUC

In generale, un test è tanto più informativo quanto più la curva ROC si avvicina al punto (0,1) o, ugualmente, tanto più essa si estende sopra la bisettrice del quadrato di lato (0,1). Per tale ragione, una delle misure di sintesi per verificare la capacità discriminatoria del test è l'AUC (*Area under the ROC Curve*), ossia l'area sottesta alla curva ROC:

$$\text{AUC} = \int_0^1 ROC(u)du$$

L'AUC può essere interpretata come la probabilità che, selezionando casualmente una coppia formata da un sano e un malato, il test diagnostico assuma un valore più alto per il soggetto malato, e quindi

$$\text{AUC} = \text{P}(\text{Y} > \text{X})$$

Tale formulazione è conosciuta come modello sollecitazione-resistenza[2](*stress-strenght model*).

L'AUC assume valori ragionevoli compresi tra 0.5 e 1. Convenzionalmente, per l'interpretazione dell'AUC si usa il criterio seguente [3] :

- AUC = 0.5: test non informativo (la curva ROC coincide con la bisettrice);
- $0.5 < \text{AUC} \leq 0.7$: test poco accurato;
- $0.7 < \text{AUC} \leq 0.9$: test moderatamente accurato;
- $0.9 < \text{AUC} < 1$: test molto accurato.

L'AUC, oltre a valutare il singolo test diagnostico, è utile anche per fare confronti tra test differenti: il test con AUC più elevato è quello preferibile. In *Figura 1.5* è riportato un confronto tra due test diagnostici: la curva rossa si estende al di sopra della blu (più vicina al punto ideale (0,1)), per cui il primo test è preferibile.

[2]Vedi Kotz e Pensky (2003)

[3]La classificazione è stata proposta da Swets (1988)

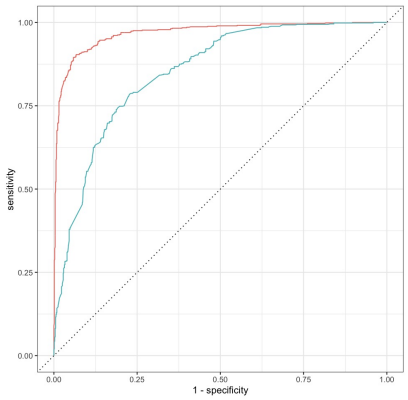

Figura 1.5: Confronto tra due curve ROC.

La letteratura presenta diversi metodi (parametrici e non) per la stima dell'AUC. Bamber (1975) propone un metodo non parametrico individuando una relazione tra l'AUC e la statistica di Mann-Whitney; tra i metodi parametrici, quello relativo al modello bi-normale è quello più noto (Faraggi e Reiser (2002)).

1.3 L'indice di Youden

Sebbene la curva ROC e l'AUC siano strumenti essenziali per la valutazione di un test, da sole non costituiscono un criterio per la scelta del cut off ottimale, ma mostrano esclusivamente la capacità del test di classificare correttamente i soggetti, al variare della soglia.

E' importante specificare, però, che la scelta della soglia non può essere dettata solamente da considerazioni probabilistiche volte a minimizzare la proporzione di classificazioni errate, ma è necessario valutare anche l'impatto medico e sanitario di tale decisione.

A seconda della malattia presa in analisi, si può preferire una sensibilità o una specificità maggiore. Ad esempio, quando la malattia ha un decorso grave e il trattamento è costoso o rischioso, si predilige avere il minor numero di falsi positivi, e dunque, una maggiore specificità per evitare di sottoporre inutilmente il paziente a una terapia nociva per la sua salute. Al contrario, quando si tratta di malattie ad alta contagiosità potrebbe essere opportuno minimizzare i falsi negativi, ossia privilegiare la sensibilità. Spostando il cut off su valori alti si migliora la specificità a

discapito della sensibilità; al contrario, spostandolo verso valori inferiori, si predilige la sensibilità alla specificità. Per tale ragione, è necessario trovare un compromesso tra le due.

Da un punto di vista statistico, esistono diversi approcci per la scelta del cut off ottimale basati sulla curva ROC. Il criterio più conosciuto è l'indice di Youden, introdotto da Youden nel 1950.

L'indice di Youden è definito come:

$$\mathrm{J} = max_c \mathrm{J(c)} = max_c [\mathrm{Se(c)} + \mathrm{Sp(c)} - 1]$$

e vuole massimizzare la somma della sensibilità e della specificità, o ugualmente, la differenza tra la sensibilità e il FPR (False Positive Rate) . Graficamente può essere interpretato come la massima distanza verticale tra un punto sulla bisettrice e un punto della curva ROC, come si vede in *Figura 1.6*.

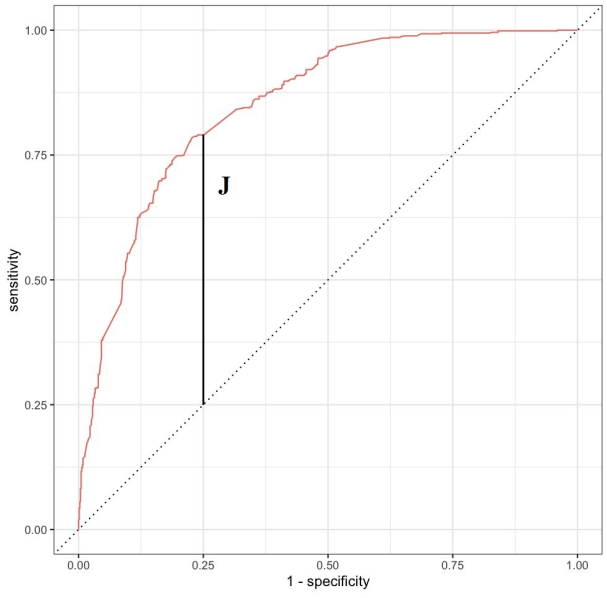

Figura 1.6: Indice J di Youden.

