UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA SPECIALISTICA IN

SCIENZE STATISTICHE ECONOMICHE FINANZIARIE E AZIENDALI

TESI DI LAUREA

VALUTAZIONE DI TEST DIAGNOSTICI IN ASSENZA DI GOLD STANDARD

Relatore:

CH.MA PROF.SSA

MONICA CHIOGNA

Laureanda:

CARLA SPESSATO

N.matricola: 530454-SEA

Anno Accademico 2010/2011

Ad Alessandro

"Questi problemi sono classificati come probabilità delle cause
e sono i più importanti di tutti per le loro applicazioni scientifiche.
Un effetto potrebbe essere prodotto dalla causa "a" o dalla causa "b".
L'effetto è appena stato osservato.
Ci domandiamo la probabilità che sia dovuto alla causa "a":
questa è una probabilità di causa a posteriori.
Ma non la potrei calcolare,
se una convenzione più o meno giustificata
non mi dicesse in anticipo
quale è la probabilità a priori che la causa "a" entri in gioco."
(Henri Poincarè).

INDICE

INTRODUZIONE

Nello scenario medico, i test diagnostici giocano un ruolo importante nella cura della salute. Nuovi test diagnostici per la scoperta di infezioni virali e batteriche sono continuamente sviluppati e necessitano di essere confrontati con quelli esistenti. Un test viene valutato in base alla sua capacità diagnostica di distinguere accuratamente in una popolazione i soggetti sani da quelli malati. Una tecnica ampiamente utilizzata per valutare il comportamento di un test diagnostico in una popolazione (in termini di sensibilità e specificità) è l'analisi ROC, in funzione di determinati valori di soglia. La capacità discriminante di un test è proporzionale all'estensione dell'area sottesa alla curva ROC (AUC), che fornisce quindi un indicatore sintetico dell'accuratezza di un test diagnostico. La differenza tra le AUC può essere usata come una misura per confrontare l'accuratezza tra due test diagnostici. La performance di un test diagnostico è idealmente valutata mediante il confronto con un *gold standard* in grado di determinare il vero stato di malattia per ogni paziente indipendentemente dal risultato del test. Per molte malattie è difficile o impossibile stabilire una diagnosi definitiva, quindi nell'usuale ambiente delle pratiche mediche un perfetto *gold standard* potrebbe non esistere, essere troppo costoso o impraticabile da realizzare. Si parla allora di *imperfect gold standard* quando il vero stato di malattia non è noto, ma è indicato dal migliore test disponibile. Tuttavia, quando un *imperfect gold standard* è usato al posto di un *gold standard*, l'accuratezza del test spesso è soggetta ad errore, che si traduce in stime distorte di cui si deve tener conto. Sebbene la letteratura offra numerosi studi di metodologie statistiche sviluppate per valutare l'accuratezza di un nuovo test diagnostico quando un *gold standard* esiste, l'inferenza statistica per l'analisi ROC senza un *gold standard* test rimane pressoché inesplorata.

Obiettivo della tesi è proprio la valutazione di test diagnostici fallibili quando non si dispone di un *gold standard*. In particolare, vengono presentati due approcci per la stima dell'accuratezza che sviluppano e definiscono per la prima volta in modo rigoroso il concetto di stato di malattia come variabile latente, che comporta l'applicazione dell'algoritmo EM per la stima dei parametri.

Nel capitolo 1 vengono introdotte la nomenclatura e le definizioni fondamentali che saranno utili alla comprensione della tesi, inerenti ai test diagnostici e alle loro caratteristiche, con riguardo all'analisi ROC.

Il capitolo 2 offre una rassegna dei metodi statistici proposti per stimare i tassi di errore e la prevalenza di test di screening o diagnostici. È una rivisitazione degli approcci esistenti più significativi quando il vero stato di malattia non è noto per qualche soggetto, che si focalizzano nella maggior parte dei casi su test diagnostici che producono risultati binari. I metodi descritti sono stati raggruppati in due classi: quelli che

assumono l'ipotesi che due test siano indipendenti, condizionatamente al vero stato della malattia e quelli che rilassano tale assunzione, ammettendo correlazione tra i test, caso che si verifica quando i test sono basati su un comune fenomeno biologico. L'assunzione di indipendenza condizionata è tipicamente adottata nei primi approcci sviluppati, per la sua semplicità, ma successivamente è stata abbandonata da vari autori perché non è realistica in molte situazioni pratiche. I modelli proposti cercano di risolvere il problema della mancata identificabilità del modello ponendo restrizioni su alcuni parametri o adottando altre strategie. Per esempio, nel caso d'indipendenza condizionata si possono applicare più di due test in una popolazione, due test in una popolazione con introduzione di un terzo test "resolver" nell'analisi discrepante, due test in due popolazioni, più di due test in s popolazioni, oppure utilizzare approcci bayesiani. Nel caso in cui si tenesse conto della dipendenza condizionata, sono stati proposti un metodo con più di due classi latenti e un modello ad effetti casuali o strategie bayesiane, laddove non sia possibile avere risultati da più test in quanto costosi, dispendiosi di tempo o invasivi.

Il capitolo 3 è il cuore della tesi e descrive due metodi che introducono l'uso del vero stato di malattia come variabile latente per la valutazione di test diagnostici in assenza di *gold standard,* mediante l'applicazione dell'algoritmo EM. La maggior parte dei metodi statistici per la stima dell'accuratezza diagnostica di uno o più nuovi test in assenza di *gold standard* è basata su stime di massima verosimiglianza per la curva ROC e usa un modello normale multivariato di mistura latente. Il limite maggiore di questo approccio è che si assume che le variabili casuali latenti per molteplici test su scala ordinale seguano una distribuzione normale multivariata, ma non sempre questa assunzione è supportata dall'evidenza empirica. Per superare questo problema, nel capitolo 3 viene allora descritto il metodo di Zhou et al.(2005) per stimare non parametricamente curve ROC e le rispettive AUC senza *gold standard* di test su scala ordinale quando il numero di test è maggiore di due. Inoltre si descrive il metodo di Hsieh et al. (2009) per la stima intervallare della differenza di due AUC appaiate in assenza di un *gold standard*, sotto l'assunzione di normalità dei risultati del test da ogni gruppo di soggetti malati, usando l'algoritmo EM congiuntamente al metodo *bootstrap*. Tale approccio si propone di superare i limiti emersi nel modello bayesiano di Choi et al. (2006), che aveva evidenziato come i metodi bayesiani potessero essere sensibili all'assunzione di distribuzione parametrica bivariata sui risultati del test e dipendessero troppo pesantemente dalla specificazione delle distribuzioni a priori e dall'area di sovrapposizione delle due curve, oltre che dall'entità della correlazione.

CAPITOLO 1

"L'errore nasce sempre dalla tendenza

dell'uomo a dedurre la causa dalla conseguenza"

(Arthur Schopenhauer)

VALUTAZIONE DI TEST DIAGNOSTICI

1.1 TEST DIAGNOSTICI

In tutti i campi della scienza vengono sistematicamente messe a punto e utilizzate procedure più o meno complesse e della più svariata natura, ma sempre ben codificate, allo scopo di verificare un'ipotesi. Tali procedure vengono comunemente chiamate "test". In particolare in epidemiologia i test rappresentano lo strumento di base nelle operazioni di screening, eseguite cioè su popolazioni presuntivamente sane (e nelle quali la *prevalenza* della malattia in studio è ignota) allo scopo di identificare precocemente la presenza di infezioni o di malattie subcliniche. Anche nell'attività diagnostica di routine i test rappresentano elementi fondamentali, e spesso determinanti, nel processo decisionale volto a confermare (o escludere) la presenza di una determinata malattia già sospettata in base ai dati clinici. Nel settore della medicina veterinaria, ai suddetti due vasti campi di azione se ne può aggiungere un terzo che spazia nell'opera di controllo degli alimenti di origine animale, dove i test vengono applicati per determinare sanità e salubrità (es. assenza di agenti patogeni, sostanze chimiche ecc.) dei suddetti prodotti.

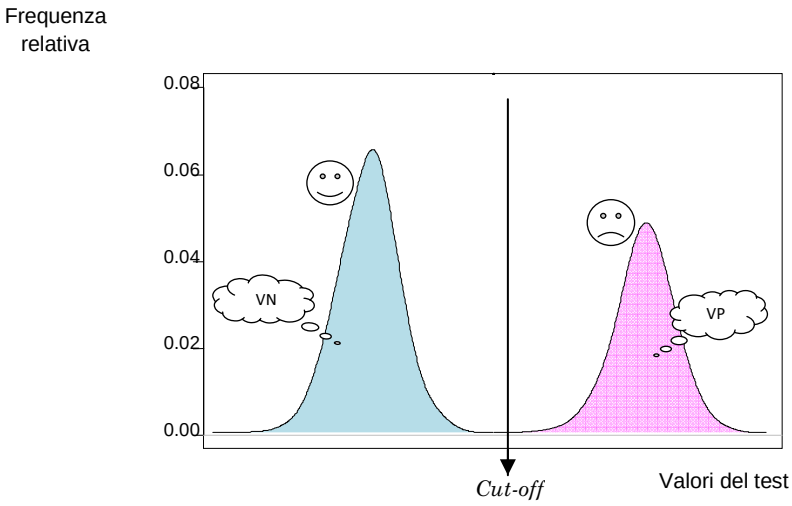
In base alla tipologia di responso fornito, i test possono essere classificati in due categorie. Alla prima appartengono i test "qualitativi", ossia che restituiscono un *output* (risposta) categoriale (es. positivo/negativo, vero/falso ecc.); la seconda, numericamente più consistente, comprende i test di tipo

quantitativo, ossia che producono risultati sotto forma di variabili numeriche "discrete" (es. titolo anticorpale su diluizioni seriali di siero) o "continue" (es. test sierologici ELISA su singola diluizione).

Due sono i requisiti che un test, sia esso clinico, strumentale o di laboratorio, dovrebbe soddisfare: affidabilità e validità. Per affidabilità generalmente s'intende la capacità di un test di offrire sempre lo stesso risultato nel corso di misurazioni ripetute. Questa è una caratteristica intrinseca al test e dipende dalla bontà dello strumento e/o dell'operatore. Ma i test sono comunemente utilizzati con l'obiettivo di riconoscere, relativamente a una qualche patologia, i soggetti malati da quelli sani. La validità (o *performance*), è proprio la capacità diagnostica del test di distinguere accuratamente in una popolazione i soggetti sani da quelli malati. In questo senso, un test si definisce ideale se distingue perfettamente tutti i soggetti sani dai malati.

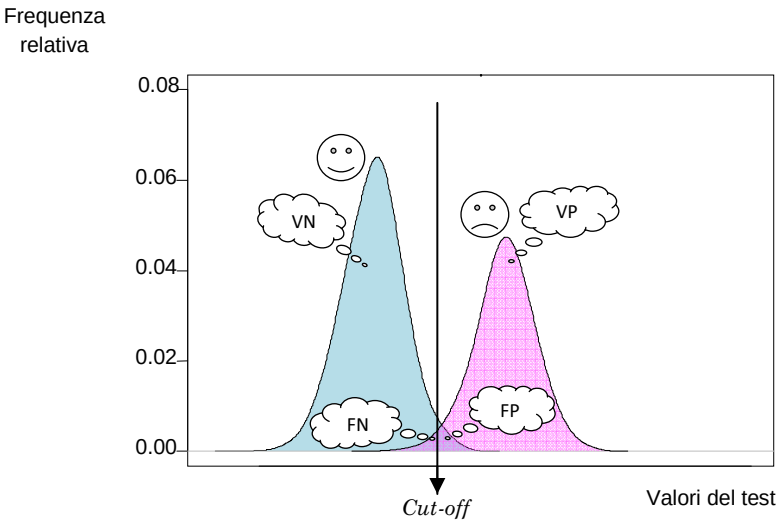
Nella situazione più semplice, un esame diagnostico fornisce un risultato che può essere espresso come "positivo" o "negativo". Per i test quantitativi (siano essi discreti o continui), occorre individuare sulla scala di lettura un valore-soglia ("*cut-off*", "*cut-point*" o "*threshold*") che discrimini i risultati da dichiarare "positivi" da quelli "negativi", cioè quel valore assunto dalla variabile misurata nel test al di sopra del quale il soggetto viene dichiarato malato e al di sotto del quale viene definito sano. Ciò consente di categorizzare in "positivi" e "negativi" la gamma di tutti i possibili risultati e di equiparare l'interpretazione di un test quantitativo a quella di un test qualitativo. La Figura 1.1 rappresenta la distribuzione di un test ideale nei sani e nei malati. In questo caso le due distribuzioni sono separate ed è facile individuare sull'asse delle ascisse il valore di *cut-off* capace di distinguere con precisione assoluta le due popolazioni: tutti i soggetti sani hanno valori del test inferiori al *cut-off* e risultano negativi al test ("veri negativi", VN), mentre tutti i soggetti malati hanno valori del test superiori al *cut-off* e risultano positivi al test ("veri positivi", VP).

Figura 1.1. Distribuzione degli esiti di un ipotetico test nelle classi di individui malati, ☹, e non malati, ☺, senza sovrapposizione inter-classe.



In realtà accade che, quando si sottopone una popolazione ad una procedura diagnostica, per un certo valore di *cut-off*, non tutti i soggetti malati risulteranno positivi al test, così come non tutti i soggetti sani risulteranno negativi. Questo problema genera incertezza nell'interpretazione di un test perché – nella grande maggioranza dei casi – esiste una zona di sovrapposizione fra le distribuzioni dei risultati del test medesimo applicato in popolazioni di individui rispettivamente sani ed ammalati (Figura 1.2). Si avrà sempre un certo numero di soggetti sani che risulteranno positivi al test ("falsi positivi", FP) e, simmetricamente, un certo numero di soggetti malati che il test non riuscirà ad identificare come tali, e pertanto saranno erroneamente classificati come "sani" ("falsi negativi", FN). Nella pratica, quindi, si verifica sempre una transvariazione più o meno ampia nelle due distribuzioni ed è perciò impossibile individuare sull'asse delle ascisse un valore di *cut-off* che consenta una classificazione perfetta, ossia tale da azzerare sia i falsi positivi che i falsi negativi. Dato che un esame non è infallibile, allora un test diagnostico è tanto più *accurato* quanto più è piccola la probabilità che esso produca dei "falsi".

Figura 1.2. Distribuzione degli esiti di un ipotetico test nelle classi di individui malati, ☹, e non malati, ☺, con sovrapposizione inter-classe.



È possibile rappresentare questo tipo di situazione utilizzando una tabella di contingenza di tipo 2×2, che confronta l'output del test in esame con il vero stato dei soggetti (Tabella 1.1). Il vero stato dei soggetti può già essere noto in partenza (ad esempio, nel caso di una malattia infettiva, saggiando gruppi di individui sicuramente esenti dall'infezione oppure *Specific Pathogen Free* ed altri sottoposti ad infezione sperimentale) oppure può essere stabilito per mezzo di un test di referenza provvisto della più alta attendibilità ("test aureo" o "*golden test*" o "*gold standard"*), possibilmente basato su un principio biologico diverso rispetto al test da valutare. In genere i *golden test* presentano alcuni svantaggi (es. difficile somministrazione, rischio per il soggetto da testare, costo elevato, ecc.) che li rendono inapplicabili di *routine* o in operazioni di *screening*.

Ai fini del raffronto con il test in studio, nell'ipotesi più semplice si assume che il *golden test* fornisca risultati perfettamente corrispondenti alla verità.
Sono possibili quattro risultati a seconda della posizione del valore di *cut-off*:

- se il risultato del test è positivo e anche il valore vero è positivo, viene chiamato "vero positivo" (VP);
- se invece il risultato del test è positivo ma il valore vero è negativo, viene chiamato "falso positivo" (FP);
- contrariamente, si ottiene un "vero negativo" (VN) quando entrambi, il risultato e il valore vero, sono negativi;
- si ha un "falso negativo" (FN), invece, quando il risultato è negativo ma il valore vero è positivo.

La validità del test può essere misurata tramite la proporzione di falsi positivi e negativi: tanto più basse saranno le loro quote, tanto più il test sarà valido. L'accuratezza del test è definita come

$$Accuratezza = \frac{(VP + VN)}{(T_P + T_N)},$$

dove T_P e T_N è il totale di individui con risultato positivo e negativo del test, rispettivamente.

Tabella 1.1. Rappresentazione su una tabella 2×2 della distribuzione di una popolazione in base ai risultati di un test.
$M+ =$ malati; $M- =$ sani; $T+ =$ test positivo; $T- =$ test negativo, $T_{M+} =$ totale dei malati, $T_{M-} =$ totale dei sani, $T_P =$ totale dei soggetti con test positivo, $T_N =$ totale dei soggetti con test negativo.

		Vero stato		
		$M+$	$M-$	*Totale*
Risultato del test	$T+$	VP	FP	T_P
	$T-$	FN	VN	T_N
	Totale	T_{M+}	T_{M-}	N

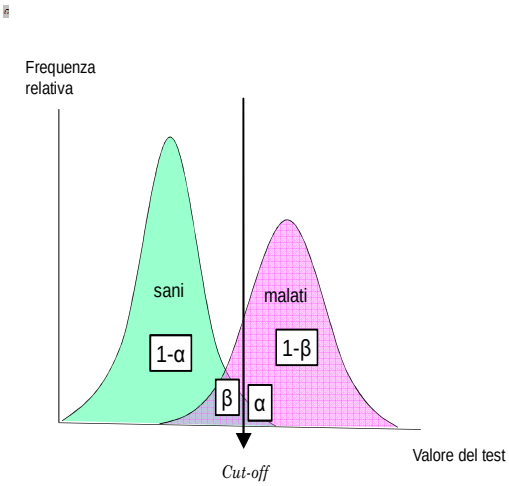
1.2 CARATTERISTICHE DI UN TEST DIAGNOSTICO

Si indichi con "D" la malattia d'interesse ("disease"). Una persona nello studio di test diagnostici può essere malata $(D = 1)$ o non malata $(D = 1)$. La prevalenza della malattia nella popolazione è $P(D = 1) = p$: essa rappresenta semplicemente la proporzione di malati nella popolazione (malati della specifica malattia nel momento in cui si utilizza il test).
Il risultato di un test T può essere positivo $(T = 1)$ o negativo $(T = 0)$.

1.2. CARATTERISTICHE DI UN TEST DIAGNOSTICO

Il tasso di falso positivo (α) è la probabilità che una persona sana abbia un test positivo, vale a dire $\alpha = P(T = 1|D = 0)$. Il tasso di falso negativo (β) è la probabilità che una persona malata abbia un risultato negativo del test: $\beta = P(T = 0|D = 1)$.

Tabella 1.3. Rappresentazione grafica della sensibilità $(1 - \beta)$, specificità $(1 - \alpha)$ e relativi errori (α, β) sulla distribuzione degli esiti di un ipotetico test nelle classi di individui malati e non malati, con sovrapposizione inter-classe.



Le misure più comunemente usate per valutare la performance di un test sono la sua *sensibilità* e *specificità*. Per *sensibilità* si intende la capacità di un test di individuare in una popolazione i soggetti malati: è la probabilità che il risultato di un test sia positivo per una persona malata. Essa è data dalla proporzione dei soggetti realmente malati e positivi al test (veri positivi) rispetto all'intera popolazione dei malati.

$$sensibilità = P(T = 1|D = 1) = \frac{VP}{T_{M+}} = \frac{VP}{VP + FN} = 1 - \beta.$$

La sensibilità è condizionata negativamente dalla quota di falsi negativi: pertanto un test molto sensibile dovrà associarsi ad una quota molto bassa di falsi negativi, ovvero di soggetti malati che "sfuggono" all'identificazione attraverso il test. Il calcolo della sensibilità considera esclusivamente la popolazione dei malati (ovvero la prima colonna della tabella 2×2), in funzione dell'identificazione come positivi o negativi al test.

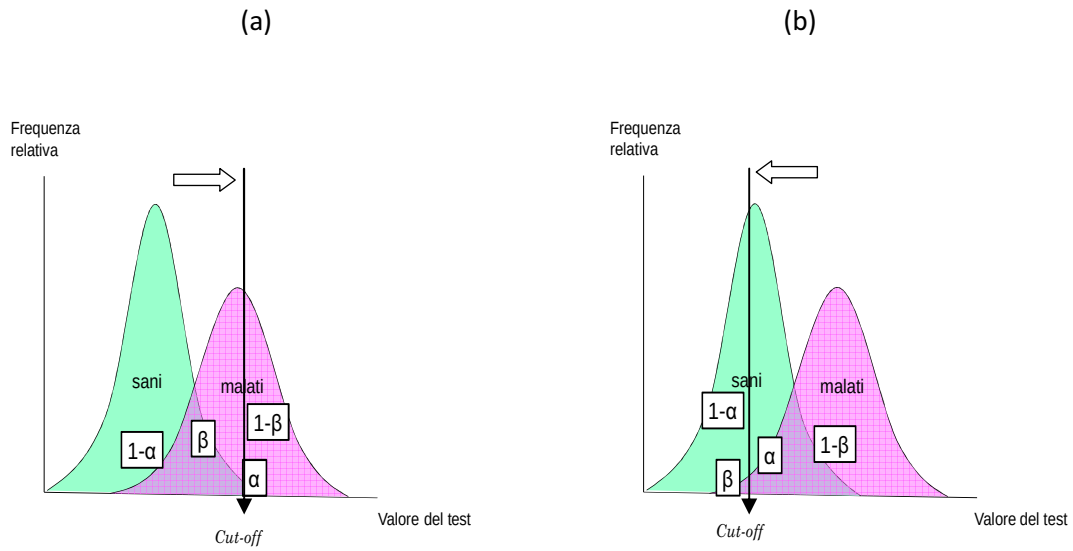
Un secondo parametro, per certi versi speculare al precedente, è dato dalla *specificità*. Per specificità si intende la capacità di un test di identificare come negativi i soggetti sani: è la probabilità che il risultato di un test sia negativo per una persona non malata.

$$specificità = P(T = 0|D = 0) = \frac{VN}{T_{M-}} = \frac{VN}{VN + FP} = 1 - \alpha.$$

La specificità è influenzata in particolare dalla quota di falsi positivi, ovvero un test sarà tanto più specifico quanto più bassa risulterà la quota di falsi positivi, cioè di soggetti sani identificati dal test come malati. Un test molto specifico ci consente di limitare la possibilità che un soggetto sano risulti positivo al test. Per calcolare la specificità dovremo fare riferimento esclusivamente al gruppo dei sani ed alla loro distribuzione fra positivi e negativi al test (ovvero la seconda colonna della tabella 2×2). Un test altamente specifico sarà dunque un test che produrrà una bassa quota di falsi positivi.

E' facile verificare che sensibilità e specificità sono due parametri reciprocamente dipendenti, in quanto sono fra loro inversamente correlati in rapporto alla scelta del valore di *cut- off.* In altre parole, l'adozione di una soglia che offre un'elevata sensibilità comporta una perdita di specificità e viceversa. Se il test in questione fosse rappresentato dalla misurazione di una variabile continua, una maniera per aumentare la specificità sarebbe quello di aumentare il limite di *cut-off*, ovvero il livello al di sopra del quale "etichettare" un soggetto come malato. In ogni caso, ovviamente, la reale distribuzione della popolazione fra malati e sani in funzione della variabile misurata non cambierebbe, pertanto spostando la soglia a destra (Figura 1.4-a) si avrebbe una riduzione globale dei soggetti positivi al test con un conseguente aumento della quota di falsi negativi, cioè di soggetti realmente malati che vengono identificati come sani. Essendo aumentata la quota di falsi negativi, diminuirebbe quindi la sensibilità. In maniera analoga, per garantirci di poter riconoscere la quota più alta possibile di soggetti malati, potremmo invece aumentare la sensibilità del test (Figura 1.4-b). In tal caso, inevitabilmente, abbassando il livello di *cut-off*, includeremmo nel gruppo dei positivi un certo numero di sani (la coda destra della curva dei sani) che rappresenterebbero i falsi positivi: diminuirebbe pertanto la specificità del test.

Figura 1.4. Sensibilità e specificità sono parametri fra loro inversamente proporzionali. (a) Aumentando la specificità, il *cut-off* viene spostato verso destra e la sensibilità si riduce; (b) aumentando la specificità, il *cut-off* viene spostato verso sinistra e la specificità si riduce.

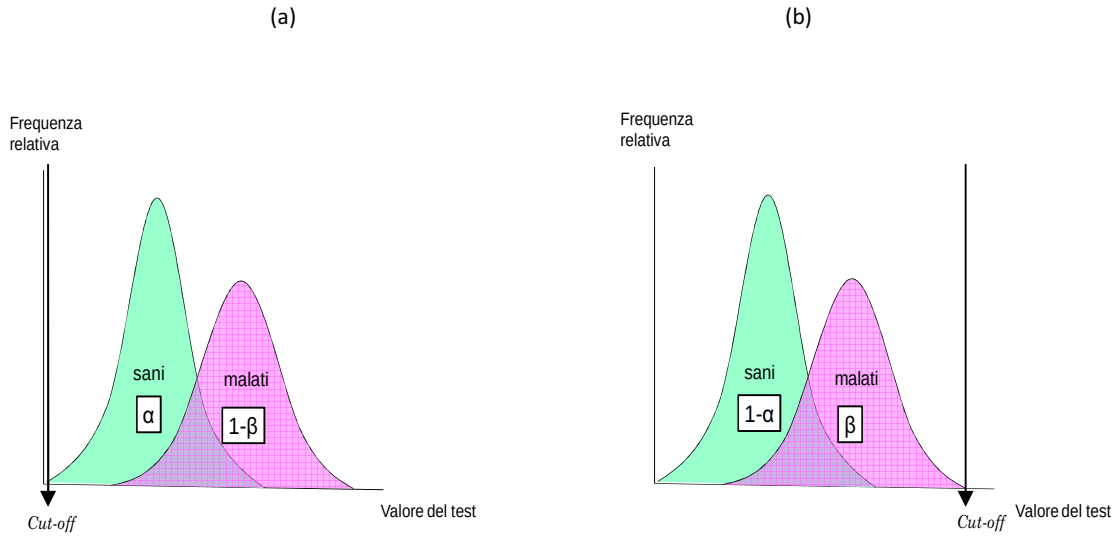


1.2. CARATTERISTICHE DI UN TEST DIAGNOSTICO

E' possibile dimostrare che, quando la distribuzione dei valori nelle due classi malati-sani è di tipo normale, la "soglia discriminante ottimale", ossia il valore di *cut-off* che minimizza gli errori di classificazione, è pari al valore in ascissa corrispondente al punto d'intersezione delle due distribuzioni (Bottarelli e Parodi, 2003).

Nei casi limite, se il *cut-off* è posto al suo valore minimo (Figura 1.5-a), la sensibilità $(1-\beta)$ sarà pari a uno, ma anche la quota di falsi negativi (α), mentre la specificità $(1-\alpha)$ e la quota di falsi negativi (β) saranno nulli. La situazione opposta si verifica nel caso in cui il livello di *cut-off* sia posto all'estremo superiore del range di valori possibili (Figura 1.5-b), con specificità uguale a uno e sensibilità nulla.

Figura 1.5. Sensibilità e specificità nei casi limite. (a) sensibilità = 1 e specificità = 0; (b) specificità = 1 e specificità = 0.



La scelta del *cut-off* non può essere dettata soltanto da considerazioni di ordine probabilistico volte a minimizzare la proporzione di classificazioni errate, ma è necessario basarsi anche sul prevedibile impatto di tipo sanitario, economico, sociale, ecc. di ciascuno dei due tipi di errata classificazione (falsi positivi e falsi negativi). Ad esempio, per malattie ad alta contagiosità potrebbe essere opportuno minimizzare la quota di falsi negativi, quindi privilegiare la sensibilità a scapito della specificità. Viceversa, in altre situazioni (es. malattie non contagiose, trattabili soltanto con una terapia molto costosa) il prezzo di un falso positivo sarà verosimilmente superiore rispetto a quello di un falso negativo, quindi il *cut-off* verrà determinato in modo da privilegiare la specificità. Un metodo empirico comunemente utilizzato per la scelta del *cut-off* consiste nel fissare a priori il valore desiderato di specificità (generalmente > 0.9) e quindi nel calcolare la corrispondente sensibilità del test nella suddetta condizione. Questo approccio genera tuttavia due effetti collaterali negativi. Il primo è rappresentato dall'evenienza che il test in questione possa produrre risultati complessivamente migliori attraverso l'adozione di un *cut-off* diverso da quello scelto. Il secondo è legato

all'impossibilità di effettuare un raffronto affidabile fra la performance di due o più test valutati in base ad un singolo valore di *cut-off*. Pertanto, si configura un evidente e non trascurabile inconveniente pratico quando si tratta di scegliere fra uno o più test e, in subordine, vengono ostacolati gli studi di meta-analisi nei quali, come è noto, vengono effettuate comparazioni qualitative fra i risultati ottenuti in studi diversi sullo stesso argomento. Alle difficoltà ora accennate, è da sovrapporre un ulteriore elemento che ostacola sia la scelta del *cut-off* ottimale per un singolo test che il raffronto tra le performance di test diversi. Tale elemento è costituito dal fatto che i valori predittivi dipendono, oltre che dalla specificità e specificità del test, anche dalla *prevalenza* della malattia nella popolazione studiata. Nel complesso, le suddette osservazioni comportano tre importanti implicazioni:

1. è possibile scegliere un valore di *cut-off* tale che risponda ad un predeterminato valore di sensibilità o di specificità, ma non è detto che tale valore sia ottimale per gli scopi contingenti;
2. la sensibilità e la specificità associate ad un singolo valore di *cut-off* non rappresentano descrittori esaurienti della performance del test potenzialmente ottenibile adottando altri valori di *cut-off*;
3. i valori predittivi in quanto dipendenti dalla prevalenza della malattia nella popolazione studiata, non sono caratteristiche intrinseche del test e quindi non possono essere utilizzati come descrittori esaurienti della performance dei test.

1.2.1 *Prevalenza della malattia e predittività*

Sensibilità e specificità sono parametri definibili a priori, perché sono caratteristiche intrinseche del test in quanto dipendono esclusivamente dalla tipologia di test adottato. Esse ci informano su qual è la probabilità di reclutare soggetti malati o sani da una certa popolazione di partenza (di malati e di sani), mentre nulla ci dicono sulla probabilità che abbiamo, di fronte ad un singolo risultato positivo, che quel soggetto sia realmente malato. Soprattutto nel campo dell'epidemiologia clinica, cioè quando i test vengono utilizzati a scopo diagnostico e non in operazioni di screening, ancor più interessanti risultano altri due parametri: il valore predittivo positivo (VPP) e il valore predittivo negativo (VPN). VPP è definita come la probabilità che un soggetto scelto casualmente dalla popolazione, risultato positivo al test, sia effettivamente malato: $P(D = 1|T = 1)$; esso si calcola come quota di soggetti veri positivi sul totale dei positivi:

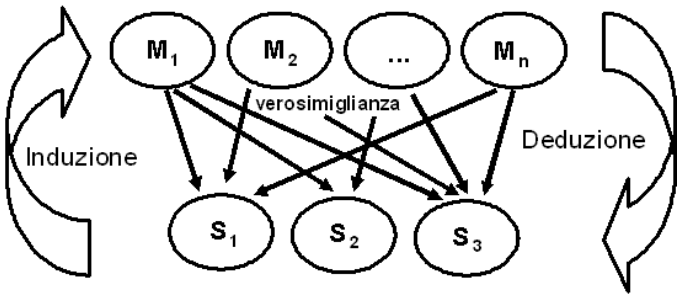
$$VPP = \frac{VP}{T_P} = \frac{VP}{VP + FP}.$$

VPN esprime la probabilità che un soggetto risultato negativo ad un test sia effettivamente sano: $P(D = 0|T = 0)$; si calcola come la quota di veri negativi sul totale dei negativi:

$$VPN = \frac{VN}{T_N} = \frac{VN}{VN + FN}.$$

A questa situazione si può applicare lo schema induzione/deduzione rappresentato nella Figura 1.6:

Figura 1.6 Schema induzione/deduzione.



Conoscendo le cause (le malattie $M_1, M_2, \dots, M_n$) possiamo da queste dedurre gli effetti corrispondenti: è quanto viene fatto con lo studio della patologia medica. La soluzione del problema inverso, ovvero l'induzione, viene mostrata con la clinica medica, che insegna come affrontare una logica ribaltata, davanti al problema cardine che si incontra nella pratica medica quotidiana: un malato con una serie di sintomi e/o segni (S_1, S_2, S_3), a partire dai quali dobbiamo risalire alla malattia. L'induzione è oltremodo più difficile, in quanto ad un sintomo e/o segno possono corrispondere più cause in termini di malattie. Il teorema di Bayes consente, a partire dai sintomi/segni osservati, di calcolare la verosimiglianza delle cause/malattie, espressa nell'unico modo possibile: in termini di probabilità. Esso mostra che sia VPP che VPN dipendono dalle caratteristiche del test (sensibilità e specificità) così come dalla prevalenza (p) della malattia nella popolazione attraverso la seguente relazione:

$$VPP = P(D = 1|T = 1) = \frac{P(T = 1|D = 1)\, P(D = 1)}{P(T = 1)}$$

$$= \frac{P(T = 1|D = 1)\, P(D = 1)}{P(T = 1|D = 1)\, P(D = 1) + P(T = 1|D = 0)\, P(D = 0)}$$

$$= \frac{(1 - \beta)p}{(1 - \beta)p + \alpha(1 - p)},$$

Analogamente,

$$VPN = P(D = 0|T = 0)$$

$$= \frac{P(T = 0|D = 0)\, P(D = 0)}{P(T = 0|D = 0)\, P(D = 0) + P(T = 0|D = 1)\, P(D = 1)}$$

$$= \frac{(1 - \alpha)(1 - p)}{(1 - \alpha)(1 - p) + \beta p}.$$

La prevalenza, o meglio, la probabilità di incontrare un paziente con una certa malattia, si definisce come "probabilità pre-test", perché è rilevata prima di eseguire il test. Il valore predittivo del test positivo è la probabilità a posteriori, dopo avere eseguito il test. La differenza tra queste due probabilità è il valore aggiunto in termini d'informazione che il test fornisce alla diagnosi. E' sconsigliato l'uso come test di screening in una popolazione non selezionata di marcatori tumorali che forniscono alla diagnosi uno scarso valore aggiunto informativo. Il teorema di Bayes rappresenta l'unico strumento che consente di fornire una misura quantitativa, quindi oggettiva, espressa in termini d'informazione, del valore aggiunto fornito da un test diagnostico. Dal punto di vista epistemologico risultano evidenti nel teorema di Bayes:

- la capacità di formalizzare il meccanismo con cui si ribalta la logica, da *l'effetto*|*data la causa* a la *causa*|*dato l'effetto*, passando dalla deduzione all'induzione (che risulta per definizione solo probabile);
- la capacità di formalizzare il meccanismo con cui l'informazione fornita dall'esperienza (lo specifico risultato del test di laboratorio) si somma all'informazione a priori, aumentando la nostra conoscenza;
- la capacità di *misurare l'informazione* che un test diagnostico fornisce alla diagnosi medica.

Dal punto di vista pratico risulta evidente dal teorema di Bayes che:

- in condizioni di bassa prevalenza diminuisce il valore predittivo del test positivo;
- in condizioni di bassa specificità del test diminuisce il valore predittivo del test positivo;
- in condizioni di bassa specificità e di bassa prevalenza aumenta il valore predittivo del test negativo, quindi un test diventa utile soprattutto per escludere la malattia.

Si può aggiungere una considerazione sulla capacità del teorema di Bayes di formalizzare il meccanismo con cui l'informazione fornita dall'esperienza (lo specifico risultato del test di laboratorio) si somma all'informazione a priori, aumentando la nostra conoscenza. Infatti, in alternativa all'interpretazione frequentista appena illustrata, nella quale la probabilità a priori è rappresentata dalla prevalenza della malattia, è possibile adottare un approccio "soggettivista", forse per alcuni aspetti più vicino al modo di ragionare del clinico.

1.2.1.1 *Odd di prevalenza*

Si indichi con OP l'odd di prevalenza, definito formalmente come rapporto tra la probabilità di osservare un soggetto malato rispetto a quella di osservare un soggetto non malato:

$$OP = \frac{P(D=1)}{P(D=0)} = \frac{p}{1-p}.$$

Tale indice viene spesso impiegato per comodità, in quanto aumenta con la prevalenza e può essere stimato molto semplicemente come rapporto tra il numero dei malati e quello dei non malati nel campione. Applicando la formula di Bayes alle definizioni sopra riportate si ricava immediatamente la relazione tra VPP, p, sensibilità $(1-\beta)$ e specificità $(1-\alpha)$:

$$VPP = \frac{sens}{sens + \frac{1-spec}{OP}} = \frac{(1-\beta)}{(1-\beta) + \frac{\alpha}{OP}},$$

$$VPN = \frac{spec}{spec\ + (1-sens)\cdot OP} = \frac{(1-\alpha)}{(1-\alpha) + \beta \cdot OP}.$$

Tale formulazione mostra che, all'aumentare della frazione dei malati nel campione sottoposto al test, la proporzione dei malati positivi aumenta nell'insieme dei positivi al test. Al contrario, per una patologia poco rappresentata tenderà ad aumentare la frazione dei falsi positivi sul totale dei positivi al test. In particolare, aumentando la prevalenza, a parità di sensibilità e specificità, VPP salirà. Dunque, i valori predittivi di un test sono influenzati pesantemente dalla prevalenza della condizione in esame; questo è particolarmente importante qualora si decidesse di avviare una campagna di screening di massa. Un test di screening, oltre a possedere alcune caratteristiche particolari (di facile esecuzione, poco costoso, accettabile dall'utente, ecc.), deve possedere senza dubbio una buona sensibilità. Ma la predittività del test sarà sempre proporzionale alla prevalenza della malattia nella popolazione sottoposta a screening: per aumentarla, pertanto, bisogna scegliere accuratamente la popolazione su cui avviare lo screening, per evitare di dovere fare i conti con una quota troppo elevata di falsi positivi. In genere, ad un primo test di screening conviene far seguire un secondo test, cosiddetto "di conferma", dotato generalmente di maggior specificità, che avrà proprio lo scopo di identificare (e quindi escludere) i falsi positivi nel gruppo dei soggetti risultati positivi al primo test. La predittività del secondo test sarà sempre molto elevata, in quanto esso verrà eseguito su una popolazione fortemente selezionata dal primo test e, quindi, ad elevata prevalenza. Quanto più la prevalenza della condizione in esame è elevata, tanto migliore sarà la performance di un test con un elevato valore predittivo. Se la prevalenza della condizione che si vuole

studiare è molto bassa, un test con un valore predittivo positivo molto vicino al 100% sarà comunque poco utile. La conseguenza diretta di questa osservazione è che lo stesso test diagnostico potrà funzionare in modo diverso secondo la popolazione che viene ad esso sottoposto. La prevalenza della condizione in esame, in effetti, è funzione dello scenario in cui si opera e può variare secondo la prevalenza nella popolazione generale, il gruppo d'età, il sesso, la presenza di sintomi clinici e, appunto, lo scenario nel quale il paziente viene osservato. Se si vuole applicare un test di screening alla popolazione generale, la probabilità d'incontrare una determinata condizione patologica sarà uguale alla prevalenza. Se invece si vuole applicare il test diagnostico ai pazienti che afferiscono ad un ambulatorio specialistico, la prevalenza di questa popolazione sarà notevolmente maggiore di quella della popolazione generale.

1.2.2 *Rapporto di verosimiglianza*

Un altro parametro spesso impiegato per valutare la performance di un test diagnostico è il rapporto di verosimiglianza (LR, dall'inglese "*likelihood ratio*"), che esprime di quante volte la probabilità di una determinata diagnosi di malattia è modificata per effetto del test. Il rapporto di verosimiglianza positivo (LR^+) esprime la probabilità di un risultato positivo in un soggetto malato rispetto alla medesima probabilità in un soggetto sano. Analogamente, il rapporto di verosimiglianza di un risultato negativo (LR^-) esprime la probabilità di un risultato negativo in un soggetto malato rispetto alla medesima probabilità in un soggetto sano. Si calcolano come segue:

$$LR^+ = \frac{proporzione\ di\ VP}{proporzione\ di\ FP} = \frac{sensibilità}{1 - specificità} = \frac{P(T=1|D=1)}{P(T=1|D=0)} = \frac{1-\beta}{\alpha},$$

$$LR^- = \frac{proporzione\ di\ FN}{proporzione\ di\ VN} = \frac{1 - sensibilità}{specificità} = \frac{P(T=0|D=1)}{P(T=0|D=0)} = \frac{\beta}{1-\alpha}.$$

In termini di rapporto di verosimiglianza, le relazioni precedentemente illustrate diventano:

$$VPP = \frac{LR^+ \cdot OP}{LR^+ \cdot OP + 1},$$

$$VPN = \frac{1}{1 + OP \cdot LR^-}.$$

Si può notare che il VPP tende ad aumentare in modo non lineare con la prevalenza e con maggiore rapidità per valori elevati di LR^+, mentre il VPN tende a diminuire con la prevalenza, tanto più

rapidamente quanto più elevato è LR^{-}. In tutti e due i casi il rapporto di verosimiglianza esprime un valore che moltiplicato per la probabilità pre-test (p) ci permetterà di calcolare la probabilità post-test in caso di risultato positivo $(LR+)$ o negativo $(LR-)$.

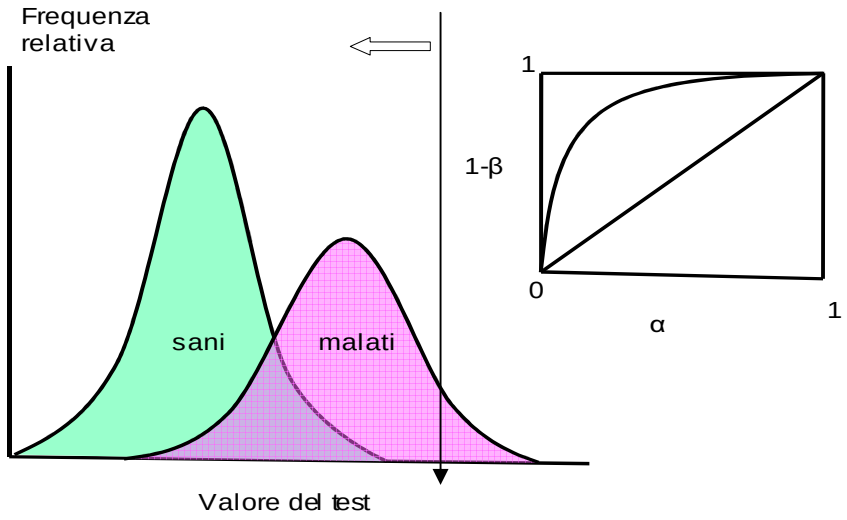
Per il rapporto di verosimiglianza positivo, valori superiori a 10 indicano che il test è molto efficace nell'aumentare la nostra probabilità pre-test. Allo stesso modo, per il rapporto di verosimiglianza negativo, valori minori di 0,1 sono da considerare tipici di test particolarmente attendibili. L'uso di questo parametro permette di eseguire valutazioni della performance di un test diagnostico del tutto indipendenti dalla prevalenza della condizione in esame e di verificarne l'utilità secondo la propria realtà specifica.