L'indice è definito per tutti i punti sulla curva ROC; massimizzando la funzione J(c) rispetto al cut off c, si individua quel punto che individua il massimo valore della somma di sensibilità e specificità.
L'indice J assume il suo valore massimo, pari a 1, quando un test diagnostico separa perfettamente la popolazione dei sani dalla popolazione dei malati, mentre assume

valore minimo, pari a 0, quando il test non porta informazione sullo stato di salute del soggetto.
Altri criteri proposti in letteratura sono:

- **Criterio della distanza minima**: si sceglie il punto sulla curva ROC che ha distanza minima dal punto (0,1) in corrispondenza del quale si ha massima sensibilità e minimo FPR. Essendo la distanza in questione una funzione del cut off c, il criterio seleziona il valore soglia c che minimizza tale distanza;

- **Punto di simmetria**: tale criterio va a massimizzare contemporaneamente le due classi di corretta classificazione;

- **Criterio dell'accuratezza**: la probabilità di corretta classificazione usata per valutare l'accuratezza del test può essere espressa come una funzione della sensibilità e della specificità per una certa soglia. La soglia che massimizza tale funzione è la soglia ottimale.

Essendo criteri differenti, individuano generalmente cut off ottimali differenti.
Per ulteriori approfondimenti si rimanda a Liu (2012).

Capitolo 2

Tre stadi della malattia

La letteratura è spesso incentrata su test diagnostici in grado di discriminare la popolazione dei sani da quella dei malati. Tuttavia, in molti casi, può essere utile classificare lo stato di salute di un paziente in tre o più livelli ordinali.
Nell'evoluzione di una malattia esiste, solitamente, uno stato di transizione che viene definito di seguito in modo generico come "stadio iniziale della malattia". Si consideri l'esempio dell'Alzheimer, una delle principali cause di demenza senile: lo stadio precoce è una fase di transizione tra i problemi dovuti all'invecchiamento e lo sviluppo di problemi più gravi, per cui non è sempre facile effettuare in tempo una diagnosi.
Di seguito, i tre stadi di avanzamento della malattia verranno indicati come: "assenza di malattia" (in cui non sono presenti sintomi), lo "stadio precoce" (in cui iniziano a comparire dei sintomi che potrebbero essere associati alla malattia) e la "malattia conclamata".
Tutti i concetti presentati nel capitolo precedentemente vengono ora generalizzati al caso di malattie con tre livelli di gravità: viene presentata la superficie ROC (metodologia introdotta da Scurfield (1996) e sviluppata negli anni successivi da Mossman (1999) e Heckerling (2001)) e i problemi di inferenza ad essa connessi.

2.1 Le TCF e FCF

Siano X, Y, Z variabili casuali che descrivono l'esito del test nei tre gruppi e F_x, F_y, F_z le rispettive funzioni di ripartizione, ipotizzando che i tre gruppi corrispondano ai tre stadi di avanzamento della malattia: assenza della malattia, stadio iniziale e malattia conclamata.
Nel caso di classificazione in tre livelli, è necessario individuare due livelli soglia c_1 e c_2 (supponendo $c_1 < c_2$) che vanno a determinare tre TCF(*True Class Fraction*, classi di corretta classificazione) e sei FCF(*False Class Fractions*, classi di errata classificazione).
Le tre **TCF** sono:

- la proporzione di sani classificati come nel primo gruppo, ossia

$$Pr(X \leq c_1)$$

- la proporzione di soggetti del secondo gruppo classificati come appartenenti al secondo gruppo, ossia:

$$Pr(c_1 < Y \leq c_2)$$

- la proporzione di soggetti con malattia conclamata classificati come appartenenti al terzo gruppo, ossia:

$$Pr(Z > c_2)$$

Le sei **FCF** sono, invece:

- la proporzione di pazienti sani che vengono classificati come appartenenti allo stadio iniziale, ossia

$$FCF_{xy} = Pr(c_1 < X \leq c_2)$$

- la proporzione di pazienti sani che vengono classificati come malati, ossia

$$FCF_{xz} = Pr(X > c_2)$$

- la proporzione di pazienti dello stadio iniziale classificati come sani

$$FCF_{yx} = Pr(Y \leq c_1)$$

- la proporzione di pazienti dello stadio iniziale classificati come malati, ossia

$$FCF_{yz} = Pr(Y > c_2)$$

- la proporzione di pazienti malati classificati come sani, ossia

$$FCF_{zx} = Pr(Z \leq c_1)$$

- la proporzione di pazienti malati classificati come appartenenti allo stadio iniziale, ossia

$$FCF_{zy} = Pr(c_1 < Z \leq c_2)$$

Tali risultati possono essere riportati nuovamente in una matrice di confusione 3x3, come quella mostrata in *Tabella 2.1.*

	Sani	**Stadio iniziale**	**malato**
Gruppo1	TCF_{xx}	FCF_{xy}	FCF_{xz}
Gruppo 2	FCF_{yx}	TCF_{yy}	FCF_{yz}
Gruppo 3	FCF_{zx}	FCF_{zy}	TCF_{zz}

Tabella 2.1: Matrice di confusione per la divisione in tre classi: in riga la classificazione ottenuta dal test, in colonna la vera condizione del paziente.

La presenza di errore di classificazione è dovuta alla sovrapposizione più o meno evidente delle tre funzioni di densità. Come si è visto nel capitolo precedente, se il test non è informativo, le tre curve risultano completamente sovrapposte, per cui il test assegna casualmente i soggetti a una delle tre classi. Quando, invece, il test è perfetto le curve sono completamente separate e ciò si traduce in una classificazione priva di errori. Il caso che si riscontra maggiormente nella realtà è, però, quello in cui esiste solo una parziale sovrapposizione, per cui sono presenti le TCF e (non necessariamente tutte) le FCF. I tre casi sono riportati in *Figura 2.1, 2.2, 2.3.*
Il concetto di accuratezza rimane invariato: un test è tanto più accurato quanto più l'errore di classificazione è ridotto.
Si può generalizzare ulteriormente al caso di una classificazione su n livelli ordinali: sono necessarie *n-1* soglie che individueranno n classi di corretta classificazione e $n^2 - n$ classi di errata classificazione.

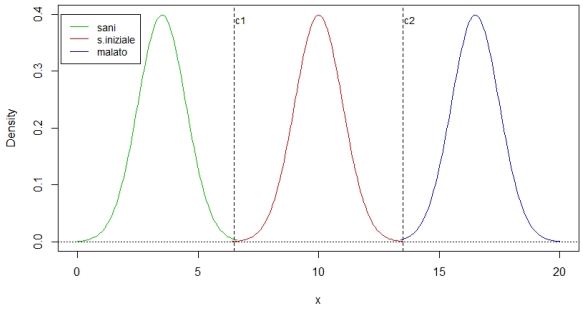

Figura 2.1: Distribuzione del test nei tre gruppi - test perfetto.

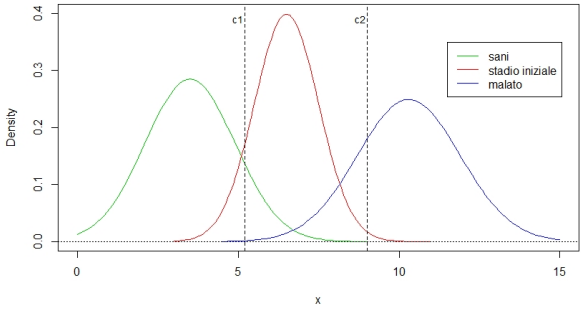

Figura 2.2: Distribuzione del test nei tre gruppi. E' presente una sovrapposizione parziale, per cui la scelta dei due cut off genera in ogni caso errori di classificazione.

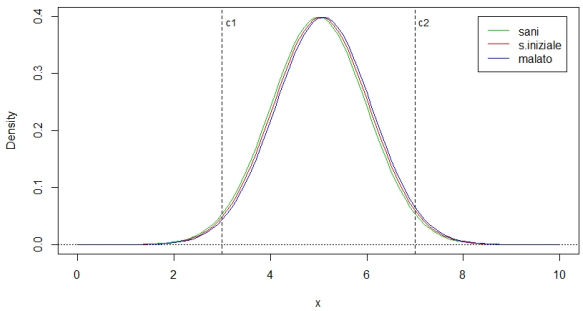

Figura 2.3: Distribuzione del test nei tre gruppi - test non informativo.

2.2 La superficie ROC

Siano nuovamente X, Y, Z le variabili casuali che descrivono i risultati del test diagnostico quando viene effettuato rispettivamente sui soggetti della classe 1,2,3. Dati i due cut off c_1 e c_2 (con $c_1<c_2$), le tre TCF si possono definire come:

$$\theta_1 = TCF1(c1) = Pr(X \leq c_1) = F_1(c_1)$$

$$\theta_2 = TCF2(c_1, c_2) = Pr(c_1 < Y \leq c_1) = F_2(c_2) - F_2(c_1)$$

$$\theta_3 = TCF3(c_2) = Pr(Z > c_2) = 1 - F_3(c_2)$$

dove $F_x(\cdot)$,$F_y(\cdot)$, $F_z(\cdot)$ sono le funzioni di ripartizione di X,Y,Z.
Nella classificazione dicotomica si è visto come, facendo variare la soglia c su tutto il suo range, le relative coppie di sensibilità e specificità vanno a delineare la curva ROC. Analogamente, facendo variare le due soglie c_1 e c_2 (sotto il vincolo di $c_1 < c_2$), si ottengono le corrispondenti TCF che consentono di costruire la superficie ROC, estensione della curva ROC al caso di una classificazione in tre livelli ordinali della malattia.
La superficie ROC è il luogo geometrico descritto da

$$\{F_x(c_1), F_y(c_2) - F_y(c_1), F_z(c_2)\} \text{ al variare di } c_1 \text{ e } c_2 \ (c_1 < c_2)$$

e quanto più la superficie si avvicina al punto di coordinate (1,1,1), tanto più la classificazione del test è accurata.

2.2.1 VUS: Volume sotto la curva ROC

Il VUS (*Volume Under the ROC Surface*), analogamente all'AUC, è un indice di valutazione del test diagnostico. Può essere visto come la probabilità che, scelti casualmente tre soggetti (ognuno da una delle tre classi), si verifichi X<Y<Z.
Esistono diversi metodi per la stima del VUS.

Un esempio di stimatore non-parametrico è:

$$\widehat{VUS} = \frac{1}{n_x n_y n_z} \sum_{i=1}^{n_x} \sum_{i=1}^{n_y} \sum_{i=1}^{n_z} I(x_i, y_i, z_i)$$

con

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

e n_1, n_2 e n_3 dimensioni campionarie dei tre gruppi.

Il VUS assume valore 1/6 quando le tre distribuzioni sono completamente sovrapposte (test non informativo), mentre assume valore massimo pari a 1 quando il test discrimina perfettamente nei tre livelli della malattia.

Per approfondimenti, si rimanda a Xiong *et al.* (2006).

2.3 Generalizzazione dell'indice di Youden

Per la classificazione in tre classi ordinali, è stata proposta una generalizzazione dell'indice di Youden per la scelta dei due livelli soglia.