1.3 CURVE ROC

I concetti finora esposti sono di estrema importanza soprattutto nella pratica clinica. Infatti, può essere molto interessante valutare come si comporti un test in una popolazione (in termini di sensibilità e specificità) in funzione di determinati valori di *cut-off*. A tale scopo sono state realizzate le cosiddette curve ROC, una tecnica per visualizzare, organizzare e selezionare classificatori in base alla loro performance.
ROC è l'acronimo di Receiver Operating Characteristic ("caratteristiche operative del ricevitore") e trae origine nell'ambito della teoria della rilevazione del segnale. Si tratta di una metodologia che è stata utilizzata per la prima volta da alcuni ingegneri elettrici durante la seconda guerra mondiale, per l'analisi delle immagini radar e lo studio del rapporto segnale/disturbo. Il problema era quello di riconoscere il segnale causato dalla presenza di oggetti nemici sui campi di battaglia (tipicamente nei cieli, ma anche in mare), distinguendolo dal rumore di fondo presente nei segnali radar. A tal proposito, si supponga di avere un filtro originale, che possa essere variato di continuo, ponendo di volta in volta una determinata soglia, e si considerino "rumore" i valori inferiori alla soglia fissata e "segnale" i valori uguali o superiori alla stessa. Ci troviamo di fronte ad una delle infinite varianti del paradosso del sorite di Zenone: "Qual è il granello che fa passare un mucchio di sabbia in un non-mucchio?", la cui forma in questo caso è: qual è il valore che segna la transizione da un segnale ad un non-segnale, ovvero qual è il valore soglia al di sotto del quale dobbiamo pensare che non si tratti di un segnale ma si tratti semplicemente di rumore di fondo?
La curva ROC venne ben presto applicata in altri campi della tecnica e, a partire dagli anni '70, trovò un ampio raggio di applicazioni in campo medico, inizialmente allo scopo di quantificare l'attendibilità dei responsi di immagini radiografiche interpretate da operatori diversi. In tempi più recenti, l'utilizzo delle curve ROC si è fatto relativamente comune per la valutazione non solo delle immagini, ma anche dei più svariati test sia nel settore medico (con particolare riguardo alla valutazione dei test clinici di laboratorio), psicologico, biologico, veterinario che in altri ambiti, quali il machine learning e data mining (Metz, 1986; Pepe, 1998; Zou, 2001).

L'analisi ROC viene effettuata attraverso lo studio della funzione che – in un test quantitativo – lega la probabilità di ottenere un risultato vero positivo nella classe dei malati-veri (ossia la sensibilità $= 1 - \beta$) alla probabilità di ottenere un risultato falso positivo nella classe dei non malati (ossia $1 -$specificità $= \alpha$). La curva ROC non è altro che la rappresentazione grafica a due dimensioni, che riporta, in un sistema di assi cartesiani e per ogni possibile valore di *cut-off*, la proporzione di falsi positivi (α) in ascissa e la proporzione di veri positivi ($1 - \beta$) in ordinata, relativamente ai valori ottenuti da un test applicato a una popolazione (Figura 1.7).

Figura 1.7. Rappresentazione grafica della curva ROC e relazione con la distribuzione della popolazione in funzione del risultato di un test di screening per tutti i possibili valori di *cut-off*.

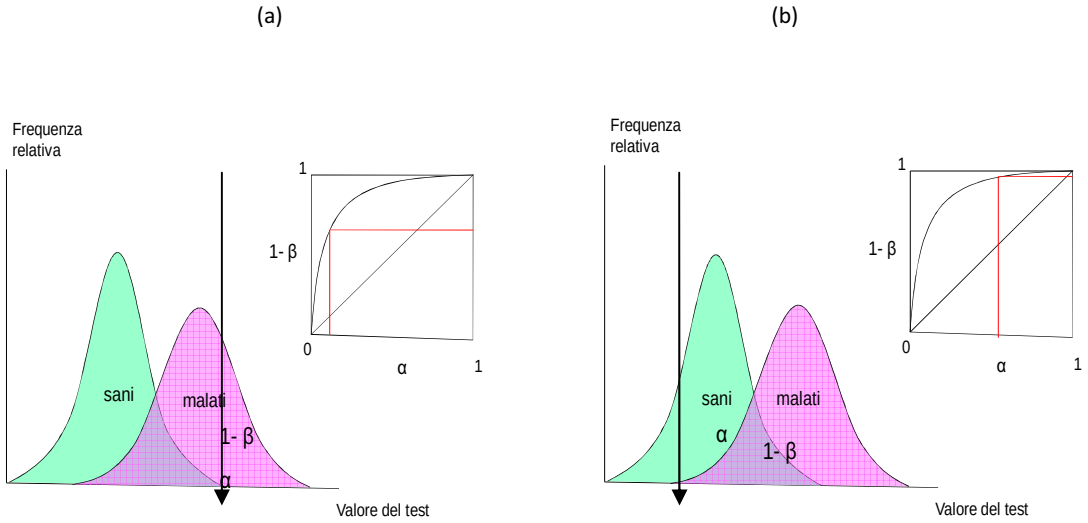


In altre parole, vengono studiati i rapporti fra allarmi veri (*hit rate*) e falsi allarmi. Questa curva raffigura l'utilità del test attraverso un intero *range* di possibili soglie e coglie i *trade-off* tra i benefici (veri positivi) e i costi (falsi positivi). La pendenza della retta che congiunge un punto della curva con l'origine degli assi è uguale al rapporto di verosimiglianza del test positivo: $\frac{sens}{1-spec} = \frac{1-\beta}{\alpha}$. Piccoli spostamenti lungo la curva informano sulle variazioni reciproche di sensibilità e specificità per piccole variazioni del *cut-off*. In questo senso è importante la pendenza locale della curva, ad esempio una forte pendenza significa un buon incremento di sensibilità con piccola perdita di specificità.

Si valuti ora la relazione esistente tra la distribuzione di una popolazione in funzione del risultato del test di screening nelle classi di individui malati e non malati e la corrispondente curva ROC.

Spostando il *cut-off* del test da destra verso sinistra attraverso l'istogramma (Figura 1.8), il grafico ROC muoverà da sinistra verso destra. Vale a dire che, in corrispondenza di un *cut-off* alto (Figura 1.8-a), ci saranno pochi FP ma anche un numero limitato di VP (α e $1-\beta$ vicini a zero); appena si sposta il *cut-off* a sinistra dell'istogramma (Figura 1.8-b), i VP aumenteranno (velocemente all'inizio), ma anche il numero di FP comincerà a salire: $(1-\beta)$ e α si avvicineranno sempre più ad 1.

Figura 1.8. Relazione esistente tra la distribuzione di una popolazione in funzione del risultato del test di screening nelle classi di individui malati e non malati e la corrispondente curva ROC.



1.3.1 *Come costruire una curva ROC*

Per costruire la curva bisogna disporre di misurazioni nella popolazione con livelli intermedi di *cut-off*: ad ogni livello corrisponderà una coppia "sensibilità/errore falso positivo", quindi un puntino sul grafico.
Si considerino due popolazioni, una di individui sani e una di individui affetti dalla malattia. Si disponga di un test per la malattia e lo si applichi ad un gruppo misto di persone, alcune con la malattia e altre sane. I valori del test sono compresi tra zero ed un numero molto grande. Si ordinino i risultati e si leggano i valori uno dopo l'altro cominciando dalla parte "*disease like*" (intendendo per "*disease like*" un valore basso, se per il test considerato i malati tendono ad avere valori più bassi, ad es. l'emoglobina nell'anemia; un valore alto, se per il test considerato i malati tendono ad avere valori più alti, ad es. la glicemia nel diabete) (Bamber, 1975; Zweig e Campbell, 1993). Nella seguente descrizione della procedura di costruzione della curva si è arbitrariamente deciso che i pazienti con valori più alti del test hanno maggior probabilità di essere malati.

I passaggi sono i seguenti.

1. Iniziare dall'angolo inferiore sinistro della curva ROC, dove sia α che $(1-\beta)$ sono pari a zero (questo corrisponde ad avere la linea del *cut-off* del test sulla destra del grafico delle due distribuzioni).
2. Esaminare il risultato più grande. Per iniziare a costruire la curva ROC, porre la soglia del test appena sotto il risultato più alto – si sposti l'indicatore leggermente a sinistra. Se il primo risultato appartiene ad un paziente con la malattia, allora si è di fronte ad un caso di vero positivo ($1-\beta$ deve essere più grande): segnare il primo punto della curva ROC. Al contrario, se la malattia è assente, si ha un caso di falso positivo (α è maggiore di zero): spostarsi a destra e tracciare il punto.
3. Impostare la soglia del test di poco inferiore, appena sotto al secondo risultato più grande, e ripetere il processo descritto in (2).
4. Continuare fino a quando si sia spostato la soglia sotto il valore del test più basso. Infine si arriva all'angolo superiore a destra della curva ROC: tutti i risultati saranno classificate come positivi, così α e $(1-\beta)$ saranno entrambi 1.0, dato che la soglia è sotto il valore più basso.

Quando tutti i risultati sono stati disegnati, il grafico somiglierà ad una curva con pendenza ripida all'inizio ma che diviene sempre più orizzontale a mano che ci si sposta verso l'angolo superiore destro (Figura 1.9). La spezzata che si ottiene è detta anche ROC empirica, avendo l'accortezza di ridimensionare gli assi e le loro scale (la O-ROC – *ordinary ROC* – è contenuta in un quadrato, i cui lati corrispondono al 100% dei malati ed al 100% dei non malati). Si può lisciare la curva mediante interpolazione, come mostrato nella Figura 1.10.

Figura 1.9. Costruzione della curva ROC in funzione della sensibilità $(1-\beta)$ e del tasso di falsi positivi (α) per tutti i possibili valori di *cut-off*.

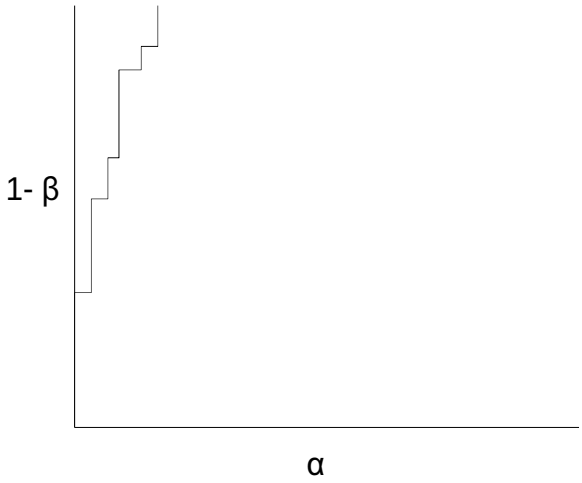
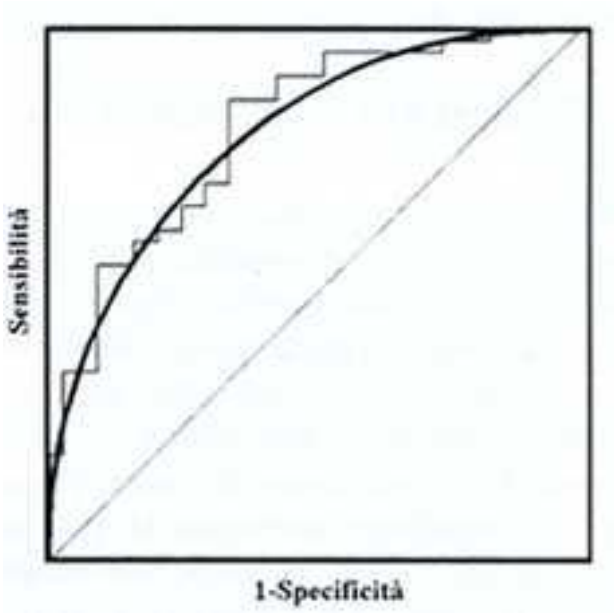


Figura 1.10. Curva ROC prima e dopo interpolazione.



1.3.2 *Trasformazione degli assi*

Esistono varie trasformazioni degli assi usualmente utilizzati per le curve ROC.

Le curve F-ROC si ottengono moltiplicando gli assi per la probabilità p di malattia e $(1-p)$ di non malattia:

$$y = p \cdot sens = p \cdot (1-\beta),$$

$$x = (1-p) \cdot (1-spec) = (1-p) \cdot \alpha.$$

Questa trasformazione degli assi restituisce la reale dimensione pratica nella quale il test viene solitamente applicato (una cosa è provare un test su 100 sani e 100 malati, un'altra è applicarlo in una situazione in cui una probabilità pre-test non è del 50%, ma del 10%, dell'1%...). La pendenza della retta che congiunge un punto della curva con l'origine degli assi è uguale al "post-test odd", o rapporto tra veri e falsi positivi. Il punto più a nord-ovest della curva ROC corrisponde al miglior *cut-off*, nel senso di massimizzazione del numero di soggetti classificati correttamente (e di minimizzazione del numero di soggetti con diagnosi falsa).

Le curve EU-ROC si ottengono moltiplicando gli assi di una F-ROC per un valore di utilità attesa (EU = expected utility) derivante dalla corretta diagnosi (o di utilità perduta nel caso di diagnosi errata):

$$y = B \cdot p \cdot sens = B \cdot p \cdot (1-\beta),$$

$$x = C \cdot (1-p) \cdot (1-spec) = C \cdot (1-p) \cdot \alpha,$$

dove B indica il beneficio medio legato alla scoperta di un caso di malattia e C rappresenta il costo medio di un falso positivo. La pendenza della retta che congiunge un punto della curva con l'origine degli assi è uguale al "post-test expected regret ratio". Il punto più a nord-ovest della curva ROC corrisponde al miglior *cut-off*, nel senso di massimizzazione dell'utilità attesa.

La prima letteratura sulle curve ROC spesso ha fatto l'assunzione che le distribuzioni sottostanti fossero curve normali (spesso si usano curve normali per ragioni di convenienza). Sotto questa assunzione, un trucco è stato creare un speciale grafico dove gli assi sono trasformati conformemente alla distribuzione normale ('*double normal probability coordinates scales*'). Usando tali coordinate, la curva ROC diventa lineare: in questo modo la pendenza e gli assi corrispondono ai due parametri che contengono la media e la deviazione standard. Curve adattate possono essere ricavate usando tecniche speciali, non minimi quadrati, per trovare la linea che meglio si adatta alle coordinate disegnate. Questi metodi sono usati nei principali studi di psicologia sperimentale.

Si noti che se si usa la rappresentazione sopra descritta, la pendenza della retta ottenuta tracciando $(1-\beta)$ contro α fornirà il rapporto delle deviazioni standard delle due distribuzioni. In altre parole, se le deviazioni standard delle popolazioni malata $(D=1)$ e non malata $(D=0)$ sono $s_{D=1}$ e $s_{D=0}$, la pendenza della retta è $s_{D=0}/s_{D=1}$. Nel caso particolare che questo valore sia 1, si può misurare la distanza tra la linea disegnata e la "*chance line*" (ottenuta congiungendo l'angolo inferiore sinistro e l'angolo superiore destro del grafico). Questa distanza è una misura normalizzata della distanza tra la media delle due distribuzioni, dove $\mu_{D=d}$, $d=0,1$, si riferisce alla media della popolazione malata e sana, e $s=s_{D=1}=s_{D=0}$ alla deviazione standard:

$$d' = \frac{(\mu_{D=1}/\mu_{D=0})}{s}.$$

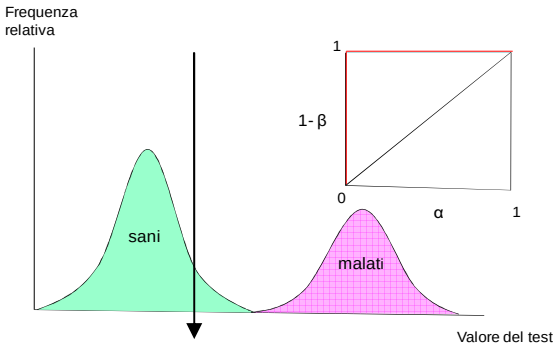
1.4 AREA SOTTESA ALLA CURVA ROC (AUC)

Il potere diagnostico di un test è espresso attraverso l'area sottesa alla curva ROC (AUC). La capacità discriminante di un test, ossia la sua attitudine a separare propriamente la popolazione in studio in malati e sani è proporzionale all'estensione dell'area sottesa alla curva ROC (*Area Under the Curve*, AUC) ed equivale alla probabilità che il risultato di un test su un individuo estratto a caso dal gruppo dei malati sia più "*disease like*" rispetto a quello di uno estratto a caso dal gruppo dei non-malati.

Nel caso di un test perfetto, ossia che non restituisce alcun falso positivo né falso negativo (capacità discriminante = 100%), la curva ROC sale perfettamente verticale sull'asse delle ordinate, poi piega ad angolo retto in orizzontale, parallela rispetto all'asse delle ascisse (Figura 1.11). Qui il valore dell'AUC equivale all'area dell'intero quadrato delimitato dai punti di coordinate $(0;0),(0;1),(1;0),(1;1)$, che

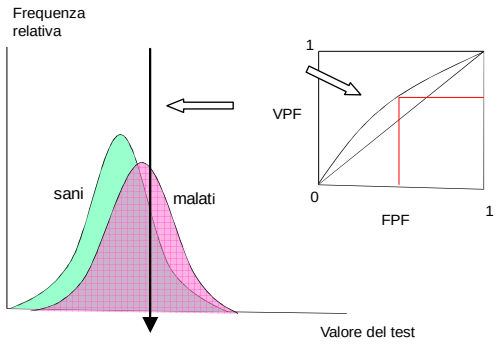
assume valore pari a 1, corrispondendo ad una probabilità del 100% di una corretta classificazione (potere informativo massimo, assenza di errori). Si noti che in tale caso limite, associato ad una distribuzione separata della variabile nei due gruppi a confronto, i valori predittivi non dipendono più dalla prevalenza.

Figura 1.11. Curva ROC per un test perfetto avente massimo potere informativo (AUC = 1).



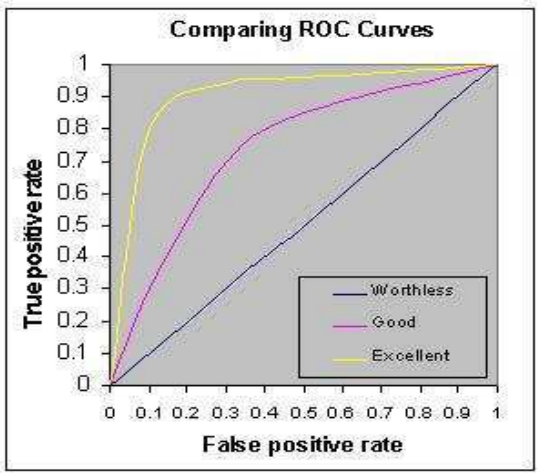
Nella situazione opposta, di completa sovrapposizione delle distribuzioni dei soggetti sani e malati, si ottiene una curva ROC che è una retta che va dall'angolo inferiore sinistro all'angolo superiore destro: la ROC per un test assolutamente privo di valore informativo è rappresentata dalla diagonale ("*chance line*") che passa per l'origine, con AUC= 0.5 (Figura 1.12). Ovviamente la linea retta ritrae una situazione poco auspicabile, in cui la sensibilità è sempre pari al valore del tasso di falsi positivi: rappresenta cioè la linea di "nessun beneficio". In questo caso i risultati del test nel gruppo dei sani e nel gruppo dei malati sono identici: una volta eseguito il test, in base al suo risultato non sappiamo se attribuire il paziente al gruppo dei sani o al gruppo dei malati. Il test non può essere utilizzato per la diagnosi della malattia in questione: l'informazione fornita dal test di laboratorio è uguale a quella che si può ricavare dal lancio di una moneta. Si tratta di un test inutile, che corrisponde a *cut-off* di utilità nulla, per il quale il rapporto di verosimiglianza è pari a 1 (diagonale di "indifferenza" del test).

Figura 1.12. Curva ROC per un test assolutamente privo di valore informativo (AUC = 0.5).



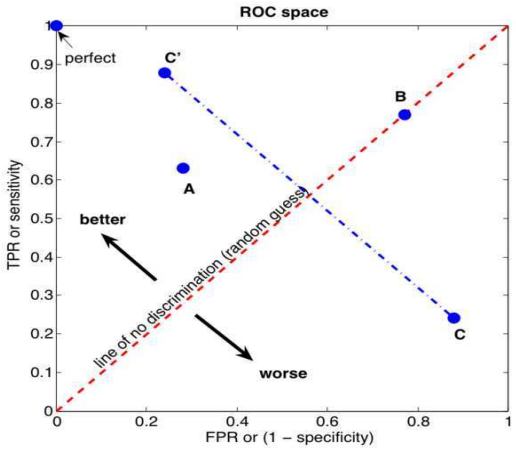
Due punti della curva situati su una retta inclinata di 45°, parallela alla diagonale di "indifferenza", hanno la stessa somma di giuste classificazioni, cioè (1-α)+(1-β) è costante, e di errate classificazioni, α+β è costante. Tanto più la curva misurata si scosta dalla linea di nessun beneficio, tanto essa sarà migliore, in quanto permetterà d'identificare un valore di *cut-off* (ovvero un punto) che dia il massimo della sensibilità con il tasso di errore falso positivo più basso possibile. Pertanto il punto più a nord-ovest della curva ROC corrisponde al miglior *cut-off*, nel senso di massimizzazione delle classificazioni corrette e di minimizzazione degli errori. Confrontando due curve ROC si deduce che il classificatore più vicino all'angolo superiore sinistro ha un'area più grande e una migliore performance media.

Figura 1.13. Confronto in termini di performance tra test diversi: il miglior classificatore è quello che più si avvicina all'angolo superiore destro.



Ogni risultato previsto o un caso della matrice di confusione rappresenta un punto nello spazio ROC (Figura 1.14). Una congettura completamente casuale dà un punto lungo la linea diagonale (B). I punti sopra la linea diagonale (A, C') indicano buoni risultati di classificazione. I punti sotto la diagonale (C) indicano risultati non corretti.

Figura 1.14. Spazio ROC.

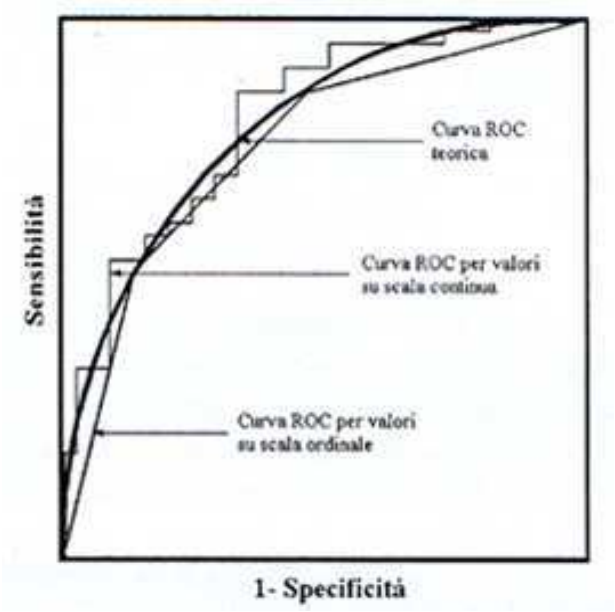


1.4. AREA SOTTESA ALLA CURVA ROC (AUC)

Il calcolo dell'AUC per una curva empirica (cioè ottenuta da un campione finito) può venire effettuato semplicemente connettendo i diversi punti del grafico ROC all'asse delle ascisse con segmenti verticali e sommando le aree dei risultanti poligoni generati nella zona sottostante. Tale curva è una lunga serie di piccoli scalini verticali ed orizzontali: ogni volta che la curva prende uno scalino orizzontale, si calcola semplicemente l'altezza della curva (cioè il numero di VP/numero di persone con la condizione) e si moltiplica questo per 1/numero di persone senza la malattia. In altre parole bisogna conoscere l'ampiezza dello scalino (di quanto ci si sposta a destra) e moltiplicare questo valore per la sua altezza ogni volta che si colpisce un risultato FP sulla curva ROC.

Questa tecnica, detta "regola trapezoidale", può fornire risultati sistematicamente distorti per difetto.
L'area sottesa ad una curva ROC (AUC) costruita su dati ordinali (ad esempio, derivati dall'aggregazione in classi di una variabile continua) costituisce una sottostima sistematica dell'area sottesa alla curva ROC teorica (ottenibile per popolazione campionaria infinita), quindi della performance del test. Tale asserzione è illustrata nella Figura 1.15.

Figura 1.15. Area sottesa ad una curva ROC teorica, empirica per dati continui ed empirica per dati ordinali.



In pratica, il calcolo della AUC può essere eseguito al calcolatore per mezzo di pacchetti statistici completi (NCSS, R, SAS, SPSS, SimStat, Stata, SYSTAT ecc) oppure di software specifici per la valutazione dei test diagnostici e delle curve ROC.

Per quanto riguarda l'interpretazione del valore di AUC, si può tenere presente la classificazione della capacità discriminante di un test proposta da Swets (1998).

Essa è basata su criteri largamente soggettivi ed avviene secondo lo schema seguente.

- $\text{AUC} = 0.5$ test non informativo
- $0.5 < \text{AUC} \leq 0.7$ test poco accurato
- $0.7 < \text{AUC} \leq 0.9$ test moderatamente accurato
- $0.9 < \text{AUC} < 1.0$ test altamente accurato
- $\text{AUC} = 1.0$ test perfetto

1.4.1 *Scelta del cut-off*

In una curva ROC esistono in genere due segmenti di scarsa o nulla importanza ai fini della valutazione dell'attitudine discriminante del test in esame. Essi sono rappresentati dalle frazioni di curva sovrapposte rispettivamente all'asse delle ascisse ed all'asse delle ordinate. Infatti, i corrispondenti valori possono essere scartati in quanto esistono altri valori di *cut-off* che forniscono una migliore specificità senza perdita di sensibilità o, viceversa, una migliore sensibilità senza perdita di specificità. Infine è da ricordare che la valutazione di un test attraverso l'AUC viene compiuta attribuendo ugual importanza alla sensibilità e alla specificità, mentre in molti casi è necessario, nella pratica, differenziare il peso da assegnare ai suddetti parametri. Nella maggioranza degli studi, l'individuazione del *cut-off* ottimale viene effettuata assumendo una distribuzione normale per la variabile in studio e si raggiunge adottando un valore pari a [media aritmetica $+ 2 \cdot$ deviazioni standard] dei risultati generati dal gruppo di individui sani di referenza. Questo approccio rigido corrisponde ad ottenere un test con specificità pari a 97.5% (Barajas-Rojas et al., 1993) e presenta lo svantaggio di trascurare completamente il valore della sensibilità. Un'altra possibilità è quella di selezionare un livello di *cut-off* sulla base dei percentili della distribuzione dei non-malati (ad esempio, il 90-esimo percentile), e di considerare come potenzialmente malati i soggetti con valori superiori. Tale metodo corrisponde a fissare a priori la specificità del test (si noti, infatti, che il 90-esimo percentile nella distribuzione dei non-malati corrisponde a fissare la specificità al 90%). Un approccio più adeguato può essere adottato tenendo in considerazione la relazione che lega sensibilità e specificità, ovvero studiando la curva ROC. L'utilizzo della curva ROC rappresenta un criterio più "flessibile", in quanto offre la possibilità di visualizzare, dato un valore a scelta di specificità, la corrispondente sensibilità e viceversa (Schäfer, 1989). Come regola generale, si può affermare che il punto sulla curva ROC più vicino all'angolo superiore sinistro rappresenta il miglior compromesso fra sensibilità e specificità. Tuttavia, in condizioni ottimali, la procedura di selezione del *cut-off* consiste in un percorso decisionale molto più complesso che deve tener conto, come già ricordato in precedenza, sia della situazione epidemiologica nella popolazione da studiare (con particolare riferimento alla prevalenza della malattia) che dell'esame comparativo delle conseguenze

pratiche derivanti dall'ottenimento di risultati falsi positivi e falsi negativi in quella particolare situazione contingente.

1.4.2 *Scelta economica del cut-off*

È possibile usare analisi economiche per supportare la decisione sul *cut-off* di un test usando curve ROC. Disegnare una curva ROC è essenzialmente disegnare "*hits*" contro "falsi allarmi". Si può limitare il numero di falsi allarmi alla spesa di pochi *hits* ma la decisione di dove collocare un *cut-off* dipenderà dai relativi costi di ognuno (monetari o altri costi). Quando si pone un *cut-off* si dovrebbero considerare i costi finanziari di trattamento della malattia (presente o assente) o i costi di fallimento nel trattare una malattia (presente o assente): costi di successive investigazioni, disagio nei confronti del paziente per le investigazioni e trattamenti e la mortalità o morbosità del trattamento o non trattamento. Se il costo di perdita di una malattia è oneroso e il trattamento è relativamente non nocivo si dovrebbe porre il *cut-off* verso la destra della curva ROC (cioè VP e FP alti). Se il rischio del trattamento è grave o l'effetto del trattamento è limitato si dovrebbero risparmiare i soggetti sani dal trattamento e porre il *cut-off* verso sinistra.

Il costo medio risultante dall'uso di un test diagnostico potrebbe essere:

$$C_{medio} = C_0 + CVP \cdot P(VP) + CVN \cdot P(VN) + CFP \cdot P(FP) + CFN \cdot P(FN),$$

dove C_0 è la spesa generale del test, CVP è il costo associato al risultato vero positivo e $P(VP)$ è la probabilità di un risultato vero positivo.

Inoltre,

$$P(VP) = P(D = 1) \cdot P(T = 1|D = 1) = P(D = 1) \cdot VPF,$$

che indica che la probabilità di un vero positivo è uguale al prodotto tra la prevalenza e la frazione di veri positivi (VPF). Comunque:

$$\begin{aligned} C_{medio} &= C_0 + CVP \cdot P(D = 1) \cdot P(T = 1|D = 1) + CVN \cdot P(D = 0) \cdot P(T = 0|D = 0) + CFP \\ &\quad \cdot P(D = 0) \cdot P(T = 1|D = 0) + CFN \cdot P(D = 1) \cdot P(T = 0|D = 1) \\ &= C_0 + CVP \cdot P(D = 1) \cdot VPF + CVN \cdot P(D = 0) \cdot VNF + CFP \cdot P(D = 0) \cdot FPF \\ &\quad + CFN \cdot P(D = 1) \cdot FNF. \end{aligned}$$

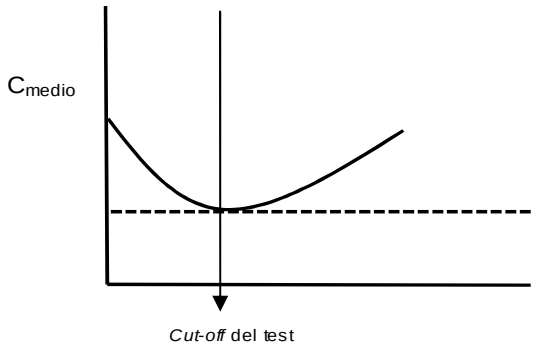
Sostituendo $VNF = 1 - FPF$ e $FNF = 1 - VPF$, si ottiene:

$$C_{medio} = C_0 + CVP \cdot P(D=1) \cdot VPF + CVN \cdot P(D=0) \cdot (1 - FPF) + CFP \cdot P(D=0) \cdot FPF + CFN \cdot P(D=1) \cdot (1 - VPF)$$

$$= VPF \cdot P(D=1) \cdot \{CVP - CFN\} + FPF \cdot P(D=0) \cdot \{CFP - CVN\} + C_0 + CVN \cdot P(D=0) + CFN \cdot P(D=1). \quad (1.1)$$

L'espressione 1.1 mostra come il C_{medio} dipenda da VPF e $FPF -$ le coordinate dei punti che costituiscono la curva ROC. Tuttavia il costo medio dipende dalla posizione del *cut-off* sulla curva ROC: variando il *cut-off* si altera il costo. Per ricavare il C_{medio} più basso possibile si uguagli a zero la derivata (gradiente) dell'equazione del costo medio. Se il *cut-off* è troppo basso, il costo è alto (FP troppo costosi). Se il *cut-off* è alto, anche il costo è alto (FN troppo costosi). La funzione del costo medio rispetto al *cut-off* ha un grafico simile a questo (Figura 1.16):

Figura 1.16. Funzione del costo medio rispetto al *cut-off*.



La linea tratteggiata demarca il punto in cui il costo medio è minimo (gradiente uguale a zero).

Usando la curva ROC si può esprimere VPF come una funzione dei FPF:

$$C_{medio} = ROC(FPF) \cdot P(D=1) \cdot \{CVP - CFN\} + FPF \cdot P(D=0) \cdot \{CFP - CVN\} + C_0 + CVN \cdot P(D=0) + CFN \cdot P(D=1).$$

Differenziando questa funzione rispetto a FPF, si ottiene:

$$\frac{\partial C_{medio}}{\partial FPF} = \frac{\partial ROC}{\partial FPF} \cdot P(D=1) \cdot \{CVP - CFN\} + P(D=0) \cdot \{CFP - CVN\}.$$

Ponendo tale derivata uguale a zero, si ricava:

$$\frac{\partial C_{medio}}{\partial FPF} = 0$$

$$\leftrightarrow \frac{\partial ROC}{\partial FPF} \cdot \ P(D = 1) \cdot \{CVP - CFN\} + P(D = 0) \cdot \{CFP - CVN\} = 0$$

$$\leftrightarrow \frac{\partial ROC}{\partial FPF} \cdot \ P(D = 1) \cdot \{CVP - CFN\} = -P(D = 0) \cdot \{CFP - CVN\}$$

$$\leftrightarrow \frac{\partial ROC}{\partial FPF} = \frac{P(D = 0) \cdot \{CFP - CVN\}}{P(D = 1) \cdot \{CFN - CVP\}}.$$

L'espressione $\frac{\partial ROC}{\partial FPF}$ è il gradiente della curva ROC, cioè la sua pendenza. Quando i costi sono ottimali (il costo medio è minimo) allora il gradiente della curva ROC $\left(\frac{\partial ROC}{\partial FPF}\right)$ è uguale a $\frac{P(D=0)\cdot\{CFP-CVN\}}{P(D=1)\cdot\{CFN-CVP\}}$.

Se una malattia è molto rara, $\frac{P(D=0)}{P(D=1)}$ sarà molto grande: si deve collocare il *cut-off* vicino al lato sinistro del grafico ROC dove il gradiente $\frac{\partial ROC}{\partial FPF}$ è grande. Questo minimizza i falsi positivi, al costo di non riuscire a cogliere i veri positivi: infatti i falsi negativi possono superare molto velocemente il numero di veri positivi (basso VPP del test a causa dell'alta quota della popolazione sana, $P(D = 0)$, rispetto a quella con la malattia, $P(D = 1)$). Al contrario, se la malattia è largamente diffusa, si pone il *cut-off* verso destra sul grafico ROC (un *cut-off* più indulgente), per evitare di ottenere un numero elevato di falsi negativi. La prevalenza della malattia non è l'unico fattore che ha un effetto profondo sulla scelta del *cut-off* ottimale. La pendenza della curva diventa ripida anche se la differenza tra i costi è maggiore per $CFP - CVN$ piuttosto che per $CFN - CVP$. Si consideri ad esempio un test per diagnosticare un cancro al cervello – se si ottiene un risultato del test positivo, si deve aprire il cranio del paziente e tagliare il cervello per trovare il presunto tumore. Se si ottiene un risultato negativo, non si esegue nessuna operazione. Si assuma inoltre che l'operazione non aiuti i pazienti che hanno il cancro, perché molti moriranno ugualmente. Allora il costo di un falso positivo (operando sul cervello di una persona sana) è molto più elevato del costo di un vero negativo (non si fa nulla) e il costo di un falso negativo (non si fa un'operazione che non potrebbe aiutare molto) è simile al costo di un vero positivo (fare un'operazione piuttosto inutile). Quindi il costo di un FP potrebbe essere molto elevato (neurochirurgia/riabilitazione e cure) e il costo di un VP relativamente basso (potrebbero morire sia se si operi che se non si operi). La pendenza della curva è ripida, così si sposta il *cut-off* verso sinistra della curva ROC. Può verificarsi anche lo scenario opposto, nel caso in cui le conseguenze di un falso positivo siano minime e se ci sia un grande beneficio nel trattare pazienti che soffrono della malattia, disponendo di un trattamento economico ed innocuo. In questa situazione, anche se si trattano falsi positivi, ha senso che il *cut-off* del test sia più indulgente: si sposta quindi il *cut-off* verso la destra della curva ROC.

1.4.3 *Relazione tra AUC e la statistica di Wilcoxon*

I metodi non parametrici forniscono un'alternativa ai metodi statistici parametrici che non richiedono né assunzioni sui dati (o ne richiedono assai limitate) nè sulle distribuzioni di probabilità della popolazione (*distribution-free methods*). Esiste una stretta relazione che lega la AUC alla statistica U di Wilcoxon e Mann-Whitney. La statistica U, ideata dal chimico Wilcoxon (1945) e perfezionata dal matematico Mann-Whitney (1947), rappresenta uno dei più noti test statistici non parametrici. È un'alternativa non parametrica al test t a due campioni basata solamente sull'ordine in cui cadono le osservazioni dai due campioni. Viene utilizzata per il confronto della distribuzione di una variabile continua tra due gruppi, X e Y, per determinare se c'è una differenza tra le due popolazioni, testando l'ipotesi nulla che i due gruppi presentino la stessa mediana. Si definisce mediana, di seguito Me, quella quantità per cui $\Pr(X \leq Me) = \Pr(X \geq Me) = 0.5$. Si considerino due campioni indipendenti di dati $x_1, x_2, \dots, x_{n_x}$ e $y_1, y_2, \dots, y_{n_y}$, provenienti da due distribuzioni continue X e Y con numerosità n_x e n_y, rispettivamente. Il principio su cui si basa il test è che se la mediana di X (Me_x) supera la mediana di Y (Me_Y), allora i ranghi del campione combinato e ordinato delle unità provenienti da X saranno prevalentemente superiori ai ranghi delle unità provenienti da Y. Si definisce rango l'intero corrispondente al posto che l'osservazione occupa quando si passa dalle realizzazioni $(x_1, \dots, x_n)$ al campione ordinato in senso crescente $\left(x_{(1)}, \dots, x_{(n)}\right)$.
Il sistema di ipotesi considerato è:

$$H_0: Me_x = Me_Y,$$
$$H_1: Me_x \neq Me_Y.$$

Per calcolare la statistica test proposta da Wilcoxon si devono seguire i seguenti passi.

- Combinare i campioni indipendenti in un unico campione $\left(n = n_x + n_y\right)$.
- Classificare i dati combinati dal valore più piccolo a quello più alto, con i valori uguali ai quali viene assegnata la media dei punteggi uguali.
- Calcolare W, la somma dei punteggi per le osservazioni nel primo campione (se le due popolazioni sono identiche, la somma dei punteggi del primo campione e quelli nel secondo campione dovrebbero essere vicini allo stesso valore):

$$W = \sum_{i=1}^{n_x} r(X_j),$$

dove $r(X_i)$ è il rango che compete a X_i nel campione ordinato.

1.4. AREA SOTTESA ALLA CURVA ROC (AUC)

Si può verificare che la distribuzione di W ha un minimo in $min(W) = \frac{n_x(n_x+1)}{2}$ e massimo in $max(W) = n_x n_y + \frac{n_x(n_x+1)}{2}$, media e varianza, rispettivamente:

$$\mu_W = E[W] = \frac{n_x(n_x + n_y + 1)}{2}; \qquad V_W = Var[W] = \frac{n_x n_y(n_x + n_y + 1)}{12}.$$

La statistica test è

$$Z = \frac{W - \mu_W}{\sqrt{V_W}}.$$

Per numerosità superiori a 12, tale statistica si distribuisce asintoticamente come una $N(0,1)$. Fissato quindi un livello di significatività, si rifiuta l'ipotesi nulla per valori della statistica test per cui $|Z| \geq z_{1-\frac{\alpha}{2}}$.

Tale ipotesi è del tutto equivalente a testare che un soggetto estratto a caso da un gruppo X abbia la stessa probabilità di presentare un valore della variabile superiore ad un valore predefinito di quello di un soggetto estratto a caso dall'altro gruppo Y. Nella sua formulazione originaria, il test si basa sul numero delle coppie di valori (X, Y), tali che $X > Y$. Risulta:

$$U = n_x n_y + \frac{n_x(n_x + 1)}{2} - R_x,$$

dove n_x e n_y sono le numerosità campionarie dei due gruppi, mentre R_x è la somma dei ranghi nel gruppo a numerosità n_x. Il valore atteso di U, sotto l'ipotesi sopra formulata, è

$$\mu_U = E[U] = \frac{n_x n_y}{2}.$$

Bamber (1975) ha dimostrato l'equivalenza tra l'area AUC sottesa ad una curva ROC, costruita per dati su scala continua, e la statistica U. La relazione che lega i due parametri è la seguente:

$$AUC = \frac{U}{n_1 n_2},$$

da cui (sempre sotto l'ipotesi sopra formulata):

$$\mu_{AUC} = E[AUC] = \frac{1}{2}.$$

1.5 VALUTAZIONE DI UN SINGOLO TEST MEDIANTE ANALISI ROC

L'area sottesa ad una curva ROC rappresenta un parametro fondamentale per la valutazione della performance di un test, in quanto costituisce una misura di accuratezza non dipendente dalla prevalenza ("*pure accuracy*"). Poiché AUC rappresenta una stima da popolazione campionaria finita, risulta quasi sempre necessario testare la significatività della capacità discriminante del test, ovvero se l'area sotto la curva eccede significativamente il suo valore atteso di 0.5 (nell'ipotesi di un test non discriminante). Tale procedura corrisponde a verificare se la proporzione dei veri positivi è superiore a quella dei falsi positivi. Dalle proprietà della statistica U, AUC può essere considerata una variabile normale, per cui si può costruire un test z nella seguente maniera:

$$z = \frac{AUC - 0.5}{\sqrt{\sigma^2_{AUC}}},$$

dove σ^2_{AUC} rappresenta la varianza di AUC. Secondo Hanley e McNeil (1983), la varianza di una curva ROC può essere stimata dalla seguente formula, spesso applicata sia a dati ordinali che, con buona approssimazione, su scala continua:

$$\hat{\sigma}^2_{AUC} = \frac{AUC \cdot (1 - AUC) + (n_x - 1) \cdot (Q_1 - AUC^2) + \big(n_y - 1\big) \cdot (Q_2 - AUC^2)}{n_x n_y},$$

dove n_x e n_y rappresentano la numerosità dei due gruppi a confronto, e Q_x e Q_y sono stimati da:

$$Q_1 = \frac{AUC}{2 - AUC},$$

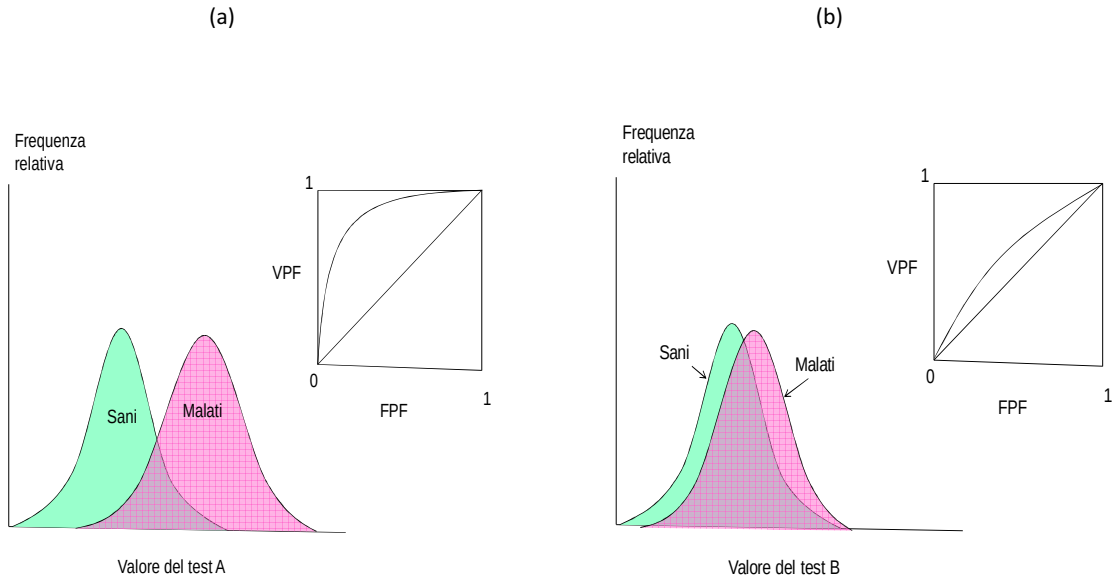
$$Q_2 = \frac{2 \cdot AUC^2}{1 + AUC}.$$

Se, ad esempio, il valore di z eccede il valore critico di 1.96, si può affermare che il test diagnostico presenta una performance significativamente superiore a quella di un test non discriminante, con un livello di significatività osservato inferiore a 0.05. Se il test z risulta invece significativamente inferiore (curva ROC al di sotto della *chance line*), occorre invertire il criterio di classificazione, in quanto il marcatore evidenziato dal test presenta valori mediamente più elevati nella popolazione dei non-malati (evenienza di difficile riscontro).

1.6 CONFRONTO DI DUE TEST MEDIANTE ANALISI ROC

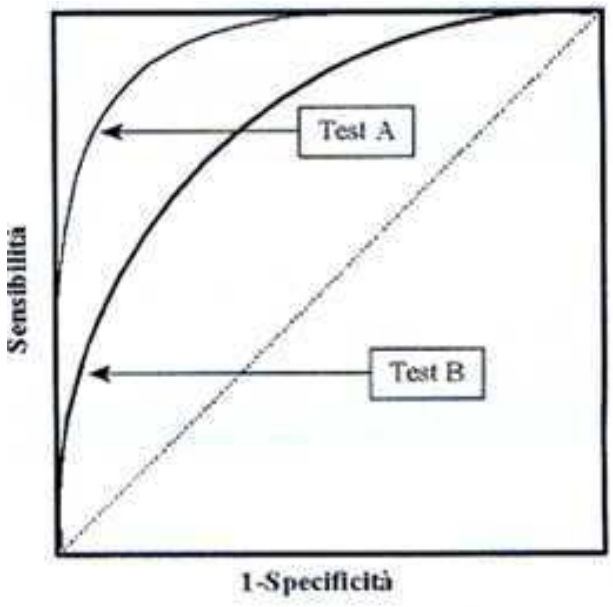
Si considerino due test A e B. Il test A è efficace nel discriminare le popolazioni con o senza una particolare malattia. Il test B è un discriminatore povero. La Figura 1.17 rappresenta le corrispondenti curve ROC per i due test. Per il test B, appena si muove il *cut-off* a sinistra cogliendo i veri positivi e i falsi positivi, la verosimiglianza di incontrare un vero positivo è approssimativamente la stessa d'incontrare un falso positivo. La curva tende a divenire una linea più o meno diagonale (Figura 1.17-a). Invece per il test A, la verosimiglianza di incontrare un vero positivo è molto più alta e la curva inizia a salire molto più rapidamente (Figura 1.17-a). Solo successivamente, appena si resta sprovvisti di veri positivi, si iniziano a cogliere i falsi positivi e la curva si avvicina alla diagonale. L'area sotto la curva (AUC) è ovviamente molto più elevata per il test A, la cui curva si trova interamente al di sopra di quella corrispondente al test B.

Figura 1.17. Confronto tra due test diagnostici mediante analisi ROC. (a) il test A ha una buona capacità di distinguere la popolazione affetta dalla malattia da quella sana e fornisce un'AUC prossima a 1; (b) il test B ha una pessima capacità discriminante e genera un'AUC prossima a 0.5.



Sotto l'ipotesi bi-normale sopra descritta, due test possono essere quindi confrontati tra di loro comparando le accuratezze stimate mediante l'area sottesa alle corrispondenti curve ROC (Figura 1.18). Un test basato sulla distribuzione normale standardizzata può essere eseguito rapportando la differenza delle due aree all'errore standard di tale differenza. Nel caso di indipendenza dei due test, tale parametro viene facilmente stimato dalla radice quadrata della somma della stima della varianza di ogni area.

Figura 1.18. Confronto tra due test diagnostici mediante analisi ROC. Sotto l'ipotesi bi-normale (curve ROC proprie), tale confronto corrisponde a testare la differenza tra le rispettive aree. Risulta evidente la superiorità del test A, la cui curva ROC teorica si trova interamente al di sopra di quella corrispondente al test B.



Il test per il confronto tra le due curve ROC indipendenti è:

$$z = \frac{\mathrm{AUC}_1 - \mathrm{AUC}_2}{\sqrt{\hat{\sigma}^2_{\mathrm{AUC}_1} + \hat{\sigma}^2_{\mathrm{AUC}_2}}}.$$

Nel caso in cui i due test non siano indipendenti (situazione che può verificarsi se vengono applicati agli stessi soggetti), l'errore standard della differenza delle due aree dipende dalla correlazione r esistente tra esse:

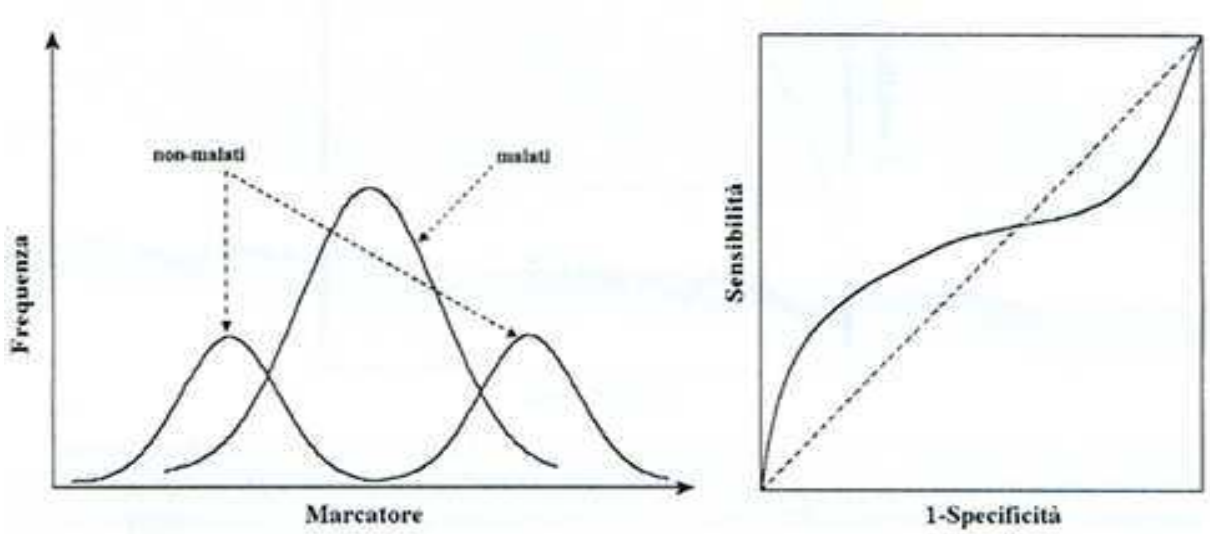
$$z = \frac{AUC_1 - AUC_2}{\sqrt{\hat{\sigma}^2_{\mathrm{AUC}_1} + \hat{\sigma}^2_{\mathrm{AUC}_2} - 2r\hat{\sigma}_{\mathrm{AUC}_1}\hat{\sigma}_{\mathrm{AUC}_2}}}.$$

La stima di r, sia nel caso di variabili continue che categoriche ordinali, è stata illustrata in dettaglio da Hanley e McNeil (1983). In pratica, il primo passaggio consiste nel calcolare il coefficiente di correlazione dei valori dei due test separatamente nel gruppo dei malati e in quello dei non malati. Questo dipende dal metodo di punteggio dei dati: se sono misurati su scala intervallare (per esempio, la pressione sanguigna in millimetri di mercurio), allora il metodo appropriato è la correlazione di Pearson. Per l'informazione ordinale (per esempio, un'immagine è "decisamente anormale" o è "probabilmente anormale"), si usa il coefficiente tau di Kendall. Entrambi possono essere derivati dalla maggior parte dei pacchetti statistici.

La correlazione media tra i due test può essere quindi stimata dalla media dei due coefficienti così ottenuti. Si calcoli poi la media tra le aree AUC_1 e AUC_2, vale a dire $(AUC_1 + AUC_2)/2$. Infine, si può ricavare la correlazione tra le due AUC, che è funzione della media delle due aree e della correlazione media tra i due test, utilizzando la tavola di Hanley e McNeil.

Lo scostamento dall'ipotesi bi-normale può produrre curve ROC non proprie, ovvero si può assistere ad una perdita della concavità oppure della simmetria rispetto alla diagonale discendente. Un caso piuttosto frequente consiste nell'incrocio della curva con la *chance line*; ciò può indicare l'esistenza di una distribuzione bimodale all'interno di uno dei due gruppi a confronto, come illustrato nella Figura 1.19.
In tal caso risulta che la popolazione dei non malati è costituita da (almeno) due diverse sottopopolazioni di cui una presenta un valore di marcatore mediamente più elevato rispetto al gruppo dei malati e l'altra mediamente più basso. In linea teorica si potrebbe quindi generare una regola di classificazione basata su due valori diversi di *cut-off*. Tuttavia, visto che la prevalenza delle diverse sotto-popolazioni nel gruppo dei non malati non è nota, in genere un simile risultato induce al rigetto del test.

Figura 1.19. Distribuzione bimodale di uno dei gruppi a confronto e corrispondente curva ROC non propria.



1.7 FONTI DI ERRORE NELL'ANALISI DI TEST MEDIANTE CURVE ROC

1.7.1 *L'effetto del rumore*

Le curve ROC sono inclini all'errore come qualsiasi altro strumento statistico. Si prenda in considerazione come il "rumore casuale" possa influenzare una curva ROC. Se si dispone di un *gold standard* capace di

confermare l'eventuale presenza della malattia, ci si può domandare che cosa accada nel caso in cui il "rumore" confonda il test utilizzato, cioè quando i risultati del test siano interessati da variazioni casuali sulle quali non si ha nessun controllo. Se si assume che il test correli perfettamente con il *gold standard*, allora l'area sotto la curva ROC sarebbe pari a 1.0. Introducendo il rumore, alcuni risultati dei test sarebbero mal classificati e si introdurrebbero falsi positivi e falsi negativi, provocando una diminuzione dell'AUC.

E se il test fosse già abbastanza scadente nel differenziare i "sani" dai "malati"? Qui le cose diventerebbero più complesse, poiché i falsi positivi e falsi negativi potrebbero accidentalmente essere classificati come valori veri. Si noti che, se la numerosità dei campioni fosse sufficiente e il test avesse un potere discriminatorio, in media il rumore complessivo peggiorerebbe la performance del test. È improbabile che il rumore casuale possa portare a credere che il test abbia prestazioni migliori di quelle che ha effettivamente. Si analizzi cosa succede quando il test e il *gold standard* non sono indipendenti. Questa interdipendenza fornirebbe falsamente un'ampia area sotto la curva ROC. Si consideri il caso estremo in cui il *gold standard* sia confrontato con se stesso. L'AUC sarebbe pari a 1.0, a prescindere. Questa situazione diventerebbe estremamente preoccupante se il *gold standard* fosse di per sé un po' sospetto: se il test, confrontato con il *gold standard,* variasse anch'esso come fa lo il test aureo, ma entrambi avessero una relazione povera con la malattia che si sta tentando di scoprire, allora si potrebbe falsamente ritenere che le diagnosi condotte siano appropriate. Al contrario, se il *gold standard* fosse un po' scadente, ma indipendente dal test, allora l'effetto sarebbe quello di "rumore" – le caratteristiche del test sarebbero sottostimate (caso denominato "errore di classificazione non differenziale").

1.7.2 *Altre fonti di errore*

Dovrebbe essere chiaro che qualsiasi distorsione inerente ad un test non è trasferito nella distorsione della curva ROC. Se un test fosse distorto a favore di una diagnosi di assenza di malattia, questo rifletterebbe meramente una posizione sulla curva ROC e non avrebbe alcun impatto sulla forma complessiva della curva. Tuttavia, altri errori potrebbero ancora insinuarsi. Un articolo interessante che esamina le fonti di errore è quello di Ransohoff e Feinstein (1978). In ogni esame di un test bisogna considerare i seguenti punti.

1. Esaminare l'intero spettro di un processo di malattia. Se fossero riportati solo i casi gravi, allora il test potrebbe essere inutile nei casi più lievi (sia le componenti patologiche che cliniche della malattia dovrebbero rappresentare il suo spettro completo). Un esempio potrebbe essere quello relativo ai tumori maligni: i tumori avanzati e di grandi dimensioni sono facilmente individuabili e

un test di screening potrebbe funzionare bene in questa cornice, ma mancherebbe la malattia allo stato iniziale;

2. Utilizzare pazienti di confronto ("di controllo"). Questi dovrebbero essere simili;
3. Considerare la co-morbidità. Ciò potrebbe influenzare la positività/negatività di un test;
4. Valutare la distorsione da verifica. Se un medico avesse a disposizione il risultato di un test, egli tenderebbe ad esaminare molto attentamente solo i pazienti con esito positivo, aumentando quindi la probabilità di scoprire la malattia per questi soggetti (che potrebbe non essere colta in altri pazienti che hanno avuto un test negativo). Begg e McNeil (1988) descrivono bene questa tendenza e mostrano come può essere corretta.
5. Valutare la distorsione da recensione diagnostica. Nel caso in cui prima si esegua il test e poi venga fatta la diagnosi definitiva, la conoscenza del risultato del test potrebbe influenzare la diagnosi "definitiva" finale. Simile è la "distorsione della rassegna del test", dove la conoscenza della diagnosi del *gold standard* potrebbe influenzare l'interpretazione del test. Studi in radiologia hanno dimostrato che la disponibilità di informazioni cliniche potrebbe spostare gli osservatori lungo la curva ROC, o addirittura interamente su una nuova curva (per controllare questa forma di distorsione si può eseguire un'"analisi covariata", verificando l'esistenza di altri fattori che influenzano i risultati).
6. Valutare la "distorsione da incorporazione". Ciò è stato già menzionato in precedenza sotto "l'indipendenza dal *gold standard*": qui il test è stato incorporato nell'evidenza usata per diagnosticare la malattia.
7. Gestire i risultati del test non interpretabili. Questi sono frequentemente riportati negli studi. Tali risultati dovrebbero essere considerati "equivoci" nel caso in cui il test non fosse ripetibile. Tuttavia, se il test fosse ripetibile, allora sarebbe possibile una correzione (e la stima della sensibilità e specificità), a condizione che la variazione sia casuale. Test non interpretabili potrebbero avere un'associazione con lo stato di malattia (o anche con la "normalità").
8. Tenere presente la variazione inter-osservatore. Negli studi in cui le capacità dell'osservatore sono importanti, osservatori diversi potrebbero eseguire differenti curve ROC o spostarsi lungo la stessa curva ROC.

1.8 UN ESEMPIO: USO DELL'AUC IN UN DISEGNO DI MISURE RIPETUTE

L'utilizzo dell'AUC per la valutazione di un test può essere non banale, specie quando il disegno di studio è complesso. La precisione di una stima dell'area sotto una curva ROC necessita di essere calcolata per condurre un test sulla significatività statistica dell'area sotto la curva ROC e per costruire gli intervalli di confidenza dell'area sotto una curva ROC. È difficile ottenere una soluzione in forma chiusa dell'area sotto

la curva ROC, a causa della sua complessità. Quindi non è disponibile neppure una soluzione in forma chiusa dell'errore standard stimato dell'area $\widehat{AUC}$ sotto la curva ROC, $\sqrt{\hat{\sigma}^2_{\widehat{AUC}}}$. Nella situazione in cui le osservazioni usate per stimare la curva ROC siano indipendenti, Dorfman (1969) e Wieand et al. (1989) hanno derivato una formula per calcolare l'errore standard dell'area sotto una curva ROC che è correlato alla stima non parametrica di Wilcoxon. Questo metodo è stato usato per stimare l'errore standard dell'area sotto la curva ROC ed è implementato in differenti software statistici. In situazioni più generali, è difficile stabilire un metodo statistico per testare la significatività di curve ROC. Per esempio, sotto un disegno di misure ripetute, la stima dell'errore standard dell'AUC può essere ottenuta solo tramite approssimazione, come l'applicazione di un metodo *boostrap* (Liu et al., 2005). Tale errore standard *bootstrap* può essere successivamente usato per condurre un test sulla significatività statistica dell'area sottesa ad una curva ROC (si veda Appendice A.2 per i dettagli sul metodo *bootstrap*).

I dati raccolti per un disegno di misure ripetute hanno parecchi vantaggi. Per prima cosa, un disegno di misure ripetute può ridurre la possibile distorsione dalla raccolta dei dati da ciascun soggetto e aumentarne la fiducia. Secondo, un disegno di misure ripetute ha un costo inferiore rispetto alla raccolta della stessa quantità di dati in cui ogni osservazione è un soggetto differente (raccogliere punti di dati aggiuntivi da un soggetto esistente probabilmente costa meno rispetto alla raccolta di dati dal reclutamento di soggetti aggiuntivi). Terzo, poiché ogni paziente ha molteplici osservazioni, il disegno di misure ripetute fornisce l'opportunità di analizzare sia la varianza intra-paziente sia il cambiamento nel corso del tempo dell'intera coorte. Sotto un disegno di misure ripetute, le osservazioni per un dato soggetto non sono più indipendenti e sono introdotte la correlazione e la varianza intra-soggetto. Comunque, l'impatto di covariate sull'accuratezza di un test diagnostico/biomaker o di un modello statistico può dipendere da un effetto fisso globale (es., la razza del paziente/etnicità) sia da effetti casuali del singolo paziente (es., un cambiamento nel tempo di una covariata che varia nel tempo). Per modellare le variabili che potrebbero essere continue o non continue, gli effetti casuali potrebbero essere presi in considerazione attraverso l'estensione del modello lineare generalizzato (GLM) al modello lineare generalizzato misto (GLMM) (Bresloe e Clayton, 1993), nel quale il predittore lineare è composto da due parti: gli effetti fissi e casuali. Si definisca con $\mu_i = E(Y_i|\gamma_i)$ la media condizionata di una variabile esito Y_i e con $\eta_i = g(\mu)$ la funzione legame che connette μ_i con il predittore lineare, che consiste di entrambi gli effetti fisso e casuale. Si può scrivere GLMM come $\eta_i = X_i\beta + Z_i\gamma_i$ (per $i = 1, \dots, n$) con varianza condizionata $Var(Y_i|\gamma_i)$, dove y_i è un vettore $n \times 1$ dei risultati di un test per l'i-esimo soggetto, η_i è il numero di misure degli esiti per l'i-esimo paziente, e X_i è una matrice $n_i \times p$, che contiene covariate note che sono associate agli effetti fissi. Il vettore dei parametri dell'effetto fisso, β, è $p \times 1$, e Z_i è una matrice $n_i \times k$ che rappresenta le covariate note che sono associate alla parte casuale del modello. Si assuma che γ_i, il vettore dei parametri ad effetto casuale, $k \times 1$, sia distribuito come $N(0, D_i)$. Ora si assuma che il vero stato di malattia sia binario

(malato/non malato) e si definisca $p_{ij,}$ $(i = 1, \dots, ; j = 1, \dots, n_i)$ la probabilità di essere malato/positivo per l'i-esimo soggetto al j-esimo punto temporale e n_{ij} la funzione legame logit per l'i-esimo soggetto al j-esimo punto temporale tra la media e il predittore lineare. Si può modellare l'impatto delle covariate sulla probabilità prevista di essere malato/non malato mediante GLMM nel seguente modo:

$$n_{ij} = x_{ij}\beta + z_{ij}\gamma_i,$$

e $n_{ij} = g(p_{ij}) = log(p_{ij}/(1 + p_{ij})) = x_{ij}\beta + z_{ij}\gamma_i$.

Si definiscano $\hat{\beta}$ e $\hat{\gamma}_i$ le stime mediante la quasi-verosimiglianza penalizzata (Breslow e Clayton, 1993) o pseudo-verosimiglianza ristretta (Wolfinger o O'Connell, 1993) e $\hat{p}_{ij}$ la corrispondente stima di p_{ij}. Si ha:

$$\hat{p}_{ij} = \frac{\exp(x_{ij}\hat{\beta} + z_{ij}\hat{\gamma}_i)}{1 + \exp(x_{ij}\hat{\beta} + z_{ij}\hat{\gamma}_i)}.$$

La probabilità stimata $\hat{p}_{ij}$, per $i = 1, \dots, n$ e $j = 1, \dots, n$, è una funzione di tutte le covariate. Se si applica un test ad un gruppo di individui seguiti per un tempo definito, si può costruire la curva ROC per discriminare un soggetto malato/positivo da un soggetto non malato/non malato. Si indichi con $ROC_{x,z}(t)$ il valore della curva ROC con tasso di falsi positivi che è associato con i predittori ad effetti fissi x e i predittori ad effetti casuali z. Dalla definizione, l'area sotto una curva ROC AUC è:

$$AUC = \int_0^1 ROC_{x,z}(t)dt,$$

dove i limiti dell'integrale variano da 0 a 1. L'area sotto una curva ROC, AUC, può essere calcolata usando il metodo non parametrico di Wilcoxon attraverso la comparazione della grandezza delle probabibilità previste di ogni coppia discordante. Nel disegno di misure ripetute, ogni soggetto ha più di un'osservazione e i valori dell'esito potrebbero variare da tempo a tempo. Tuttavia, la classificazione di un caso malato/positivo e non malato/negativo necessita di essere scelta a livello di osservazione piuttosto che a livello del singolo soggetto. Si definisca $\hat{p}_{ij(D)}$ $(i = 1, \dots, n \; e \; j = 1, \dots, s_i)$ la probabilità prevista di un malato/positività per l'i-esimo soggetto al j-esimo punto temporale che ha un valore osservato malato (positivo), e si indichi con $\hat{p}_{kl(\bar{D})}(k = 1, \dots, n \; e \; l = 1, \dots, t_k)$ la probabilità prevista di un malato/positività per il k-esimo paziente all'l-esimo punto temporale che ha avuto un valore osservato non malato (negativo). Si indichi con $N_D = \sum_{i=1}^{n} s_i$ e $N_{\bar{D}} = \sum_{k=1}^{n} t_k$ il numero totale di osservazioni con valori osservati positivi e negativi, rispettivamente; il numero totale di coppie discordanti allora è pari a $N = N_D \cdot N_{\bar{D}}$. L'area sotto una curva ROC può essere calcolata comparando le probabilità previste di ogni coppia discordante che è definita al livello di osservazione (Liu e Wu, 2003). Si indichi con $A(\cdot)$ un indicatore tale che:

$$A\left(\hat{p}_{ij(D)}\right) > \hat{p}_{kl(\bar{D})} = \begin{cases} 1 & se\ \hat{p}_{ij(D)} > \hat{p}_{ij(\bar{D})}, \\ 0 & se \hat{p}_{ij(D)} < \hat{p}_{ij(\bar{D})}, \end{cases}$$

e

$$A(\hat{p}_{ij(D)}) = \hat{p}_{kl(\bar{D})} = \begin{cases} 1 & se\ \hat{p}_{ij(D)} = \hat{p}_{ij(\bar{D})}, \\ 0 & altrimenti. \end{cases}$$

L'area sottesa alla curva ROC, AUC,è poi stimata mediante il rapporto:

$$\widehat{AUC} = \frac{\sum_i \sum_j \sum_k \sum_l \left[\left(A(\hat{p}_{ij(D)} > \hat{p}_{kl(\bar{D})})\right) + \frac{1}{2} A(\hat{p}_{ij(D)} = \hat{p}_{kl(\bar{D})})\right]}{N_D\ N_{\bar{D}}}$$

Per generare la vera curva ROC, si calcoli una serie di coppie di sensibilità e 1-specificità basata su predizioni dai modelli lineari generalizzati misti. Per ottenere una curva liscia si può usare un incremento di 0.005 nella probabilità prevista per definire la positività. Per esempio, per generare la curva ROC, si possono calcolare 200 coppie di sensibilità e 1-specificità, relative ai corrispondenti valori di *cut-off* $c(1), c(2), \dots, c(200)$. Si rappresentino poi tali punti nel grafico e si tracci la curva ROC.

Per testare statisticamente la statistica riassuntiva di una curva ROC e misurare il grado di fiducia di una stima dell'area sotto alla curva, è essenziale calcolare la varianza dell'area sottesa ad una curva ROC, cioè $\sigma^2_{AUC} = Var(AUC) = Var\left(\int_0^1 ROC_{x,z}(t)dt\right)$. Data la natura della complessità nella stima dell'area sotto una curva ROC, è difficile ottenere una soluzione in forma chiusa per la varianza. Se l'area sotto una curva ROC è stimata dal metodo non parametrico di Wilcoxon e le osservazioni sono indipendenti, la stima della varianza dell'area sottesa alla curva ROC è data da (Hanley e McNeil, 1982):

$$\hat{\sigma}^2_{\widehat{AUC}} = \widehat{Var}\left(\widehat{AUC}\right) = \frac{\widehat{AUC}\left(1 - \widehat{AUC}\right) + (n_D - 1)\left(Q_1 - \widehat{AUC}^2\right) + (\ n_{\bar{D}} - 1)(Q_2 - \widehat{AUC}^2)}{n_D\ n_{\bar{D}}}, \qquad (1.2)$$

dove Q_1 è la probabilità che due osservazioni positive scelte casualmente siano entrambe ordinate con maggior sospetto di un'osservazione negativa scelta a caso e Q_2 è la probabilità che un'osservazione positiva scelta casualmente sia ordinata con maggior sospetto di due osservazioni negative scelte a caso. n_D è il numero di osservazioni del gruppo positivo e $n_{\bar{D}}$ è il numero di osservazioni del gruppo negativo.

Sfortunatamente, a causa dei coefficienti da un modello di regressione usato per calcolare i valori previsti, le stima basate sul modello delle probabilità previste non sono sempre indipendenti. Quando le probabilità previste sono stimate da un modello con misure ripetute, questa dipendenza sarà ancora più forte poiché in questo caso non solo i coefficienti comuni ma anche la correlazione intra-soggetto genera dipendenza sulle probabilità previste. Ciò comporta che la formula (1.2) non è sempre valida. La dipendenza sulle

osservazioni o sulle coppie discordanti non può essere ignorata e necessita di essere presa in considerazione nei calcoli per stimare la varianza di una curva ROC. Per ottenere una stima dell'errore standard dell'area sotto una curva ROC, un metodo *bootstrap* può essere utilizzato per stimare la varianza dell'area sotto una curva ROC, calcolata col metodo non parametrico di Wilcoxon (1.2) (Efron, 1979; Efron, Tibshirani, 1986). Si assuma che ci sia un totale di n soggetti, l'i-esimo soggetto ha n_i osservazioni (per $1 \leq i \leq n$) e che ci sia un totale di $N = \sum_{i=1}^{n} n_i$ osservazioni nel dataset. Per preservare la varianza originale intra-paziente e la struttura dei dati, l'algoritmo di ri-campionamento *bootstrap* è disegnato per campionare i dati a livello del soggetto piuttosto che a livello di osservazione. Vale a dire che dall'insieme degli n soggetti, è stato disegnato un campione casuale di n soggetti con rimpiazzo. Per ogni dato soggetto disegnato dal piano, si dica soggetto k, tutte le n_k osservazioni che appartengono a questo soggetto nel dataset originale saranno automaticamente incluse nel campione *bootstrap*. Tuttavia, un campione *bootstrap* consisterà ancora di n soggetti ma con solo s soggetti unici, dove s $\leq n$ (si ottiene l'uguaglianza se e solo se il campione *bootstrap* è identico al dataset originale). Un dataset *bootstrap* consisterà di $\widetilde{N}$ osservazioni con $\widetilde{N} = \sum_{i=1}^{n} s_i$, dove s_i è il numero di osservazioni dell'i-esimo soggetto disegnato. Usualmente $\widetilde{N}$ non uguaglia N, il numero originale di osservazioni.

Per ogni campione *bootstrap*, la statistica dell'area sotto la curva ROC è stimata usando l'algoritmo non parametrico di Wilcoxon (Liu e Wu, 2003). Supponiamo che siano generati r campioni *bootstrap*, allora r statistiche dell'area sotto le curve ROC $\widehat{AUC}_1, \widehat{AUC}_2, \ldots, \widehat{AUC}_r$ saranno stimate attraverso il metodo non parametrico di Wilcoxon. In base ai valori dell'area stimata sotto la curva, l'errore standard dell'area stimata originariamente sotto la curva è calcolato come:

$$\sqrt{\hat{\sigma}^2_{\widehat{AUC}}} = \widehat{SE}\left(\widehat{AUC}\right) = \sqrt{\frac{\sum_{i=1}^{r}(\widehat{AUC}_i - \overline{\widehat{AUC}})^2}{r-1}},$$

dove $\overline{\widehat{AUC}} = \sum_{i=1}^{r} \frac{\widehat{AUC}_i}{r}$ è la media delle r aree stimate sotto le curve ROC. La stabilità e accuratezza di $\widehat{SE}\left(\widehat{AUC}\right)$ sono determinate dalle stime di $\widehat{AUC}$ e il numero di r campioni ripetuti *bootstrap*. Più è grande il numero di campioni *bootstrap*, migliore è la stima dell'errore standard. L'errore standard $\widehat{SE}\left(\widehat{AUC}\right)$ tende ad essere stabile quando il numero di repliche r è grande.

CAPITOLO 2

"Non abbiamo un criterio che ci porti alla verità e quindi alla certezza,
ma abbiamo solo apparenze, che ci danno la probabilità.
Noi alla percezione certa del vero oggettivo non perveniamo,
ma ad essa ci avviciniamo con l'evidenza del probabile"
(Filone di Larissa, II sec a.C.)

VALUTAZIONE DI TEST DIAGNOSTICI SENZA GOLD STANDARD

2.1 INTRODUZIONE

Questo capitolo riassume brevemente gli sviluppi storici, con particolare attenzione verso quelli più recenti, dei metodi statistici proposti per stimare la sensibilità e la specificità di test di screening o diagnostici quando test fallibili non sono valutati contro un *gold standard*.

La performance di un test diagnostico è giudicata da quanto accuratamente il risultato del test può identificare una persona sana o malata; la sensibilità e la specificità sono utili misure per valutare l'accuratezza di test binari (esiti positivi e negativi). Il vero stato di malattia è "*gold standard*" contro il quale un test dovrebbe essere misurato. Comunque, ci sono molte malattie per le quali la diagnosi definitiva è difficile o costosa da stabilire. Questo è particolarmente vero per la diagnosi di una complessa condizione clinica nell'usuale scenario della pratica medica. Per esempio, è difficile stabilire una diagnosi definitiva di infarto miocardico per tutti i pazienti ammessi all'ospedale e la diagnosi del morbo di Alzheimer può essere definitiva solo dopo la morte del paziente e dopo aver effettuato l'esame neuropatologico. La biopsia e le

analisi istologiche sono di solito accettate come *gold standard* per i problemi di classificazione, ma resta sempre un margine di errore anche quando è disponibile un report patologico; la biopsia può fornire informazione sulla frazione di falsi positivi ma nel caso in cui una lesione non sia stata scoperta da un particolare studio e successivamente non sia stata eseguita la biopsia, il suo contributo alla frazione dei falsi negativi rimarrà non nota.

Così, dato che un test *gold standard* spesso non è disponibile, o è costoso in maniera proibitiva, o non etico o patologicamente invasivo da realizzare, sono necessari metodi alternativi di stima della sensibilità e specificità. Si parla di "*imperfect gold standard*" quando il vero stato di malattia non è noto, ma è indicato dal migliore test disponibile. Sebbene imperfetto, esso è tipicamente usato come uno standard contro il quale nuovi test sono valutati. Ci sono, comunque, situazioni nelle quali molteplici test sono valutati l'uno contro l'altro senza una designazione a priori di uno che possa essere l'*imperfect gold standard*. Una soluzione comune per l'*imperfect gold standard* è l'applicazione del *gold standard* ad un piccolo sotto-campione di soggetti. Questo tipo di disegno è stato usato per valutare un nuovo test contro un *imperfect gold standard* sotto vari schemi di campionamento. Quando la probabilità campionaria del sotto-campione dipende dai risultati di test imperfetti, si ha un problema di distorsione della verifica. Spesso, comunque, un vero *gold standard* semplicemente non esiste. Vengono di seguito trattati metodi statistici che stimano l'accuratezza diagnostica di uno o più nuovi test, con o senza l'utilizzo di un *imperfect gold standard*, quando il vero stato di malattia non è noto per qualche soggetto, che si focalizzano nella maggior parte dei casi su test diagnostici che producono risultati binari.

2.2 ERRORI DI UN IMPERFECT GOLD STANDARD

Si indichi con D e T le variabili casuali che rappresentano lo stato di malattia e i risultati del test, rispettivamente. Si assuma che lo stato di malattia sia binario, $D = d$, con $d = 0, 1$, dove $D = 1$ indica uno stato di malattia e $D = 0$ uno stato di assenza di malattia. Inoltre si consideri un test con due livelli di risultati: $T = t$, con $t = 0, 1$. In particolare, $T = 1$ se l'esito è positivo e $T = 0$ se l'esito è negativo.

Il soggetto studiato è sempre indicizzato da $i = 1, \dots, n$, il test diagnostico è indicizzato da $k = 1, \dots, K$.

Se uno studio coinvolge $S(> 1)$ popolazioni, la popolazione è indicizzata da $s = 1, \dots, S$.

La prevalenza dello stato di malattia (o incidenza) in una popolazione è denotato da $p = P(D = 1)$.

L'accuratezza diagnostica e i tassi di errore per il test k sono denotati dalle probabilità condizionate:

$$\eta_{kt}^{(d)} = P(T_k = t|D = d), \qquad k = 1, \dots, K \quad d = 0, 1 \quad t = 0, 1$$

Si noti che $\eta_{k1}^{(0)} = \alpha_k$ e $\eta_{k0}^{(1)} = \beta_k$, dove α_k e β_k si riferiscono alla terminologia utilizzata nel capitolo 1 (cfr. §1.3) per indicare i tassi di falso positivo e falso negativo, rispettivamente, relativi al test k.

Per stimare tali quantità, si fa ricorso a tabelle di contingenza che classificano gli individui in base al loro stato e all'esito dei test.

La classificazione degli individui nel caso in cui fosse noto il loro vero stato di malattia e si eseguisse un *imperfect gold standard* (detto T_1) sarebbe quella rappresentata nella tabella 2.1. Qui, n_{ik} si riferisce al numero di soggetti che cadono nella classe ik, dove i si riferisce al test T_1 e k al vero stato di malattia.

Tabella 2.1. Risultato del test mediante il vero stato di malattia.

		D		
		1	0	*Totale*
T_1	1	n_{1D}	$n_{1\bar{D}}$	n_{1+}
	0	n_{0D}	$n_{0\bar{D}}$	n_{0+}
	Totale	n_{+D}	$n_{+\bar{D}}$	n_{++}

Dalla Tabella 2.1 si possono ricavare le sensibilità e specificità del test T_1, vale a dire rispetto al vero stato di malattia:

$$P(T_1 = 1|D = 1) = \eta_{11}^{(1)} = \frac{n_{1D}}{n_{+D}},$$

$$P(T_1 = 0|D = 0) = \eta_{10}^{(0)} = \frac{n_{0\bar{D}}}{n_{+\bar{D}}}.$$

Si noti che tali quantità sono stimate sulla base degli n_{++} individui campionati, ma fotografando le capacità di T_1 con riferimento al vero stato di malattia, verranno interpretate come le "vere" quantità da studiare. Sotto un modello di campionamento binomiale, si possono ottenere anche le stime degli errori standard, necessarie per la costruzione di intervalli di confidenza.

$$SE\left(\eta_{11}^{(1)}\right) = \sqrt{\frac{n_{1D}\, n_{0D}}{n_{+D}^3}},$$

$$SE\left(\eta_{10}^{(0)}\right) = \sqrt{\frac{n_{1\bar{D}}\, n_{0\bar{D}}}{n_{+\bar{D}}^3}}.$$

Quando il vero stato di malattia dei soggetti non è noto, si possono sottoporre i soggetti in esame ad un *imperfect gold standard* test (T_1) e ad un secondo test (T_2). I dati osservati possono essere riassunti in una tabella di contingenza "osservata" (Tabella 2.2), in cui n_{ij} indica il numero di soggetti sottoposti ai due test, dove i si riferisce al primo test e j al secondo. Questa classificazione rappresenta la contaminazione della

Tabella 2.1 con l'introduzione del secondo test, per soccombere alla mancanza di informazione sul vero stato di malattia.

Tabella 2.2. Confronto tra i test T_1 (l'*imperfect gold standard*) e T_2 (un secondo test).

		T_2		
		1	***0***	***Totale***
$\boldsymbol{T_1}$	***1***	n_{11}	n_{10}	$\boldsymbol{n_{1+}}$
	0	n_{01}	n_{00}	$\boldsymbol{n_{0+}}$
	Totale	$\boldsymbol{n_{+1}}$	$\boldsymbol{n_{+0}}$	$\boldsymbol{n_{++}}$

Dalla Tabella 2.2 si evince che la sensibilità e la specificità dell'*imperfect gold standard* (T_1), stimate mediante T_2, sono:

$$\hat{\eta}_{11}^{(1)} = \frac{n_{11}}{n_{+1}},$$

$$\hat{\eta}_{10}^{(0)} = \frac{n_{00}}{n_{+0}}. \qquad (2.1)$$

La contaminazione della Tabella 2.2 attraverso l'introduzione del secondo test T_2 provoca delle stime distorte della sensibilità e specificità dell'*imperfect gold standard*. La difficoltà nella determinazione della distorsione potenziale può essere illustrata suddividendo i soggetti mediante il risultato del test T_2, che può essere visualizzata in una tabella 2×2×2 (Tabella 2.3).

Tabella 2.3. Introduzione della distorsione delle stime attraverso il confronto tra i test T_1, T_2 e il vero stato di malattia.

		$D=1$			$D=0$			
		T_2			T_2			
		1	***0***		***1***	***0***		***Totale***
$\boldsymbol{T_1}$	***1***	$\mathrm{n_{1D}} - \mathrm{r}$	r	$\mathrm{n_{1D}}$	t	$\mathrm{n_{1\bar{D}}} - \mathrm{t}$	$\mathrm{n_{1\bar{D}}}$	$\mathbf{n_{1+}}$
	0	$\mathrm{n_{0D}} - \mathrm{s}$	s	$\mathrm{n_{0D}}$	u	$\mathrm{n_{0\bar{D}}} - \mathrm{u}$	$\mathrm{n_{0\bar{D}}}$	$\mathbf{n_{0+}}$
	Totale	$\mathbf{n_{+D} - (r+s)}$	**r+s**		**t+u**	$\mathbf{n_{+\bar{D}} - (t+u)}$	$\mathbf{n_{+\bar{D}}}$	$\mathbf{n_{++}}$

Le lettere r, s, t e u rappresentano gli errori del test T_2; r e s sono i falsi negativi; t e u sono i falsi positivi. Sommando sul vero stato di malattia si ottiene la Tabella 2.4.

Tabella 2.4. Distorsioni introdotte con l'utilizzo dei test T_1 e T_2 in assenza di informazione sul vero stato di malattia.

		T_2		***Totale***
		1	***0***	
$\boldsymbol{T_1}$	***1***	$\mathrm{n_{11}} = \mathrm{n_{1D}} - \mathrm{r} + \mathrm{t}$	$\mathrm{n_{10}} = \mathrm{n_{1\bar{D}}} + \mathrm{r} - \mathrm{t}$	$\mathbf{n_{1+}}$
	0	$\mathrm{n_{01}} = \mathrm{n_{0D}} - \mathrm{s} + \mathrm{u}$	$\mathrm{n_{00}} = \mathrm{n_{0\bar{D}}} + \mathrm{s} - \mathrm{u}$	$\mathbf{n_{0+}}$
	Totale	$\mathbf{n_{+1} = n_{+D} - (r+s) + (t+u)}$	$\mathbf{n_{+0} = n_{+\bar{D}} + (r+s) - (t+u)}$	$\mathbf{n_{++}}$

Usando i risultati della Tabella 2.4, la sensibilità e la specificità del test T_1, relativamente ai risultati prodotti dal test T_2, possono essere scritte come:

$$\hat{\eta}_{11}^{(1)} = \frac{n_{1D} - r + t}{[n_{+D} - r + t - s + u]},$$

$$\hat{\eta}_{10}^{(0)} = \frac{n_{0D} + s - u}{[n_{+\bar{D}} + r - t + s - u]}.$$

Queste ultime due espressioni illustrano che le connessioni tra la sensibilità e la specificità "vere" e stimate non sono chiare. La sensibilità e specificità stimata potrebbe essere distorta sia verso l'alto che verso il basso, a seconda della quota di errori nel test T_2 distribuita sulle 4 celle della Tabella 2.4. È anche possibile che gli errori t - r e s - u siano pari a zero, in questo caso le stime della sensibilità e specificità sarebbero corrette. Tuttavia, generalizzazioni circa la distorsione nella sensibilità e specificità stimata richiedono una considerevole cautela.

2.3 IDENTIFICABILITÀ DEL MODELLO STATISTICO

Nonostante le considerazioni conclusive della sezione precedente, per valutare gli errori di un *imperfect gold standard* (T_1) per un campione casuale di soggetti da una singola popolazione, i primi studi usualmente hanno applicato un secondo test (T_2).

Sotto le ipotesi del disegno precedentemente descritto, in cui si è assunto che il vero stato di malattia abbia due livelli, i parametri coinvolti riguardano la prevalenza della malattia nella popolazione, p, i tassi di falsi positivi, $\eta_{11}^{(0)} = P(T_1 = 1|D = 0)$ *dell'imperfect gold standard* test e $\eta_{21}^{(0)} = P(T_2 = 1|D = 0)$ del secondo test e il tasso di falsi negativi $\eta_{10}^{(1)} = P(T_1 = 0|D = 1)$ dell'*imperfect gold standard* e $\eta_{20}^{(1)} = P(T_2 = 0|D = 1)$ del secondo test, oltre alla correlazione tra i due test condizionatamente al vero stato di malattia (ρ_D e $\rho_{\bar{D}}$). La sensibilità e la specificità per il k-esimo test sono dati da $\eta_{k1}^{(1)}$ e $\eta_{k0}^{(0)}$, rispettivamente, per $k = 1, 2$. Si indichi con $\boldsymbol{\theta} = \left(p, \eta_{11}^{(0)}, \eta_{21}^{(0)}, \eta_{10}^{(1)}, \eta_{20}^{(1)}, \rho_D, \rho_{\bar{D}}\right)$ il vettore dei parametri non noti da stimare.

Per stimare tali parametri, molti metodi statistici fanno l'assunzione che i due test siano indipendenti, condizionatamente al vero stato della malattia, cioè:

$$P(T_1 \cap T_2|D) = P(T_1|D)\, P(T_2|D) \tag{2.2}$$

L'indipendenza condizionata è un'assunzione piuttosto forte, tanto da indurre vari autori ad evidenziare che non è realistica in molte situazioni pratiche. Per esempio, quando c'è uno spettro di gravità della malattia, i casi più acuti sono colti quasi certamente da qualunque test mentre i casi meno gravi possono risultare negativi in più di un test.

Assumendo l'indipendenza condizionata tra i test, si ha che $\boldsymbol{\theta} = \left(p, \eta_{11}^{(0)}, \eta_{21}^{(0)}, \eta_{10}^{(1)}, \eta_{20}^{(1)}, 0{,}0\right)$. La dimensione parametrica ora è ridotta a 5. Il numero di gradi di libertà del modello statistico (df) è 3, ricordando che $df = \zeta\varrho - 1$, dove ζ e ϱ indicano il numero di righe e colonne, rispettivamente. Essendo $dim(\boldsymbol{\theta}) > df$, il modello non è identificabile, a meno che non si pongano dei vincoli o non si utilizzino tecniche quali l'introduzione di altri test, stime bayesiane, ecc.

Se poi si rilassa l'ipotesi di indipendenza condizionata tra i test, il numero dei parametri da stimare aumenta in quanto s'introduce una correlazione tra i test diversa da zero; è necessario porre ulteriori restrizioni o proporre strategie alternative per accrescere i gradi di libertà ed ottenere così l'identificabilità del modello. Infatti, quando un dato non è indipendente, l'informazione che esso fornisce è già contenuta implicitamente negli altri.

Numerosi autori hanno affrontato questo problema di identificazione del modello utilizzando tecniche diverse. Il lavoro prodotto da Gart e Buck (1966) è stato il primo tentativo comprensibile di trattare un *imperfect gold standard* in questa situazione. Si introdussero parecchi importanti concetti che furono ampiamente sviluppati più tardi da altri autori, come l'uso della malattia come una variabile latente, idea ampiamente utilizzata nella maggior parte dei metodi successivi. I metodi pubblicati sulla valutazione di test diagnostici imperfetti possono essere raggruppati approssimativamente in due classi. Le tecniche più semplici spesso assumono che i tassi di errore dell'*imperfect gold standard* (o altri sottoinsiemi di parametri) siano noti (sezione 2.4.1 e 2.5.1). I metodi più recenti generalmente non richiedono una conoscenza a priori dei tassi d'errore di qualche test studiato (sezione 2.4.2 e 2.5.2).