Definenendo nuovamente X, Y, Z le distribuzioni del test nel gruppo dei sani, stadio iniziale e malattia e F_x, F_y, F_z le rispettive funzioni di ripartizione, c_1 e c_2 (sotto il vincolo di $c_1 < c_2$), l'indice di Youden è:

$$J = max_{c_1,c_2}\{TCF_x + TCF_y + TCF_z - 1\}$$

$$= \max_{c1,c\,2}\{F_x(c_1) + F_y(c_2) - F_y(c_1) - F_z(c_2)\}$$

L'indice J assume valori in $[0, 2]$.

Capitolo 3

Inferenza sulla sensibilità allo stadio iniziale della malattia

Tornando all'esempio dell'Alzheimer, i sintomi iniziali della patologia sono simili a quelli della demenza senile. Dal momento in cui nessun trattamento farmacologico per la malattia in questione è efficace a uno stadio avanzato, si può comprendere l'attenzione rivolta allo stadio iniziale, momento cruciale per la diagnosi della malattia: diagnosticarla in tempo vuol dire aumentare l'efficacia del percorso terapeutico scelto dal medico, per evitare peggioramenti irrimediabili.

In questo capitolo si considera il problema dell'inferenza sulla sensibilità allo stadio iniziale della malattia (ossia su P_2, il *True Positive Rate* della seconda classe). La stima di P_2 consente, infatti, di valutare la capacità di un test diagnostico di individuare la presenza di malattia quando essa non si è ancora manifestata con sintomi evidenti.

Si definiscono nuovamente le quantità

$$P_1 = F_1(c_1)$$

$$P_2 = F_2(c_2) - F_2(c_1) = F_2[F_3^{-1}(1 - P_3)] - F_2[F_1^{-1}(P_1)]$$

$$P_3 = 1 - F_3(c_2)$$

ossia la specificità, la sensibilità allo stadio iniziale e la sensibilità del test nei malati (o ugualmente, i TPR nelle tre classi). P_2 può essere descritta come funzione di P_1 e P_3: $P_2(P_1, P_3)$.

Quando P_1 e P_3 sono note, si può risalire al valore delle soglie c_1 e c_2 tramite semplice trasformazione inversa della funzione di ripartizione: $c_1{=}F_1^{-1}(P_1)$ e $c_2 = F_3^{-1}(1-P_3)$. Solitamente, si scelgono arbitrariamente dei valori per P_1 e P_3 (quali 80%, 90%, etc) in modo da ricavare i valori dei cut off per la classificazione nei tre gruppi e diventa possibile calcolare P_2 come funzione di quantità note.

Tale soluzione è analoga a quella utilizzara per i test con esito binario, in cui si fa inferenza sulla sensibilità definita per un valore del cut off che garantisca una specificità desiderata.[4]

Vengono presentati di seguito diversi approcci per la stima intervallare della sensibilità del test allo stadio iniziale: quelli descritti nei paragrafi §3.1 e §3.2 sono stati introdotti da Dong *et al.* (2011), i successivi (§3.3 e §3.4) da Dong e Tian (2015).

3.1 Metodi parametrici

3.1.1 GI: con assunzione di normalità

Il primo metodo parametrico preso in analisi si basa sull'ipotesi di distribuzione normale del test nei tre livelli considerati. Le osservazioni y_{ij} sono realizzazione di variabili casuali Y_{ij} indipendenti, con $Y_{ij} \sim \mathrm{N}(\mu_i\ ,\sigma_i^2)$.

Y_{1j} $(j = 1, ..., n_1)$, Y_{2j} $(j = 1, ..., n_2)$ e Y_{3j} $(j = 1, ..., n_3)$ sono i vettori casuali che descrivono l'esito del test nei gruppi 1,2,3 di numerosità n_1, n_2, n_3.

La quantità P_2 può essere espressa in funzione di P_1 e P_3 (fissati) come segue:

$$P_2 = \Phi\left[\frac{\mu_3 - \mu_2 + \Phi^{-1}(1-P_3)\sigma_3}{\sigma_2}\right] - \Phi\left[\frac{\mu_1 - \mu_2 + \Phi^{-1}(P_1)\sigma_1}{\sigma_2}\right] \qquad (3.1)$$

dove Φ sta a indicare la funzione di ripartizione della normale standard. Le distribuzioni del test nei gruppi 1 e 3 non sono note, perché non lo sono i veri valori delle rispettive medie e delle varianze. É possibile, però, stimarle tramite i valori dalle

[4] Greenhouse e Mantel (1950) presentano procedure inferenziali con e senza assunzione di distribuzione normale. Zhou e Qin (2005) introducono intervalli di confidenza non parametrici per la sensibilità. Qin *et al.* (2011) propongono un metodo basato sulla funzione di verosimiglianza empirica.

quantità campionarie: siano $\bar{Y}_i$ e S_i^2 la media e la varianza campionaria e $\bar{y}_i$ e s_i^2 i valori osservati corrispondenti. P_2 può essere così stimata:

$$\hat{P}_2 = \Phi\left[\frac{\hat{Y}_3 - \hat{Y}_2 + \Phi^{-1}(1-P_3)s_3}{s_2}\right] - \Phi\left[\frac{\hat{Y}_2 - \hat{Y}_1 + \Phi^{-1}(P_1)s_1}{s_2}\right] \tag{3.2}$$

Dal momento in cui, sotto ipotesi di normalità, vale:

$$V_i = \frac{(n_i-1)S_i^2}{\sigma_i^2} \sim \chi^2_{n_i-1}, i = 1,2,3$$

si può utilizzare la quantità pivotale generalizzata [5] per la varianza σ_i^2, ossia:

$$R_{\sigma_i^2} = \frac{(n_i-1)s_i^2}{Vi} \sim \frac{(n_i-1)s_i^2}{\chi^2_{n_i-1}}, i = 1,2,3 \tag{3.3}$$

Ugualmente, sapendo che:

$$Z_i = \frac{\bar{Y}_i - \mu_i}{\sqrt{\sigma_i^2/n_i}} \sim N(0,1)$$

la quantità pivotale geralizzata per la media μ_i è:

$$R_{\mu_i} = \bar{y}_i - Z_i\sqrt{R_{\sigma_i^2}/n_i}, i = 1,2,3 \tag{3.4}$$

Dunque, una quantità pivotale generalizzata per P_2 è:

$$R_{P_2} = \Phi\left[\frac{R_{\mu_3} - R_{\mu_2} + \Phi^{-1}(1-P_3)R_{\sigma_3}}{R_{\sigma_2}}\right] - \Phi\left[\frac{R_{\mu_1} - R_{\mu_2} + \Phi^{-1}(P_1)R_{\sigma_1}}{R_{\sigma_2}}\right] \tag{3.5}$$

Generando valori casuali V_i dalla distribuzione $\chi^2_{n_i-1}$ per ottenere le quantità pivotali generalizzate per la varianza R_{σ_i} e generando valori casuali Z_i da una distribuzione normale standard per ricavare le quantità pivotali generalizzate R_{μ_i} per la media, si genera un vettore di valori R_{P_2}. A partire da tale vettore, si individuano i quantili $R_{P_2}(\alpha/2)$ e $R_{P_2}(1-\alpha/2)$, che vanno a costituire gli estremi dell' intervallo di confidenza di livello $1-\alpha$ per P_2.

[5] Tsui e Weerahandi (1989) e Weerahandi (1995) introducono i concetti di variabili generalizzate e quantità pivotali generalizzate.

3.1.2 BCGI: inferenza generalizzata con trasformazione di Box-Cox

Nella maggior parte dei casi, la condizione di normalità non è rispettata. Qualora essa non sussista, è possibile ricorrere alla trasformazione di Box-Cox così definita:

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^{(\lambda)}-1}{\lambda}, \lambda \neq 0 \\ log(Y_i), \lambda = 0 \end{cases} \tag{3.6}$$

con $Y_i^{(\lambda)} \sim N(\mu_i, \sigma_i^2)$.
Un modo per determinare il valore ottimale del parametro λ consiste nel massimizzare la funzione di log-verosimiglianza costruita a partire dalle osservazioni nei tre campioni:

$$\sum_{i=1}^{3}\sum_{j=1}^{n_i}\left[-\frac{1}{2}log(2\pi) - \frac{Y_{ij}^{\lambda}-\mu_i}{2\sigma_i^2} + (\lambda-1)logY_{ij}\right] \tag{3.7}$$

Una volta individuato il valore di λ, si applica la trasformazione 3.6 e si procede con il metodo discusso nel paragrafo 3.1.1.

3.2 Metodi non parametrici

I metodi parametrici descritti sopra richiedono che sia rispettata la condizione di normalità o quanto meno che la trasformazione di Box-Cox sia applicabile. I metodi non parametrici consentono, invece, di evitare tale problema poiché non richiedono ipotesi sulla distribuzione del test.

BTP

Supponiamo che le funzioni di distribuzione del test nel primo e nel terzo gruppo F_1 e F_3 siano note. Si può allora costruire l'intervallo di confidenza di Wald di livello 1-α per P_2

$$(\bar{P}_2 - z_{1-\alpha/2}\sqrt{\bar{P}_2(1-\bar{P}_2)/n_2}, \bar{P}_2 + z_{1-\alpha/2}\sqrt{\bar{P}_2(1-\bar{P}_2)/n_2})$$

in cui

$$\bar{P}_2 = \frac{\sum_{j=1}^{n_2} I_{[F_1^{-1}(P_1) \leq Y_{2j} \leq F_3^{-1}(1-P_3)]}}{n_2} \tag{3.8}$$

presenta al numeratore una somma di Bernoulli con probabilità di successo $P_2 = P[F_1^{-1}(P_1) \leq Y_i \leq F_3^{-1}(1-P_3)]$. $\bar{P}_2$ rappresenta la quota di osservazioni da Y_2 che cadono tra i due quantili $F_1^{-1}(P_1)$ e $F_3^{-1}(1-P_3)$.

Dato che le distribuzioni nei due gruppi non sono note, $F_1^{-1}(P_1)$ e $F_3^{-1}(1-P_3)$ devono essere sostituite dalle loro stime $\hat{F}_1^{-1}(P_1)$ e $\hat{F}_3^{-1}(1-P_3)$. La stima per P_2 è ora:

$$\hat{\bar{P}}_2 = \frac{\sum_{j=1}^{n_2} I_{[\hat{F}_1^{-1}(P_1) \leq Y_{2j} \leq \hat{F}_3^{-1}(1-P_3)]}}{n_2} \tag{3.9}$$

e il nuovo intervallo alla Wald di livello 1-α è

$$(\hat{\bar{P}}_2 - z_{1-\alpha/2}\sqrt{\hat{\bar{P}}_2(1-\hat{\bar{P}}_2)/n_2}, \hat{\bar{P}}_2 + z_{1-\alpha/2}\sqrt{\hat{\bar{P}}_2(1-\hat{\bar{P}}_2)/n_2})$$

Tuttavia, gli intervalli di Wald non sono particolarmente indicati specialmente per campioni di piccole dimensioni. Si può ricorrere alternativamente a una procedura bootstrap: si sfruttano i percentili della distribuzione di $\hat{\bar{P}}_2^b$ (quantità analoga a $\hat{\bar{P}}_2$ nel "mondo bootstrap") e

$$(\hat{\bar{P}}_2^b(\alpha/2), \hat{\bar{P}}_2^b(1-\alpha/2))$$

è l'intervallo di confidenza ottenuto con il metodo **BTP**.