Si elencano ora alcune strategie proposte in letteratura, suddividendole in due grandi classi: quelle che assumono l'indipendenza condizionata dei test come definiti dall'equazione 2.2, che riguardano generalmente i primi sviluppi (cfr. §2.4), e quelli che, invece, la rilassano (cfr. §2.5).

2.4 ASSUNZIONE DI INDIPENDENZA CONDIZIONATA

2.4.1 *Parametri noti*

Quando c'è indipendenza condizionata, assumere che due parametri siano noti permette la stima dei tre parametri rimasti. Per esempio, Staquet et al. (1981) hanno proposto stimatori puntuali dei parametri ignoti, senza fornire però stimatori della varianza per $\eta_{21}^{(0)}$ e $\eta_{20}^{(1)}$, in tre casi speciali:

1) $\eta_{11}^{(0)}$ e $\eta_{10}^{(1)}$ noti: tassi di falsi positivi e falsi negativi dell'*imperfect gold standard* (T_1) noti;
2) $\eta_{10}^{(1)} = \eta_{20}^{(1)} = 0$: tassi di falsi negativi dei due test noti;
3) $\eta_{10}^{(1)} = 0$: tasso di falsi negativi dell'*imperfect gold standard* (T_1) pari a zero.

$\boxed{CASO\ 1: \eta_{11}^{(0)} e\ \eta_{10}^{(1)} noti}$

Questo caso è equivalente alla soluzione proposta da Gart e Buck (1966). Quando $\eta_{11}^{(0)}$ e $\eta_{10}^{(1)}$ sono noti, le stime di $\eta_{21}^{(0)}$ e $\eta_{20}^{(1)}$ sono linearmente dipendenti, ma le stime individuali non possono essere ottenute. Comunque, se si assume l'indipendenza condizionata degli errori tra due test dato il vero stato di malattia, il numero di parametri non noti ($p, \eta_{21}^{(0)} e\ \eta_{20}^{(1)}$) è ridotto a 3, che è uguale al numero di gradi di libertà nei dati ($df = 2 \cdot 2 - 1 = 3$). Le frequenze attese di ogni cella nella Tabella 2.2 sono date da:

$$f_{11} = \frac{n_{11}}{n_{++}} = p\left(1 - \eta_{10}^{(1)}\right)\left(1 - \eta_{20}^{(1)}\right) + (1-p)\eta_{11}^{(0)}\eta_{21}^{(0)},$$

$$f_{12} = \frac{n_{12}}{n_{++}} = p\left(1 - \eta_{10}^{(1)}\right)\eta_{20}^{(1)} + (1-p)\eta_{11}^{(0)}\left(1 - \eta_{21}^{(0)}\right),$$

$$f_{21} = \frac{n_{21}}{n_{++}} = p\,\eta_{10}^{(1)}\left(1 - \eta_{20}^{(1)}\right) + (1-p)\left(1 - \eta_{11}^{(0)}\right)\eta_{21}^{(0)},$$

$$f_{22} = \frac{n_{22}}{n_{++}} = p\,\eta_{10}^{(1)}\eta_{20}^{(1)} + (1-p)\left(1 - \eta_{11}^{(0)}\right)\left(1 - \eta_{21}^{(0)}\right).$$

Per stimare la sensibilità e specificità del nuovo test e la prevalenza della malattia, bisogna massimizzare una verosimiglianza multinomiale. Si possono derivare le formule per la stima della varianza approssimata degli stimatori usando il metodo delta. In assenza di alcuna conoscenza dei tassi di errore dell'*imperfect gold standard*, Gart e Buck (1966) propongono di minimizzare la distorsione nei tassi di errore stimati del test T_2, $\left(\eta_{21}^{(0)} - \hat{\eta}_{21}^{(0)}\right)$ e $\left(\eta_{20}^{(1)} - \hat{\eta}_{20}^{(1)}\right)$, stimando la sensibilità in una popolazione caratterizzata da un'alta prevalenza della malattia e la specificità in una popolazione con bassa prevalenza. Gart e Buck (1966) hanno notato anche l'utilità di informazione aggiuntiva derivata dai dati su due o più sotto-popolazioni con la stessa sensibilità e specificità del test ma differente prevalenza. Sfortunatamente nei primi anni la massimizzazione della verosimiglianza congiunta da molteplici popolazioni era difficile dal punto di vista computazionale. Questo approccio, comunque, getta le fondamenta per molti studi successivi.

$\boxed{CASO\ 2: \eta_{10}^{(1)} = \eta_{20}^{(1)} = 0}$

Il secondo caso è stato discusso in un contesto più generale da Mantel (1951).

Assumendo specificità perfette, per esempio $\eta_{k0}^{(1)} = 0$ per i test, con $k = 1, 2$, e indipendenza condizionata tra i test, Mantel ha derivato una stima di massima verosimiglianza della sensibilità e prevalenza per le situazioni dove:

(1) lo stesso test è applicato molteplici volte;

(2) test differenti sono applicati agli stessi individui;

(3) il numero di test per persona varia tra gli individui;

(4) l'applicazione di test successivi dipende dai risultati di test precedenti.

Per tutti gli stimatori sono fornite stime delle varianze approssimate delle varianze degli stimatori.

$\boxed{CASO\ 3: \eta_{10}^{(1)} = 0}$.

Relativamente al terzo caso, Staquet et al. (1981) hanno derivato una stima puntuale per $\eta_{21}^{(0)}$ e un intervallo di confidenza per $\eta_{20}^{(1)}$.

2.4.2 *Parametri non noti*

Quando nessun parametro è assunto noto, uno studio basato sulla Tabella 2.2 ha più parametri non noti che gradi di libertà per la stima. Si può ottenere l'identificazione solo quando il disegno dello studio sopra è modificato in modo da produrre più gradi di libertà per la stima. I gradi di libertà aggiuntivi possono essere ottenuti:

a) applicando più di due test agli stessi individui in una popolazione (è permesso utilizzare più volte lo stesso test);
b) applicando i test a più di una popolazione;
c) applicando lo stesso test ripetutamente nel tempo agli stessi individui nella popolazione, permettendo la presenza di casi incidentali di malattia tra i test.

Le prime due tecniche (a e b) si riferiscono a studi di prevalenza, mentre l'ultima (c) a studi di incidenza.

Come già visto nel capitolo 1, la prevalenza misura la proporzione di eventi presenti in una popolazione *in un dato momento*. Poiché il fattore tempo non è importante nel calcolo della prevalenza, questa misura è di tipo statico e quindi non è un tasso; si tratta invece di una proporzione.

L'incidenza, invece, misura la proporzione di nuovi eventi che si verificano in una popolazione in un dato lasso di tempo; per esempio, essa rappresenta la proporzione di individui che vengono colpiti dalla malattia *in un determinato periodo di tempo*. L'incidenza misura allora il numero di nuovi casi nel periodo di tempo ed individua il rischio (cioè la probabilità) di contrarre la malattia cui è soggetto un individuo esposto in quella popolazione. Può essere vista come un modo per misurare la velocità di transizione dallo stato di

salute (assenza di malattia) allo stato di malattia in una popolazione. Essa rappresenta la variazione di una quantità (i nuovi ammalati) rispetto alla variazione di un'altra quantità (il tempo); quindi è una misura dinamica e costituisce un vero tasso. Metaforicamente parlando, la prevalenza è come una fotografia, mentre l'incidenza è come un film.

I metodi affrontati nel caso di parametri non noti, assumendo l'indipendenza condizionata, si distinguono in studi di prevalenza (sezione 2.4.2.1) e studi d'incidenza (sezione 2.4.2.2).

2.4.2.1 *Studi di prevalenza*

- *Piu' di due test in una popolazione*

Dawid e Skene (1979) hanno proposto un approccio generale alla stima dei tassi di errore di molteplici test applicati simultaneamente ad un campione casuale per una singola popolazione quando può essere assunta l'indipendenza condizionata tra test. Senza designare alcun test come l'*imperfect gold standard*, essi considerano uno studio nel quale più test con differenti tassi di errore sono applicati agli stessi individui che appartengono ad una singola popolazione.

Si assuma che ci siano $K\ (>2)$ test, ognuno dei quali ha 2 possibili risultati corrispondenti agli stati della malattia. Non tutti i test necessitano di essere applicati ad ogni soggetto e alcuni test potrebbero essere applicati più di una volta allo stesso soggetto.

Si denoti con n_{itk} il numero di volte che il test T_k ha come risultato t nel soggetto i, $i = 1, \dots, n$.

Si indichi con δ_{id}, per $d = 0, 1$, un insieme di variabili indicatrici per il soggetto i, tali che, se lo stato della malattia del soggetto i-esimo è d^* $(d_i = d^*)$, allora $\delta_{id^*} = 1$ e $\delta_{id} = 0$ $(se\ d \neq d^*)$.

Se d_i, e di qui δ_{id}, fossero noti, la verosimiglianza sarebbe data da:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \left[p \prod_{\mathrm{k}=1}^{\mathrm{K}} \prod_{\mathrm{t}=0}^{1} \left(\eta_{kt}^{(1)}\right)^{n_{itk}} \right]^{\delta_{i1}} \left[(1-p) \prod_{\mathrm{k}=1}^{\mathrm{K}} \prod_{\mathrm{t}=0}^{1} \left(\eta_{kt}^{(0)}\right)^{n_{itk}} \right]^{\delta_{i0}} \right\}, \tag{2.3}$$

dove $\boldsymbol{\theta} = \left(p, \eta_{k1}^{(0)}, \eta_{k1}^{(1)}\right)$, $k = 1, \dots, K$, che può essere massimizzata con rispetto ai valori di p e $\eta_{kt}^{(d)}$, con

$k = 1, \dots, K, t = 0, 1,\ d = 0, 1,$ per produrre stime per le prevalenze dello stato di malattia della popolazione e i tassi d'errore di tutti i test.

Se il vero stato di malattia dei soggetti non è noto, la verosimiglianza è:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}\left[p\prod_{\mathrm{k=1}}^{\mathrm{K}}\prod_{\mathrm{t=0}}^{1}\left(\eta_{kt}^{(1)}\right)^{n_{itk}} + (1-\mathrm{p})\prod_{\mathrm{k=1}}^{\mathrm{K}}\prod_{\mathrm{t=0}}^{1}\left(\eta_{kt}^{(0)}\right)^{n_{itk}}\right]. \tag{2.4}$$

Quando il vero stato di malattia non è noto, Dawid e Skene (1979) suggeriscono di usare l'algoritmo EM (cfr. Appendice §A.1) per massimizzare la verosimiglianza (2.4), trattando il vero stato di malattia, d, come dato mancante in (2.3) e massimizzando (2.3).

- *Due test in una popolazione con introduzione di un terzo test "resolver" nell'analisi discrepante*

La presenza di distorsione nelle stime della sensibilità e specificità conduce all'idea di usare un ulteriore test per localizzare e correggere gli errori nel test T_1. L'analisi discrepante è una tecnica che cerca di risolvere la discordanza di esito di due test (un *imperfect gold standard*, T_1, e un secondo test, T_2) mediante l'introduzione di un terzo test, detto "*resolver*". L'analisi discrepante convenzionale (Hadgu, 1997) sottopone i soggetti nelle celle dove il test T_1 è risultato positivo e il test T_2 negativo al test *resolver* e usa questo per calcolare una sensibilità "*resolved*" $(n_{11}+r)/(n_{+1}+r)$. Una risoluzione simile, laddove il test T_1 fosse negativo e il test T_2 positivo, è stato usato nella stessa maniera per ottenere una specificità "*resolved*" $(n_{00}+u)/(n_{+0}+u)$. Come è stato notato (Hadgu, 1996, 1997, 1998, Miller, 1998), questo approccio non rimuove le distorsioni nella sensibilità e specificità stimata. Usare un test "*resolver"* perfettamente specifico per risolvere lo stato dei soggetti nelle celle in alto a destra della Tabella 2.4 corregge solo i casi r nei quali il test T_2 ha generato un falso positivo, o i casi u dove il test T_1 ha comportato un falso negativo. Poiché la distorsione coinvolge tutte le 4 quantità r, s, t e u, conoscere il valore di solo una di queste quantità può condurre ad una stima non corretta della sensibilità. Analogamente, l'uso di un test "*resolver"* perfettamente sensibile sulle celle in basso a sinistra della Tabella 2.4 non conduce ad una stima corretta della specificità. Mentre l'analisi discrepante conduce a stime della sensibilità e specificità "*resolved"* che sono sempre più alte delle sensibilità e specificità stimate in (2.1), non è automaticamente vero che sia pure più distorto. In circostanze differenti entrambe queste stime potenzialmente distorte potrebbero essere più vicine alla sensibilità e specificità "vere" e nessuno dei due è garantito essere il meno distorto.

2.4. ASSUNZIONE DI INDIPENDENZA CONDIZIONATA

L'analisi discrepante convenzionale non può essere certa di produrre stime corrette della vera sensibilità e specificità del test T_1. Come mostra la tabella estesa (Tabella 2.3), produrre stime corrette della sensibilità e/o specificità del test T_1 richiede o i valori esatti delle differenze $(t-r)$ e $(s-u)$ o almeno le loro stime. Questi valori potrebbero essere trovati direttamente dall'intera tabella 2×2×2 (Tabella 2.3), che mostra il vero stato del soggetto insieme alla classificazione mediante i test T_1 e T_2, se questo fosse disponibile. Si definisca allora con il termine "*resolver* perfetto" un test in grado di individuare il vero stato di malattia, capace di ottenere il "vero" valore delle celle discordanti e concordanti nell'originale tabella 2×2 (Tabella 2.1). Si indichi con $\hat{p}_{ijk}$ le stime delle 8 probabilità p_{ijk} che rappresentano le 8 condizioni nella tabella 2×2×2 (Tabella 2.3), dove il pedice i si riferisce al test T_1, j al test T_2 e k al vero stato di malattia. Tali stime possono essere ottenute considerando un totale di n soggetti e sottoponendoli ai test T_1 e T_2 ed anche al test "*resolver*" perfetto. Si scriva n_{ijk} per il numero di soggetti nella cella i,j,k della Tabella 2.3. Le stime $\hat{p}_{ijk}$ e il loro errore standard sono date da:

$$\hat{p}_{ijk} = \frac{n_{ijk}}{n},$$

$$SE(\hat{p}_{ijk}) = \sqrt{\frac{\hat{p}_{ijk}(1-\hat{p}_{ijk})}{n}}.$$

Si ottengono stime puntuali ed errori standard consistenti della "vera" sensibilità e specificità del test T_1 sommando attraverso i livelli del test T_2.

Tuttavia questo scenario non è realistico. Se il "*resolver*" perfetto fosse applicato su tutti i soggetti, esso potrebbe essere usato al posto del test T_2. Spesso non è usato come test di riferimento perché è difficile o costoso da applicare. In questo caso, non può essere utilizzato su tutti gli n soggetti ma potrebbe ragionevolmente essere applicato su un loro sottoinsieme. Questo solleva la possibilità di campionare casualmente dalle celle della tabulazione 2×2 del test T_1 mediante il test T_2 (Tabella 2.4), e valutare questi soggetti con il test "*resolver"* perfetto. Le frequenze osservate della tabella collassata 2×2 definita mediante i test T_1 e T_2 sono:

Tabella 2.5. Confronto tra i test T_1 e T_2 con l'introduzione di un terzo test "resolver".

		T_2		
		1	*0*	***Totale***
$\boldsymbol{T_1}$	*1*	n_{11+}	n_{10+}	$\boldsymbol{n_{1++}}$
	0	n_{01+}	n_{00+}	$\boldsymbol{n_{0++}}$
	Totale	$\boldsymbol{n_{+1+}}$	$\boldsymbol{n_{+0+}}$	$\boldsymbol{n}$

Le stime delle "vere" probabilità marginali 2×2 p_{ij+}, $\hat{p}_{ij+}$, e degli errori standard corrispondenti sono date da:

$$\hat{p}_{ij+} = \frac{n_{ij+}}{n}, \tag{2.5}$$

$$SE(\hat{p}_{ij+}) = \sqrt{\frac{\hat{p}_{ij+}(1-\hat{p}_{ij+})}{n}}.$$

Si supponga di testare dei numeri possibilmente più piccoli m_{ij} di questi n_{ij+} soggetti usando il test "*resolver*" perfetto, trovando che una proporzione $\hat{\pi}_{ij}$ di questi sono malati: $\hat{\pi}_{ij}$ stima la probabilità condizionata che un soggetto sia malato dato che il soggetto è classificato nella cella i,j mediante il test T_1 e T_2. Il suo errore standard stimato è dato da:

$$SE(\hat{\pi}_{ij}) = \sqrt{\frac{\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}{m_{ij}}}. \tag{2.6}$$

Questa stima della probabilità condizionata della malattia data la classificazione i,j e la stima della probabilità marginale $\hat{p}_{ij+}$ può essere moltiplicata per ottenere una stima della probabilità congiunta p_{ij1}.

$$\hat{p}_{ij1} = \hat{p}_{ij+}\hat{\pi}_{ij}. \tag{2.7}$$

L'errore standard di questa stima (2.7) può essere approssimativamente trovata usando il metodo delta come:

$$SE(\hat{p}_{ij+}\hat{\pi}_{ij}) = \sqrt{(\hat{p}_{ij+})^2\,\frac{\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}{m_{ij}} + (\hat{\pi}_{ij})^2\,\frac{\hat{p}_{ij+}(1-\hat{p}_{ij+})}{n}}.$$

Dalle stime 2.5, 2.6 e 2.7 e dai loro corrispondenti errori standard, si ricavano le stime della sensibilità e specificità del test T_1. La probabilità p_{1+1} del test T_1 che dà un risultato positivo vero è stimata come:

$$\hat{p}_{1+1} = \hat{p}_{11+}\hat{\pi}_{11} + \hat{p}_{10+}\hat{\pi}_{10.}$$

La probabilità p_{0+1} del test T_1 che dà un risultato falso negativo è stimata come:

$$\hat{p}_{0+1} = \hat{p}_{01+}\hat{\pi}_{11} + \hat{p}_{00+}\hat{\pi}_{00}.$$

La vera sensibilità può poi essere stimata da (Begg e Greenes, 1983):

$$\hat{p}_{i+1} = (\hat{p}_{11+}\hat{\pi}_{11} + \hat{p}_{10+}\hat{\pi}_{10})/(\hat{p}_{11+}\hat{\pi}_{11} + \hat{p}_{10+}\hat{\pi}_{10} + \hat{p}_{01+}\hat{\pi}_{11} + \hat{p}_{00+}\hat{\pi}_{00}).$$

Mentre le stime della vera sensibilità e specificità possono essere derivate in una maniera chiara, la derivazione dei loro errori standard è più complicata perché è necessario tenere in considerazione la correlazione tra le celle differenti nella tabella. Per semplicità si definisca con

$$X := \hat{p}_{11+}\hat{\pi}_{11} + \hat{p}_{10+}\hat{\pi}_{10}.$$

la probabilità stimata di un risultato vero positivo mediante il test T_1 e con

$$Y := \hat{p}_{11+}\hat{\pi}_{11} + \hat{p}_{10+}\hat{\pi}_{10} + \hat{p}_{01+}\hat{\pi}_{11} + \hat{p}_{00+}\hat{\pi}_{00}.$$

la prevalenza stimata della malattia. Per derivare la varianza campionaria della sensibilità stimata X/Y (espressa come una proporzione, non una percentuale), bisogna considerare l'originale tabella 2×2 (Tabella 2.5). Le 4 frequenze delle celle seguono una distribuzione congiunta multinomiale, con la covarianza tra due celle qualunque data da:

$$Cov(n_{ij+}, n_{km+}) = -n\hat{p}_{ij+}\hat{p}_{km+.}$$

Le frequenze del test "*resolver*" $\hat{\pi}_{ij}$ coinvolgono test separati delle 4 celle e ci si aspetta che tali test siano statisticamente indipendenti gli uni dagli altri. Se si considerano due termini generici coinvolti nelle stime della vera sensibilità e specificità, $\hat{p}_{ij+}\hat{\pi}_{ij}$ e $\hat{p}_{km+}\hat{\pi}_{km}$, la loro covarianza è stimata da:

$$Cov(\hat{p}_{ij+}\hat{\pi}_{ij}, \hat{p}_{km+}\hat{\pi}_{km}) = -\frac{\hat{p}_{ij+}\hat{p}_{km+}\hat{\pi}_{ij}\hat{\pi}_{km}}{n}.$$

Usando queste covarianze a coppie, si può calcolare l'errore standard delle probabilità stimate di veri positivi, o veri negativi, e delle loro covarianze. Possono essere calcolate le varianze della prevalenza e (1 −prevalenza). In generale, la forma della varianza di una somma delle $\hat{p}_{ij+}\hat{\pi}_{ij}$ è data da:

$$Var\left(\sum_i\sum_j \hat{p}_{ij+}\hat{\pi}_{ij}\right) = \sum_i\sum_j Var(\hat{p}_{ij+}\hat{\pi}_{ij}) + 2\sum_{i\le k}\sum_{j\le m} Cov(\hat{p}_{ij+}\hat{\pi}_{ij}\,, \hat{p}_{km+}\hat{\pi}_{km}).$$

Specificatamente, la stima della varianza della probabilità di vero positivo è data da:

$$Var(\hat{p}_{11+}\hat{\pi}_{11} + \hat{p}_{10+}\hat{\pi}_{10}) = Var(\hat{p}_{11+}\hat{\pi}_{11}) + Var(\hat{p}_{10+}\hat{\pi}_{10}) + 2Cov(\hat{p}_{11+}\hat{\pi}_{11}, \hat{p}_{10+}\hat{\pi}_{10})$$

$$= (\hat{p}_{11+})^2\,\frac{\hat{\pi}_{11}(1-\hat{\pi}_{11})}{m_{11}} + (\hat{\pi}_{11})^2\,\frac{\hat{p}_{11+}(1-\hat{p}_{11+})}{n} + (\hat{p}_{10+})^2\,\frac{\hat{\pi}_{10}(1-\hat{\pi}_{10})}{m_{10}} +$$

$$+(\hat{\pi}_{10})^2\,\frac{\hat{p}_{10+}(1-\hat{p}_{10+})}{n} - 2\frac{\hat{p}_{11+}\hat{p}_{10+}\hat{\pi}_{11}\hat{\pi}_{10}}{n}.$$

Questi termini sono necessari per calcolare gli errori standard per la sensibilità e la specificità, che sono usati per costruire intervalli di confidenza per le stime. Le varianze per la sensibilità e specificità possono essere calcolate usando il metodo delta per la varianza del rapporto delle due variabili correlate (X, Y).

$$Var\left(\frac{X}{Y}\right) = \frac{Var(X)}{[E(Y)]^2} + \frac{[E(X)]^2}{[E(Y)]^4}Var(Y) - 2\frac{E(X)}{[E(Y)]^3}Cov(X,Y).$$

Un intervallo di confidenza approssimato può essere generato per la sensibilità (o specificità) usando l'usuale approssimazione normale 'stima ± 2 errore standard'. Come notato in Agresti e Coull (1998) questo intervallo di confidenza comune non ha una copertura ben controllata per campioni piccoli, particolarmente se la sensibilità è vicina a 1. Per migliorare il controllo di tale copertura, essi raccomandano la procedura "Wald aggiustata", in cui, in ogni calcolo di una proporzione o un errore standard, si aggiungono al campione due positivi artificiali e due negativi artificiali.

- *Due test in due popolazioni*

Per il caso con due test applicati simultaneamente a due popolazioni, i dati come nella Tabella 2.1 sono disponibili per due campioni indipendenti con differenti prevalenze della malattia nella popolazione. Assumendo l'indipendenza condizionata tra i test, Hui e Walter (1980) forniscono una forma stretta di stimatori di massima verosimiglianza delle prevalenze della malattia nella popolazione e i tassi d'errore di entrambi i test.

- *Piu' di due test in s popolazioni*

Quando ci sono due livelli di stato della malattia e ciascuno dei due o più test è applicato esattamente una volta per ogni soggetto, i due disegni degli studi della prevalenza appena descritti possono essere posti nel lavoro generale di Walter e Irwig (1988). Il modello generale include di applicare K test differenti con risultati binari agli stessi individui in S popolazioni che hanno differenti prevalenze della malattia (p_s), con $s = 1, \dots, S$. Quando si può assumere l'indipendenza condizionata tra i test, la log-verosimiglianza per tutti questi disegni può essere posta in un'espressione generale:

$$l\left(t(k); p_s, \eta_{ks1}^{(0)}, \eta_{ks0}^{(1)}\right) =$$

$$\sum_{s=1}^{S}\sum_{t=0}^{1} n_s\,(t)\; ln\left[p_s \prod_{k=1}^{K}\left(\eta_{ks0}^{(1)}\right)^{1-t(k)}\left(1-\eta_{ks0}^{(1)}\right)^{t(k)} + (1-p_s)\prod_{k=1}^{K}\left(\eta_{ks1}^{(0)}\right)^{t(k)}\left(1-\eta_{ks1}^{(0)}\right)^{1-t(k)}\right]$$

dove $\eta_{ks1}^{(0)}$ e $\eta_{ks0}^{(1)}$ denotano i tassi di falsi positivi e falsi negativi per il test k nella popolazione s e $t(k)$ denota il risultato del test k. La seconda sommatoria è su tutte le combinazioni di risultati osservati $\boldsymbol{t} = (t(1), \dots, t(K))$, e $n_s(t)$ è il numero di individui nella popolazione s con risultato t.

Per ogni valore dato di K e S in tali disegni, il dato fornisce $S(2^K - 1)$ gradi di libertà per $S(2K + 1)$ parametri se la prevalenza e il tasso d'errore del test variano con le popolazioni.

Per il disegno comune con una popolazione, cioè $S = 1$, il numero di parametri è $(2K + 1)$ e il numero di gradi di libertà è $2^K - 1$. Esso impiega almeno 3 test per individuo affinché i parametri possano essere stimati. Quando il numero di parametri eccede il numero di gradi di libertà, è necessario porre alcune restrizioni sui tassi d'errore affinché possano essere stimati. Per esempio, quando $K = 2$, il numero di parametri è $5S$, che supera i $3S$ gradi della libertà. Il numero di parametri può essere ridotto a $3S$ se le sensibilità e le specificità sono assunte essere le stesse tra le popolazioni. Walter e Irwig (1988) presentano molti esempi specifici dalla letteratura con varie combinazioni di K e S. Essi coprono alcuni disegni speciali irregolari, che includono differenti schemi di test sequenziali.

2.4.2.2 *Approcci bayesiani*

Come descritto nei paragrafi precedenti, avendo solo 3 gradi di libertà ma almeno cinque parametri da stimare (la sensibilità e la specificità di ogni test e la prevalenza), se si usa un approccio frequentista che permetta l'indipendenza condizionata, due dei cinque parametri ignoti devono essere assunti noti per poter stimare gli altri tre. Forzare qualche parametro ad un valore fisso dà una soluzione per un problema semplice e identificabile, ma non è ovvio quale di questi parametri dovrebbe essere tenuto costante o a quale valore dovrebbero essere ristretti. Ignorando la possibile dipendenza tra i risultati del test, Joseph et al. (1995) hanno dimostrato che un approccio bayesiano per la stima dei tassi di errore diagnostico in una singola popolazione può essere usato per ottenere distribuzioni a posteriori interpretabili per ogni parametro ignoto relativo ad una data distribuzione a priori. Il caso più semplice prevede di applicare un singolo test diagnostico in una popolazione. I parametri non noti della prevalenza (p), sensibilità $\eta_{k1}^{(1)}$ e specificità $\eta_{k0}^{(0)}$ sono dati da distribuzioni a priori con differenti parametri. Le distribuzioni a priori sono scelte essere distribuzioni beta, coniugate naturali per la binomiale, e la distribuzione a posteriori congiunta di $\left\{p, \eta_{k1}^{(1)}, \eta_{k0}^{(0)}\right\}$ è ottenuta mediante un algoritmo campionario di Gibbs. Lo stesso approccio è poi applicato al caso con due test diagnostici condizionatamente indipendenti applicati agli stessi individui in una popolazione. I parametri per la distribuzione a priori di qualche parametro del test sono scelti per soddisfare la media e la deviazione standard dei valori di parametri del test dedotti da opinioni di esperti e dalla letteratura. Il campionamento di Gibbs può anche essere usato negli studi con test ripetuti senza *gold standard.*

- *MODELLO DI CHONG ET AL. (2007)*

Chong et al. (2007) hanno sviluppato una metodologia bayesiana per stime non parametriche delle curve ROC usate per valutare l'accuratezza della procedura diagnostica, stimando congiuntamente le coppie

(sensibilità, 1-specificità). Il metodo è basato su un modello multinomiale per la distribuzione congiunta delle osservazioni del test positivo e del test negativo, usando un approccio bayesiano che assicura la naturale proprietà della monotonicità della stima della risultante curva ROC. L'obiettivo è stimare la curva ROC di una procedura diagnostica (T_2) misurata su una scala ordinale o continua comparandolo ad un *imperfect gold standard* binario di riferimento (T_1). Si assuma che i due tipi di test siano applicati simultaneamente ad ogni individuo in campioni da S popolazioni (o S strati nella popolazione). Senza perdita di generalità, si assuma che alti valori per il test T_2 siano associati con un'alta verosimiglianza di presenza della malattia. Per il j-esimo valore di *cut-off* $(1 \leq j \leq J+1)$ del test T_2, si denoti con $T_{2,j}$ il corrispondente test dicotomizzato. Si indichi con $\eta_{21,j}^{(0)}$ e $\eta_{20,j}^{(1)}$ il corrispondente tasso non noto di falsi positivi (1-specificità) e il tasso di falsi negativi (1-sensibilità) al j-esimo valore di *cut-off*, rispettivamente, di questo test $T_{2,j}$. Si definisca

$$a_j = \eta_{21,j-1}^{(0)} - \eta_{21,j}^{(0)}; \qquad b_j = \eta_{20,j-1}^{(1)} - \eta_{20,j}^{(1)}.$$

per $j = 1, 2 \ldots, J+1$, dove $\sum_{j=1}^{J+1} a_j = \sum_{j=1}^{J+1} b_j = 1$. Si noti che vi è una mappatura uno-a-uno tra $\left\{\eta_{21,j}^{(0)}\right\}$ e $\{a_j\}$ e tra $\left\{\eta_{20,j}^{(1)}\right\}$ e $\{b_j\}$. Si definiscano i vettori $\boldsymbol{a} = (a_1, \ldots, a_{J+1})$ e $\boldsymbol{b} = (b_1, \ldots, b_{J+1})$. Si indichi con $\eta_{11}^{(0)}$ e $\eta_{10}^{(1)}$ i tassi non noti di falsi positivi e falsi negativi del test T_1, rispettivamente. Si assuma che questi tassi di errore per i test T_1 e T_2 non dipendano dalla popolazione s.

Nel modello così specificato, ci sono un totale di $2J + S + 2$ parametri ignoti, $\eta_{21,1}^{(0)}, \ldots, \eta_{21,J}^{(0)}, \eta_{20,1}^{(1)}, \ldots, \eta_{20,J}^{(1)}, p_1, \ldots, p_S, \eta_{11}^{(0)}, \eta_{10}^{(1)}$, e $S(2J+1)$ gradi di libertà, dove il vettore $\boldsymbol{p} = (p_1, \ldots, p_s, \ldots, p_S)$ indica le prevalenze nelle S popolazioni. Data la situazione di assenza di *gold standard,* l'identificabilità richiede che i dati derivino da almeno due popolazioni ($S \geq 2$) con almeno due prevalenze p_s distinte. Non è necessaria nessuna assunzione sulla forma della distribuzione dei valori del test dei malati e non malati in queste popolazioni. In aggiunta, si deve assumere $\eta_{11}^{(0)} + \eta_{10}^{(1)} \neq 1$ affinchè il modello sia identificabile, altrimenti la verosimiglianza delle frequenze osservate dipenderebbe da $p_s, \{a_j\}, \{b_j\}$ solo attraverso $\{p_s b_j + (1 - p_s)\, a_j\}$. Intuitivamente, $\eta_{11}^{(0)} + \eta_{10}^{(1)} = 1$ implica che il test T_1 non è meglio di una supposizione casuale e così non contribuisce a fornire informazione. Per l'identificabilità si assume quindi che $\eta_{11}^{(0)} + \eta_{10}^{(1)} < 1$. Per garantire la monotonicità della curva ROC, $\boldsymbol{a}$ e $\boldsymbol{b}$ devono essere fissati negativi. Si può implementare questo utilizzando un approccio bayesiano e specificando delle a priori sui parametri $\boldsymbol{a}$ e $\boldsymbol{b}$.

Le stime bayesiane dei parametri sono date dalle medie a posteriori. Sono utilizzati i metodi MCMC (Catena di Markov Monte Carlo) per calcolare le stime a posteriori della sensibilità e specificità che forniscono le basi per l'inferenza concernente l'accuratezza della procedura diagnostica.

Il modello è basato su una distribuzione multinomiale di dimensioni $2(J+1)$ delle frequenze osservate, che assume una forma complicata da utilizzare. Per risolvere questa difficoltà, gli autori consigliano di introdurre nuovi vettori casuali, in modo da mantenere la distribuzione marginale delle frequenze osservate del modello originale, ma rendere più trattabile la verosimiglianza con i dati aumentati nella nuova distribuzione multinomiale, permettendo l'uso di a priori coniugate, dove le a priori coniugate naturali per $\boldsymbol{a}$ e $\boldsymbol{b}$ sono distribuzioni di Dirichlet e per $\eta_{21}^{(0)},\ \eta_{20}^{(1)}, p_1, \dots, p_S$ distribuzioni beta. Si aggiunga la restrizione di identificabilità $\eta_{11}^{(0)} + \eta_{10}^{(1)} < 1$ nell'a priori. A priori informative saranno preferibili ogni volta che sia disponibile una conoscenza precedente per tutti i parametri (tassi di errore o prevalenze). L'analisi della sensibilità per differenti a priori è stata affrontata da Gustafson (2005) nel contesto della valutazione di test binari senza *gold standard*.

Nel modello appena descritto si è assunto che test di riferimento (T_1) sia un *imperfect gold standard* con risultati binari. Tuttavia, si potrebbe usare una generalizzazione di questo approccio, analizzando un test imperfetto di riferimento con risultati su scala continua, nel quale si è considerato $J'(\geq 2)$ *cut-off* per la classificazione piuttosto che una semplice dicotomizzazione dei risultati del test T_1. In questo caso, il modello sarebbe basato su una distribuzione multinomiale di dimensioni $(J'+1)(J+1)$ al posto della multinomiale di dimensioni $2(J+1)$. Il modello assume l'indipendenza tra i test T_1 e T_2 condizionatamente al vero stato di malattia. Questa assunzione è ragionevole se i test sono basati su fenomeni biologici non correlati; questo argomento è stato discusso in dettaglio da Albert e Dodd (2004) e Toft, Jϕrgensen e Hϕjsgaard (2005) nel contesto di valutazione di test binari.

2.4.2.4 *Studi di incidenza*

- *Due punti temporali*

L'*imperfect gold standard* potrebbe essere applicato più volte. Il primo lavoro di Yanagawa e Gladen (1984) stima l'incidenza della malattia con un test fallibile applicato a due punti temporali. I dati possono essere organizzati come nella Tabella 2.2, con T_1 che denota il test al tempo 1 e T_2 che denota il test al tempo 2. Il metodo richiede che T_1 sia indipendente da T_2. Questa assunzione è criticabile perchè è una forzatura pretendere che non vi sia dipendenza tra T_1 e T_2, dato che si tratta dello stesso test. Con solo tre gradi di libertà nei dati, la stima della prevalenza (al tempo T_1), incidenza e tasso di remissione della malattia richiede che i tassi di errore del test siano noti. In un esempio differente, la specificità è assunta perfetta e il tasso di remissione della malattia pari a zero, così la sensibilità del test può essere stimata assieme alla prevalenza e incidenza della malattia. Più parametri non noti possono essere stimati in studi longitudinali con punti temporali aggiuntivi o più di un test ad ogni tempo. In un approccio differente ad uno studio

d'incidenza con due punti temporali, Espeland et al. (1988) assumono che la malattia sia irreversibile. Essi propongono di porre il problema nella struttura generale di un modello LLP (*Linear, Log-linear e Product*) per stimare sia l'incidenza della malattia che i tassi di errore diagnostico. Per illustrare il concetto si può usare un esempio riguardante la diagnosi delle carie dentali. Ad ogni superficie dentale è fatta una diagnosi del suono o carie in due istanti temporali differenti, le carie sono assunte essere irreversibili. È assunto anche che lo stato di carie è indipendente tra le superfici dentali e che gli errori diagnostici sono indipendenti condizionatamente al vero stato di carie. I dati possono essere classificati nella Tabella 2.2, dove T_1e T_2 denotano le diagnosi al tempo 1 e al tempo 2, rispettivamente. Se il vero stato di malattia (carie) D è noto, allora i dati nella Tabella 2.2 potrebbero essere ulteriormente classificati utilizzando i dati completamente osservati. Si indichi con $p_{t_1 t_2 l}$ la proporzione osservata dove t_1 è il risultato di T_1, t_2 è il risultato di T_2 e l può assumere uno dei quattro valori che denotano l'errore o la correttezza delle diagnosi ai due punti temporali, cioè $l = 1$ se T_1 e T_2 sono entrambi corretti, $l = 2$ se T_1 è corretto e T_2 è incorretto, $l = 3$ se T_1 è non corretto e T_2 è corretto, e $l = 4$ se T_1 e T_2 sono entrambi non corretti. Si denoti con ϑ_1 la vera proporzione di soggetti senza la malattia in entrambi i tempi, ϑ_2 è la proporzione di soggetti senza la malattia al tempo 1 che sviluppano la malattia al tempo 2, e ϑ_3 è la proporzione di soggetti con malattia ad entrambi i tempi. Semplificando i pedici della proporzione osservata p in g, che denota t_1, t_2, l, le proporzioni delle celle attese può essere scritto nella forma di un modello generale LLP:

$$E(p_g) = \vartheta_g \prod_{h=1}^{H} \eta_h^{a(g,h)}.$$

I termini η_h, $h = 1, \dots, H$, formano l'insieme di tutte le probabilità condizionate di errore/accuratezza, per esempio $\eta_{k1}^{(0)}, \left(1 - \eta_{k1}^{(0)}\right)$, contenute in un modello. Assieme alle potenze note $a(g,h)$, i valori η_h, $h = 1, \dots, H$, mappano i valori ϑ_g ai valori attesi delle proporzioni osservate p.

Il parametro ϑ_g è assunto essere log-linerare, vale a dire che $log\, \vartheta_g$ è una combinazione lineare di covariate, che possono essere usate per modellare complessi disegni di studio. I valori η_h sono modellati come funzioni dei tassi di errore diagnostico. La scelta di parametri specifici è illustrata nel seguente esempio in Espeland et al. (1988). Si considerino i dati come nella Tabella 2.2. Con solo tre gradi di libertà disponibili, due di questi sono disponibili per stimare i valori di ϑ_g, mentre solo un grado di libertà è lasciato per la stima dei tassi d'errore diagnostico. Si assuma che la sensibilità e specificità siano uguali e costanti ad entrambi i punti temporali, vale a dire $\eta_{11}^{(0)} = \eta_{10}^{(1)} = \eta_{21}^{(0)} = \eta_{20}^{(1)} = \alpha$. Assumendo che il vero stato di malattia sia noto per tutti i soggetti, l'ipotetica tabella dei dati completi ha le seguenti probabilità attese delle celle:

$$E(p_{111}) = \vartheta_1(1-\alpha)^2, \qquad E(p_{112}) = 0,$$

$$E(p_{121}) = \vartheta_2(1-\alpha)^2, \qquad E(p_{122}) = \vartheta_3\alpha(1-\alpha),$$

$$E(p_{211}) = 0, \qquad E(p_{212}) = \vartheta_1\alpha(1-\alpha),$$

$$E(p_{221}) = \vartheta_3(1-\alpha)^2, \qquad E(p_{222}) = \vartheta_2\alpha(1-\alpha),$$

$$E(p_{113}) = \vartheta_2\alpha(1-\alpha), \qquad E(p_{114}) = \vartheta_3\alpha^2,$$

$$E(p_{123}) = \vartheta_1\alpha(1-\alpha), \qquad E(p_{124}) = 0,$$

$$E(p_{213}) = \vartheta_3\alpha(1-\alpha), \qquad E(p_{214}) = \vartheta_2\alpha^2,$$

$$E(p_{223}) = 0, \qquad E(p_{224}) = \vartheta_1\alpha^2.$$

Modellando tali probabilità nella forma di un generale LLP, la parte log-lineare è:

$$log\boldsymbol{\vartheta} = \begin{bmatrix} log\vartheta_1 \\ log\vartheta_2 \\ log\vartheta_3 \end{bmatrix} = \boldsymbol{Zb} \qquad \text{dove} \qquad \boldsymbol{Z} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \text{e} \qquad \boldsymbol{b} = (b_1, b_2, b_3)'.$$

La matrice di errore/accuratezza è parametrizzata come:

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \alpha \\ 1-\alpha \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \alpha^* + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \frac{1}{2} \quad \text{con } \alpha^* = \alpha - 0.5.$$

e la matrice delle potenze con elementi $a(g,h)$, è:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 \\ 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}'.$$

Altre parametrizzazioni possono essere usate per altri disegni di studio e strutture di errore diagnostico. È suggerito un algoritmo di punteggio di Fisher per ottenere stime di massima verosimiglianza dei parametri e delle loro varianze.

- *Molteplici punti temporali*

Espeland et al. (1988) hanno generalizzato l'approccio LLP per stimare l'incidenza e il tasso d'errore diagnostico in uno studio longitudinale con misure ripetute irregolari. Esso permette di applicare i test diagnostici molteplici volte agli stessi individui ad intervalli irregolari. Si assuma ancora che il processo di malattia sia irreversibile. Si suddivida i periodi d'osservazione in M epoche con tassi d'incidenza della malattia φ_1,…, φ_m,…, φ_M in ciascuna epoca. Se l'inizio della malattia D è osservato nell'epoca m', allora

$D_{im'} = 1$, e $D_{im} = 0$ per tutti i $m \neq m'$. Le epoche sono sufficientemente piccole in modo che non più di una diagnosi possa essere fatta in un'epoca. Se l'i-esimo soggetto è osservato in ogni epoca, il vettore delle diagnosi $\boldsymbol{t}_i = (t_{i1}, \dots, t_{im}, \dots, t_{iM})$ è una serie di 0 e 1, che indicano uno stato di non malattia (stadio di Tanner$= 1$) o malato (stadio di Tanner > 1), rispettivamente. Ogni t_{im} osservato può sorgere da strutture differenti della vera incidenza, $\boldsymbol{d}_i = (d_{i1}, \dots, d_{im}, \dots, d_{iM})$, ciascuno accoppiato con una struttura distinta di errori diagnostici, indicizzati da g.