BTI

Un'altra soluzione sfrutta l'intervallo di confidenza proposto da Agresti e Coull (1998), che gode di buone proprietà per l'inferenza su proporzioni. Un intervallo di confidenza per P_2 di questo tipo è:

$$(\tilde{P}_2 - z_{1-\alpha/2}\sqrt{\widehat{Var_{AC}}(\tilde{P}_2)}, \tilde{P}_2 + z_{1-\alpha/2}\sqrt{\widehat{Var_{AC}}(\tilde{P}_2)})$$

dove

$$\tilde{P}_2 = \frac{\sum_{j=1}^{n_2} I_{[F_1^{-1}(P_1) \leq Y_{2j} \leq F_3^{-1}(1-P_3)]} + z_{1-\alpha/2}^2/2}{n_2 + z_{1-\alpha/2}^2} \tag{3.10}$$

e

$$\widehat{Var}_{AC}(\tilde{P}_2) = \frac{\tilde{P}_2(1-\tilde{P}_2)}{n_2 + z^2_{1-\alpha/2}/2} \tag{3.11}$$

Non è possibile, però, definire $\tilde{P}_2$ non essendo note le funzioni di ripartizione F_1 e F_3. Sostituendole con le rispettive quantità empiriche, si stima P_2 come:

$$\hat{\tilde{P}}_2 = \frac{\sum_{j=1}^{n_2} I_{[\hat{F}_1^{-1}(P_1) \leq Y_{2j} \leq \hat{F}_3^{-1}(1-P_3)]} + z^2_{1-\alpha/2}/2}{n_2 + z^2_{1-\alpha/2}} \tag{3.12}$$

Il nuovo intervallo di confidenza è:

$$(\hat{\tilde{P}}_2 - z_{1-\alpha/2}\sqrt{\widehat{Var}_{AC}(\hat{\tilde{P}}_2)}, \hat{\tilde{P}}_2 + z_{1-\alpha/2}\sqrt{\widehat{Var}_{AC}(\hat{\tilde{P}}_2)})$$

La varianza può essere stimata semplicemente sostituendo $\tilde{P}_2$ con $\hat{\tilde{P}}_2$ nella 3.11, oppure tramite metodo bootstrap.

La varianza bootstrap era già stata utilizzata da Zhou e Qin (2005) per il caso di un test su scala continua che classifica sani e malati, quando si vuole calcolare l'intervallo di confidenza per la sensibilità quando la specificità è fissata. Si estende, quindi, al caso di classificazione in tre stadi, quando si vuole fare inferenza sulla sensibilità allo stadio iniziale, fissati i livelli di specificità e di sensibilità allo stadio di malattia conclamata. La stima della varianza bootstrap è definita come:

$$\widehat{Var}^{boot}(\hat{\tilde{P}}_2) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tilde{P}}_2^b - \bar{\hat{\tilde{P}}}_2^b)^2 \tag{3.13}$$

con

$$\bar{\hat{\tilde{P}}}_2^b = \frac{1}{B}\sum_{b=1}^{B}\hat{\tilde{P}}_2^b \tag{3.14}$$

dove B è il numero di campioni bootstrap. L'intervallo di confidenza con metodo BTI è dunque:

$$(\hat{\tilde{P}}_2 - z_{1-\alpha/2}\sqrt{\widehat{Var^{boot}}(\hat{\tilde{P}}_2)}, \hat{\tilde{P}}_2 + z_{1-\alpha/2}\sqrt{\widehat{Var^{boot}}(\hat{\tilde{P}}_2)})$$

BTII

La terza soluzione non parametrica sfrutta nuovamente una procedura bootstrap. Sostituendo $\hat{\tilde{P}}_2$ con la media $\bar{\hat{\tilde{P}}}_2^b$, l'intervallo di confidenza **BTII** di livello 1-α è:

$$(\bar{\hat{\tilde{P}}}_2^b - z_{1-\alpha/2}\sqrt{\widehat{Var^{boot}}(\hat{\tilde{P}}_2)}, \bar{\hat{\tilde{P}}}_2^b + z_{1-\alpha/2}\sqrt{\widehat{Var^{boot}}(\hat{\tilde{P}}_2)})$$

3.3 Metodo basato sulla normalità asintotica

Il metodo descritto di seguito è una generalizzazione dei risultati presentati da Linnet (1987), il quale fornisce una formula per la varianza della stima della sensibilità (fissata la specificità).

APV

In presenza dei tre stadi della malattia, la varianza di $\hat{\hat{P}}_2$ (definita come in 3.9) è:

$$\sigma^2_{\hat{\hat{P}}_2} = \frac{P_2(1-P_2)}{n_2} + \frac{P_1(1-P_1)}{n_1}\frac{f_2^2[F_1^{-1}(P_1)]}{f_1^2[F_1^{-1}(P_1)]} + \frac{P_3(1-P_3)}{n_3}\frac{f_2^2[F_3^{-1}(1-P_3)]}{f_3^2[F_3^{-1}(1-P_3)]} \quad (3.15)$$

Tuttavia, come detto altre volta in precedenza, difficilmente le distribuzioni del test nei gruppi 1 e 3 sono note, per cui devono essere stimate a partire dalle osservazioni. La funzione di densità f_i viene sostituita dalla stima kernel di densità e la funzione di ripartizione F_i dalla funzione di distribuzione cumulata empirica. La stima per la varianza è:

$$\widehat{\sigma^2_{\hat{\hat{P}}_2}} = \frac{\hat{\hat{P}}_2(1-\hat{\hat{P}}_2)}{n_2} + \frac{P_1(1-P_1)}{n_1}\frac{\hat{f}_2^2[\hat{F}_1^{-1}(P_1)]}{\hat{f}_1^{\,2}[\hat{F}_1^{-1}(P_1)]} + \frac{P_3(1-P_3)}{n_3}\frac{\hat{f}_2^{\,2}[\hat{F}_3^{-1}(1-P_3)]}{\hat{f}_3^{\,2}[\hat{F}_3^{-1}(1-P_3)]} \quad (3.16)$$

É possibile dimostrare che per n_1, n_2 e n_3 grandi, $\hat{\hat{P}}_2$ ha approssimativamente una distribuzione normale di media P_2 e varianza definita in 3.15 (stimabile tramite la 3.16).