Ci sono quattro probabilità di errore diagnostico/accuratezza, $\eta_{m1}^{(0)}, \left(1 - \eta_{m1}^{(0)}\right), \eta_{m0}^{(1)}, \left(1 - \eta_{m0}^{(1)}\right)$, associato con ogni epoca, e un totale di $4M$ probabilità denotate da $\eta_h, h = 1, \dots, 4M$, associate con tutte le epoche. Per una data configurazione sottostante g, la probabilità congiunta non condizionata di incidenza nell'epoca m per il soggetto i e l'insieme di probabilità dell'errore/accuratezza diagnostica può essere scritto come:

$$L(\boldsymbol{\theta}) = Pr[d_{im}(g) = 1, \boldsymbol{a}(g,i)] = \varphi_m \prod_{h=1}^{4M} \eta_h^{a(g,h,i)}, \tag{2.8}$$

dove η_h denota una delle combinazioni delle $4M$ probabilità di $\eta_{m1}^{(0)}, \left(1 - \eta_{m1}^{(0)}\right), \eta_{m0}^{(1)}, \left(1 - \eta_{m0}^{(1)}\right)$ nelle M epoche e $\boldsymbol{\theta}$ è il vettore dei parametri da stimate. Il vettore noto delle potenze $\boldsymbol{a}(g,i) = [a(g,h,i), h = 1, \dots, 4M]$, dove $a(g,h,i) = 0$ o 1, è usato per scegliere una particolare struttura seriale di tassi di accuratezza/errore. In un tipico studio longitudinale con periodi di osservazione incompleta e irregolare, $\boldsymbol{d}_i$ è osservabile solo nella forma di intervalli-censurati $\boldsymbol{d}_i^*$. I veri periodi di osservazione possono essere espressi come combinazioni lineari di epoche basati sulla struttura d'osservazione. Così $\boldsymbol{d}_i$ e $\boldsymbol{d}_i^*$ sono legati. Se non c'è errore diagnostico, l'incidenza osservata nei periodi d'osservazione possono essere usati prontamente per stimare l'incidenza in tutte le epoche. Quando sono presenti gli errori diagnostici, è presentato un algoritmo EM per la massimizzazione della verosimiglianza basata su (2.8) mediante i trattamento del vero $\boldsymbol{d}_i$ come dato mancante. Strutture differenti di errore diagnostico possono essere modellate usando questo approccio. Per esempio, il Espeland et al. (1989), i tassi di errore sono permessi variare con l'età dei soggetti e il peso del corpo in modi differenti. Gilks et al. (1993) propongono un modello simile per test di screening sequenziali del cancro della cervice senza gold standard simultanei per i test. Loro suggeriscono di usare i metodi della Catena di Markov Monte Carlo (MCMC) per la stima dei parametri dei tassi d'errore.

2.5 DIPENDENZA CONDIZIONATA TRA I TEST

L'assunzione d'indipendenza condizionata potrebbe essere violata nella pratica, specialmente nelle situazioni in cui la diagnosi della malattia è basata sull'informazione ottenuta da molteplici test diagnostici,

nessuno dei quali è un *gold standard* perfettamente accurato in grado di fornire ottime sensibilità e specificità. In tale situazione, due o più test diagnostici potrebbero essere condizionatamente dipendenti a causa di un fattore oltre che per lo stato di malattia, che sorge, per esempio da un comune fenomeno biologico sul quale due test sono basati. Parecchi autori hanno dimostrato che è importante tenere conto di questa dipendenza condizionata quando si analizzano i risultati dei test diagnostici al fine di ottenere stime non distorte della prevalenza della malattia e dell'accuratezza dei test (Fryback, 1978; Vacek, 1985). Approcci frequentisti che affrontano questo problema richiedono che si disponga di risultati da almeno quattro test differenti per ottenere una soluzione identificabile, cioè per avere sufficienti gradi di libertà per stimare ciascun parametro di interesse unicamente (Walter e Irwig, 1988).

Nella sezione 2.5.1 si accenna ad alcune strategie proposte che assumono parametri noti, mentre nella sezione 2.5.2 si affrontano metodi che rilassano tali restrizioni. In particolare, sono proposti un metodo con più di due classi latenti (2.5.2.1) e un modello ad effetti casuali (2.5.2.2).
Nella pratica, però, non è sempre possibile avere risultati da quattro differenti test, particolarmente quando i test sono costosi, impiegano tanto tempo, o sono invasivi. Nella sezione 2.5.2.3 si affrontano alcune strategie bayesiane per ovviare a questo problema.

2.5.1 *Parametri noti*

Quando l'indipendenza condizionata non è un'assunzione ragionevole, Mantel (1951) illustra una soluzione parziale mediante Neyman (1947) che assume due classi di individui malati, la più severa delle quali ha una sensibilità del 100%. Una soluzione più generale è data da Thibodeau (1981) quando i tassi di errore dell'*imperfect gold standard* (T_1), $\eta_{11}^{(0)}$ e $\eta_{10}^{(1)}$, sono noti. Assumendo che gli errori dell'*imperfect gold standard* e il test T_2 siano correlati positivamente dato lo stato di malattia e che il test T_2 non sia più accurato dell'*imperfect gold standard*, Thibodeau (1981) ha derivato intervalli per le stime della sensibilità e specificità per il test T_2.

2.5.2 *Parametri non noti*

2.5.2.1 *Più di due classi latenti*

Molti metodi che permettono la dipendenza condizionata tra i test sono basati su analisi di classi latenti. Nel caso in cui la bontà di adattamento di un modello con una variabile latente non fosse soddisfatto, Rindskopf e Rindskopf (1986) suggeriscono di usare più di due classi latenti (malati e non malati) per

migliorare l'adattamento. Formann (1994) dimostra in un esempio che la bontà di adattamento può essere migliorata usando tre o più classi latenti, e che le classi latenti aggiuntive fanno invece corrispondere le dipendenze tra test. Alvord et al. (1988) propongono un modello con 3 classi latenti per migliorare l'adattamento ad un insieme di risultati da quattro casi di anticorpi HIV applicati agli stessi soggetti senza un *gold standard* test.

2.5.2.2 *Modello ad effetti casuali*

Un modello ad effetti casuali è stato proposto da Qu et al. (1996), supponendo inizialmente l'assunzione di indipendenza condizionata in un modello a due classi latenti ($2LC$), per poi rilassarla successivamente. Si assuma di disporre di K test diagnostici. Si indichi con $\boldsymbol{\theta} = \left(p, \eta_{k1}^{(0)}, \eta_{k1}^{(1)}\right)$, $k = 1, \dots, K$, il vettore dei parametri ignoti, dove si richiami che $\eta_{k1}^{(d)} = P(T_k = 1|D = d)$ denota la probabilità che il k-esimo test ($k = 1, \dots, K$) sia positivo dato il vero stato della malattia $D = d$ ($d = 0, 1$). Allora $\eta_{k1}^{(1)}$ è la sensibilità e $(1 - \eta_{k1}^{0})$ è la specificità del k-esimo test. Indicando i risultati del test per l' i-esimo soggetto con $\boldsymbol{t}_i = (t_{i1}, \dots, t_{ik}, \dots, t_{iK})$, la funzione di verosimiglianza per ogni soggetto sotto il modello $2LC$ è:

$$L(\boldsymbol{\theta}; t_i) = p \prod_{k=1}^{K} \left(\eta_{k1}^{(1)}\right)^{t_{ik}} \left(1 - \eta_{k1}^{(1)}\right)^{1-t_{ik}} + (1 - p) \prod_{k=1}^{K} \left(\eta_{k1}^{(0)}\right)^{t_{ik}} \left(1 - \eta_{k1}^{(0)}\right)^{1-t_{ik}}.$$

Si può estendere il modello $2LC$ ad un modello a classe latente con effetti casuali ($2LCR$) che include correlazioni tra i test, aggiungendo un effetto casuale dovuto ad alcune caratteristiche non osservate del soggetto, denotato da una variabile continua H, che varia da soggetto a soggetto e ha una distribuzione normale standard, di cui h è una sua determinazione. Per il k-esimo test, il tasso di risposta positivo dell'i-esimo soggetto è modellato come:

$$P(t_{ik} = 1|D = d, H = h) = \Phi(a_{kd} + b_{kd}h), \qquad d = 0, 1, \quad H \sim N(0,1)$$

dove Φ è la funzione di distribuzione cumulata di una covariata normale standard.
Integrando i risultati casuali dell'effetto h nei tassi di errore per il k-esimo test si ottiene:

$$\eta_{k1} = \Phi\left(\frac{a_{k1}}{\sqrt{1 + b_{k1}^2}}\right) \qquad \text{e} \qquad 1 - \eta_{k0} = \Phi\left(\frac{-a_{k0}}{\sqrt{1 + b_{k0}^2}}\right).$$

Assumendo che i test siano indipendenti condizionatamente al vero stato di malattia d e ad h, la probabilità condizionata della risposta $\boldsymbol{t}_i = (t_{i1}, \dots, t_{iK})$ dato $D = d$, è data da:

$$Pr(T_i|D=d) = \int_{-\infty}^{+\infty} \prod_k \Phi(a_{kd} + b_{kd}h)^{t_{ik}} \left(1 - \Phi(a_{kd} + b_{kd}h)\right)^{1-t_{ik}} d\Phi(h), \qquad d = 0, 1$$

La verosimiglianza marginale per l'i-esimo soggetto è:

$$L(\boldsymbol{\theta}; \boldsymbol{t}_i) = pPr(T_i|D=1) + (1-p)Pr(T_i|D=0).$$

Le stime di massima verosimiglianza dei parametri sono ottenute usando l'algoritmo EM o l'algoritmo di minimizzazione multivariata senza restrizioni di Powell, con la matrice di varianza-covarianza stimata attraverso la matrice d'informazione di Fisher. Le restrizioni sono usualmente poste su b_{kd} per casi speciali dei modelli di classe $2LCR$. Per esempio, in un modello $2LCR1$ con componenti ad uguale varianza tra tutti i test, $b_{kd} = b_d$ per tutti i k, $d = 0, 1$. Questo modello ha due parametri in più, b_0 per i non malati e b_1 per le popolazioni malate, rispetto al modello $2LC$ con indipendenza condizionata. Per un modello a classe latente con effetto diretto ($2LCD$), il modello è riparametrizzato per riflettere le correlazioni solo tra coppie specifiche di test. Se solo il test u è correlato con il test v, la probabilità condizionata $Pr(T_i|D=d)$ nel modello $2LCD$ è:

$$Pr(T_i|D=d) = \left\{1 + \rho_{vd}\frac{(t_{iu} - \eta_{ud})(t_{iv} - \eta_{vd})}{\eta_{ud}(1-\eta_{ud})}\right\} \prod_{k=1}^{K} \eta_{kd}^{t_{ik}} (1-\eta_{kd})^{1-t_{ik}}, \qquad d = 0, 1,$$

dove $\rho_{vd} = Pr(T_{iv} = 1|D=d, T_{iu}=1) - Pr(T_{iv}=1|D=d, T_{iu}=0)$.

Qu et al. (1996) propongono procedure di controllo del modello usando entrambi i test chi-quadrato di bontà di adattamento del modello e una procedura grafica basata sulla correlazione dei residui.

2.5.2.3 *Approccio bayesiano*

Tutti i metodi visti precedentemente (inclusione di variabili latenti e modelli ad effetti casuali) usano un approccio frequentista per stimare i parametri coinvolti. Anche se essi propongono criteri differenti alla dipendenza del modello, tutti richiedono almeno quattro o più test per ottenere una soluzione identificabile e quindi stimare in modo unico tutti i parametri d'interesse. Dato che non è sempre fattibile ottenere risultati da un numero così ampio di test, si può proporre un approccio bayesiano per disegnare inferenze sulla prevalenza della malattia e le proprietà del test con aggiustamenti per la possibilità di dipendenza condizionata tra i test, in particolare quando si dispone solo di due test (Dendukuri a Joseph, 2001), estendendo quindi il lavoro di Joseph et al. (1995) al caso di test dipendenti condizionatamente al vero stato di malattia. Con i dati da una singola popolazione, ci sono solo 3 gradi di libertà per la stima di 7 e 9 parametri in modelli ad effetti fissi e casuali, rispettivamente. Il modello di Dendukuri e Joseph (2001) con

effetti casuali include la specificazione di a priori sulla prevalenza e sulla combinazione "intercetta-pendenza" per individui malati e non malati, ma ciò può essere complicato. D'altro canto, la specificazione a priori proposta da Dendukuri e Joseph (2001) per il modello ad effetti fissi è relativamente semplice. Per entrambi i modelli, la maggior parte delle condizionate piene non sono riconoscibili e richiedono un meccanismo generale per la simulazione. L'aggiunta di una seconda popolazione nel contesto di un modello ad effetti fissi aumenta i gradi di libertà a 6 per 8 parametri da stimare e da qui decresce l'affidamento sull'informazione aggiuntiva (a priori). Basato sull'esperienza dei suddetti autori, l'uso di $S > 2$ popolazioni non riduce ulteriormente il problema dell'identificazione. Tuttavia, non c'è difficoltà nell'aumentare il numero di popolazioni, aumentando così l'informazione per questi parametri che sono stimabili dai soli dati.
Se il modello non è identificabile, i risultati dipendono pesantemente dalle distribuzioni a priori introdotte.
Spesso, è disponibile un'effettiva informazione a priori per le caratteristiche di un test e/o le prevalenze della popolazione campionata con informazione limitata sui restanti parametri. In altre situazioni, i dati sulla prevalenza potrebbero essere disponibili per alcune popolazioni, permettendo così un'adeguata specificazione a priori di questi parametri. Se si dispone di un'informazione ragionevolmente accurata su uno dei due test di screening (forse l'*imperfect gold standard* usato correntemente) o la prevalenza della popolazione testata, sono possibili inferenze accurate su tutti i parametri, inclusa la correlazione dei test.

Si considerino di seguito alcuni esempi di approcci bayesiani.

- *MODELLO DI DENDUKURI ET AL. (2001)*

Dendukuri et al. (2001) presentano due modelli, un modello ad effetti fissi e un modello ad effetti casuali, per disegnare inferenze simultanee sulla prevalenza della malattia e tutti i parametri nella situazione in cui sono usati molteplici test diagnostici, con aggiustamento della dipendenza condizionata tra i test. In particolare si considera la situazione di non identificabilità con meno di 4 test e si propone un approccio bayesiano per la sua soluzione.

Per il modello ad effetti fissi si modella la dipendenza condizionata tra due test T_1 e T_2 , da un campione di n soggetti, usando la covarianza tra i test nella popolazione malata e non malata. Si può scrivere la funzione di verosimiglianza multinomiale dei dati osservati condizionatamente ai dati latenti.
Si usino le famiglie di distribuzione coniugate per rappresentare l'informazione a priori. La scelta delle distribuzioni non è unica e potrebbero essere rimpiazzate da altre densità adeguate, a seconda del bisogno. Una proposta può essere la seguente.

(1) si assume che la prevalenza segua una distribuzione a priori beta;

(2) si assume anche che le sensibilità e le specificità abbiano densità a priori beta;

(3) si assume che i parametri della covarianza abbiano distribuzioni a priori beta generalizzata.

Quando la funzione di verosimiglianza è combinata con le distribuzioni a priori sopra indicate, si ottiene l'espressione per la distribuzione congiunta a posteriori dei parametri ignoti (prevalenza, due sensibilità, due specificità e due covarianze). Data la complessità di questo modello, non è possibile ottenere le distribuzioni marginali per i parametri analiticamente. Tuttavia, si può usare l'algoritmo del campionamento di Gibbs (Gelfand e Smith, 1990) per ottenere i campioni dalla distribuzione marginale a posteriori di ogni parametro.

Per il modello ad effetti casuali si può usare un approccio bayesiano simile a quello frequentista di Qu et al. (1996), considerando la variazione nei parametri del test sulla popolazione e modellando la dipendenza condizionata tra molteplici test mediante effetti casuali. Le sensibilità e le specificità dei test sono modellati come funzioni di una variabile casuale latente, specifica per soggetto. Applicando lo stesso valore della variabile latente in ogni paziente tra tutti i test, si ottiene una dipendenza tra i test senza riferimento esplicito al parametro della covarianza. Questa situazione può essere concettualizzata nel modo in cui la performance di un test in un dato soggetto è una funzione di una variabile casuale continua, detta intensità, H_i. Questa intensità può essere considerata, ad esempio, come una misura riassuntiva della severità della malattia del soggetto che concerne la facilità di scoperta della malattia nel soggetto. La sensibilità e la specificità di un test per ogni soggetto sono funzioni di un'intensità sottostante, di forma $f(h_i)$, dove f è una funzione continua e monotona crescente, che assume i valori compresi tra zero e uno. Qui H_i è considerata una variabile casuale che segue una distribuzione $N(0,1)$. Si possono usare varianze uguali oppure, più generalmente, varianze differenti per i soggetti malati e non malati. Si introducono alcuni parametri a livello di soggetto individuale. I risultati di test differenti sono assunti essere indipendenti tra loro condizionatamente allo stato di malattia D_i e alla variabile latente H_i. Si derivi la funzione di verosimiglianza e si specifichino le distribuzioni a priori per i parametri. Come nel caso di modelli ad effetti fissi, si può usare un campionamento di Gibbs per ottenere campioni da distribuzioni marginali a posteriori dei parametri da stimare. In entrambi i modelli ad effetti fissi o casuali, la mancanza di identificabilità significa che la distribuzione a posteriori non necessariamente si concentra sui valori del vero parametro, anche se la numerosità campionaria tende all'infinito. Piuttosto, si concentra sull'insieme di valori del parametro consistenti con i dati e le distribuzioni a priori sono usate per delineare quali insiemi di valori del parametro sono più plausibili di altri. Tuttavia, l'influenza delle distribuzioni a priori non svanisce anche con un campione di numerosità infinita.

- *MODELLO DI CHOI ET AL. (2006)*

Nel caso in cui i valori del test di entrambi gli individui malati e non malati siano misurati su scala continua e siano normalmente distribuiti, Choi et al. (2006) hanno proposto un metodo bayesiano parametrico per la stima della curva ROC e della prevalenza della malattia in assenza di un *gold standard* basato sulla differenza tra AUC. Essi hanno sviluppato un'analisi ROC senza *gold standard* applicata a due test diagnostici correlati (T_1 e T_2) che sono usati sugli stessi individui per stabilire l'accuratezza di test diagnostici da un campione di valori del test ottenuti da una popolazione d'interesse. I punteggi dei test potrebbero richiedere una trasformazione adeguata per conformarsi all'assunzione di bi-normalità, generalmente utilizzando i dati della pratica, se disponibili, in congiunzione con i dati correnti o l'opinione di esperti scientifici.

Si indichino con X_{1i} e X_{2i} due valori di test diagnostici raccolti sull'i-esimo individuo in un campione casuale di m individui che hanno la malattia, e con Y_{1j} e Y_{2j} i valori del test diagnostico raccolti dal j-esimo individuo in un campione casuale di n individui che sono sani. In assenza di *gold standard* non si ha informazione sullo stato di malattia degli individui testati, quindi è necessario definire con D_i la variabile latente che indica lo stato della malattia dell'i-esimo individuo (1 se l'individuo ha la malattia, 0 altrimenti), dove $i = 1, \dots, n$ e n è il numero degli individui testati. Si denoti con T_{ik} il risultato per l'individuo i usando il test T_k, dove $k = 1, 2$ e $i = 1, \dots, n$. Si assuma che siano state raccolte n misure bivariate $i.i.d.$ e che D_i abbia una distribuzione di Bernoulli, così definita:

$$D_i \sim Bernoulli\ (p), \tag{2.9}$$

dove $p = P(D_i = 1) = 1 - P(D_i = 0)$.

Si assuma inoltre che i risultati del test di un soggetto malato e sano siano distribuiti come una normale bivariata, rispettivamente:

$$\boldsymbol{X} \sim N_2\left(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D\right) \quad \text{e} \quad \boldsymbol{Y} \sim N_2\left(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}}\right).$$

dove

$$\boldsymbol{\mu}_D = \begin{bmatrix}\mu_{1D} \\ \mu_{2D}\end{bmatrix}, \quad \boldsymbol{\Sigma}_D = \begin{bmatrix}\sigma^2_{11D} & \sigma^2_{12D} \\ \sigma^2_{12D} & \sigma^2_{22D}\end{bmatrix}, \quad \boldsymbol{\mu}_{\bar{D}} = \begin{bmatrix}\mu_{1\bar{D}} \\ \mu_{2\bar{D}}\end{bmatrix}, \quad \boldsymbol{\Sigma}_{\bar{D}} = \begin{bmatrix}\sigma^2_{11\bar{D}} & \sigma^2_{12\bar{D}} \\ \sigma^2_{12\bar{D}} & \sigma^2_{22\bar{D}}\end{bmatrix}.$$

La correlazione è data da:

$$\rho_D = \frac{\sigma^2_{12D}}{\sqrt{\sigma^2_{11D}}\sqrt{\sigma^2_{22D}}} \quad \text{e} \quad \rho_{\bar{D}} = \frac{\sigma^2_{12\bar{D}}}{\sqrt{\sigma^2_{11\bar{D}}}\sqrt{\sigma^2_{22\bar{D}}}}.$$

La distribuzione dei risultati per l'individuo i usando il test k è data da:

$$T_{ki} \sim f_{\boldsymbol{X}}(\cdot\,|\mu_{1D}, \mu_{2D}, \sigma^2_{11D}, \sigma^2_{22D}, \rho_D)^{D_i} f_{\boldsymbol{Y}}\big(\cdot\,|\mu_{1\bar{D}}, \mu_{2\bar{D}}, \sigma^2_{11\bar{D}}, \sigma^2_{22\bar{D}}, \rho_{\bar{D}}\big)^{(1-D_i)},$$

per $k = 1, 2$, dove $f_{\boldsymbol{X}}(\cdot)$ e $f_{\boldsymbol{Y}}(\cdot)$ sono le funzioni di densità dei punteggi di individui malati e sani, rispettivamente. Per l'identificabilità, si assuma che $\mu_{k\bar{D}} < \mu_{kD} = 1.2$, che potrebbe valere per la maggior parte dei test diagnostici.

Per i valori di *cut-off* $c \in (-\infty, \infty)$, le curve ROC possono essere costruite disegnando i punti di coordinate (1-specificità, sensibilità), che si calcolano come:

$$\left[1 - \Phi\left(\frac{c - \mu_{k\bar{D}}}{\sqrt{\sigma^2_{kk\bar{D}}}}\right), 1 - \Phi\left(\frac{c - \mu_{kD}}{\sqrt{\sigma^2_{kkD}}}\right)\right]$$

per il test k $(k = 1, 2)$; $\Phi(\cdot)$ è la funzione di ripartizione di una variabile normale standard. Le AUC per i due corrispondenti test diagnostici possono essere calcolati come:

$$AUC_k = \Phi\left(-\frac{\mu_{k\bar{D}} - \mu_{kD}}{\sqrt{\sigma^2_{kk\bar{D}} + \sigma^2_{kkD}}}\right)$$

per il test k. La differenza tra le AUC è $\Delta = AUC_1 - AUC_2$, compara l'accuratezza complessiva del test T_1 rispetto al test T_2. Si consideri una misura di quanto sia "vicina" la distribuzione di $\mathbf{Y}$ alla distribuzione di $\mathbf{X}$ per un dato test. Questo è definito come l'area sotto la densità per $\mathbf{X}$ a sinistra del 95-esimo percentile di $\mathbf{Y}$. Se la media di $\mathbf{X}$ uguaglia il 95-esimo percentile di $\mathbf{Y}$, la misura sarà 0.5. Si può definire il parametro Δ_k, per il test k, nel seguente modo:

$$\Delta_k = \Phi\left(\frac{\delta_k - \mu_{kD}}{\sqrt{\sigma^2_{kkD}}}\right),$$

dove δ_k è il 95-esimo percentile per i valori del test k per il gruppo dei non malati. Un'altra interpretazione di Δ_k è che rappresenta la proporzione di falsi negativi (1 – sensibilità) quando la specificità è 0.95.

La funzione di verosimiglianza è basata sulla descrizione formale (2.9) per i dati osservati $\{T_{ki}; k = 1, 2 \text{ e } i = 1, \dots, n\}$. Si ricorre all'inferenza bayesiana per determinare la stima del vettore dei parametri ignoti $\boldsymbol{\theta} = \left(p, \mu_{Dk}, \mu_{\bar{D}k}, \frac{1}{\sigma^2_{Dkk}}, \frac{1}{\sigma^2_{\bar{D}kk}}, \rho_D, \rho_{\bar{D}}\right)$, per $k = 1, 2$. Si assuma per p una distribuzione beta, in particolare, in assenza di informazioni, si può scegliere una distribuzione beta $(1,1)$. Altri potrebbero

preferire un'a priori di Jeffrey (beta $(0.5,0.5)$). Naturalmente sono preferibili le a priori informative quando è disponibile un input scientifico. Si assuma per μ_{Dk} e $\mu_{\bar{D}k}$ una distribuzione normale e per $\frac{1}{\sigma^2_{Dkk}}$ e $\frac{1}{\sigma^2_{\bar{D}kk}}$ una gamma. Queste a priori approssimano le a priori di Jeffrey per questi parametri quando non è presente la correlazione ($\rho_D = \rho_{\bar{D}} = 0$). Inoltre si assuma per ρ_D e $\rho_{\bar{D}}$ una distribuzione a priori uniforme. La scelta di tale a priori presume uguale plausibilità di tutti i possibili valori di ciascuna correlazione. Una scelta alternativa potrebbe essere una distribuzioni beta generalizzata o l'a priori di riferimento discussa da Bernardo e Smith (1994). Una scelta alternativa delle a priori per le matrici di covarianza ($\boldsymbol{\Sigma}$) potrebbe essere le a priori indipendenti Wishart, che possono essere attribuite a Bernardo e Smith (1994).

L'inferenza bayesiana può essere condotta applicando il campionamento di Gibbs (German e German, 1984; Gelfand e Smith, 1990), che funziona appropriatamente per modelli complessi con alte dimensioni basati su assunzioni di indipendenza condizionata. Questa tecnica campiona iterativamente la funzione di densità di probabilità per le condizionate piene, $p_j\,(\theta_j|\boldsymbol{\theta}_{(j)}, data)$, dove $\boldsymbol{\theta} = (\theta_1, \dots, \theta_j, \dots, \theta_m)$ è un generico vettore casuale e $\boldsymbol{\theta}_{(i)}$ è lo stesso vettore senza l'elemento θ_j. Le condizionate piene potrebbero essere campionate utilizzando un adeguato campionamento di rifiuto (Gilks, 1992), nel caso in cui la condizionata piena fosse ignota ma soddisfacesse un controllo per log-concavità. Tuttavia, se la condizionata piena non soddisfacesse la log-concavità, si potrebbe applicare un campionamento di Metropolis (Tierney, 1994) o il campionamento a fetta (Neal, 1997). La convergenza è controllata monitorando storie di quantità campionate oppure usando il metodo di Brooks e Gelman (1998). L'inferenza bayesiana con il campionamento di Gibbs può essere implementata mediante software statistici, per esempio Winbugs .

In assenza di *gold standard* questo metodo funziona bene nelle situazioni in cui la sovrapposizione tra i gruppi dei malati e non malati non è troppo ampia. Quando le distribuzioni dei valori del test si sovrappongono troppo, il metodo probabilmente incontra difficoltà ad assegnare il corretto stato di malattia nelle regioni sovrapposte, fornendo ampi intervalli e risultati che potrebbero essere lontani dal loro obiettivo. Naturalmente, questo è un classico problema nella stima di misture di distribuzioni. Se fosse difficile risolvere la mistura, allora sarebbe necessariamente complicato stimare funzionali che coinvolgono le distribuzioni separate. Gli intervalli per le differenze delle AUC sono più stretti nel caso in cui la correlazione sia più elevata. Un limite della procedura è che si richiede l'assunzione di normalità bivariata su qualche scala. In assenza di un *gold standard* test, questo potrebbe essere difficoltoso da determinare. Un'estensione aggiuntiva di interesse potrebbe essere quella di permettere delle covariate che potrebbero riguardare l'abilità discriminatoria di una misura diagnostica. Per esempio, una particolare misura sierologica potrebbe avere un'abilità discriminatoria più potente sugli individui più vecchi rispetto a quelli più giovani.

- *MODELLO DI GEORGIADIS ET AL. (2006)*

Un altro approccio bayesiano per fare inferenza sulla prevalenza e l'accuratezza di due test di screening (T_1 e T_2) sotto ipotesi di dipendenza, quando sono testate $S \geq 1$ popolazioni, è quello di assumere che le prevalenze per ogni popolazione siano distinte, mentre l'accuratezza diagnostica e i tassi di errore siano gli stessi per tutte le popolazioni (Georgiadis et al., 2006). I dati sono definiti come $\{y_{t_1 t_2 s}\}$, per $t_1, t_2 = 0, 1, s = 1, \dots, S$, che corrisponde al numero di individui tra gli n_s campionati dalla popolazione s, che è risultato t_1 nel test T_1 e t_2 nel T_2. Per esempio, y_{11s} corrisponde al numero di individui tra gli n_s campionati dalla popolazione s, che è risultato positivo in entrambi i test. Si indichi con $\{p_{kjs}\}$ probabilità delle celle corrispondenti. Si assuma che questi dati generino S campioni indipendenti con distribuzione multinomiale. Le probabilità condizionate, che indicano l'accuratezza diagnostica e i tassi di errore dei due test condizionatamente al vero stato di malattia, sono definite da:

$\omega_{11} = P(T_1 = 1, T_2 = 1|D = 1)$, $\omega_{10} = P(T_1 = 1, T_2 = 0|D = 1)$, $\omega_{01} = P(T_1 = 0, T_2 = 1|D = 1)$, $\omega_{00} = P(T_1 = 0, T_2 = 0|D = 1)$,

$\phi_{11} = P(T_1 = 1, T_2 = 1|D = 0)$, $\phi_{10} = P(T_1 = 1, T_2 = 0|D = 0)$, $\phi_{01} = P(T_1 = 0, T_2 = 1|D = 0)$, $\phi_{00} = P(T_1 = 0, T_2 = 0|D = 0)$,

che sono assunte essere le stesse per tutte le popolazioni. Per la legge di probabilità totale, la sensibilità e la specificità per il test T_1, per esempio, sono rispettivamente, $\omega_1 = \omega_{11} + \omega_{10}$ e $\phi_1 = \phi_{01} + \phi_{00}$ ecc. Le probabilità delle celle sono date da $p_{kjs} = p_s\omega_{kj} + (1 - p_s)\phi_{kj}$. Il modello per due popolazioni è così completamente specificato. C'è un totale di $6 + S$ parametri, poiché ciascuna sommatoria degli ω_{kj} e ϕ_{kj} è uguale a 1. Poi si definisca la correlazione condizionata tra gli esiti dei test come:

$$\rho_D = \frac{\delta_D}{\sqrt{\omega_1\,(1-\omega_1)\omega_0\,(1-\omega_0)}}, \qquad \delta_D = \omega_{11} - \omega_1\omega_0$$

$$\rho_{\bar{D}} = \frac{\delta_{\bar{D}}}{\sqrt{\phi_1(1-\phi_1)\phi_0(1-\phi_0)}}, \qquad \delta_{\bar{D}} = \phi_{00} - \phi_1\phi_0$$

Queste correlazioni sono zero se e solo se i test sono condizionatamente indipendenti, per esempio se $\omega_{11} = \omega_1\omega_0$, $\omega_{10} = \omega_1(1 - \omega_0)$, ecc. Si definisca con $\boldsymbol{\theta} = (p_1, \dots, p_S, \omega_1, \phi_1)$ il vettore dei parametri da stimare. La funzione di verosimiglianza per i dati osservati multinomiali, $\{y_{kjs}\}$, è:

$$L(\boldsymbol{\theta}; y_{kjs}) = \prod_{t_1=0}^{1}\prod_{t_2=0}^{1}\prod_{s=1}^{S}\{p_s\omega_{kj} + (1 - p_s)\,\phi_{kj}\}^{y_{kjs}}$$

Poiché il modello così specificato non è identificabile, si richiede una distribuzione a priori informativa per almeno alcuni parametri componenti. Il vettore dei parametri non noti è $\boldsymbol{\theta} = (p_1,\dots, p_S, \omega_1,\phi_1)$, poiché si è

ipotizzata una situazione in cui vi sia un'adeguata informazione sulle caratteristiche del test T_1 e una conoscenza sulla prevalenza prontamente disponibile. Si è scelta una distribuzione a priori beta indipendente, come in Johnson e Gastwirth (1991), Gastwirth et al. (1991), Joseph et al. (1995), Mendoza-Blanco et al. (1996), e Johnson et al. (2001). Hanson et al. (1995) hanno considerato il caso in cui il numero delle popolazioni è ampio e le prevalenze sono scambiabili (si veda l'esempio successivo: "*Modello di Hanson et al. (2003)*". In generale, ci si aspetta che ci sia meno informazione disponibile per le correlazioni e per l'accuratezza del nuovo test (test T_2). Si consideri la seguente ri-parametrizzazione per i parametri rimasti, che facilita la specificazione a priori e l'implementazione del campionamento di Gibbs.

$$\lambda_D = P(T_2 = 1|T_1 = 1, D = 1) = \frac{\omega_{11}}{\omega_1},$$

$$\gamma_D = P(T_2 = 1|T_1 = 0, D = 1) = \frac{\omega_{01}}{(1 - \omega_1)},$$

$$\lambda_{\bar{D}} = P(T_2 = 0|T_1 = 0, D = 0) = \frac{\phi_{00}}{\phi_1},$$

$$\gamma_{\bar{D}} = P(T_2 = 0|T_1 = 1, D = 0) = \frac{\phi_{10}}{(1 - \phi_1)}.$$

Così, la precedente scrittura in congiunzione con (p, ω_1, ϕ_1) definisce una trasformazione uno a uno dei parametri. Se i test sono condizionatamente indipendenti, $\lambda_D = \gamma_D = \omega_0$ e $\lambda_{\bar{D}} = \gamma_{\bar{D}} = \phi_0$, mentre se i test sono positivamente correlati $\lambda_D > \omega_0 > \gamma_D$ e $\lambda_{\bar{D}} > \phi_0 > \gamma_{\bar{D}}$.

Un semplice approccio all'inferenza coinvolge la scelta di distribuzioni a priori beta indipendenti per i $6 + S$ parametri. Il nuovo vettore dei parametri da stimare è $\boldsymbol{\theta}' = (p_1, \dots, p_S,\ \omega_1,\ \phi_1,\ \lambda_D, \lambda_{\bar{D}},\ \gamma_D,\ \gamma_{\bar{D}})$. Generalmente ci si aspetta di scegliere un'a priori informativa per η_1, ϕ_1 ed alcune o tutte le prevalenze, e a priori non informative per i rimanenti quattro. Comunque, si può usare un'a priori informativa beta per parametri $\lambda_D, \lambda_{\bar{D}}, \gamma_D, \gamma_{\bar{D}}$. Quando lo stato di malattia non è noto, la verosimiglianza diventa complicata. Conviene indicare con $\{z_{t_1 t_2 s}\}$ la collezione di punteggi che corrispondono agli individui che sono malati. Per esempio, z_{11s} conta il numero non osservato di individui che sono malati da y_{11s} dalla popolazione s, che sono testati positivi su entrambi i test. Se il dato latente $\{z_{t_1 t_2 s}\}$ fosse noto, la verosimiglianza con i "dati-completati", basata sulla nuova parametrizzazione, sarebbe fattorizzata in termini che assomigliano ai contributi binomiali indipendenti. La verosimiglianza con i dati completati e le a priori beta indipendenti determinano l'a posteriori come un prodotto di 8 posteriori beta indipendenti, che possono essere facilmente campionate mediante i seguenti passaggi del campionamento di Gibbs (Tanner, 1996):

1. selezionare i valori di partenza dei parametri in $\boldsymbol{\theta}'$;

2. usando il valore iniziale per $\boldsymbol{\theta}'$, campionare dalle distribuzioni condizionate dei dati latenti condizionatamente ai valori osservati e $\boldsymbol{\theta}'$;
3. usando i dati completati nuovamente campionati, campionare dalle a posteriori con i dati completati, per ottenere una nuova stima del vettore dei parametri $\boldsymbol{\theta}'$.
4. ripetere i passi 2 e 3 iterativamente molte volte per ottenere, dopo un periodo "*burn in*", un campione Monte Carlo (MC), che può essere considerato come un campione dipendente dall'a posteriori congiunta otto-dimensionale (Tanner, 1996). Si controlli la convergenza del campione di Gibbs attraverso la rappresentazione grafica dei risultati per ogni covariata campionata.

- *MODELLO DI HANSON ET AL. (2003)*

Hanson, Johnson e Gardner (2003) hanno proposto una classe di modelli gerarchici con lo scopo di stimare la distribuzione del livello di prevalenza e le accuratezze di due test (T_1 e T_2) in assenza di un *gold standard* quando sono disponibili per il campionamento parecchie popolazioni scambiabili con differente prevalenza della malattia, rilassando l'assunzione d'indipendenza condizionata tra i test. I modelli presentati possono essere estesi a tre o più test o modificati per l'uso con solo un test.

Si assuma che le popolazioni siano scelte casualmente e che, in ogni popolazione, un numero fisso di soggetti scelti a caso siano stati classificati secondo i risultati dei test. Si disponga di $S \geq 2$ popolazioni scambiabili e si assuma che la sensibilità e la specificità siano le stesse in ogni popolazione, anche se tale assunzione di accuratezza costante del test attraverso sotto-popolazioni potrebbe essere non sempre ragionevole. Un campione di numerosità n_s dall's-esima popolazione è classificato in base ai risultati dei test. Per la popolazione s, la distribuzione congiunta di tali osservazioni è multinomiale (n_s, p_s), dove p_s indica la prevalenza della malattia nella popolazione s.

Si assegnino alle sensibilità e specificità distribuzioni a priori di Dirichlet, dato che tale distribuzione è coniugata naturale ai dati multinomiali e generalizza la beta. Un'a priori di Dirichlet dà un peso approssimativamente uguale all'informazione a priori per entrambi i test e le correlazioni condizionate.

Tuttavia si potrebbe notare che il modello descritto senza una distribuzione su p_s non è identificabile. In questo caso, l'ulteriore assunzione di indipendenza condizionata tra i test condizionatamente allo stato di malattia fornirebbe l'identificabilità, ma spesso al costo della distorsione. L'inferenza a posteriori può ancora essere eseguita senza l'assunzione di indipendenza tra i test, ma deve essere fornita un'informazione a priori fortemente accurata sulla prevalenza della malattia in ogni popolazione per ottenere un'inferenza affidabile. Inoltre, assumere che le popolazioni siano interscambiabili, permette alla prevalenza di essere modellata come disegni indipendenti e identicamente distribuiti; una scelta ovvia e altamente flessibile per la distribuzione a priori della prevalenza è la distribuzione beta. Le a posteriori

risultanti piene condizionate sono riconosciute come Dirichlet e sono campionate facilmente ed efficientemente nel campionamento di Gibbs. Se è data un'informazione a priori molto precisa – per esempio, la sensibilità del test T_1 è nota con grande accuratezza ma la sensibilità del test T_2 è relativamente non nota – si deve preferire l'a priori descritta da Dendukuri e Joseph (2001). Nel modello di Hanson et al. questo comporta l'uso dell'algoritmo di Metropolis-Hastings (Tierney, 1994). Essi hanno dimostrato che l'approccio nella sotto-popolazione malata è simile a quello nella sottopopolazione dei non malati.
Assumere una distribuzione beta per la prevalenza, sebbene flessibile, potrebbe non essere una scelta adeguata, perché non si riuscirebbero a catturare alcune caratteristiche quali la multi-modalità; per esempio, parecchie sotto-popolazioni potrebbero avere una prevalenza estremamente bassa, mentre le rimanenti una prevalenza moderata. Un'estensione della distribuzione beta è assegnare a p_s una mistura a priori di processi di Dirichlet (MPD) con una mistura di base beta.

2.6 CONCLUSIONE

La valutazione dei test diagnostici imperfetti è progredita molto negli ultimi decenni, specialmente nell'area di stima della sensibilità e specificità senza fare l'assunzione d'indipendenza condizionata tra i test sul vero stato di malattia. Questi metodi sono necessariamente basati su modelli piuttosto complessi che coinvolgono variabili latenti. Una ragione per la complessità è che devono essere applicati almeno parecchi test agli stessi individui per ottenere abbastanza gradi di libertà per stimare tutti i parametri. Per ogni metodo devono essere sviluppati software specializzati e le procedure di stima numeriche sono generalmente calcoli intensivi. È necessario esplorare più approcci quando l'assunzione di dipendenza condizionata ha un impatto sostanziale sulla stima della sensibilità e specificità dei test diagnostici, per apprendere dagli esperti e dall'analisi empirica di situazioni simili a quella che si sta studiando.

CAPITOLO 3

"Si ritiene la cosa non spiegata e oscura

più importante di quella spiegata e chiara"

(Friedrich Nietzsche)

VALUTAZIONE DI TEST DIAGNOSTICI SENZA GOLD STANDARD MEDIANTE L'INTRODUZIONE DI VARIABILE LATENTE

3.1 INTRODUZIONE

Uno dei primi obiettivi in qualunque studio di valutazione del test diagnostico è confrontare l'accuratezza diagnostica di una nuova procedura diagnostica con quella di una procedura corrente. Nei capitoli 1 e 2 si è più volte esposto che una delle misure comuni per misurare l'accuratezza diagnostica di un test è l'area sottesa alla curva ROC (AUC). La differenza tra le AUC può essere usata come una misura per comparare l'accuratezza diagnostica tra due test diagnostici. Quando il *gold standard* sullo stato di malattia è disponibile, sono stati proposti parecchi metodi per comparare la differenza tra le AUC. Tuttavia, un *gold standard* non sempre potrebbe esistere o potrebbe essere troppo costoso o non realizzabile. Quindi, in molti studi di accuratezza diagnostica viene usato un *imperfect gold standard*. L'inferenza statistica per l'analisi ROC senza un *gold standard* test rimane pressoché inesplorata. Hui e Zhou (1998) hanno elencato i metodi statistici per la stima dell'accuratezza diagnostica di uno o più nuovi test in assenza di *gold standard*. Essi hanno evidenziato che la maggior parte di questi metodi è basata su modelli di mistura e assume l'indipendenza condizionata tra i test, cioè che i due test diagnostici siano indipendenti, condizionatamente al vero stato di malattia. Solo pochi articoli pubblicati hanno trattato la stima di curve

ROC di test su scala continua o ordinale in assenza di un *gold standard*. Henkelman et al. (1990) hanno proposto un metodo di stima di massima verosimiglianza per la curva ROC usando un modello normale multivariato di mistura latente. Il limite maggiore di questo approccio è che si assume che le variabili casuali latenti per molteplici test su scala ordinale seguano una distribuzione normale multivariata, ma non sempre questo è il caso. Beiden et al (2000) hanno proposto una stima di massima verosimiglianza (SMV) di curve ROC di test su scala continua usando l'algoritmo EM. Hall e Zhou (2003) hanno proposto un metodo non parametrico di stima delle curve ROC di test su scala continua sotto l'assunzione d'indipendenza condizionata quando il numero di test è maggiore di due. L'assunzione d'indipendenza condizionata è necessaria perché altrimenti le distribuzioni che compongono un modello multivariato di classe latente non sarebbero identificabili non parametricamente. Zhou et al. (2005) hanno sviluppato l'idea di Hall e Zhou (2003) per stimare in modo non parametrico curve ROC e le loro aree di test su scala ordinale quando il numero di test è maggiore di due. Choi et al. (2006) hanno presentato un metodo bayesiano per la costruzione della differenza tra le AUC di due test correlati in assenza di un *gold standard*, basato sull'assunzione che i dati osservati derivino da una mistura di due distribuzioni normali. Branscum et al. (2008) hanno proposto un altro approccio bayesiano per la stima di curve ROC, basato su un miscuglio di alberi di Polya, che permettono più flessibilità, specialmente se le distribuzioni sottostanti dei risultati del test sono multi-modali. I metodi bayesiani richiedono un'attenta scelta sulle distribuzioni a priori per i parametri del modello. Branscum et al. (2008) ammoniscono sull'uso di a priori non informative nell'analisi bayesiana di problemi di test diagnostici in assenza di un *gold standard* e sostengono l'uso di un'a priori reale e informativa. In aggiunta, i metodi bayesiani potrebbero essere sensibili all'assunzione di distribuzione parametrica bivariata sui risultati del test, come notato da Choi et al. (2006). Hsieh et al. (2009) propongono una nuova procedura basata sulla verosimiglianza per la costruzione di intervalli di confidenza per la differenza di AUC appaiate in assenza di un *gold standard*, sotto l'assunzione di normalità dei risultati del test da ogni gruppo di soggetti malati, usando l'algoritmo EM congiuntamente al metodo *bootstrap*.

Nel paragrafo 3.2 si descriverà il metodo proposto da Zhou et al. (2005) per stimare non parametricamente curve ROC e le rispettive AUC senza *gold standard*. Nel paragrafo 3.3 si descriverà il metodo proposto da Hsieh et al. (2009) per la stima intervallare della differenza di due AUC.