L'intervallo di confidenza di livello approssimato 1-α di $\hat{\hat{P}}_2$ è quindi:

$$(\hat{\hat{P}}_2 - z_{1-\alpha/2}\sqrt{\widehat{\sigma^2_{\hat{\hat{P}}_2}}}, \hat{\hat{P}}_2 + z_{1-\alpha/2}\sqrt{\widehat{\sigma^2_{\hat{\hat{P}}_2}}})$$

3.4 Metodi basati sulla verosimiglianza empirica

Il metodo ora descritto sfrutta invece la verosimiglianza empirica, introdotta da Owen (2001). Si definiscono di seguito la funzione ϕ come

$$\phi(Y_1, Y_2, Y_3) = \begin{cases} 1 & \text{se } Y_1 < Y_2 < Y_3 \\ \frac{1}{2} & \text{se } Y_1 = Y_2 < Y_3 \text{ o } Y_1 < Y_2 = Y_3 \\ \frac{1}{6} & \text{se } Y_1 = Y_2 = Y_3 \\ 0 & \text{altrimenti} \end{cases}$$

e la variabile casuale $U = \phi[F_1^{-1}(P_1), Y, F_3^{-1}(1 - P_3)]$, dove Y $= (Y_1, Y_2, Y_3)$.
Si individua la relazione tra U e la sensibilità del test allo stadio iniziale P_2: $E[U] = P_2$.
Infatti, $E(U) = E\{\phi[F_1^{-1}(P_1), Y, F_3^{-1}(1-P_3)]\} = P[F_1^{-1}(P_1) < Y_2 < F_3^{-1}(1-P_3)] = P[F_1^{-1}(P_1) < Y_2 \leq F_3^{-1}(1 - P_3)] = P_2$.
Per cui, $E[U(Y) - P_2]$ è una funzione non distorta di stima per P_2 che può essere utilizzata per costruire la funzione di verosimiglianza empirica dalla quale ottenere la stima di massima verosimiglianza empirica di P_2. Dalle osservazioni del secondo gruppo, si ottiene il vettore $\mathbf{p} = (p_1, p_2, ..., p_{n_2})$ delle probabilità, tali che $\sum_{i=1}^{n_2} p_i = 1$ e $p_i \geq 0$ per ogni i .
Si può, dunque, scrivere la funzione di verosimiglianza empirica per P_2 come segue:

$$\tilde{L}(P_2) = sup \left\{ \prod_{i=1}^{n_2} p_i : \sum_{i=1}^{n_2} p_i = 1, \sum_{i=1}^{n_2} p_i(U_i - P_2) = 0 \right\}$$

dove $U_i = \phi[F_1^{-1}(P_1), Y_i, F_3^{-1}(1 - P_3)]$, i=1,2,...,n_2.
É possibile dimostrare che il log-rapporto di verosimiglianza empirica $l(P_2)$ segue una distribuzione χ_1^2. Tuttavia, le distribuzioni del test nel gruppo 1 e 3 non sono note, per cui non si possono definire le quantità U_i ma è necessario stimarle, a partire dalle funzioni di ripartizioni empiriche nei due gruppi interessati.
La stima di U_i è $\hat{U}_i = \phi[\hat{F}_1^{-1}(P_1), Y, \hat{F}_3^{-1}(1 - P_3)]$ e la fuzione di verosimiglianza empirica profilo per P_2 diventa:

$$L(P_2) = sup \left\{ \prod_{i=1}^{n_2} p_i : \sum_{i=1}^{n_2} p_i = 1, \sum_{i=1}^{n_2} p_i(\hat{U}_i - P_2) = 0 \right\}$$

Il log-rapporto di verosimiglianza empirica per P_2 è definito come:

$$l(P_2) \equiv -2logr(P_2) = 2\sum_{i=1}^{n_2} log\{1 + \tilde{\lambda}(\hat{U}_i - P_2)\}$$

dove $\tilde{\lambda}$ è soluzione di

$$\frac{1}{n_2}\sum_{i=1}^{n_2}\left\{\frac{\hat{U}_i - P_2}{1 + \tilde{\lambda}(\hat{U}_i - P_2)}\right\} = 0$$

Il log-rapporto di verosimiglianza empirica $l(P_2)$ non segue più una distribuzione χ^2_1, ma la sostituzione delle funzioni di ripartizione con le corrispondenti quantità empiriche comporta solamente una trasformazione scalare della distribuzione χ^2_1 secondo un parametro r_{P_1,P_2,P_3}. Dunque

$$r_{P_1,P_2,P_3} l(P_2) \overset{L}{\longmapsto} \chi^2_1$$

Il parametro di scala è definito come:

$$r_{P_1,P_2,P_3} = \frac{\sigma^2_{\hat{U}_i}}{n_2 \sigma^2_{\hat{\bar{P}}_2}}$$

con $\sigma^2_{\hat{U}_i} = P_2(1 - P_2)$ e $\sigma^2_{\hat{\bar{P}}_2}$ quantità non note e da stimare.
$\sigma^2_{\hat{U}_i} = P_2(1 - P_2)$ può essere stimata come $\hat{\bar{P}}_2(1 - \hat{\bar{P}}_2)$, mentre $\sigma^2_{\hat{\bar{P}}_2}$ può essere stimata come in 3.16 o con metodo bootstrap.