3.2 METODO DI ZHOU ET AL. (2005)

Si consideri la situazione in cui K test diagnostici con un punteggio su scala ordinale da 1 a J siano applicati ad N pazienti. Si assume che lo stato di malattia non sia noto per tutti gli N pazienti. Si denotino $T_1, T_2, \dots, T_n$ le risposte ottenute da K test diagnostici per un particolare paziente di cui lo stato di malattia

non noto è indicato da D, dove $D = 1$ se il paziente è malato e $D = 0$ se il paziente è sano. Dato che ogni test ha un punteggio da 1 a J, allora si può definire la sua curva ROC in due modi:

1) curva ROC non parametrica basata su valori discreti di sensibilità e specificità,
2) curva ROC continua di una variabile latente sottostante ai dati ordinali osservabili.

Si focalizzi l'attenzione su curve ROC non parametriche (caso 1). Per tracciare una curva ROC discreta da dati ordinali, si varia il *cut-off* per un test positivo e poi si calcolano le $(j + 1)$ coppie di frazioni di veri positivi (VPF) e frazioni di falsi positivi (FPF). Nello specifico, per il k-esimo test, se si definisce un test positivo come quello per cui $T_k \geq j$, la corrispondente coppia di VPF e FPF è data da:

$$VPF_k(j) = P(T_k \geq j|D = 1),$$

$$FPF_k(j) = P(T_k \geq j|D = 0),$$

rispettivamente, per $j = 1, 2, \dots, j + 1$.

In particolare, per $j = 1, VPF_k(1) = FPF_k(1) = 1$, e per $j = J + 1, VPF_k(J + 1) = FPF_k(J + 1) = 0$.

Una curva ROC discreta è definita come una funzione discreta di $\big(FPF_k(j), VPF_k(j)\big)$, $j = 1, 2, \dots, j + 1$. Congiungendo le coordinate con segmenti lineari, si ottiene la curva ROC non parametrica. Usando la regola trapezoidale per l'integrazione (Bamber, 1975), si ottiene l'area sotto la curva ROC non parametrica del k-esimo test:

$$AUC_k = \sum_{j=1}^{J-1}\left[P(T_k = j|D = 0)\sum_{l=j+1}^{J} P(T_k = l|D = 1)\right] + \frac{1}{2}\sum_{j=1}^{J} P(T_k = j|D = 0)\ P(T_k = j|D = 1). \quad (3.1)$$

Definendo $\eta_{kj}^{(0)} = P(T_k = j|D = 0)$ e $\eta_{kj}^{(1)} = P(T_k = j|D = 1)$, si può esprimere la curva ROC e la sua area in funzione di $\eta_{kj}^{(0)}$ e $\eta_{kj}^{(1)}$. Le coordinate della curva ROC non parametrica di T_k sono $\big(FPF_k(j), VPF_k(j)\big)$, che sono associate ai parametri $\eta_{kj}^{(0)}$ e $\eta_{kj}^{(1)}$ nella seguente forma:

$$FPF_k(j) = P(T_k \geq j|D = 0)$$

$$= P(T_k = j|D = 0) + P(T_k = j + 1|D = 0) \dots + P(T_k = J|D = 0)$$

$$= \sum_{l=j}^{J} P(T_k = l|D = 0) = \sum_{l=j}^{J} \eta_{kl}^{(0)}.$$

Analogamente,

$$VPF_k(j) = \sum_{l=j}^{J} P(T_k = l|D = 1) = \sum_{l=j}^{J} \eta_{kl}^{(1)}.$$

L'area sotto alla curva ROC per il k-esimo test formulata in (3.1) può quindi essere scritta come segue:

$$AUC_k = \sum_{j=1}^{J-1} \left[\eta_{kj}^{(0)} \sum_{l=j+1}^{J} \eta_{kj}^{(1)} \right] + \frac{1}{2} \sum_{j=1}^{J} \eta_{kj}^{(0)} \eta_{kj}^{(1)}. \tag{3.2}$$

Si desidera formulare la verosimiglianza per questo particolare problema; a tal proposito, si può notare come $\eta_{kj}^{(0)}$ e $\eta_{kj}^{(1)}$ giochino un ruolo centrale. Nello specifico, bisogna essere in grado di trovare una stima di massima verosimiglianza (SMV) per questi parametri e impiegarli per calcolare la SMV per la curva ROC e la sua area sotto ciascuno dei K test. Si definisca y_{ikj} come una variabile binaria, dove $i = 1, 2, \dots, N$, $k = 1, 2, \dots K$ e $j = 1, 2, \dots J$, in modo che $y_{ikj} = 1$ se la risposta del k-esimo test è j per l'i-esimo paziente e $y_{ikj} = 0$ altrimenti.

Si costruisca un vettore di dimensioni $K \times J$ di variabili binarie, $\boldsymbol{y}_i$, in modo che:

$$\boldsymbol{y}_i = (\, y_{i11}, \dots, \, y_{i1J}, \dots, y_{iK1}, \dots, \, y_{iKJ}).$$

Si definisca con $\boldsymbol{y}_i$ il vettore punteggio per l'i-esimo paziente, mentre D_i è lo stato di malattia dell'i-esimo soggetto, dove $D_i = 1$ se l' i-esimo paziente è malato e $D_i = 0$ se l' i-esimo paziente è sano.

Si denoti la funzione con $g_d(\boldsymbol{y}_i) = P(\boldsymbol{y}_i|D_i = d)$ la probabilità condizionata del vettore dei punteggi del test dell'i-esimo paziente $(\boldsymbol{y}_i)$ dato il loro stato di malattia $D_i = d$, con $d = 0, 1$. Assumendo l'indipendenza condizionata dei K test, si può scrivere che:

$$g_d(\boldsymbol{y}_i) = \prod_{k=1}^{K} \prod_{j=1}^{J} [P(T_k = j|D = d)]^{\, y_{ikj}} = \prod_{k=1}^{K} \prod_{j=1}^{J} \left(\eta_{kj}^{(d)} \right)^{y_{ikj}},$$

dove $\eta_{kj}^{(d)} = P(T_k = j|D = d)$.

Ora si impieghi la proprietà "1/0" del vettore $\boldsymbol{y}_i$ per "accendere/spegnere" il $\eta_{kj}^{(d)}$ rilevante.

Si assuma una distribuzione di Bernoulli con $p_d = P(D = d),\ d = 0, 1,\ p_d \in (0, 1)$.

Applicando la formula della probabilità totale, si ottiene che il contributo alla verosimiglianza da parte dell'i-esimo paziente ha la seguente forma:

$$P(\boldsymbol{y}_i) = \sum_{d=0}^{1} P(\boldsymbol{y}_i|D_i = d)P(D_i = d) = \sum_{d=0}^{1} g_d(\boldsymbol{y}_i)P(D_i = d)$$

$$= g_1(\boldsymbol{y}_i)P(D_i = 1) + g_0(\boldsymbol{y}_i)P(D_i = 0) = p_1 g_1(\boldsymbol{y}_i) + p_0 g_0(\boldsymbol{y}_i).$$

Si definiscano i vettori delle probabilità condizionata con

$$\boldsymbol{\eta}_0 = \left(\eta_{11}^{(0)}, \dots, \eta_{1J}^{(0)}, \dots, \eta_{K1}^{(0)}, \dots \eta_{KJ}^{(0)}\right)$$

$$\boldsymbol{\eta}_1 = \left(\eta_{11}^{(1)}, \dots, \eta_{1J}^{(1)}, \dots, \eta_{K1}^{(1)}, \dots \eta_{KJ}^{(1)}\right).$$

La funzione di verosimiglianza di tutti gli N pazienti è data da:

$$L(p_1, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1) = \prod_{i=1}^{N} P(\boldsymbol{y}_i) = \prod_{i=1}^{N} (p_1 g_1(\boldsymbol{y}_i) + p_0 g_0(\boldsymbol{y}_i)).$$

Quindi la log-verosimiglianza risulta:

$$l(p_1, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1) = \sum_{i=1}^{N} log\big(p_1 g_1(\boldsymbol{y}_i) + p_0 g_0(\boldsymbol{y}_i)\big), \tag{3.3}$$

dove $p_0 = 1 - p_1$.

L'obiettivo è trovare le stime di massima verosimiglianza per $p_1, \boldsymbol{\eta}_0 = \left(\eta_{11}^{(0)}, \dots \eta_{KJ}^{(0)}\right)$ e $\boldsymbol{\eta}_1 = \left(\eta_{11}^{(1)}, \dots \eta_{KJ}^{(1)}\right)$, soggette alle condizioni di normalizzazione $\sum_{j=1}^{J} \eta_{kj}^{(d)} = 1$, per $d = 0, 1$ e $k = 1, 2, \dots, K$. Questi sono i parametri necessari per stimare le curve ROC e le loro rispettive aree per i K test. A questo proposito si impiega l'algoritmo EM (*Expectation-Maximization*) per trovare le stime di MV trattando D come un dato mancante.

3.2.1 *Applicazione dell'algoritmo EM*

I dati completi consistono di $(\boldsymbol{y}, D)$. Sia $\boldsymbol{\theta} = (p_1, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1)$.

La funzione di verosimiglianza dei dati completi per tutti gli N pazienti è data da:

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left[\left(p_1 g_1(\boldsymbol{y}_i)\right)^{D_i} \left(p_0 g_0(\boldsymbol{y}_i)\right)^{(1-D_i)} \right]$$

La log-verosimiglianza dei dati completi è data da:

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left[D_i log\left(p_1 g_1(\boldsymbol{y}_i)\right) + (1 - D_i) log\left(p_0 g_0(\boldsymbol{y}_i)\right) \right]$$

Sia $\boldsymbol{\theta}^{(t)}$ la stima di $\boldsymbol{\theta}$ dopo la t-esima iterazione dell'algoritmo EM. L'algoritmo prevede i seguenti passi.

- **Passo E**: la fase E calcola il valore atteso di $l_c(\boldsymbol{\theta})$ dato il valore osservato $\boldsymbol{y}$ e la stima corrente del parametro $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.

$$E\left[l_c(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}\right] = \sum_{i=1}^{N} \sum_{d=0}^{1} \left[P\left(D_i = d|\boldsymbol{y}_i, \boldsymbol{\theta}^{(t)}\right) log\left(p_d g_d(\boldsymbol{y}_i)\right) \right]$$

$$= \sum_{i=1}^{N} \left[P\left(D_i = 1|\boldsymbol{y}_i, \boldsymbol{\theta}^{(t)}\right) log\left(p_1 g_1(\boldsymbol{y}_i)\right) \right] + \sum_{i=1}^{N} \left[P\left(D_i = 0|\boldsymbol{y}_i, \boldsymbol{\theta}^{(t)}\right) log\left(p_0 g_0(\boldsymbol{y}_i)\right) \right] \qquad (3.4)$$

Si denoti con $z_{id}^{(t)}$ la probabilità condizionata di $D_i = d$, dato il valore osservato $\boldsymbol{y}_i$ e la stima corrente del parametro $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$:

$$z_{id}^{(t)} = P\left(D_i = d|\boldsymbol{\theta}^{(t)}\right) = P\left(D_i = d|\boldsymbol{y}_i, p_1^{(t)}, \boldsymbol{\eta}_0^{(t)}, \boldsymbol{\eta}_1^{(t)}\right), \qquad (3.5)$$

per $d = 0, 1$.

Inoltre si indichi con $g_d^{(t)}(\boldsymbol{y}_i = P(\boldsymbol{y}_i|D_i = d)$ la funzione di probabilità condizionata del vettore dei punteggi del test dell'i-esimo paziente ($\boldsymbol{y}_i$) dato il loro stato di malattia $D_i = d$, per la stima corrente del parametro, $\boldsymbol{\theta}^{(t)}$:

$$g_d^{(t)}(\boldsymbol{y}_i) = \prod_{k=1}^{K} \prod_{j=1}^{J} \left[{\eta_{kj}^{(d)}}^{(t)} \right]^{y_{ikj}}, \qquad (3.6)$$

per $d = 0, 1$.

Si può dimostrare che la (3.5) può essere calcolata come:

$$z_{id}^{(t)} = \frac{p_d^{(t)} g_d^{(t)}(\boldsymbol{y}_i)}{p_0^{(t)} g_0^{(t)}(\boldsymbol{y}_i) + p_1^{(t)} g_1^{(t)}(\boldsymbol{y}_i)}. \tag{3.7}$$

Allora il valore atteso (3.4) di $l_c(\boldsymbol{\theta})$ dato il valore osservato $\boldsymbol{y}$ e la stima corrente del parametro $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ può essere espresso come:

$$E\left[l_c(\boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}\right] = \sum_{i=1}^{N}\sum_{d=0}^{1}\left(\frac{p_d^{(t)} g_d^{(t)}(\boldsymbol{y}_i)}{p_0^{(t)} g_0^{(t)}(\boldsymbol{y}_i) + p_1^{(t)} g_1^{(t)}(\boldsymbol{y}_i)}\right) log\left(p_d g_d(\boldsymbol{y}_i)\right)$$

$$= \sum_{i=1}^{N}\sum_{d=0}^{1} z_{id}^{(t)} log\left(\, p_d g_d(\boldsymbol{y}_i)\right)$$

$$= \sum_{i=1}^{N} z_{i0}^{(t)} log\left(\, p_0 g_0(\boldsymbol{y}_i)\right) + \sum_{i=1}^{N} z_{i1}^{(t)} log\left(p_1 g_1(\boldsymbol{y}_i)\right) \tag{3.8}$$

Passo M: nella fase M si calcola la stima aggiornata $\boldsymbol{\theta}^{(t+1)}$ per $\boldsymbol{\theta}$ massimizzando $E\left[l_c(\boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}\right]$ rispetto a $\boldsymbol{\theta}$.

Per esempio, l stima di $p_1^{(t+1)}$ si ottiene derivando la (3.8) rispetto a p_1, ricordando che $p_0 = 1 - p_1$, e ponendo l'espressione risultante uguale a zero.

È

$$\frac{\partial \mathrm{E}\left(l_c(\boldsymbol{\theta})\middle|\mathbf{y}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}\right)}{\partial p_1} = \sum_{i=1}^{N} z_{i0}^{(m)} \frac{1}{(1-p_1) g_0(\boldsymbol{y}_i)}\left(-g_0(\boldsymbol{y}_i)\right) + \sum_{i=1}^{N} z_{i1}^{(m)} \frac{1}{p_1 g_1(\boldsymbol{y}_i)} g_1(\boldsymbol{y}_i),$$

da cui:

$$p_1^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} z_{i1}^{(t)}.$$

Analogamente si ricavano le stime degli altri due parametri, $p_0^{(t+1)}$ e $\eta_{kj}^{(d)^{(t+1)}}$, per $d = 0, 1$.

Dunque si ottiene che $\boldsymbol{\theta}^{(t+1)}$ ha la seguente espressione esplicita:

$$p_1^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} z_{i1}^{(t)}.$$

$$p_0^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} z_{i0}^{(t)},$$

$$\eta_{kj}^{(d)^{(t+1)}} = \frac{\sum_{i=1}^{N} z_{id}^{(t)} y_{ikj}}{\sum_{i=1}^{N} z_{id}^{(t)}},$$

per $d = 0, 1$.

E'utile notare che, sostituendo le precedenti espressioni, si ottiene:

$$p_0^{(t+1)} \eta_{kj}^{(0)^{(t+1)}} + p_1^{(m+1)} \eta_{kj}^{(1)^{(t+1)}}$$

$$= \frac{1}{N} \sum_{i=1}^{N} z_{i0}^{(t)} \frac{\sum_{i=1}^{N} z_{i0}^{(t)} y_{ikj}}{\sum_{i=1}^{N} z_{i0}^{(t)}} + \frac{1}{N} \sum_{i=1}^{N} z_{i1}^{(t)} \frac{\sum_{i=1}^{N} z_{i1}^{(t)} y_{ikj}}{\sum_{i=1}^{N} z_{i1}^{(t)}}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left[\left(z_{i0}^{(t)} + z_{i1}^{(t)} \right) y_{ikj} \right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} y_{ikj} \equiv \overline{y}_{*kj}. \tag{3.9}$$

Anche se non richiesto per le stime iniziali del parametro, la condizione (3.9) vale ad ogni interazione dell'algoritmo. Perciò, questa è una condizione necessaria per ogni insieme di SMV sotto il nostro modello non parametrico. Questa condizione:

$$\hat{p}_0 \hat{\eta}_{kj}^{(0)} + \hat{p}_1 \hat{\eta}_{kj}^{(1)} = \overline{y}_{*kj}. \tag{3.10}$$

verrà chiamata condizione di mistura. Grazie alla proprietà (3.10), si è dimezzato l'effettivo spazio parametrico.

Si ottiene la matrice di covarianza stimata per $\boldsymbol{\theta}$ usando la matrice d'informazione di Fisher:

$$E\left[-\frac{\partial^2 l(p_1, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1)}{\partial (p_1, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1)^2} \right].$$

Si definisca con $\pi_d(j_1, \dots, j_K)$, $d = 0, 1$, la probabilità congiunta di $D = d$ ai *cut-off* j_l, $l = 1, \dots, K$, dei K test. Inoltre, si indichi con $\pi_d(j_k = j)$, $d = 0, 1$, $k = 1, \dots, K$, $j = 1, \dots, J$ la probabilità di $D = d$ quando il *cut-off* j_k relativo al test K è uguale a j. Si definisca con $n(j_1, \dots, j_K)$ la numerosità relativa ai *cut-off* j_l, $l = 1, \dots, K$, dei K test.

Gli elementi della matrice sono così definiti:

$$E\left[-\frac{\partial^2 l(p_1,\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)}{\partial {p_1}^2}\right]\sum_{j_1=1}^{J}\dots\sum_{j_K=1}^{J}\left[E[n(j_1,\dots,j_K)]\times\left(\frac{\pi_1(j_1,\dots,j_K)}{p_1}-\frac{\pi_0(j_1,\dots,j_K)}{p_0}\right)^2\right],$$

$$E\left[-\frac{\partial^2 l(p_1,\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)}{\partial p_1\partial \eta_{kj}^{(0)}}\right]=\sum_{j_1=1}^{J}\dots\sum_{j_{k-1}=1}^{J}\sum_{j_{k+1}=1}^{J}\dots\sum_{j_K=1}^{J}\times$$
$$\left[E[n(j_k=j)]\left(\frac{\pi_0(j_k=j)\,\pi_1(j_k=j)}{p_0p_1\eta_{kj}^{(0)}}\right)-E[n(j_k=J)]\times\left(\frac{\pi_0(j_k=J)\,\pi_1(j_k=J)}{p_0p_1\eta_{kJ}^{(0)}}\right)\right],$$

$$E\left[-\frac{\partial^2 l(p_1,\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)}{\partial p_1\partial \eta_{kj}^{(1)}}\right]=\sum_{j_1=1}^{J}\dots\sum_{j_{k-1}=1}^{J}\sum_{j_{k+1}=1}^{J}\dots\sum_{j_K=1}^{J}\times$$
$$\left[-E[n(j_k=j)]\left(\frac{\pi_0(j_k=j)\,\pi_1(j_k=j)}{p_0p_1\eta_{kj}^{(1)}}\right)+E[n(j_k=J)]\times\left(\frac{\pi_0(j_k=J)\pi_1(j_k=J)}{p_0p_1\eta_{kJ}^{(1)}}\right)\right],$$

$$E\left[-\frac{\partial^2 l(p_1,\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)}{\partial \eta_{kj}^{(0)}\partial \eta_{kj}^{(0)}}\right]=\sum_{j_1=1}^{J}\dots\sum_{j_{k-1}=1}^{J}\sum_{j_{k+1}=1}^{J}\dots\sum_{j_K=1}^{J}\times$$
$$\left[E[n(j_k=j)]\left(\frac{\pi_0(j_k=j)}{\eta_{kj}^{(0)}}\right)^2+E[n(j_k=J)]\times\left(\frac{\pi_0(j_k=J)}{\eta_{kJ}^{(0)}}\right)^2\right]$$

$$E\left[-\frac{\partial^2 l(p_1,\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)}{\partial \eta_{kj_1}^{(0)}\partial \eta_{kj_2}^{(0)}}\right]=\sum_{j_1=1}^{J}\dots\sum_{j_{k-1}=1}^{J}\sum_{j_{k+1}=1}^{J}\dots\sum_{j_K=1}^{J}\times\left[E[n(j_k=J)]\left(\frac{\pi_0(j_k=J)}{\eta_{kJ}^{(0)}}\right)^2\right],$$

$$E\left[-\frac{\partial^2 l(p_1,\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)}{\partial \eta_{k_1j_1}^{(0)}\partial \eta_{k_2j_2}^{(0)}}\right]=\sum_{j_1=1}^{J}\dots\sum_{j_{k_1-1}=1}^{J}\sum_{j_{k_1+1}=1}^{J}\dots\sum_{j_{k_2-1}=1}^{J}\sum_{j_{k_2+1}=1}^{J}\sum_{j_K=1}^{J}\times$$
$$\left[E\left[n\left(j_{k_1}=j_1,j_{k_2}=J\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=j_1,j_{k_2}=J\right)\pi_1\left(j_{k_1}=j_1,j_{k_2}=J\right)}{\eta_{k_1j_1}^{(0)}\eta_{k_2J}^{(0)}}\right)\right.$$
$$+E\left[n\left(j_{k_1}=J,j_{k_2}=j_2\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=J,j_{k_2}=j_2\right)\pi_1\left(j_{k_1}=J,j_{k_2}=j_2\right)}{\eta_{k_1J}^{(0)}\eta_{k_2j_2}^{(0)}}\right)$$
$$\left.-E\left[n\left(j_{k_1}=j_1,j_{k_2}=j_2\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=j_1,j_{k_2}=j_2\right)\pi_1\left(j_{k_1}=j_1,j_{k_2}=j_2\right)}{\eta_{k_1j_1}^{(0)}\eta_{k_2j_2}^{(0)}}\right)\right.$$

$$-E\left[n\left(j_{k_1}=J,j_{k_2}=J\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=J,j_{k_2}=J\right)\pi_1\left(j_{k_1}=J,j_{k_2}=J\right)}{\eta_{k_1J}^{(0)}\eta_{k_2J}^{(0)}}\right)\Bigg],$$

$$E\left[-\frac{\partial^2 l(p_1,\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)}{\partial\eta_{kj}^{(0)}\partial\eta_{kj}^{(1)}}\right]=\sum_{j_1=1}^{J}\dots\sum_{j_{k-1}=1}^{J}\sum_{j_{k+1}=1}^{J}\dots\sum_{j_K=1}^{J}\times$$

$$\left[E[n(j_k=j)]\left(\frac{\pi_0(j_k=j)\,\pi_1(j_k=j)}{\eta_{kj}^{(0)}\eta_{kj}^{(1)}}\right)+E[n(j_k=J)]\times\left(\frac{\pi_0\,(j_k=J)\pi_1(j_k=J)}{\eta_{kJ}^{(0)}\eta_{kJ}^{(1)}}\right)\right],$$

$$E\left[-\frac{\partial^2 l(p_1,\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)}{\partial\eta_{kj_1}^{(0)}\partial\eta_{kj_2}^{(1)}}\right]=\sum_{j_1=1}^{J}\dots\sum_{j_{k-1}=1}^{J}\sum_{j_{k+1}=1}^{J}\dots\sum_{j_K=1}^{J}\times\left[E[n(j_k=J)]\left(\frac{\pi_0\,(j_k=J)\pi_1(j_k=J)}{\eta_{kJ}^{(0)}\eta_{kJ}^{(1)}}\right)\right],$$

$$E\left[-\frac{\partial^2 l(p_1,\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)}{\partial\eta_{k_1j_1}^{(0)}\partial\eta_{k_2j_2}^{(1)}}\right]=\sum_{j_1=1}^{J}\dots\sum_{j_{k_1-1}=1}^{J}\sum_{j_{k_1+1}=1}^{J}\dots\sum_{j_{k_2-1}=1}^{J}\sum_{j_{k_2+1}=1}^{J}\sum_{j_K=1}^{J}\times$$

$$\left[-E\left[n\left(j_{k_1}=j_1,j_{k_2}=J\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=j_1,j_{k_2}=J\right)\pi_1\left(j_{k_1}=j_1,j_{k_2}=J\right)}{\eta_{k_1j_1}^{(0)}\eta_{k_2J}^{(1)}}\right)\right]$$

$$-E\left[n\left(j_{k_1}=J,j_{k_2}=j_2\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=J,j_{k_2}=j_2\right)\pi_1\left(j_{k_1}=J,j_{k_2}=j_2\right)}{\eta_{k_1J}^{(0)}\eta_{k_2j_2}^{(1)}}\right)$$

$$+E\left[n\left(j_{k_1}=j_1,j_{k_2}=j_2\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=j_1,j_{k_2}=j_2\right)\pi_1\left(j_{k_1}=j_1,j_{k_2}=j_2\right)}{\eta_{k_1j_1}^{(0)}\eta_{k_2j_2}^{(1)}}\right)$$

$$+E\left[n\left(j_{k_1}=J,j_{k_2}=J\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=J,j_{k_2}=J\right)\pi_1\left(j_{k_1}=J,j_{k_2}=J\right)}{\eta_{k_1J}^{(0)}\eta_{k_2J}^{(1)}}\right),$$

$$E\left[-\frac{\partial^2 l(p_1,\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)}{\partial\eta_{kj}^{(1)}\partial\eta_{kj}^{(1)}}\right]=\sum_{j_1=1}^{J}\dots\sum_{j_{k-1}=1}^{J}\sum_{j_{k+1}=1}^{J}\dots\sum_{j_K=1}^{J}\times$$

$$\left[E[n(j_k=j)]\left(\frac{\pi_1(j_k=j)}{\eta_{kj}^{(1)}}\right)^2E[n(j_k=J)]\left(\frac{\pi_1(j_k=J)}{\eta_{kJ}^{(1)}}\right)^2\right],$$

$$E\left[-\frac{\partial^2 l(p_1,\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)}{\partial\eta_{kj_1}^{(1)}\partial\eta_{kj_2}^{(1)}}\right]=\sum_{j_1=1}^{J}\dots\sum_{j_{k-1}=1}^{J}\sum_{j_{k+1}=1}^{J}\dots\sum_{j_K=1}^{J}\times\left[E[n(j_k=J)]\left(\frac{\pi_1(j_k=J)}{\eta_{kJ}^{(1)}}\right)^2\right],$$

e

$$E\left[-\frac{\partial^2 l(p_1,\phi_0\,\phi_1)}{\partial\eta_{k_1j_1}^{(1)}\partial\eta_{k_2j_2}^{(1)}}\right]=\sum_{j_1=1}^{J}\dots\sum_{j_{k_1-1}=1}^{J}\sum_{j_{k_1+1}=1}^{J}\dots\sum_{j_{k_2-1}=1}^{J}\sum_{j_{k_2+1}=1}^{J}\sum_{j_K=1}^{J}\times$$

$$\left[E\left[n\left(j_{k_1}=j_1,j_{k_2}=J\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=j_1,j_{k_2}=J\right)\pi_1\left(j_{k_1}=j_1,j_{k_2}=J\right)}{\eta_{k_1j_1}^{(1)}\eta_{k_2J}^{(1)}}\right)\right.$$

$$+E\left[n\left(j_{k_1}=J,j_{k_2}=j_2\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=J,j_{k_2}=j_2\right)\pi_1\left(j_{k_1}=J,j_{k_2}=j_2\right)}{\eta_{k_1J}^{(1)}\eta_{k_2j_2}^{(1)}}\right)$$

$$-E\left[n\left(j_{k_1}=j_1,j_{k_2}=j_2\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=j_1,j_{k_2}=j_2\right)\pi_1\left(j_{k_1}=j_1,j_{k_2}=j_2\right)}{\eta_{k_1j_1}^{(1)}\eta_{k_2j_2}^{(1)}}\right)$$

$$\left.-E\left[n\left(j_{k_1}=J,j_{k_2}=J\right)\right]\times\left(\frac{\pi_0\left(j_{k_1}=J,j_{k_2}=J\right)\pi_1\left(j_{k_1}=J,j_{k_2}=J\right)}{\eta_{k_1J}^{(1)}\eta_{k_2J}^{(1)}}\right)\right].$$

3.2.2 *Uguale probabilità condizionata*

Se si selezionano come parametri iniziali $\eta_{kj}^{(0)}=\eta_{kj}^{(1)}$ per ogni k e j, si ottiene che $g_0^{(t=0)}(y_i)=g_1^{(t=0)}(y_i)$. Dall'equazione (3.7) si evince che $z_{i1}^{(t)}$ non dipende dai dati $\boldsymbol{y}$ e rimane costante per tutti i pazienti $i=1,\dots,N$.

Infatti:

$$z_{i1}^{(t)}=\frac{p_1^{(t)}g_1^{(t)}(\boldsymbol{y}_i)}{p_0^{(t)}g_0^{(t)}(\boldsymbol{y}_i)+p_1^{(t)}g_1^{(t)}(\boldsymbol{y}_i)}=$$

$$\frac{p_1^{(t)}g_1^{(t)}(\boldsymbol{y}_i)}{p_0^{(t)}g_1^{(t)}(\boldsymbol{y}_i)+p_1^{(t)}g_1^{(t)}(\boldsymbol{y}_i)}=$$

$$\frac{p_1^{(t)}g_1^{(t)}(\boldsymbol{y}_i)}{g_1^{(t)}(\boldsymbol{y}_i)\left(p_0^{(t)}+p_1^{(t)}\right)}=$$

$$\frac{p_1^{(t)}}{p_0^{(t)}+p_1^{(t)}}=p_1^{(t)}.$$

Così, durante ogni interazione, si ottiene $p_1^{(t+1)}=p_1^{(t)}=\dots=p_1^{(0)}$, quindi la prevalenza p_1 resta fissa al valore selezionato per la stima iniziale.

Inoltre si può dimostrare che la relazione (3.9) diviene:

$$p_0^{(t+1)}\eta_{kj}^{(0)^{(t+1)}} + p_1^{(t+1)}\eta_{kj}^{(1)^{(t+1)}} = \eta_{kj}^{(1)^{(t+1)}}\left(p_0^{(t+1)} + p_1^{(t+1)}\right)$$

$$= \eta_{kj}^{(1)^{(t+1)}} = \frac{1}{N}\sum_{i=1}^{N} y_{ikj} \equiv \overline{y}_{*kj}$$

Quindi, $\eta_{kj}^{(d)^{(t+1)}} = \frac{1}{N}\sum_{i=1}^{N} y_{ikj} \equiv \overline{y}_{*kj}$ per $d = 0, 1,\ 1 \le k \le K$ e $1 \le j \le J$.

Così, se si scelgono i parametri iniziali tali che $\eta_{kj}^{(0)} = \eta_{kj}^{(1)}$ per ogni k e j e un valore per $p_1^{(0)}$, la procedura iterativa si fermerà dopo appena un'iterazione. In questo caso, la funzione punteggio della log-verosimiglianza è zero; ne deriva che ogni caso è un massimo locale di log-verosimiglianza. Quindi, si nota l'esistenza di un infinito numero di massimi locali di log-verosimiglianza. Se l'insieme di massimi locali comprende tutti i massimi locali, significa che p_1 è indeterminato e $\eta_{kj}^{(0)} = \eta_{kj}^{(1)}$ per ogni k e j. Questo implica che ogni test è privo di valore per determinare lo stato di malattia perché ogni risultato di qualunque test per un paziente è ugualmente probabile, indifferentemente dallo stato di malattia. Ovviamente, questo problema è sensibile alla selezione delle stime iniziali del parametro e si possono scartare quelli che probabilmente sarebbe opportuno evitare. Tuttavia nel cercare il massimo globale usando l'approccio non parametrico proposto, si fanno queste raccomandazioni:

(1) evitare valori uguali di $\eta_{kj}^{(0)} = \eta_{kj}^{(1)}$ per ogni k e j; questa non dovrebbe essere una decisione difficile poiché nella pratica è spesso evidente una certa asimmetria nei punteggi del test;
(2) provare un insieme di stime iniziali ragionevoli del parametro e paragonare i massimi locali di log-verosimiglianza ottenuti;
(3) ottenere valori iniziali ragionevoli da studi simili con lo stato di malattia noto;
(4) studiare la superficie di verosimiglianza usando tecniche di esplorazione e simulazione, come l'algoritmo EM stocastico.

3.2.3 *Proprietà di invarianza della funzione di log-verosimiglianza*

Considerando l'equazione di log-verosimiglianza (3.3), si può notare come essa sia invariante rispetto ad una riclassificazione dell'insieme di parametri $(p_0, \boldsymbol{\eta}_0, p_1, \boldsymbol{\eta}_1)$ in $(p_1, \boldsymbol{\eta}_1, p_{0,}\boldsymbol{\eta}_0)$. Questo implica che non può mai esistere un unico massimo globale soluzione di verosimiglianza, perché ciascun massimo, diciamo $(\hat{p}_1\hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\eta}}_1)$, potrebbe implicare l'esistenza di un massimo "speculare" di uguale verosimiglianza in $(1 - \hat{p}_1\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_0)$. Ciò che si può fare è arrivare al presunto massimo globale mediante l'uso dell'algoritmo EM

e distinguere tra le due possibilità ordinando ragionevolmente il tasso di prevalenza $\hat{p}_1$o $\hat{p}_0$ o considerando la plausibilità delle aree risultanti sotto le K curve ROC. Ciò potrebbe essere possibile perché l'area sotto qualunque test T_k, diciamo $AUC_k\left(\boldsymbol{\eta}_{k\bullet}^{(0)}, \boldsymbol{\eta}_{k\bullet}^{(1)}\right)$, è uguale a $1-AUC_k\left(\boldsymbol{\eta}_{k\bullet}^{(1)}, \boldsymbol{\eta}_{k\bullet}^{(0)}\right)$, dove $\boldsymbol{\eta}_{k\bullet}^{(d)}=\left(\eta_{k1}^{(0)}, \ldots, \eta_{kJ}^{(0)}\right)$.

Dall'espressione (3.2) su un'area della curva ROC, si ottiene infatti che:

$$AUC_k\left(\boldsymbol{\eta}_{k\bullet}^{(0)}, \boldsymbol{\eta}_{k\bullet}^{(1)}\right) + AUC_k\left(\boldsymbol{\eta}_{k\bullet}^{(1)}, \boldsymbol{\eta}_{k\bullet}^{(0)}\right)$$

$$= \sum_{j=1}^{J-1}\left[\eta_{kj}^{(0)} \sum_{l=j+1}^{J} \eta_{kl}^{(1)}\right] + \frac{1}{2}\sum_{j=1}^{J} \eta_{kj}^{(0)}\eta_{kj}^{(1)} + \sum_{l=1}^{J-1}\left[\eta_{kl}^{(1)} \sum_{j=l+1}^{J} \eta_{kj}^{(0)}\right] + \frac{1}{2}\sum_{j=1}^{J} \eta_{kj}^{(1)}\eta_{kj}^{(0)}$$

$$= \sum_{j=1}^{J-1}\left[\eta_{kj}^{(0)} \sum_{l=j+1}^{J} \eta_{kl}^{(1)}\right] + \sum_{l=1}^{J-1}\left[\eta_{kl}^{(1)} \sum_{j=l+1}^{J} \eta_{kj}^{(0)}\right] + \sum_{j=1}^{J} \eta_{kj}^{(0)}\eta_{kj}^{(1)}.$$

Si noti che:

$$\sum_{l=1}^{J-1}\left[\eta_{kl}^{(1)} \sum_{j=l+1}^{J} \eta_{kj}^{(0)}\right] = \sum_{j=2}^{J}\left[\eta_{kj}^{(0)} \sum_{l=1}^{j-1} \eta_{kl}^{(1)}\right]$$

Allora si ricava che:

$$AUC_k\left(\boldsymbol{\eta}_{k\bullet}^{(0)}, \boldsymbol{\eta}_{k\bullet}^{(1)}\right) + AUC_k\left(\boldsymbol{\eta}_{k\bullet}^{(1)}, \boldsymbol{\eta}_{k\bullet}^{(0)}\right)$$

$$= \sum_{j=1}^{J-1}\left[\eta_{kj}^{(0)} \sum_{l=j+1}^{J} \eta_{kl}^{(1)}\right] + \sum_{j=2}^{J}\left[\eta_{kj}^{(0)} \sum_{l=1}^{j-1} \eta_{kl}^{(1)}\right] + \sum_{j=1}^{J} \eta_{kj}^{(0)}\eta_{kj}^{(1)}$$

$$= \eta_{k1}^{(0)}\left(\sum_{l=2}^{J} \eta_{kl}^{(1)}\right) + \sum_{j=2}^{J-1}\left[\eta_{kj}^{(0)} \sum_{l=j+1}^{J} \eta_{kl}^{(1)}\right] + \eta_{kJ}^{(0)}\left(\sum_{l=1}^{J-1} \eta_{kl}^{(1)}\right) + \sum_{j=2}^{J-1}\left[\eta_{kj}^{(0)} \sum_{l=1}^{j-1} \eta_{kl}^{(1)}\right] + \sum_{j=1}^{J} \eta_{kj}^{(0)}\eta_{kj}^{(1)}$$

$$= \eta_{k1}^{(0)}\left(\sum_{l=2}^{J} \eta_{kl}^{(1)}\right) + \eta_{kJ}^{(0)}\left(\sum_{l=1}^{J-1} \eta_{kl}^{(1)}\right) + \sum_{j=2}^{J-1}\left[\eta_{kj}^{(0)}\left(\sum_{l=1}^{j-1} \eta_{kl}^{(1)} + \sum_{l=j+1}^{J} \eta_{kl}^{(1)}\right)\right] + \sum_{j=1}^{J} \eta_{kj}^{(0)}\eta_{kj}^{(1)}$$

$$= \eta_{k1}^{(0)}\left(1-\eta_{k1}^{(1)}\right) + \eta_{kJ}^{(0)}\left(1-\eta_{kJ}^{(1)}\right) + \sum_{j=2}^{J-1} \eta_{kj}^{(0)}\left(1-\eta_{kj}^{(1)}\right) + \sum_{j=1}^{J} \eta_{kj}^{(0)}\eta_{kj}^{(1)}$$

$$= \sum_{j=1}^{J} \eta_{kj}^{(0)}\left(1-\eta_{kj}^{(1)}\right) + \sum_{j=1}^{J} \eta_{kj}^{(0)}\phi_{1kj}$$

$$= \sum_{j=1}^{J} \eta_{kj}^{(0)} - \sum_{j=1}^{J} \eta_{kj}^{(0)} \eta_{kj}^{(1)} + \sum_{j=1}^{J} \eta_{kj}^{(0)} \eta_{kj}^{(1)}$$

$$= \sum_{j=1}^{J} \eta_{kj}^{(0)} = 1,$$

dove si è ripetutamente usato la nozione che $\sum_{j=1}^{J} \eta_{kj}^{(0)} = \sum_{j=1}^{J} \eta_{kj}^{(1)} = 1$.

Quindi si può concludere che:

$$AUC_k \boldsymbol{\eta}_{k\bullet}^{(0)}, \boldsymbol{\eta}_{k\bullet}^{(1)} + AUC_k \boldsymbol{\eta}_{k\bullet}^{(1)}, \boldsymbol{\eta}_{k\bullet}^{(0)} = 1.$$

Tuttavia, per ogni test credibile si potrebbe presumibilmente scegliere il massimo globale per il quale $AUC_k > 0.5$ per ogni $k = 1, \dots, K$, se ne esiste almeno uno.

3.2.4 *Somma dei quadrati dei residui*

La somma dei quadrati dei residui è definita da:

$$SS = \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} \left(y_{ikj} - E\left[y_{ikj}\right]\right)^2.$$

Dato che per ogni paziente i, $E\left[y_{ikj}\right] = \hat{p}_0 \, \hat{\phi}_{0kj} + \hat{p}_1 \, \hat{\phi}_{1kj}$, dalla condizione di mistura di MVS si ottiene che $E\left[y_{ikj}\right] = \overline{y}_{*kj}$. Quindi, ad ogni massimo locale e per ogni iterazione dell'algoritmo EM la somma dei quadrati dei residui rimane costante al valore:

$$SS = \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} \left(y_{ikj} - \bar{y}_{*kj}\right)^2.$$

Quindi, ad ogni interazione dell'algoritmo EM, i parametri si aggiornano nella stessa maniera in cui aumenta la log-verosimiglianza e la somma dei quadrati dei residui rimane fissa. Esattamente come per le stime di massima verosimiglianza, essendoci un numero infinito di scelte adeguate sull'insieme di parametri, dalla somma dei quadrati non c'è modo di distinguere tra le scelte possibili per un set "migliore".

3.3 METODO DI HSIEH ET AL. (2009)

Si denoti con T_1 e T_2 i risultati del test di due test diagnostici sullo stesso paziente il cui stato di malattia è denotato da D. Se il paziente è malato, allora $D = 1$, se il paziente è sano, allora $D = 0$. Si indichino i risultati dei due test su un paziente malato con X_1 e X_2 rispettivamente, mentre quelli su un paziente sano con Y_1 e Y_2, rispettivamente. Inoltre si definisca con $P(X_k > j) = S_{X,k}(j)$ e $P(Y_k > j) = S_{Y,k}(j)$ la frazione di veri positivi e falsi positivi al *cut-off j* per il test diagnostico T_k, rispettivamente, per $k = 1, 2$. Per il test diagnostico T_k, una curva ROC è ottenuta dai punti di coordinate $\{S_{Y,k}(j),\ S_{X,k}(j)\}$ per tutti i valori possibili del *cut-off* j. Si può anche esprimere la curva ROC come una funzione di $t = S_{Y,k}(j)$, data da $ROC_k(j) = S_{X,k}\left(S_{Y,k}^{-1}(j)\right)$, dove $S_{Y,k}^{-1}(j)$ è la funzione inversa di $S_{Y,k}(j)$. L'AUC per il test diagnostico T_k è:

$$AUC_k = \int_0^1 ROC_k(t)dt,$$

che si dimostra essere uguale a $AUC_k = P(X_k \geq Y_k)$.

Si assuma che i risultati dei due test di un soggetto malato, X_1 e X_2, seguano una distribuzione normale bivariata,

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N_2(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D),$$

e che anche i risultati del test di un soggetto sano, Y_1 e Y_2, seguano una distribuzione normale bivariata,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N_2(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}}),$$

dove

$$\boldsymbol{\mu}_D = \begin{bmatrix} \mu_{1D} \\ \mu_{2D} \end{bmatrix}, \quad \boldsymbol{\Sigma}_D = \begin{bmatrix} \sigma_{1D}^2 & \rho_D \\ \rho_D & \sigma_{2D}^2 \end{bmatrix}, \quad \boldsymbol{\mu}_{\bar{D}} = \begin{bmatrix} \mu_{1\bar{D}} \\ \mu_{2\bar{D}} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\bar{D}} = \begin{bmatrix} \sigma_{1\bar{D}}^2 & \rho_{\bar{D}} \\ \rho_{\bar{D}} & \sigma_{2\bar{D}}^2 \end{bmatrix}.$$

Per semplicità, si è usata la notazione che usa il pedice D per indicare gli individui malati e $\bar{D}$ per i soggetti sani. Il vettore dei parametri in questo insieme è dato da:

$$\boldsymbol{\theta}' = (p, \mu_{1D}, \mu_{2D}, \mu_{1\bar{D}}, \mu_{2\bar{D}}, \sigma_{1D}^2, \sigma_{2D}^2, \sigma_{1\bar{D}}^2, \sigma_{2\bar{D}}^2, \rho_D, \rho_{\bar{D}}).$$

dove $p = P(D = 1)$. Sotto questo modello, l'assunzione di indipendenza condizionata tra test è un caso speciale con $\rho_D = \rho_{\bar{D}} = 0$.

L'AUC per un test diagnostico T_k può essere espressa come $AUC_k = \Phi(\psi_k)$, dove:

$$\psi_k = \frac{\mu_{kD} - \mu_{k\bar{D}}}{\sqrt{\sigma_{kD}^2 + \sigma_{k\bar{D}}^2}},$$

con $k = 1, 2$, dove $\Phi(\cdot)$ è la funzione di distribuzione della nomale standard.

Per ottenere una stima di verosimiglianza per $\boldsymbol{\theta}$ e una stima intervallare per $\Delta = AUC_1 - AUC_2$, bisogna osservare se si disponga di un *gold standard*. Interessa studiare il problema nel caso di assenza di *gold standard*. Tuttavia, per completezza, verrà prima richiamato il caso in cui si abbia a disposizione un *gold standard* sul vero stato di malattia.

3.3.1 *Disponibilità di un gold standard*

Se esiste un *gold standard* sul vero stato di malattia, allora $\boldsymbol{X}$ e $\boldsymbol{Y}$ sono disponibili. Così, la stima di massima verosimiglianza per $\boldsymbol{\theta}$ può essere facilmente derivata e la stima intervallare per $\Delta = AUC_1 - AUC_2$ può essere ottenuta usando il metodo *bootstrap*.

Si supponga che $\boldsymbol{X}_1, \dots, \boldsymbol{X}_{n_1}$ e $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{n-n_1}$ siano due campioni casuali da $N_2(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ e $N_2(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}})$, rispettivamente. Si ottengono i seguenti stimatori per $(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ e $(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}})$:

$$\left(\widehat{\boldsymbol{\mu}}_D, \widehat{\boldsymbol{\Sigma}}_D\right) = \left(\frac{1}{n_1}\sum_{i=1}^{n_1} \boldsymbol{X}_i, \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})\ (\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T\right)$$

$$= \left(\begin{bmatrix}\bar{X}_1 \\ \bar{X}_2\end{bmatrix}, \frac{1}{n_1 - 1}\begin{bmatrix}SX_1 & SX_{12} \\ SX_{12} & SX_2\end{bmatrix}\right),$$

$$\left(\widehat{\boldsymbol{\mu}}_{\bar{D}}, \widehat{\boldsymbol{\Sigma}}_{\bar{D}}\right) = \left(\frac{1}{n - n_1}\sum_{i'=1}^{n-n_1} \boldsymbol{Y}_{i'}, \frac{1}{(n - n_1) - 1}\sum_{i'=1}^{n-n_1} (\boldsymbol{Y}_{i'} - \overline{\boldsymbol{Y}})\ (\boldsymbol{Y}_{i'} - \overline{\boldsymbol{Y}})^T\right)$$

$$= \left(\begin{bmatrix}\bar{Y}_1 \\ \bar{Y}_2\end{bmatrix}, \frac{1}{(n - n_1) - 1}\begin{bmatrix}SY_1 & SY_{12} \\ SY_{12} & SY_2\end{bmatrix}\right).$$

Da queste stime si ottiene la stime di massima verosimiglianza di Δ, $\hat{\Delta}$, che è la differenza nelle AUC appaiate dei due test. Si usi la seguente procedura per ottenere un intervallo di confidenza *bootstrap* per Δ con livello di confidenza $100(1 - \alpha)\%$.

Passo 1. Calcolare le stime di massima verosimiglianza per Δ, $\hat{\Delta}$, basate sui dati osservati, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2,\ldots, \boldsymbol{x}_{n_1})$ e $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2,\ldots, \boldsymbol{y}_{n-n_1})$.

Passo 2. Generare campioni casuali *bootstrap*, $\boldsymbol{x}^\star = (\boldsymbol{x}_1^\star, \boldsymbol{x}_2^\star, \ldots, \boldsymbol{x}_{n_1}^\star)$ e $\boldsymbol{y}^\star = (\boldsymbol{y}_1^\star, \boldsymbol{y}_2^\star, \ldots, \boldsymbol{y}_{n-n_1}^\star)$, con una numerosità pari a n, da un campionamento con reinserimento dai dati osservati, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2,\ldots, \boldsymbol{x}_{n_1})$ e $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2,\ldots, \boldsymbol{y}_{n-n_1})$, dove $B = 200$.

Passo 3. Stimare $\Delta = AUC_1 - AUC_2$ per ognuno dei B campioni casuali *bootstrap*, ottenendo $\hat{\Delta}_{boot}$. Poi si calcoli la varianza campionaria di queste B stime e la si denoti con $\widehat{var}(\hat{\Delta}_{boot})$.

Passo 4. Usare il risultato $\hat{\Delta}$ del passo 1 e $\widehat{var}(\hat{\Delta}_{boot})$ del passo 3 per costruire l'intervallo di confidenza per Δ di livello $100(1-\alpha)$% nel seguente modo:

$$\left(\hat{\Delta} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\Delta}_{boot})}, \hat{\Delta} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}\left(\hat{\Delta}_{boot}\right)}\right).$$

3.3.2 *Assenza di un gold standard*

Se non si possiede un *gold standard*, allora $\boldsymbol{X}$ e $\boldsymbol{Y}$ non sarebbero disponibili e la derivazione non sarebbe così semplice. Si proponga una stima intervallare basata sulla stima di massima verosimiglianza per $\Delta = AUC_1 - AUC_2$ usando l'algoritmo EM e il metodo *bootstrap* nella situazione di assenza di *gold standard.*

Si denoti con t_{ik} il risultato osservato del k-esimo test sull'i-esimo soggetto, per $k = 1, 2$, con D_i lo stato di malattia non osservato dell'i-esimo soggetto, e con $p = (D_i = 1)$ la prevalenza della malattia. Si indichi con $\boldsymbol{t}_i = (t_{i1}, t_{i2})$ il vettore del risultato osservato sull'i-esimo soggetto, con $\boldsymbol{t} = (\boldsymbol{t}_1,\ldots, \boldsymbol{t}_n)$ il vettore del risultato osservato per gli n soggetti e con $\boldsymbol{D} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n)$ il vettore dei veri stati di malattia non osservati di tutto il campione di n soggetti. Si ricordi che $\boldsymbol{X}$ e $\boldsymbol{Y}$ seguono distribuzioni normali bivariate, tali che $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ e $\boldsymbol{Y} \sim N_2(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}})$. Se $\boldsymbol{D}$ fosse osservato, allora la funzione di verosimiglianza con i dati completi sarebbe data dalla seguente formula:

$$L_c(\boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{D}) = \prod_{i=1}^{n}\left[\left(p f_{\boldsymbol{X}}(\boldsymbol{t}_i)\right)^{D_i} \cdot \left((1-p) f_{\boldsymbol{Y}}(\boldsymbol{y}_i)\right)^{(1-D_i)}\right],$$

dove $f_{\boldsymbol{X}}(\boldsymbol{t})$ è la funzione di densità di $N_2(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ e $f_{\boldsymbol{Y}}(\boldsymbol{t})$ è la funzione di densità di $N_2(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}})$.

Dunque la funzione di log-verosimiglianza è:

$$l_c(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{D}) = \sum_{i=1}^{n}\left[D_i\ log\left(pf_{\boldsymbol{X}}(\boldsymbol{t}_i)\right) + \ (1-D_i\ )\, log\left((1-p)f_{\boldsymbol{Y}}(\boldsymbol{t}_i)\right)\right].$$

Indichiamo con $\boldsymbol{\theta}^{(t)}$ la stima di $\boldsymbol{\theta}$ dopo la t-esima iterazione dell'algoritmo EM. Le seguenti fasi, passo E e passo M, sono usate per trovare $\boldsymbol{\theta}^{(t+1)}$, una stima aggiornata di $\boldsymbol{\theta}$.

- **Passo E**. Il passo E calcola il valore atteso condizionato di $l^c(\boldsymbol{\theta})$ sotto i dati osservati $\boldsymbol{t}$ e la stima del parametro corrente, $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. Questo è:

$$E\left[l_c(\boldsymbol{\theta})|\boldsymbol{t},\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}\right] =$$

$$= \sum_{i=1}^{n}\left[P(D_i = 1\left|\boldsymbol{t}_i,\boldsymbol{\theta}^{(t)}\right)\ log\left(p\ f_{\boldsymbol{X}}(\boldsymbol{t}_i)\right) + P(D_i\ = 0\left|\boldsymbol{t}_i,\boldsymbol{\theta}^{(t)}\right)\ log\left((1-p)\ f_{\boldsymbol{Y}}(\boldsymbol{t}_i)\right)\right].$$

Definendo $z_{id}^{(t)}$ come

$$z_{id}^{(t)} = P\left(D_i\ = d|\boldsymbol{t}_i, p^{(t)}, \mu_{D1}^{(t)}\mu_{,2D}^{(t)}, \mu_{1\bar{D}}^{(t)}, \mu_{2\bar{D}}^{(t)}, \sigma_{1D}^{2^{(t)}}, \sigma_{2D}^{2^{(t)}}, \sigma_{1\bar{D}}^{2^{(t)}}, \sigma_{2\bar{D}}^{2^{(t)}}, \rho_D^{(t)}, \rho_{\bar{D}}^{(t)}\right),$$

Si può dimostrare che:

$$z_{i1}^{(t)} = \frac{p^{(t)}f_{\boldsymbol{X}}^{(t)}(\boldsymbol{t}_i)}{p^{(t)}f_{\boldsymbol{X}}^{(t)}(\boldsymbol{t}_i) + (1-p)^{(t)}f_{\boldsymbol{Y}}^{(t)}(\boldsymbol{t}_i)},$$

$$z_{i0}^{(t)} = \frac{(1-p)^{(t)}\ f_{\boldsymbol{Y}}^{(t)}(\boldsymbol{t}_i)}{p^{(t)}f_{\boldsymbol{X}}^{(t)}(\boldsymbol{t}_i) + (1-p)^{(t)}f_{\boldsymbol{Y}}^{(t)}(\boldsymbol{t}_i)},$$

$$E\left[l_c(\boldsymbol{\theta})|\boldsymbol{t},\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}\right] = \sum_{i=1}^{n}\left[z_{i1}^{(t)}\log\left(p\ f_{\boldsymbol{X}}(\boldsymbol{t}_i)\right) + z_{i0}^{(t)}\log\left((1-p)f_{\boldsymbol{Y}}(\boldsymbol{t}_i)\right)\right].$$

- **Passo M**. Il passo M trova la stima aggiornata $\boldsymbol{\theta}^{(t+1)}$ per $\boldsymbol{\theta}$ massimizzando $E[l^c(\boldsymbol{\theta})|\boldsymbol{t},\boldsymbol{\theta} = \boldsymbol{\theta}^{t}]$ rispetto a $\boldsymbol{\theta}$. Gli elementi di $\boldsymbol{\theta}^{(t+1)}$ sono:

$$\hat{p}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{i1}^{(t)},$$

$$\hat{\mu}_{1D}^{(t+1)} = \frac{\sum_{i=1}^{n} z_{i1}^{(t)} t_{i1}}{\sum_{i=1}^{n} z_{i1}^{(t)}},$$

$$\hat{\mu}_{2D}^{(t+1)} = \frac{\sum_{i=1}^{n} z_{i1}^{(t)} t_{i2}}{\sum_{i=1}^{n} z_{i1}^{(t)}},$$

$$\hat{\sigma}^{2}{}_{1D}^{(t+1)} = \frac{\sum_{i=1}^{n} z_{i1}^{(t)} (t_{i1} - \hat{\mu}_{1D})^2}{\sum_{i=1}^{n} z_{i1}^{(t)}},$$

$$\hat{\sigma}^{2}{}_{2D}^{(t+1)} = \frac{\sum_{i=1}^{n} z_{i1}^{(t)} (t_{i2} - \hat{\mu}_{2D})^2}{\sum_{i=1}^{n} z_{i1}^{(t)}},$$

$$\hat{\rho}_{D}^{(t+1)} = \frac{\sum_{i=1}^{n} z_{i1}^{(t)} (t_{i1} - \hat{\mu}_{1D})(t_{i2} - \hat{\mu}_{2D})}{\sum_{i=1}^{n} z_{i1}^{(t)}}.$$

E anche:

$$\hat{\mu}_{1\bar{D}}^{(t+1)} = \frac{\sum_{i=1}^{n} z_{i0}^{(t)} t_{i1}}{\sum_{i=1}^{n} z_{i0}^{(t)}},$$

$$\hat{\mu}_{2\bar{D}}^{(t+1)} = \frac{\sum_{i=1}^{n} z_{i0}^{(t)} t_{i2}}{\sum_{i=1}^{n} z_{i0}^{(t)}},$$

$$\hat{\sigma}^{2}{}_{1\bar{D}}^{(t+1)} = \frac{\sum_{i=1}^{n} z_{i0}^{(t)} (t_{i1} - \hat{\mu}_{1\bar{D}})^2}{\sum_{i=1}^{n} z_{i0}^{(t)}},$$

$$\hat{\sigma}^{2}{}_{2\bar{D}}^{(t+1)} = \frac{\sum_{i=1}^{n} z_{i0}^{(t)} (t_{i2} - \hat{\mu}_{2\bar{D}})^2}{\sum_{i=1}^{n} z_{i0}^{(t)}},$$

$$\hat{\rho}_{\bar{D}}^{(m+1)} = \frac{\sum_{i=1}^{n} z_{i0}^{(m)} (t_{i1} - \hat{\mu}_{1\bar{D}})(t_{i2} - \hat{\mu}_{2\bar{D}})}{\sum_{i=1}^{n} z_{i0}^{(m)}}.$$

Il valore a cui converge di $\boldsymbol{\theta}^{(t+1)}$ nell'algoritmo EM è la stima di massima verosimiglianza di $\boldsymbol{\theta}$. Infine, sostituendo la stima di massima verosimiglianza di $\boldsymbol{\theta}$ in $\Delta = AUC_1 - AUC_2$, si ottiene la stima di massima verosimiglianza di Δ, $\hat{\Delta}$.

A causa della forma complicata della varianza di $\hat{\Delta}$, per ottenere una stima della sua varianza si può usare il metodo *bootstrap*. Successivamente, $\hat{\Delta}$ e la sua varianza stimata sono usati per costruire l'intervallo di confidenza della differenza di AUC appaiate in assenza di un *gold standard*.

Un intervallo di confidenza di livello $100(1-\alpha)\%$ *boostrap* con code uguali per $\Delta = AUC_1 - AUC_2$ può essere ottenuto dalla seguente procedura.

Passo 1. Definire un insieme iniziale di valori per $p, \boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D$ *e* $\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}}$.

Passo 2. Usare l'algoritmo EM per ottenere $\hat{\Delta}$, basati sui dati osservati, $\boldsymbol{t} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n)$.

Passo 3. Generare B campioni *bootstrap*, $\boldsymbol{t}^\star = (\boldsymbol{t}_1^\star, \boldsymbol{t}_2^\star, \ldots, \boldsymbol{t}_n^\star)$, dai dati osservati, $\boldsymbol{t}$, senza reinserimento, in modo che ogni campione *bootstrap* abbia numerosità n, dove $B = 200$.

Passo 4. Usare l'algoritmo EM per stimare $\Delta = AUC_1 - AUC_2$ per ogni campione *bootstrap*. Poi, da queste B stime *bootstrap* di Δ, si può ottenere la stima della varianza campionaria per la varianza di $\hat{\Delta}$, denotata da $\widehat{var}(\hat{\Delta}_{boot})$.

Passo 5. Usare il risultato $\hat{\Delta}$ del passo 2 e $\widehat{var}(\hat{\Delta}_{boot})$ del passo 4 per costruire l'intervallo di confidenza per Δ di livello $100(1-\alpha)\%$ nel seguente modo:

$$\left(\hat{\Delta} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\Delta}_{boot}}), \hat{\Delta} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\Delta}_{boot})}\right)$$

3.4 CONCLUSIONE

Nella sezione 3.2, si è dimostrato che il metodo non parametrico presentato è caratterizzato dall'esistenza di una stima globale di massima verosimiglianza che ha una soluzione "speculare" con lo stesso valore della log-verosimiglianza, risultato consistente con il problema di distorsione dell'*imperfect gold standard* su scala binaria come osservato da Hui e Walter (1980). Una complicazione aggiuntiva nella ricerca di una stima globale di massima verosimiglianza è che vi sono molteplici soluzioni di stima locale di massima verosimiglianza. Inoltre il metodo descritto comporta stime locali di verosimiglianza che, includendo quelle globali, danno tutte la stessa somma dei quadrati dei residui. Un vantaggio dell'approccio non parametrico

di massima verosimiglianza proposto è che non richiede specifiche assunzioni sul modello sottostante, perciò può essere più robusto. Un altro vantaggio è che c'è una soluzione esplicita al passo M, così è molto più facile implementare il corrispondente algoritmo EM. Gli studi di simulazione eseguiti dagli autori mostrano che le stime ricavate non parametricamente sono comparabili a quelle ottenute dai modelli parametrici. In questo modo si può raggirare la verosimiglianza parametrica che, a causa dello stato latente di malattia, spesso comporta una forma di mistura così complicata da rendere difficoltosi i calcoli. Inoltre i calcoli per i modelli parametrici sono più sensibili ai valori iniziali e meno stabili quando le celle delle tabelle di contingenza sono vicine alla degenerazione.

Nella sezione 3.3 è stata presentata una procedura per la costruzione di intervalli di confidenza per la differenza di AUC appaiate ($\Delta = AUC_1 - AUC_2$) in assenza di un *gold standard* sul vero stato di malattia di un paziente, sotto l'assunzione di normalità per i risultati di test diagnostici da ciascun gruppo di soggetti malati, usando l'algoritmo EM assieme al metodo *bootstrap*. Il metodo descritto si propone di superare i limiti derivanti dall'approccio bayesiano presentato da Choi et al. (2006) (cfr.§2.5.2.3, "*MODELLO DI CHOI ET AL. (2006)*") per la stima di intervalli di confidenza a posteriori di Δ, abbandonando la logica bayesiana basata sulla specificazione delle distribuzioni a priori dei parametri e sul campionamento di Gibbs per l'inferenza a posteriori, la quale risentiva fortemente della specificazione a priori, dell'ampiezza della sovrapposizione tra i gruppi dei malati e non malati, dalla correlazione tra i test e dall'assunzione di normalità bivariata. Esso propone invece di applicare l'algoritmo EM per trovare le stime aggiornate dei parametri e ricavare stime *bootstrap* di Δ, $\hat{\Delta}_{boot}$, ottenendo così anche la stima della varianza campionaria per la varianza di $\hat{\Delta}$, $\widehat{var}(\hat{\Delta}_{boot})$.

Gli studi di simulazione condotti da Hsieh et al. (2009) hanno mostrato che il metodo proposto funziona bene con campioni di numerosità finita. Il metodo presentato è basato sul metodo del percentile *bootstrap* (BP) (si veda l'Appendice, A.2.3.1). Obuchowsky et al. (1998) avevano stabilito l'adeguatezza di vari intervalli di confidenza *bootstrap* per l'AUC quando i risultati del test sono continui e quando le numerosità campionarie sono ridotte, ricavando che è preferibile utilizzare l'intervallo di confidenza percentile *t bootstrap* (si veda l'Appendice, A.2.3.1). Uno studio di simulazione è stato condotto da Hsieh et al. (2009) per valutare se l'uso del metodo *t bootstrap* abbia davvero prestazioni migliori del metodo presentato, generando risultati del test di soggetti malati e non malati da una distribuzione normale bivariata con parametri diversi. Essi hanno osservato che la probabilità empirica di copertura dell'intervallo di confidenza *t bootstrap* e l'intervallo di confidenza del metodo *bootstrap* proposto sono entrambi prossimi al livello di confidenza nominale del 95% sia nel caso di presenza che di assenza di un *gold standard*. In generale, sembra che questi due metodi *bootstrap* abbiano performance simili sia nel caso di disponibilità che di assenza di un *gold standard*. Quando i risultati dei test non seguono una distribuzione normale, la performance della probabilità empirica di copertura del metodo proposto basato sulla verosimiglianza è

ancora robusto, non come il metodo bayesiano proposto da Choi et al. (2006), che è sensibile all'allontanamento dall'assunzione di normalità bivariata.

Si potrebbe osservare che l'algoritmo EM usato in questo capitolo non conduce sempre a stime globali di massima verosimiglianza. Per ovviare a questo problema, Zhou et al. (2005) suggeriscono di "mescolare" casualmente i punti iniziali o ricalcolare le stime di massima verosimiglianza basate su un insieme di plausibili valori iniziali. Così, si usano differenti punti di partenza per i parametri. Si è trovato che le stime dei parametri convergono sempre agli stessi valori.

CONCLUSIONE

I metodi proposti in letteratura per la valutazione di test diagnostici in assenza di *gold standard* cercano di stimare l'accuratezza di un test diagnostico, vale a dire la capacità di un test di separare propriamente la popolazione in studio in malati e sani, che è generalmente misurata dall'area sottesa alla curva ROC (AUC). Quando il vero stato di malattia non è noto, alcuni autori (cfr. Dawid e Skene, 1979) avevano suggerito di usare l'algoritmo EM per massimizzare la verosimiglianza con dati incompleti, trattando il vero stato di malattia come dato mancante nella log-verosimiglianza completa. Lo scopo di questa tesi era quello di valutare test diagnostici in assenza di *gold standard* con riguardo a due metodi che sviluppano il concetto di vero stato di malattia come variabile latente e definiscono per la prima volta una struttura rigorosa al problema, mediante applicazione dell'algoritmo EM.

In particolare, il primo metodo affrontato stima non parametricamente curve ROC e le rispettive AUC senza *gold standard* di test su scala ordinale quando il numero di test è maggiore di due. È emerso che lo stimatore globale di massima verosimiglianza non è unico, perché, per la proprietà d'invarianza della log-verosimiglianza rispetto ad una riclassificazione dell'insieme di parametri, esiste una soluzione "speculare". Ciò significa che se $(\hat{p}_1\widehat{\boldsymbol{\eta}}_0, \widehat{\boldsymbol{\eta}}_1)$ è uno stimatore di massima verosimiglianza per $(p_1, \boldsymbol{\eta}_1, \boldsymbol{\eta}_0)$, anche $(1-\hat{p}_1\widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_0)$ è uno stimatore di massima verosimiglianza. Si può arrivare al presunto massimo globale mediante l'applicazione dell'algoritmo EM e distinguere poi tra le due possibilità ordinando ragionevolmente il tasso di prevalenza $\hat{p}_1$o $\hat{p}_0$ o considerando la plausibilità delle aree risultanti sotto le K curve ROC. Tuttavia, per ogni test credibile si potrebbe presumibilmente scegliere il massimo globale per il quale $AUC_k > 0.5$ per ogni $k = 1, \dots, K$, se ne esiste almeno uno. È emerso anche che vi sono molteplici soluzioni di stima locale di massima verosimiglianza. Questo implica che ogni test è privo di valore per determinare lo stato di malattia perché ogni risultato di qualunque test per un paziente è ugualmente probabile, indifferentemente dallo stato di malattia. Ovviamente, questo problema è sensibile alla selezione delle stime iniziali del parametro e quindi si possono scartare quelli che probabilmente sarebbe opportuno evitare. Nella ricerca del massimo globale usando l'approccio non parametrico proposto, si raccomanda di (1) evitare valori uguali della sensibilità o specificità, scelta ragionevole poiché nella pratica è spesso evidente una certa asimmetria nei punteggi del test, (2) provare un insieme di stime iniziali ragionevoli del parametro e paragonare i massimi locali di log-verosimiglianza ottenuti; (3) ottenere valori iniziali ragionevoli da studi simili con lo stato di malattia noto e infine (4) studiare la superficie di verosimiglianza usando tecniche di esplorazione e simulazione, come l'algoritmo EM stocastico. Inoltre, ad

ogni interazione dell'algoritmo EM, i parametri si aggiornano nella stessa maniera in cui aumenta la log-verosimiglianza e la somma dei quadrati dei residui rimane fissa. Esattamente come per le stime di massima verosimiglianza, essendoci un numero infinito di scelte adeguate sull'insieme di parametri, dalla somma dei quadrati non si riesce a distinguere tra le scelte possibili per un set "migliore"di parametri. Un vantaggio dell'approccio non parametrico di massima verosimiglianza proposto è che, non richiedendo specifiche assunzioni sul modello sottostante, può essere più robusto. Inoltre è molto più facile implementare il corrispondente algoritmo EM, perché al passo M è disponibile una soluzione esplicita. Gli studi di simulazione eseguiti dagli autori mostrano che le stime ricavate non parametricamente sono comparabili a quelle ottenute dai modelli parametrici. In questo modo si può evitare la verosimiglianza parametrica che, a causa dello stato latente di malattia, spesso comporta una forma di mistura così complicata da rendere difficoltosi i calcoli, i quali tendono ad essere più sensibili ai valori iniziali e meno stabili quando le celle delle tabelle di contingenza sono vicine alla degenerazione.

Il secondo metodo descrive una procedura per la stima intervallare della differenza di due AUC appaiate (Δ) in assenza di un *gold standard*, sotto l'assunzione di normalità dei risultati del test da ogni gruppo di soggetti malati, usando l'algoritmo EM congiuntamente al metodo *bootstrap.* Questo approccio si propone di superare i limiti dei metodi bayesiani (cfr. Choi et al., 2006), sensibili all'allontanamento dall'assunzione di normalità bivariata, ricavando stime aggiornate dei parametri mediante l'applicazione dell'algoritmo EM, ottenendo anche la stima *bootstrap* della varianza campionaria di $\hat{\Delta}$, $\widehat{var}(\hat{\Delta}_{boot})$. Si è dimostrato che, quando i risultati dei test non seguono una distribuzione normale, la performance della probabilità empirica di copertura del metodo proposto basato sulla verosimiglianza è ancora robusto. Gli studi di simulazione condotti da Hsieh et al. (2009) hanno mostrato anche che il metodo descritto funziona bene con campioni di numerosità finita. Inoltre, il modello evidenzia come il metodo del percentile *bootstrap* (BP) adottato abbia prestazioni simili all'intervallo *t bootstrap*, ritenuto più adeguato da Obuchowsky et al. (1998) per intervalli di confidenza *bootstrap* per l'AUC quando i risultati del test sono continui e le numerosità campionarie sono ridotte.

Si potrebbe osservare che l'algoritmo EM usato in questo capitolo non conduce sempre a stime globali di massima verosimiglianza. Per ovviare a questo problema, Zhou et al. (2005) suggeriscono di "mescolare" casualmente i punti iniziali o ricalcolare le stime di massima verosimiglianza basate su un insieme di plausibili valori iniziali. Così, si usano differenti punti di partenza per i parametri. Si è trovato che le stime dei parametri convergono sempre agli stessi valori.

Una possibile area di ricerca futura potrebbe essere lo sviluppo di metodi che rilassano l'assunzione d'indipendenza condizionata, ad esempio modelli di classe latente con effetti casuali, come proposto da Hadgu et Qu (1998).

APPENDICE

A.1 ALGORITMO EM

A.1.1 *Introduzione*

L'algoritmo EM (*Expectation-Maximization*) è un metodo iterativo ampiamente utilizzato per l'individuazione di stime parametriche di massima verosimiglianza o del massimo a posteriori (MAP) in modelli statistici, dove il modello dipende da variabili latenti non osservate o nel caso di dati incompleti.
È uno strumento che viene tendenzialmente usato quando la funzione di verosimiglianza assume forme particolarmente complicate e diventa necessario ricorrere a metodi numerici (ad esempio l'algoritmo di Newton-Raphson) che, però, possono essere molto onerosi a livello computazionale, in particolare se la funzione di verosimiglianza ha molteplici estremi o il parametro $\boldsymbol{\theta}$ è multi-dimensionale. Il successo dell'algoritmo è dovuto alla semplicità di programmazione, al pregio di porre il problema di massimizzazione in termini statistici e alla sua generalità: infatti, le situazioni in cui può essere applicato comprendono non solo i casi evidenti di dati incompleti, ma anche una grande varietà di situazioni in cui l'incompletezza dei dati non è così palese (variabili latenti, modelli log-lineari, ecc).

A.1.2 *Cenni storici*

L'algoritmo EM è stato proposto da Dempster, Lair e Rubin (1977). Essi hanno sottolineato che il metodo "è stato proposto molte volte in circostanze particolari" da autori precedenti. In particolare, un trattamento molto dettagliato del metodo EM per le famiglie esponenziali è stato pubblicato da Rolf Sundberg nel 1972 nella sua tesi e in vari documenti a seguito della sua collaborazione con la pro-Martin Löf e Anders Martin-Löf. Il lavoro di Dempster, Laird, Rubin del 1977 ha generalizzato il metodo e ha imbastito un'analisi di convergenza per una più ampia classe di problemi. Tuttavia, l'analisi di convergenza dell'articolo di Dempster-Laird-Rubin era imperfetta. Una corretta analisi di convergenza è stata pubblicata da CF Jeff Wu nel 1983; egli è andato oltre l'assunto di Dempster, Laird e Rubin, dimostrando la convergenza del metodo EM all'esterno della famiglia esponenziale.

A.1.3 *La logica dell'algoritmo EM*

L'algoritmo EM formalizza un'idea elementare per trattare i dati mancanti che consiste nel:

1. sostituire i valori mancanti con dei valori stimati;
2. stimare i parametri;
3. ri-stimare i dati mancanti, assumendo che le nuove stime dei parametri siano corrette;
4. ri-stimare i parametri, ripetendo la procedura fino alla convergenza.

Ogni iterazione dell'algoritmo EM consiste in un passo E (Expectation step) ed in un passo M (Maximization step). Il passo E trova i valori attesi condizionati dei 'dati mancanti', condizionatamente ai valori osservati e alle stime correnti dei parametri, quindi sostituisce i valori mancanti con quelli attesi. Il passo M è particolarmente semplice da descrivere: calcola le stime di massima verosimiglianza (SMV) del parametro su dati 'completati' (come se non fossero presenti dati mancanti). Quindi, il passo M sfrutta gli stessi metodi computazionali utilizzati per dati completi. 'Dati mancanti' è stato scritto tra apici in quanto l'algoritmo EM non sostituisce direttamente i valori mancanti con valori attesi trovati al passo E, ma le funzioni di variabili latenti opportunamente introdotte che compaiono nella log-verosimiglianza dei dati completi $l(\theta; \boldsymbol{x})$. È proprio per questo motivo che si ritiene che l'algoritmo EM tratti il problema di dati mancanti a livello statistico e non semplicemente numerico. Nonostante l'algoritmo sia applicabile ad una vasta classe di modelli, è particolarmente utile quando i dati completi provengono da una famiglia esponenziale: in questa situazione il passo E si riduce al calcolo del valore atteso condizionato delle statistiche sufficienti per i dati completi e il passo M è, spesso, molto semplice a livello numerico.

A.1.4 *Tre esempi introduttivi*

Si introducono tre casi per esemplificare la logica dell'algoritmo EM.

- **Esempio 1**. *Dati mancanti da distribuzione normale*

Sia $x_1, x_2, \dots, x_n$ un campione casuale semplice da una variabile iid con distribuzione $N(\mu, \sigma^2)$. La funzione di densità di probabilità è data da:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

La funzione di log-verosimiglianza per i dati completi è data da:

$$l(\mu, \sigma^2; \boldsymbol{x}) = -\frac{n}{2} log\sigma^2 - \frac{n}{2} log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2} \qquad (A.1)$$

Massimizzando la (A.1), risolvendo l'equazione di verosimiglianza rispetto a $\boldsymbol{\theta} = (\mu, \sigma^2)$, si ottiene la SMV di $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$, nel caso di dati completi, data da:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}^2.$$

Nel caso di dati mancanti, il precedente campione è così modificato:

c.c.s. $\quad x_1, x_2, \dots, x_m, \underbrace{x_{m+1}, \dots, x_n}_{\text{dati mancanti}}.$

Per la SMV dei parametri, si ha:

$$\hat{\mu} = \frac{1}{n}\left(\sum_{i=1}^{m} x_i + \boxed{\sum_{j=m+1}^{n} x_j}\right),$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^{m} x_i^2 + \boxed{\sum_{j=m+1}^{n} x_j{}^2}\right) - \hat{\mu}^2.$$

Si utilizzi l'algoritmo EM per trovare una stima di $\boldsymbol{\theta} = (\mu, \sigma^2)$ nel caso di dati incompleti, mediante l'applicazione dei seguenti due passi (E e M).

Passo E: definendo $\mu^{(t)}$ e $\sigma^{2(t)}$ come i parametri stimati all'inizio della delle t-esima iterazione, si calcoli il valore atteso rispetto alla distribuzione dei dati mancanti della log-verosimiglianza per dati completi, condizionatamente ai valori osservati di X e $\boldsymbol{\theta}^{(t)}$:

$$E_{\mu^{(t)}, \sigma^{2(t)}}\left[\sum_{j=m+1}^{n} x_j \,|\boldsymbol{x}\right] = (n-m)\mu^{(t)},$$

$$E_{\mu^{(t)}, \sigma^{2(t)}}\left[\sum_{j=m+1}^{n} x_j^2 \,|\boldsymbol{x}\right] = (n-m)\left(\mu^{(t)^2} + \sigma^{2(t)}\right).$$

Passo M: si trovi il massimo della funzione di log-verosimiglianza per dati completi, (A.1), con $\hat{\mu} = \mu^{(t)}$ e $\hat{\sigma}^2 = \sigma^{2(t)}$:

$$\mu^{(t+1)} = \frac{s_1^{(t)}}{n}, \qquad (A.2)$$

$$\sigma^{2(t+1)} = \frac{s_2^{(t)}}{n} - \mu^{(t+1)^2}. \qquad (A.3)$$

dove

$$s_1^{(t)} = \sum_{i=1}^{m} x_i + (n-m)\mu^{(t)}, \qquad s_2^{(t)} = \sum_{i=1}^{m} x_i^2 + (n-m)\left(\mu^{(t)^2} + \sigma^{2(t)}\right).$$

L'algoritmo EM itera la (A.2) e la (A.3) fino a convergenza.

- **Esempio 2.** *Attributi misti da popolazioni multinomiali*

Si definisca con $\boldsymbol{x} = (x_1, x_2, x_3, x_4)$ la determinazione di una v.c. multinomiale X con probabilità:

$$\boldsymbol{p} = (p_1, p_2, p_3, p_4) = \left(\frac{1}{2} - \frac{\theta}{2}, \frac{\theta}{4}, \frac{\theta}{4}, \frac{1}{2}\right),$$

tale che

$$\sum_{i=1}^{4} p_i = 1.$$

Il vettore dei valori osservati $\boldsymbol{y} = (y_1, y_{2,} y_3)$ corrisponde invece all'osservazione della variabile di interesse $\boldsymbol{x} = (x_1, x_2, x_3, x_4)$ con:

$$x_1 = y_1,$$

$$x_2 = y_2,$$

$$x_3 + x_4 = y_{3.}$$

Il valore mancante si può quindi identificare come la parte di $x_3 + x_4$ corrispondente a x_3 (o x_4).
L'obiettivo è trovare la SMV di θ. A tal fine, si suppone che il vettore di dati osservati $\boldsymbol{y} = (y_1, y_{2,} y_3)$ provenga da una variabile casuale con distribuzione multinomiale con probabilità di celle:

$$\boldsymbol{p} = (p_1, p_2, p_3) = \left(\frac{1}{2} - \frac{\theta}{2}, \frac{\theta}{4}, \frac{\theta}{4} + \frac{1}{2}\right).$$

Se fosse stato osservato X, la SMV di θ si sarebbe trovata massimizzando la funzione di verosimiglianza dei dati completi:

$$L(\theta;\boldsymbol{x}) = p(\boldsymbol{x};\theta) = \frac{(x_1+x_2+x_3+x_4)!}{x_1!\,x_2!\,x_3!\,x_4!}\cdot p_1^{x_1}\cdot p_2^{x_2}\cdot p_3^{x_3}\cdot p_4^{x_4} \propto \left(\frac{1}{2}-\frac{\theta}{2}\right)^{x_1}\left(\frac{\theta}{4}\right)^{x_2}\left(\frac{\theta}{4}\right)^{x_3}\left(\frac{1}{2}\right)^{x_4}.$$

Quindi la log-verosimiglianza per i dati completi sarebbe stata:

$$l(\theta;\boldsymbol{x}) = logL(\theta;\boldsymbol{x}) \propto x_1 \ln \mathrm{p}_1 + x_2 \ln \mathrm{p}_2 + x_3 \ln \mathrm{p}_3$$

$$\propto x_1 \ln\left(\frac{1}{2}-\frac{\theta}{2}\right) + x_2 \ln\left(\frac{\theta}{4}\right) + x_3 \ln\left(\frac{\theta}{4}\right) \propto x_1 \ln(1-\theta) + x_2 \ln(\theta) + x_3 \ln(\theta).$$

Risolvendo l'equazione di verosimiglianza rispetto a θ,

$$\frac{d}{d\theta} l(\theta;\boldsymbol{x}) = 0,$$

si sarebbe ottenuta la SMV di θ nel caso dei dati completi:

$$\hat{\theta} = \frac{x_2 + x_3}{x_1 + x_2 + x_3}.$$

Nel caso di dati incompleti, data la presenza di x_3 in $\hat{\theta}$, la stima non è reperibile.

Si utilizzi quindi l'algoritmo EM.

Il **passo E** comporta la sostituzione dei dati mancanti stessi con delle stime:

$$E_{\theta^{(t)}}[x_3|\boldsymbol{x}] = (x_3 + x_4)\frac{\frac{\theta^{(t)}}{4}}{\frac{1}{2}+\frac{\theta^{(t)}}{4}} = \hat{x}_3^{(t)}.$$

Il **passo M** consiste nel trovare il massimo della funzione di log-verosimiglianza per dati completi, dato dalla con $x_3 = \hat{x}_3^{(t)}$

$$\theta^{(t+1)} = \frac{x_2 + \hat{x}_3^{(t)}}{x_1 + x_2 + \hat{x}_3^{(t)}}. \qquad (A.4)$$

Un esempio numerico di questo problema è stato proposto da Dempster, Laird e Rubin (1977) nel loro articolo. Essi hanno mostrato la convergenza partendo da $\theta^{(0)} = 0.5$ e hanno visualizzato in forma tabulare le iterazioni dell'algoritmo. Inoltre hanno risolto il loro problema di stima utilizzando anche l'algoritmo di Newton-Raphson, partendo sempre da un valore iniziale $\theta^{(0)} = 0.5$. Confrontando i valori ottenuti con le iterazioni dell'algoritmo EM, si può notare che partendo dallo stesso valore iniziale, dopo solo due iterazioni l'algoritmo di Newton-Raphson è già abbastanza vicino al valore della SMV mentre con l'algoritmo EM si deve aspettare la quinta iterazione, mettendo in luce uno dei problemi dell'algoritmo EM, vale a dire la sua lentezza di convergenza.

- **Esempio 3.** *Mistura binomiale/Poisson*

Sia M una variabile casuale distribuita come una binomiale elementare $Bi(1,1-\xi)$, dove $1-\xi$ rappresenta la probabilità di successo, tale che $P(M)=1-\xi$ e $P(M^c)=\xi$, $M^c=1-M$.
Sia $X|M$ una variabile casuale con distribuzione di Poisson (λ), definita per $x=1,\dots n$. Si assuma che la probabilità che $X=0$, dato lo stato di natura M^c, sia pari a uno.

$$X|M\sim P(\lambda), \qquad P(X=x|M)=\frac{\lambda^x e^{-\lambda}}{x!}, \qquad P(X=0|M^c)=1.$$

Si indichi con n_x il numero di casi con la caratteristica $X=x$, per $x=1,\dots,m$. Inoltre si definisca con n_A il numero di casi con la caratteristica $X=0$ per M^c e con n_B numero di casi con la caratteristica $X=0$ per M. Si supponga che n_A e n_B siano dati non osservati e di avere a disposizione solo l'informazione $n_0=n_A+n_B$. Si ricavino le seguenti probabilità:

$$p_A=\xi, \qquad p_B=e^{-\lambda}(1-\xi), \qquad \text{per } x=0$$

$$p_x=\frac{\lambda^x e^{-\lambda}}{x!}(1-\xi), \qquad \text{per } x=1,2,\dots,m.$$

La distribuzione dei dati completi è data da:

$$\boldsymbol{n}=(n_A,n_B,n_1,n_2,\dots,n_m)^T,$$

mentre quella dei dati incompleti:

$$\boldsymbol{n}_{obs}=(n_0,n_1,n_2,\dots,n_m)^T,$$

dove si ricordi che $n_0=n_A+n_B$.

Se n_A e n_B fossero osservati, la SMV di $\boldsymbol{\theta}=(\xi,\lambda)$ si otterrebbe massimizzando la funzione di verosimiglianza multinomiale dei dati completi:

$$L(\xi,\lambda\,;\boldsymbol{n})=p(\boldsymbol{n};\,\xi,\lambda)=\frac{(n_A+n_B+\,n_1+\cdots+n_m)!}{n_A!\,n_B!\,n_1!\cdots n_m!}p_A^{n_A}p_B^{n_B}p_1^{n_1}\cdots p_m^{n_m}$$

$$=\frac{(n_A+n_B+\,n_1+\cdots+n_m)!}{n_A!\,n_B!\,n_1!\cdots n_m!}\xi^{n_A}\left[e^{-\lambda}(1-\xi)\right]^{n_B}\prod_{x=1}^{m}\left[\frac{\lambda^x e^{-\lambda}}{x!}(1-\xi)\right]^{n_x}.$$

Quindi la log-verosimiglianza per i dati completi è:

$$l(\xi,\lambda\,;\boldsymbol{n})\propto\ n_A\log\xi-n_B\,\lambda+n_B\log(1-\xi)+\sum_{x=1}^{m}n_x\,[-\lambda+x\,log\lambda+log(1-\xi)]. \qquad (A.5)$$

Risolvendo l'equazione di verosimiglianza rispetto $\boldsymbol{\theta} = (\xi, \lambda)$, si ottiene la stima di massima verosimiglianza nel caso di dati completi:

$$\frac{\partial}{\partial \xi}\, l(\xi, \lambda \mid \boldsymbol{n}) = \frac{n_A}{\xi} - \frac{n_B + n_1 + \cdots + n_m}{1-\xi} = \frac{n_A}{\xi} - \frac{N - n_A}{1-\xi}.$$

$$\frac{\partial}{\partial \xi}\, l(\xi, \lambda \mid \boldsymbol{n}) = 0 \rightarrow \hat{\xi} = \frac{n_A}{N},$$

dove si è posto $N = n_0 + \cdots + n_m$

$$\frac{\partial}{\partial \lambda}\, l(\xi, \lambda \mid \boldsymbol{n}) = -(n_B + n_1 + \cdots + n_m) + \frac{1}{\lambda}\sum_{x=1}^{m} x n_x = -(N - n_A) + \frac{1}{\lambda}\sum_{x=1}^{m} x n_x.$$

$$\frac{\partial}{\partial \lambda}\, l(\xi, \lambda \mid \boldsymbol{n}) = 0 \rightarrow \hat{\lambda} = \frac{\sum_{x=1}^{m} x n_x}{N - n_A}.$$

Data la presenza di n_A in $\widehat{\boldsymbol{\theta}}$, la sua stima non può essere ottenuta. Si utilizzi allora l'algoritmo EM.

Passo E: si calcoli $n_A^{(t)}$, dato n_{obs} e $\boldsymbol{\theta}^{(t)} = \left(\xi^{(t)}, \lambda^{(t)}\right)$.

$$n_A^{(t)} = E_{\xi^{(t)}, \lambda^{(t)}}[n_A | \boldsymbol{n}_{obs}] = \frac{n_0 \xi^{(t)}}{\xi^{(t)} + (1 - \xi^{(t)}) e^{-\lambda^{(t)}}}.$$

Passo M: si trovi il massimo della funzione di log-verosimiglianza per dati completi, dato dalla (A.5) con $n_A = n_A^{(t)}$:

$$\xi^{(t+1)} = \frac{n_A^{(t)}}{N}. \qquad (A.6)$$

$$\lambda^{(t+1)} = \frac{\sum_{x=1}^{m} x n_x}{N - n_A^{(t)}}. \qquad (A.7)$$

L'algoritmo EM itera la (A.6) e la (A.7) fino a convergenza.