ELP

Quando la varianza $\sigma^2_{\hat{\bar{P}}_2}$ è stimata come in 3.16, l'intervallo di confidenza di livello $1 - \alpha$ è così costruito:

$$CI_\alpha(P_2) = \{P_2 : r^*_{P_1,P_2,P_3} l(P_2) \leq \chi^2_1(1 - \alpha)\}$$

con

$$r^*_{P_1,P_2,P_3} = \frac{\hat{\bar{P}}_2(1 - \hat{\bar{P}}_2)}{n_2 \widehat{\sigma^2}_{\hat{\bar{P}}_2}}$$

ELB

In alternativa, la varianza per $\hat{\bar{P}}_2$ può essere stimata con metodo bootstrap come:

$$\widehat{\sigma^2}^b_{\hat{\bar{P}}_2} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\bar{P}}^b_2 - \bar{\hat{\bar{P}}}^b_2)^2$$

dove $\hat{\hat{P}}_2^{\,b}$ è la versione bootstrap di $\hat{\hat{P}}_2$.

L'intervallo di confidenza è così definito:

$$CI_\alpha(P_2) = \{P_2 : r^*_{P_1,P_2,P_3} l(P_2) \leq \chi^2_1(1-\alpha)\}$$

dove

$$r^*_{P_1,P_2,P_3} = \frac{\hat{\hat{P}}_2(1-\hat{\hat{P}}_2)}{n_2 \widehat{\sigma^2}^{\,b}_{\hat{\hat{P}}_2}}$$

.

Conclusioni

L'obiettivo perseguito in questa tesi è stato quello di approfondire lo studio di test diagnostici in grado di classificare i pazienti in tre diversi stadi di avanzamento della malattia. In particolare, la tesi è una rassegna di alcuni strumenti proposti di recente per fare inferenza sulla sensibilità allo stadio iniziale di una malattia, quando quest'ultima si manifesta attraverso tre livelli di gravità.
Per malattie di questo tipo, la sensibilità allo stadio iniziale (fissate la specificità e la sensibilità nel terzo gruppo) è un'importante misura di accuratezza del test: maggiore è P_2 (sensibilità allo stadio iniziale), più il test è in grado di identificare la malattia a uno stadio primitivo. Per tale ragione, un'accurata stima intervallare per P_2 aiuta i ricercatori anche a identificare dei buoni biomarcatori quando l'obiettivo è quello di studiare le caratteristiche dei soggetti per i quali la malattia si presenta allo stadio iniziale.
I metodi utilizzati sono in sintesi:

- **GI** e **BCGI**: metodi parametrici che sfruttano l'assunzione di normalità del test nei tre gruppi e quando essa non è rispettata, viene applicata preliminarmente una trasformazione di Box-Cox ai dati.

- **BTP**, **BTI**, **BTII**: metodi non parametrici che non richiedono dunque assunzioni sulla distribuzione del test nei tre gruppi. I tre metodi sfruttano delle procedure bootstrap; gli ultimi due, in particolare, utilizzano l'intervallo di confidenza proposto da Agresti e Coull (1998), con varianza stimata a partire da campioni bootstrap.

- **APV**: metodo parametrico che sfrutta la normalità asintotica della stima della sensibilità; nello specifico, si utilizza il risultato presentato da Linnet (1987)

per la stima della varianza necessaria per la costruzione dell'intervallo.

- **ELP** e **ELB**: metodi basati sulla costruzione della funzione di verosimiglianza empirica.

Alcuni studi di simulazione per il confronto tra i diversi metodi discussi nella tesi sono riportati in Dong e Tian (2015).

Bibliografia

Agresti A.; Coull B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **52**(2), 119–126.

Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, **12**(4), 387–415.

Dong T.; Tian L. (2015). Confidence interval estimation for sensitivity to the early diseased stage based on empirical likelihood. *Journal of biopharmaceutical statistics*, **25**(6), 1215–1233.

Dong T.; Tian L.; Hutson A.; Xiong C. (2011). Parametric and non-parametric confidence intervals of the probability of identifying early disease stage given sensitivity to full disease and specificity with three ordinal diagnostic groups. *Statistics in medicine*, **30**(30), 3532–3545.

Faraggi D.; Reiser B. (2002). Estimation of the area under the roc curve. *Statistics in medicine*, **21**(20), 3093–3106.

Greenhouse S. W.; Mantel N. (1950). The evaluation of diagnostic tests. *Biometrics*, **6**(4), 399–412.

Heckerling P. S. (2001). Parametric three-way receiver operating characteristic surface analysis using mathematica. *Medical Decision Making*, **21**(5), 409–417.

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

Linnet K. (1987). Comparison of quantitative diagnostic tests: type i error, power, and sample size. *Statistics in Medicine*, **6**(2), 147–158.

Liu X. (2012). Classification accuracy and cut point selection. *Statistics in medicine*, **31**(23), 2676–2686.

Mossman D. (1999). Three-way rocs. *Medical Decision Making*, **19**(1), 78–89.

Owen A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC.

Pepe M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.

Qin G.; Davis A. E.; Jing B.-Y. (2011). Empirical likelihood-based confidence intervals for the sensitivity of a continuous-scale diagnostic test at a fixed level of specificity. *Statistical Methods in Medical Research*, **20**(3), 217–231.

Scurfield B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*, **40**(3), 253–269.

Shapiro D. E. (1999). The interpretation of diagnostic tests. *Statistical methods in medical research*, **8**(2), 113–134.

Swets J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, **240**(4857), 1285–1293.

Tsui K.-W.; Weerahandi S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, **84**(406), 602–607.

Weerahandi S. (1995). Generalized confidence intervals. In *Exact statistical methods for data analysis*, pp. 143–168. Springer.

Xiong C.; van Belle G.; Miller J. P.; Morris J. C. (2006). Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Statistics in medicine*, **25**(7), 1251–1273.

Zhou X.-H.; Qin G. (2005). Improved confidence intervals for the sensitivity at a fixed level of specificity of a continuous-scale diagnostic test. *Statistics in medicine*, **24**(3), 465–477.