A.1.5 Formalizzazione dell'algoritmo EM

A.1.5.*1* *La struttura principale*

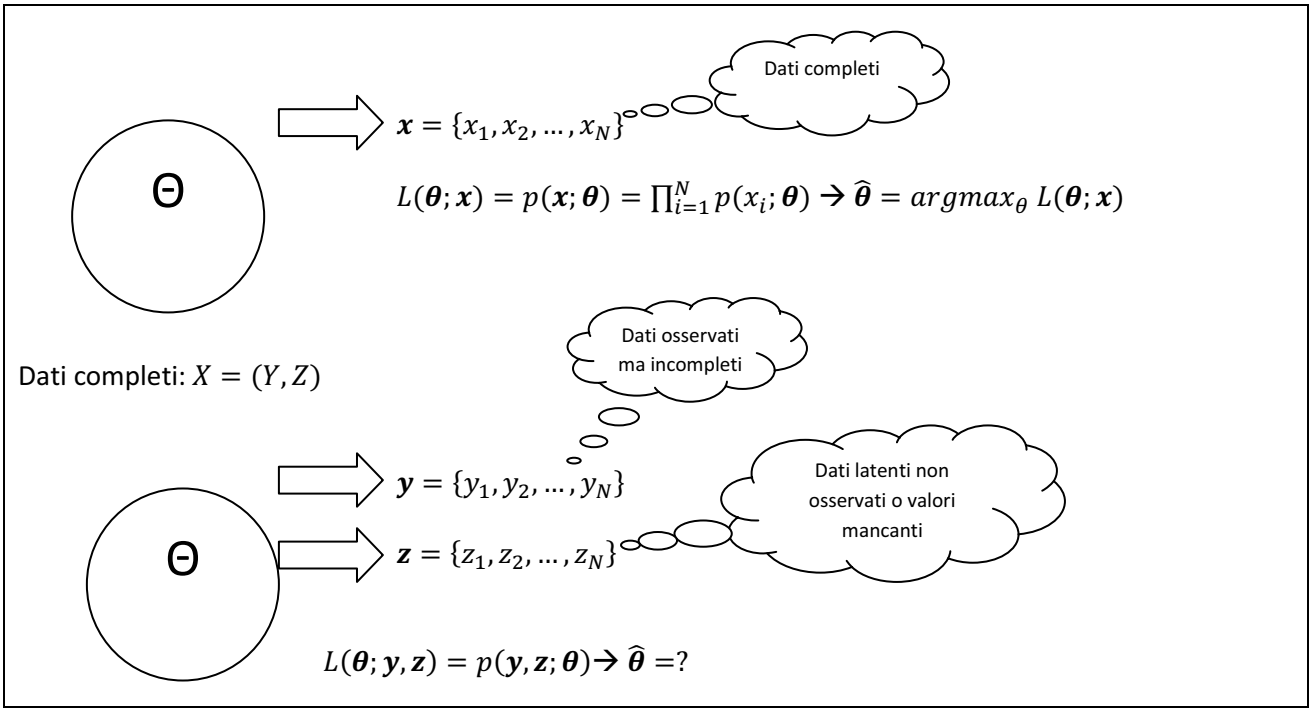
Dato un modello statistico, si definisca $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ un vettore casuale che corrisponde ad un insieme di n osservazioni da altrettante variabili casuali con funzione di densità di probabilità $f(\boldsymbol{y}; \boldsymbol{\theta})$, dove $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ è un vettore di parametri non noti. Si indichi con $\boldsymbol{z} = \{z_1, z_2, \ldots, z_n\}$ un insieme di n valori

di una variabile latente Z e con $\boldsymbol{x}$ un vettore di dati completi, $\boldsymbol{x} = [\boldsymbol{y}, \boldsymbol{z}]$, dove $\boldsymbol{z}$ sono i dati "aggiuntivi". Si denoti con $f(\boldsymbol{x}; \boldsymbol{\theta})$ la funzione di densità di probabilità del vettore casuale che corrisponde all'insieme di dati completi $\boldsymbol{x}$. La log-verosimiglianza per $\boldsymbol{\theta}$, se $\boldsymbol{x}$ fosse pienamente osservato, sarebbe

$$l(\boldsymbol{\theta}; \boldsymbol{x}) = logL(\boldsymbol{\theta}; \boldsymbol{x}) = logp(\boldsymbol{x}; \boldsymbol{\theta}).$$

La struttura logica del modello appena decritto è rappresentata nella Figura A.1. Il vettore di dati incompleti $\boldsymbol{y}$ proviene dallo spazio campionario "incompleto" $\mathcal{Y}$. C'è una corrispondenza 1 a 1 tra lo spazio campionario completo $\mathcal{X}$ e lo spazio campionario incompleto $\mathcal{Y}$- vale a dire che $x_{(i)}$ corrisponde a $y_{(i)}$. Così, per $\boldsymbol{x} \in \mathcal{X}$, si può trovare unicamente l'"incompleto" $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}) \in \mathcal{Y}$.

Figura A.1. Schema del modello statistico considerato: dati completi, osservati e latenti



La verosimiglianza con i dati completi può essere fattorizzata in

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}) = p(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta}) = p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})p(\boldsymbol{y}; \boldsymbol{\theta}).$$

La log-verosimiglianza diviene quindi

$$l(\boldsymbol{\theta}; \boldsymbol{x}) = l(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}) = l(\boldsymbol{\theta}; \boldsymbol{y}) + log[p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})], \qquad (A.8)$$

dove $l(\boldsymbol{\theta}; \boldsymbol{x})$ è la verosimiglianza dei dati completi, $l(\boldsymbol{\theta}; \boldsymbol{y})$ è la verosimiglianza dei dati osservati, mentre l'ultimo termine $log[p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})]$ è il logaritmo della funzione di densità dei dati latenti (o valori mancanti), dati i valori osservati e $\boldsymbol{\theta}$. L'obiettivo è stimare $\boldsymbol{\theta}$ massimizzando la verosimiglianza dei dati osservati rispetto a $\boldsymbol{\theta}$, per $\boldsymbol{y}$ fissato. Infatti, quando vi sono dati mancanti la funzione di verosimiglianza che viene

massimizzata è quella dei dati osservati. Però, come è già stato detto, questa procedura può essere molto laboriosa. Si cerca così di semplificare il problema, 'accontentandosi' di calcolare il valore atteso della log-verosimiglianza rispetto alla distribuzione condizionata di Z dato Y e $\boldsymbol{\theta}$. Si scriva la (A.8) nel seguente modo:

$$l(\boldsymbol{\theta};\boldsymbol{y}) = l(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z}) - \ log[p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})]. \qquad (A.9)$$

Il valore atteso della (A.9) rispetto alla distribuzione di Z dato Y e $\boldsymbol{\theta}$ è:

$$E[l(\boldsymbol{\theta};\boldsymbol{y})] = Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big) - H\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big).$$

dove, indicando con $\boldsymbol{\theta}^{(t)}$ il vettore dei parametri ottenuto al t-esimo passo, $Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big)$ è il valore atteso della log-verosimiglianza completa ed è data da (caso continuo):

$$Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big) = E\big[l(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z})|\boldsymbol{y},\boldsymbol{\theta}^{(t)}\big] = \int_{z\in Z} [l(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z})] \cdot p\big(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}\big)\, dz,$$

e $H\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big)$ è:

$$H\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big) = \int_{z\in Z} log[p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})] \cdot p\big(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}\big)\, dz.$$

Se la variabile Z è discreta, l'integrazione è sostituita da una somma.

Il termine $H\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big)$ è detto "entropia" della distribuzione $p\big(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}\big)$.
Considerando che $E[l(\boldsymbol{\theta};\boldsymbol{y})]$ rispetto a Z è $l(\boldsymbol{\theta};\boldsymbol{y})$ si ha:

$$l(\boldsymbol{\theta};\boldsymbol{y}) = \ Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big) - H\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big).$$

Se $\boldsymbol{\theta}^{(t)}$ è SMV per $l(\boldsymbol{\theta};\boldsymbol{y})$, anche $H\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big)$ è massimizzato quando $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ in quanto, per la disuguaglianza di Jensen, si ha che $H\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big) \leq H\big(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}\big)$.
Di conseguenza, anche $Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big)$ è massimizzato quando $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.
Di qui abbiamo l'algoritmo EM: poiché $l(\boldsymbol{\theta};\boldsymbol{y})$ è difficile da massimizzare, si preferisce massimizzare il valore atteso condizionato della log-verosimiglianza per dati completi $Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big)$. La massimizzazione di $Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big)$ assicura, per quanto detto sopra, la massimizzazione anche di $l(\boldsymbol{\theta};\boldsymbol{y})$. Partendo da una stima iniziale $\boldsymbol{\theta}^{(0)}$ e detta $\boldsymbol{\theta}^{(t)}$ la stima corrente di $\boldsymbol{\theta}$, i due passi dell'algoritmo EM sono:

Passo E: calcola il valore atteso $Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big)$ rispetto alla distribuzione di Z della log-verosimiglianza per dati completi, dato Y e $\boldsymbol{\theta}$;

Passo M: calcola $\boldsymbol{\theta}^{(t+1)}$ massimizzando $Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big)$:

$$\boldsymbol{\theta}^{(t+1)} = argmax_{\theta} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right),$$

vale a dire, cerca il valore $\boldsymbol{\theta}^{(t+1)}$ tale che

$$Q\left(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}\right) \geq Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) \quad \forall \boldsymbol{\theta}.$$

Cominciando da un valore iniziale dei parametri, $\boldsymbol{\theta}^{(0)}$, inizia il ciclo tra i passi E e M finchè $\boldsymbol{\theta}^{(t)}$ converge ad un massimo locale, vale a dire finché la differenza

$$L\left(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{x}\right) - L\left(\boldsymbol{\theta}^{(k)}; \boldsymbol{x}\right)$$

diventa piccola in valore assoluto.

A.1.5.2 *Esempio: caso discreto*

La funzione di log-verosimiglianza per i dati osservati nel caso discreto è:

$$l(\boldsymbol{\theta}) = log\, p(\boldsymbol{y}; \boldsymbol{\theta}) = log \sum_{Z} p(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta}).$$

Si deve stimare non solo $\boldsymbol{\theta}$ ma anche Z. Si può massimizzare direttamente $l(\boldsymbol{\theta}) = log \sum_z p\,(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})$ usando un metodo gradiente ma spesso è difficile da implementare. Per ovviare a tale problema, si definisca allora con $p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})$ la distribuzione di probabilità sui dati mancanti Z. Si consideri la seguente disuguaglianza:

$$l(\boldsymbol{\theta}) = log p(\boldsymbol{y}; \boldsymbol{\theta}) = log \sum_{Z} p(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})$$

$$= log \sum_{Z} p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}) \frac{p(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})}$$

$$\geq \sum_{Z} p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}) log \frac{p(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})}$$

$$= \sum_{Z} p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})\, l(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}) + \sum_{Z} p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}) log \frac{1}{p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})} \equiv F(p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}), \boldsymbol{\theta}), \qquad (A.10)$$

dove $F(p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}), \boldsymbol{\theta})$ indica una funzione a due variabili, $p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})$ e $\boldsymbol{\theta}$. La disuguaglianza sussiste per la disuguaglianza di Jensen. Questo significa che $F(p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}), \boldsymbol{\theta})$ è un limite inferiore su $l(\boldsymbol{\theta})$. Invece di massimizzare $l(\boldsymbol{\theta})$ direttamente, EM massimizza il limite inferiore $F(p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}), \boldsymbol{\theta})$. L'algoritmo EM si alterna tra la massimizzazione di $F(\cdot)$ con rispetto a $p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})$ ($\boldsymbol{\theta}$ fissato) e la massimizzazione di $F(\cdot)$ con rispetto a $\boldsymbol{\theta}$ ($p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})$ fissato). Inoltre, applicando una proprietà dei logaritmi, si può esprimere $F(p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}), \boldsymbol{\theta})$ nel seguente modo:

$$F(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}),\boldsymbol{\theta}) = \sum_{Z} p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})\, l(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z}) - \sum_{Z} p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}) log p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})$$

$$= Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big) - H\big(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})\big), \qquad (A.11)$$

dove si è posto

$$Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big) = \sum_{Z} p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}) l(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z}),$$

$$H\big(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})\big) = \sum_{Z} p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}) log p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}).$$

E-step: $p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})^{(t)} = argmax_{p(z|y,\theta)} F\big(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}),\boldsymbol{\theta}^{(t)}\big).$ $(A.12)$

Calcolare l'equazione (A.12) direttamente implica fissare $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ e ottimizzare sullo spazio di distribuzioni. Tuttavia, posto $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$, $p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})$ non è più una variabile sullo spazio di distribuzioni, poiché l'ottimo di $p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})$ è una distribuzione che dipende da $\boldsymbol{\theta}^{(t)}$.
Questo significa che $l\big(\boldsymbol{\theta}^{(t)}\big) = F\big(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}),\boldsymbol{\theta}^{(t)}\big)$, quindi si dimosta che $p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})^{(t)} = p\big(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}\big)$.

Inoltre, per la (A.11) il passo E equivale a calcolare il seguente valore atteso:

$$Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big) = E_{p(z|y,\theta^{(t)})}[l(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z})] = \sum_{z\in Z} [l(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z})] p\big(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}\big).$$

M-step: $\boldsymbol{\theta}^{(t+1)} = argmax_{\theta} F\big(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})^{(t)},\boldsymbol{\theta}\big).$

Il passo M si riduce a massimizzare il primo termine della somma in (A.10) con rispetto a $\boldsymbol{\theta}$ dato che non c'è $\boldsymbol{\theta}$ nel secondo termine. Così massimizzare $F(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}),\boldsymbol{\theta})$ è equivalente a massimizzare la log-verosimiglianza completa attesa, per cui si ottiene:

$$\boldsymbol{\theta}^{(t+1)} = argmax_{\theta} E_{p(z|y,\theta^{(t)})}[l(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z})] = argmax_{\theta} \sum_{z\in Z} [l(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z})] p\big(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}\big).$$

L'algoritmo EM alterna questi due passa fino a convergenza.

A.1.6 Convergenza dell'algoritmo EM

Grazie al lavoro di Dempster, Laird e Rubin (1977) si ha la sicurezza che ad ogni iterazione dell'algorito EM la log-verosimiglianza è non decrescente. Infatti, si consideri una sequenza di iterazioni

$\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(t)}$, …dove $\boldsymbol{\theta}^{(t+1)} = M\left(\boldsymbol{\theta}^{(t)}\right)$ per qualche funzione $M(\cdot)$. Si riporta di seguito l'enunciato del teorema che afferma che ogni iterazione aumenta o lascia invariata la verosimiglianza (Little e Rubin, 2002).

Teorema 1: *Ad ogni iterazione un algoritmo EM aumenta la log-verosimiglianza* $l(\boldsymbol{\theta}|\boldsymbol{y})$*, cioè* $l\left(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}\right) \geq l\left(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}\right)$. *Vale l'uguaglianza se e solo se* $Q\left(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}\right) = Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)$.
Infatti,

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) - H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right).$$

Da qui

$$l\left(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}\right) - l\left(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}\right) = \left[Q\left(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}\right) - Q\left(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}\right)\right] + \left[H\left(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}\right) - H\left(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}\right)\right].$$

Il primo termine è non negativo per la definizione di $\boldsymbol{\theta}^{(t+1)}$ come il massimizzatore di $Q\left(\cdot\,|\boldsymbol{\theta}^{(t)}\right)$.
Il secondo termine prende la forma:

$$H\left(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}\right) - H\left(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}\right) = \sum_{z \in Z} p\left(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}\right) log \frac{p\left(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}\right)}{p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}^{(t+1)})} = D\left(p_{Z|Y=y,\theta^{(t)}} || p_{Z|Y=y,\theta^{(t+1)}}\right).$$

che è anche non negativo; quindi anche la sequenza di log-verosimiglianza $l\left(\theta^{(t)}; y\right)$ è non decrescente.

Corollario. La sequenza $l\left(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}\right)$ converge se la funzione di verosimiglianza è limitata.

Inoltre, Dempster, Laird e Rubin (1977) dimostrano che, se $\theta^{(t)}$ converge, allora converge ad un punto stazionario. Questo avviene solo sotto alcune restrizioni applicate alla funzione di densità (quali l'appartenenza ad una famiglia esponenziale regolare e $l(\boldsymbol{\theta}; \boldsymbol{y})$ limitata), ma non dovrebbe stupire in quanto nessun algoritmo iterativo assicura la convergenza ad un punto stazionario. È necessario specificare tuttavia che, quando la log-verosimiglianza ha diversi punti stazionari, la convergenza dell'algoritmo EM dipende dalla scelta del valore iniziale. Per questo motivo si raccomanda di prevedere diverse iterazioni dell'algoritmo da più punti iniziali. In ogni caso, nella maggioranza dei problemi pratici rilevanti, si è visto che l'algoritmo EM converge quasi sempre ad un massimo locale.
L'algoritmo EM potrebbe essere formulato come un algoritmo di minimizzazione alternativo, come studiato da Csiszàr e Tusnady (1984). Questa interpretazione porta ad una famiglia di varianti dell'algoritmo EM che in alcuni casi sono più semplici da implementare e hanno buone proprietà di convergenza.

A.1.7 L'algoritmo EM per famiglie esponenziali

L'algoritmo EM assume una forma particolarmente semplice quando i dati completi da una v.c. X hanno una distribuzione appartenente alla famiglia esponenziale regolare, vale a dire se è esprimibile come:

$$p(x_i|z_i,\boldsymbol{\theta}) = c(\boldsymbol{\theta})h(x_i)\exp\left\{\sum_{j=1}^{k}\psi_j(\boldsymbol{\theta})t_j(x_i)\right\}.$$

dove θ è parametro ignoto, $c(\theta)$, $h(\cdot)$ e $\psi(\cdot)$ sono funzioni note la cui scelta individua una particolare distribuzione, $\boldsymbol{t} = \left[t_j\right] = \left[\sum_{i=1}^{n} t_j(x_i)\right], j = 1,\dots,k$, è statistica sufficiente per l'inferenza su $\boldsymbol{\theta}$ per i dati completi, $\boldsymbol{\theta}$ è un vettore di parametri. Il passo E è relativamente semplice da implementare quando X condizionato a Z proviene da una famiglia esponenziale, poichè

$$Q\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)}\right) = E\left[lnp(\boldsymbol{x};\boldsymbol{\theta})|Z,\widehat{\boldsymbol{\theta}}^{(t)}\right] = c'(\boldsymbol{\theta})h'(x_i)\sum_{j=1}^{k}\psi_j(\boldsymbol{\theta})E\left[t_j(x_i)|Z,\widehat{\boldsymbol{\theta}}^{(t)}\right].$$

Passo E. Si riduce alla stima della statistica sufficiente per dati completi $\boldsymbol{t}(\boldsymbol{x})$ con:

$$\boldsymbol{t}^{(t+1)} = E\left[\boldsymbol{t}(X)|Y,\boldsymbol{\theta}^{(t)}\right].$$

Cioè, il problema è ridotto a valutare il valore atteso condizionato di k statistiche.

Passo M. Il passo M determina la nuova stima di $\boldsymbol{\theta}^{(t+1)}$ di $\boldsymbol{\theta}$ risolvendo le equazioni di verosimiglianza

$$E\left[\boldsymbol{t}(X)|\boldsymbol{\theta}\right] = \boldsymbol{t}^{(t+1)},$$

Che sono semplicemente le equazioni di verosimiglianza per dati completi con $\boldsymbol{t}(\boldsymbol{x})$ sostituito da $\boldsymbol{t}^{(t+1)}$.

A.1.8 Modello di mistura

A.1.8.1 *Stima dei pesi di distribuzioni pienamente note*

Se un insieme di dati è composto da parecchie popolazioni distinte, può essere usato un modello di mistura. Stimare le misture di distribuzioni è un'importante questione statistica.

Si supponga di voler stimare i pesi di un numero fisso di distribuzioni pienamente note, illustrando l'approccio EM che introduce indicatori non osservati con l'obiettivo di semplificare la verosimiglianza. I pesi sono stimati con il metodo di massima verosimiglianza. Si assuma che un campione $x_1, x_{2,\dots,} x_n$ provenga dalla seguente mistura:

$$p(\boldsymbol{x};\boldsymbol{\omega}) = \sum_{j=1}^{k}\omega_j p_j(\boldsymbol{x}),$$

dove i pesi $0 \leq \omega_j \leq 1$ sono ignoti e costituiscono un vettore $(k-1)$-dimensionale: $\boldsymbol{\omega} = (\omega_1, \dots, \omega_{k-1})$, dato che $\omega_k = 1 - \omega_1 - \cdots - \omega_{k-1}$. Le densità della classe $p_j(\boldsymbol{x})$ sono pienamente specificate. Anche in questo caso più semplice, quando solo i pesi $\boldsymbol{\omega}$ sono i parametri da stimare, la log-verosimiglianza assume una forma abbastanza complicata:

$$logL_c(\boldsymbol{\omega}; \boldsymbol{x}) = \sum_{i=1}^{n} log\, p(x_i; \boldsymbol{\omega}) = \sum_{i=1}^{n} log\left(\sum_{j=1}^{k} \omega_j p_j(x_i)\right).$$

Le derivate rispetto a ω_j conducono al sistema di equazioni, non risolvibile in una forma chiusa.

Qui interviene l'EM. Si aumentino i dati completi $\boldsymbol{x} = (x_1, \dots, x_n)$ con una matrice non osservabile $\boldsymbol{z}$, con elemento z_{ij}, per $i = 1, \dots, n$ e $j = 1, \dots, k$. I valori z_{ij} sono variabili indicatrici, dove $z_{ij} = 1$ se l'osservazione x_i deriva dalla distribuzione p_j e $z_{ij} = 0$ altrimenti.

La matrice non osservabile $\boldsymbol{z}$ indica da dove proviene l'i-esima osservazione x_i. Si noti che ogni riga di $\boldsymbol{z}$ contiene solo un 1 e $(k-1)$ zeri. Con i dati completi, $\boldsymbol{x} = [\boldsymbol{y}, \boldsymbol{z}]$ la verosimiglianza (completa) assume una forma abbastanza semplice:

$$L_c(\boldsymbol{\omega}; \boldsymbol{x}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \big(\omega_j p_j(x_i)\big)^{z_{ij}}.$$

La log-verosimiglianza completa è

$$l_c(\boldsymbol{\omega}; \boldsymbol{x}) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} log\, \omega_j + c,$$

dove $c = \sum_i \sum_j z_{ij} log\, p_j(x_i)$ non dipende da $\boldsymbol{\omega}$.

Si assuma che si sia ottenuta la t-esima stima parziale dei pesi, $\boldsymbol{\omega}^{(t)}$.

Il t-esimo **passo E** è:

$$E_{\omega^{(t)}}\big(z_{ij}|\boldsymbol{x}\big) = P_{\omega^{(t)}}\big(z_{ij} = 1|\boldsymbol{x}\big) = z_{ij}^{(t)},$$

dove $z_{ij}^{(t)}$ è la probabilità a posteriori dell'i-esima osservazione che proviene dalla j-esima componente di mistura, p_j, nel passo iterativo t.

$$z_{ij}^{(t)} = \frac{\omega_j^{(t)} p_j(x_i)}{p(x_i; \boldsymbol{\omega}^{(t)})}.$$

Poiché $l_c(\boldsymbol{\omega}; \boldsymbol{x})$ è lineare nelle z_{ij}, $Q\big(\boldsymbol{\omega}|\boldsymbol{\omega}^{(t)}\big)$ è semplicemente $\sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij}^{(t)} log\, \boldsymbol{\omega}_j + c$.

Il successivo **passo M** consiste nel calcolare $\omega_j^{(t+1)}$ massimizzando $Q\big(\boldsymbol{\omega}|\boldsymbol{\omega}^{(t)}\big)$. Si ottiene:

$$\omega_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t)}}{n}.$$

A.1.8.2 *Modello di mistura a due componenti binomiali*

Si abbiano due monete con probabilità non nota di ottenere da un lancio il risultato "testa", denotata p e q rispettivamente. La prima moneta è scelta con probabilità π e la seconda con probabilità $1-\pi$. Si lanci una volta la moneta e si registri il risultato. Si definisca $\boldsymbol{x} = (x_1, \dots, x_n)$ il vettore "0/1" dei risultati, dove "testa" $= 1$ e "croce" $= 0$. Si indichi $Z_i \in \{0,1\}$ la variabile che denota quale moneta è stata usata in ogni lancio. I parametri che si vogliono stimare sono $\boldsymbol{\theta} = (p, q, \pi)$. Un criterio per la stima puntuale è la massima verosimiglianza:

$$\widehat{\boldsymbol{\theta}} = argmax_{\theta} \log(x; \boldsymbol{\theta}).$$

Si voglia ottenere la stima di massima verosimiglianza utilizzando l'algoritmo EM.

La log-verosimiglianza dei dati completi è:

$$l(\boldsymbol{\theta}; \boldsymbol{x}) = l(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}) = l(\boldsymbol{\theta}; \boldsymbol{y}) + log[p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})].$$

Da cui:

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = l(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}) - \ log[p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})]. \qquad (A.13)$$

Si calcoli poi il valore atteso della (A.13) rispetto alla distribuzione di Z dato Y e $\boldsymbol{\theta}$:

$$E[l(\boldsymbol{\theta}; \boldsymbol{y})] = Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big) - H\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big).$$

dove $Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big) = \sum_{z \in Z} [l(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z})] \cdot p\big(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}\big)$.

Passo E. Si calcoli il valore atteso $Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big)$.

$$Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big) = E_{p\left(z|y,\boldsymbol{\theta}^{(t)}\right)}[log\, p(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})]$$

$$= E\left[log \prod_{i=1}^{n} [\pi p^{y_i}(1-p)^{1-y_i}]^{z_i}[(1-\pi)q^{y_i}(1-q)^{1-y_i}]^{1-z_i}\right]$$

$$= \sum_{i=1}^{n} \Big\{E\big[z_i|y_i, \boldsymbol{\theta}^{(t)}\big][log\pi + y_i logp + (1-y_i)\log(1-p)]$$

$$+ \big(1 - E\big[z_i|y_i, \boldsymbol{\theta}^{(t)}\big]\big)[\log(1-\pi) + y_i logq + (1-y_i)\log(1-q)]\Big\}.$$

Si calcoli $E\left[z_i|y_i,\boldsymbol{\theta}^{(t)}\right]$:

$$\mu_i^{(t)} = E\left[z_i|y_i,\boldsymbol{\theta}^{(t)}\right] = p\left(z_i = 1|y_i,\boldsymbol{\theta}^{(t)}\right) = \frac{p\left(y_i|z_i,\boldsymbol{\theta}^{(t)}\right)p\left(z_i = 1|\boldsymbol{\theta}^{(t)}\right)}{p\left(y_i|\boldsymbol{\theta}^{(t)}\right)}$$

$$= \frac{\pi^{(t)}\left[p^{(t)}\right]^{y_i}\left[\left(1-p^{(t)}\right)\right]^{1-y_i}}{\pi^{(t)}[p^{(t)}]^{y_i}[(1-p^{(t)})]^{1-y_i} + (1-\pi^{(t)})[q^{(t)}]^{y_i}[(1-q^{(t)})]^{1-y_i}}.$$

Passo M. Massimizzando $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)$ rispetto a $\boldsymbol{\theta}$, si ottengono equazioni aggiornate.

$$\frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)}{\partial \pi} = 0 \rightarrow\ \pi^{(t+1)} = \frac{1}{n}\sum_i \mu_i^{(t)},$$

$$\frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)}{\partial p} = 0 \rightarrow\ p^{(t+1)} = \frac{\sum_i \mu_i^{(t)} y_i}{\sum_i \mu_i^{(t)}},$$

$$\frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)}{\partial q} = 0 \rightarrow\ q^{(t+1)} = \frac{\sum_i \left(1-\mu_i^{(t)}\right) y_i}{\sum_i \left(1-\mu_i^{(t)}\right)}.$$

A.1.9 Modello di mistura gaussiana (GMM)

A.1.9.1 *Stima dei pesi, medie e varianze in un modello di mistura a due componenti gaussiane d-dimensionali*

Si intende ora analizzare l'applicazione dell'algoritmo EM su un modello di mistura a due componenti gaussiane. Sia $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ un campione di osservazioni indipendenti da una mistura di due distribuzioni normali multivariate di dimensione n, e sia $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_n)$ il vettore delle variabili latenti che determinano la componente da cui proviene l'osservazione.
La distribuzione di $X_i|Z_i$ è normale d-variata:

$$X_i|(Z_i = 1) \sim N_d(\mu_1, \Sigma_1) \quad \text{e} \quad X_i|(Z_i = 2) \sim N_d(\mu_2, \Sigma_2).$$

dove

$P(Z_i = 1) = \tau_1$ e $P(Z_i = 2) = \tau_2 = 1 - \tau_1$.

L'obiettivo è stimare il vettore dei parametri non noti $\boldsymbol{\theta} = (\tau_1, \tau_2, \mu_1, \mu_2, \Sigma_1, \Sigma_2)$.

La funzione di verosimiglianza è:

$$L(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{n} \sum_{j=1}^{2} I(z_i = j)\, \tau_j\, f\left(x_i; \mu_j, \Sigma_j\right),$$

dove I è una funzione indicatrice e $f(\cdot)$ è la funzione densità di probabilità di una normale multivariata. Questo può essere riscritto in forma di famiglia esponenziale:

$$L(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{z}) = \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{2} I(z_i = j)\left[log\tau_j - \frac{1}{2}\log\left|\Sigma_j\right| - \frac{1}{2}\left(x_i - \mu_j\right)^T \Sigma_j^{-1}\left(x_i - \mu_j\right) - \frac{n}{2} log(2\pi)\right]\right\}.$$

Passo E. Data la stima attuale del parametri $\boldsymbol{\theta}^{(t)}$, la distribuzione condizionata di Z_i è determinata dal teorema di Bayes*:*

$$T_{j,i}^{(t)} = P\left(Z_i = j | X_i = x_i; \boldsymbol{\theta}^{(t)}\right) = \frac{\tau_j^{(t)} f\left(x_i; \mu_j^{(t)}, \Sigma_j^{(t)}\right)}{\tau_1^{(t)} f\left(x_i; \mu_1^{(t)}, \Sigma_1^{(t)}\right) + \tau_2^{(t)} f\left(x_i; \mu_2^{(t)}, \Sigma_2^{(t)}\right)}.$$

Quindi, il passo E è dato dal calcolo del seguente valore atteso:

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) = E[logL(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{Z})] = \sum_{i=1}^{n}\sum_{j=1}^{2} T_{j,i}^{(t)}\left[log\tau_j - \frac{1}{2}\log\left|\Sigma_j\right| - \frac{1}{2}\left(x_i - \mu_j\right)^T \Sigma_j^{-1}\left(x_i - \mu_j\right) - \frac{n}{2} log(2\pi)\right].$$

Passo M. La forma quadratica $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)$ significa che è relativamente semplice determinare la massimizzazione dei valori di $\boldsymbol{\theta}$. In primo luogo si noti che $\boldsymbol{\tau} = (\tau_1, \tau_2)$, (μ_1, Σ_1) e (μ_2, Σ_2) possono essere tutti massimizzati indipendentemente l'uno dall'altro in quanto appaiono separati in termini lineari.
In secondo luogo si consideri $\boldsymbol{\tau}$*,* che ha il vincolo $\tau_1 + \tau_2 = 1$:

$$\tau^{(t+1)} = argmax_\tau\, Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) = argmax_\tau \left\{\left[\sum_{i=1}^{n} T_{1,i}^{(t)}\right] log\tau_1 + \left[\sum_{i=1}^{n} T_{2,i}^{(t)}\right] log\tau_2\right\}.$$

Questo ha la stessa forma della SMV per la distribuzione binomiale, quindi:

$$\tau_j^{(t+1)} = \frac{\sum_{i=1}^{n} T_{j,i}^{(t)}}{\sum_{i=1}^{n}\left(T_{1,i}^{(t)} + T_{2,i}^{(t)}\right)} = \frac{1}{n}\sum_{i=1}^{n} T_{j,i}^{(t)}.$$

Per le ulteriori stime (μ_1, Σ_1):

$$\left(\mu_1^{(t+1)}, \Sigma_1^{(t+1)}\right) = argmax_{\mu_1,\Sigma_1} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)$$
$$= argmax_{\mu_1,\Sigma_1} \sum_{i=1}^{n} T_{1,i}^{(t)} \left\{-\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x_i - \mu_1)^T \Sigma_1^{-1}(x_i - \mu_1)\right\}.$$

Avendo la stessa forma della SMV ponderata per una distribuzione normale, si ricava:

$$\mu_1^{(t+1)} = \frac{\sum_{i=1}^{n} T_{1,i}^{(t)} x_i}{\sum_{i=1}^{n} T_{1,i}^{(t)}},$$

$$\Sigma_1^{(t+1)} = \frac{\sum_{i=1}^{n} T_{1,i}^{(t)} \left(x_i - \mu_1^{(t+1)}\right)\left(x_i - \mu_1^{(t+1)}\right)^T}{\sum_{i=1}^{n} T_{1,i}^{(t)}}.$$

e, per simmetria:

$$\mu_2^{(t+1)} = \frac{\sum_{i=1}^{n} T_{2,i}^{(t)} x_i}{\sum_{i=1}^{n} T_{2,i}^{(t)}},$$

$$\Sigma_2^{(t+1)} = \frac{\sum_{i=1}^{n} T_{2,i}^{(t)} \left(x_i - \mu_2^{(t+1)}\right)\left(x_i - \mu_2^{(t+1)}\right)^T}{\sum_{i=1}^{n} T_{2,i}^{(t)}}.$$

A.1.10 Pregi e difetti dell'algoritmo EM

Si è visto che l'algoritmo EM ha diversi pregi, ma anche alcuni difetti. Riassumendo, si possono elencare tra i pregi:

- non prevede il calcolo e l'inversione di matrici d'informazione (usate nell'algoritmo di Newton-Raphson);
- è facile da costruire poiché il passo E e il passo M sono basati su calcoli compiuti sui dati completi;
- è di facile implementazione;
- è concettualmente semplice, in quanto pone il problema di massimizzazione della funzione di verosimiglianza in presenza di dati mancanti in termini statistici: il passo E completa i dati mentre il passo M calcola la stima di massima verosimiglianza sui dati completi;
- ad ogni interazione aumenta la log-verosimiglianza. Inoltre, nella maggioranza dei problemi pratici converge ad un massimo locale.

Tra i difetti si possono invece riportare:

- la soluzione generalmente dipende dal valore iniziale.
- è conveniente solo quando il passo E può essere calcolato direttamente, per questo viene usato frequentemente con variabili appartenenti alla famiglia esponenziale;
- il tasso di convergenza può essere molto lento, soprattutto se vi sono molti dati mancanti: nei primi passi è tipicamente abbastanza buono, ma può diventare molto lento quando si avvicina all'ottimo locale;
- c'è una stretta relazione tra il tasso di convergenza dell'algoritmo EM e l'informazione di Fisher. Generalmente EM lavora meglio quando la frazione di informazione mancante è piccola (che può essere quantificata usando l'informazione di Fisher) e la dimensione dei dati non è troppo ampia. EM può richiedere molte iterazioni: un'alta dimensione può rallentare drammaticamente il passo E.
- la convergenza ad una SMV (cioè ad un massimo globale) non è sempre garantita. La convergenza della sequenza $l\left(\widehat{\boldsymbol{\theta}}^{(k)}\right)$ non garantisce la convergenza di $\widehat{\boldsymbol{\theta}}^{(k)}$, vale a dire che l'algoritmo EM potrebbe saltare avanti e indietro tra due attrattori. Anche se l'algoritmo ha un punto stabile, non c'è in generale garanzia che questo sia un massimo globale o locale della funzione di verosimiglianza. Se la funzione $Q(\boldsymbol{\theta}|\boldsymbol{\theta})$ è continua, è garantita la convergenza verso un punto stazionario della verosimiglianza. Per molti problemi, la convergenza ad un massimo locale deve essere dimostrata.
- l'algoritmo EM non fornisce automaticamente gli errori standard delle stime. Bisogna infatti ricorrere all'algoritmo SEM o al metodo di Louis.
- sono state sviluppate parecchie varianti dell'algoritmo EM. Uno tra i migliori è l'algoritmo di Fessler e Hero (space-alternating generalized EM) il cui tasso di convergenza è spesso molto superiore di quelli dell'algoritmo EM.

A.1.11 Varianti

La presentazione teorica dell'algoritmo è ora conclusa. Si noti che in alcuni casi esistono alternative migliori dell'algoritmo EM. Per esempio, in alcuni problemi, la funzione di log-verosimiglianza è concava e l'insieme realizzabile per il parametro $\boldsymbol{\theta}$ è convesso. Allora possono essere sviluppati efficienti algoritmi di ottimizzazione convessa che in genere convergono più velocemente dell'algoritmo EM. L'algoritmo EM ha moltissime versioni, tra cui le più note sono: l'algoritmo SEM (Supplemented EM), che fornisce l matrice di varianza e covarianza delle stime, l'algoritmo ECM (Expectation/Conditional Maximization), che semplifica il passo M quando la massimizzazione non è diretta, l'algoritmo MCEM (Monte Carlo EM), che cerca di valutare numericamente il passo E quando questo è difficile da calcolare. In particolare, l'algoritmo ECM

sostituisce ogni passo M con una sequenza di passi di massimizzazione condizionata (CM) in cui ogni parametro θ_i, è massimizzato singolarmente, condizionatamente agli altri parametri restanti fissi. Questo concetto è ulteriormente esteso nell'algoritmo GEM (Generalized Expectation Maximization), in cui si cerca solo un aumento della funzione obiettivo $F(\cdot)$ sia per il passo E che M, sotto la descrizione alternativa. L'algoritmo MCEM cerca di risolvere la difficoltà dell'algoritmo EM generale di trovare la log-verosimiglianza attesa quando i dati completi contribuiscono alla verosimiglianza completa in un modo non lineare. Wei e Tanner (1990) hanno proposto un approccio Monte Carlo per trovare la log-verosimiglianza attesa $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)$. Il loro obiettivo consiste nel generare le covariate $Z_1, Z_2, \dots, Z_m$ dalla distribuzione condizionata $h(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})}{g(\boldsymbol{y}|\boldsymbol{\theta})}$ (si richiami che $\boldsymbol{x} = [\boldsymbol{y},\boldsymbol{z}]$ è il vettore dei dati completi, dove $\boldsymbol{y}$ è stato osservato e $\boldsymbol{z}$ è il completamento). Allora nel passo M, si massimizza un'approssimazione a $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)$:

$$\hat{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right) = \frac{1}{m}\sum_{i=1}^{m} logL(\boldsymbol{\theta};\boldsymbol{y},z_i).$$

Poiché

$$\frac{1}{m}\sum_{i=1}^{m} logL(\boldsymbol{\theta};\boldsymbol{y},z_i). \rightarrow\ Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right),$$

allora il metodo Monte Carlo EM converge al regolare EM quando $m \rightarrow \infty$.

In molte situazioni di interesse il modello con dati mancanti è debolmente identificabile, cioè la funzione di verosimiglianza con i dati osservati esibisce parecchi massimi locali di grandezza comparabile anche per elevate dimensioni campionarie n. In tali situazioni l'algoritmo Monte Carlo EM potrebbe condurre a procedure inefficienti. L'algoritmo EM stocastico (SEM) incorpora un passo S che simula una realizzazione $\boldsymbol{z}^*$ dell'insieme di dati mancanti dalla densità a posteriori $\pi\left(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}^{(k)}\right)$ basata sulla stima corrente, la quale è successivamente aggiornata massimizzando la funzione di verosimiglianza dell'insieme di dati ricostruito $(\boldsymbol{y},\boldsymbol{z}^*)$ e non c'è bisogno di calcolare $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)$. Il passo S può essere completato attraverso il campionamento di Gibbs o Hasting-Metropolis (Diebolt et al., 1996).

L'algoritmo EM ha una relazione con i metodi variazionali di Bayes., dato che può essere considerato un metodo di massima verosimiglianza parzialmente non-bayesiano. Il suo risultato finale dà una distribuzione di probabilità sulle variabili latenti (in stile bayesiano) insieme ad una stima puntuale di $\boldsymbol{\theta}$ (una stima di massima verosimiglianza o una moda a posteriori). Una versione pienamente bayesiana conferisce una distribuzione di probabilità per $\boldsymbol{\theta}$ così come per le variabili latenti. In realtà l'approccio bayesiano all'inferenza è semplicemente quello di trattare $\boldsymbol{\theta}$ come un'altra variabile latente. In questo paradigma, la distinzione tra i passi E ed M scompare. Se si usa l'approssimazione fattorizzata $Q(\cdot)$ come descritto in

precedenza (Bayes variazionale), si può iterare su ogni variabile latente (ora includendo $\boldsymbol{\theta}$) e ottimizzarle una alla volta. Ora ci sono k passi per iterazione, dove k indica il numero di variabili latenti.
Molti sono stati i tentativi in letteratura di rendere più rapido l'algoritmo EM, si veda per esempio Louis (1982), Horng (1987). Negli ultimi anni le estensioni dell'algoritmo EM si sono moltiplicate, tanto che risulta difficile elencare tutte le novità: ogni estensione cerca di rimediare agli svantaggi dell'algoritmo esistenti (McLachlan, Geoffrey, 2008). Sono stati proposti alcuni metodi per accelerare la convergenza a volte lenta dell'algoritmo EM, come quelle che utilizzano il gradiente coniugato e le tecniche modificate di Newton-Raphson. Inoltre L'algoritmo EM può essere utilizzato con tecniche di stima vincolata.

A.2 IL METODO BOOTSTRAP

A.2.1 *Introduzione*

Il *bootstrap* è una tecnica statistica di ricampionamento per approssimare la distribuzione campionaria di una statistica. Permette perciò di approssimare media e varianza di uno stimatore, costruire intervalli di confidenza e calcolare livelli di significatività osservati di test quando, in particolare, non si conosce la distribuzione della statistica di interesse. L'idea alla base del *bootstrap* è quella di utilizzare la distribuzione empirica del campione, che è l'unica informazione che abbiamo sulla distribuzione $F^0(\cdot)$ della popolazione, generando numerosi campioni con una procedura di ricampionamento con ripetizione di n elementi dagli n dati campionari; in questo modo si ottengono diverse stime del parametro d'interesse con le quali, grazie all'aiuto del computer e senza utilizzare formule matematiche particolarmente complicate, si è in grado di ottenere misure di variabilità dello stimatore quali errore standard, distorsione e intervalli di confidenza e quindi di usare le stime *bootstrap* per ogni metodo statistico. Nel caso semplice di campionamento casuale semplice, il funzionamento è il seguente: si consideri un campione effettivamente osservato di numerosità pari ad n, $\boldsymbol{x} = (x_1, \dots, x_n)$. Da $\boldsymbol{x}$ si ricampionano B altri campioni di numerosità costante pari ad n, $x_1^*, \dots, x_B^*$; in ciascuna estrazione *bootstrap*, i dati provenienti dal primo elemento del campione, cioè x_1, possono essere estratti più di una volta e ciascun dato ha probabilità pari a $1/n$ di essere estratto. Sia T lo stimatore di θ, diciamo $T(\boldsymbol{X}) = \hat{\theta}$. Si calcola tale quantità per ogni campione *bootstrap*, $T(x_1^*), \dots, T(x_B^*)$. In questo modo si hanno a disposizione B stime di θ, dalle quali è possibile calcolare la media *bootstrap*, la varianza *bootstrap*, i percentili *bootstrap* ecc. che sono approssimazioni dei corrispondenti valori ignoti e portano informazioni sulla distribuzione di $T(X)$.

A.2.2 *Cenni sulla metodologia bootstrap*

A.2.2.1 Il problema statistico

Si assuma che i dati $\boldsymbol{x} = (x_1, \dots, x_n)$ siano n realizzazioni indipendenti di una variabile casuale univariata X, rilevata su una data popolazione costituita da N unità. Sia $(x_1, \dots, x_n)$ un campione C, di ampiezza n, estratto senza reinserimento dalla stessa popolazione. Frequentemente gli analisti che lavorano su dati campionari non conoscono nulla sulle caratteristiche della popolazione da cui viene estratto il campione stesso, quindi si assume che X abbia funzione di ripartizione ignota $F^0(\cdot)$. Ciò nonostante, essi hanno la necessità di ottenere una stima (puntuale o intervallare) di uno o più parametri incogniti della popolazione e, allo stesso tempo, hanno bisogno di poter valutare le proprietà dello stimatore utilizzato per ricavare tali grandezze. Il contesto nel quale gli analisti operano può essere *parametrico*, se sono note la funzione di distribuzione della popolazione da cui si estrae il campione e la funzione di distribuzione dello stimatore; *non parametrico* quando tali funzioni sono incognite. In termini più specifici, si vuole stimare una caratteristica della popolazione, poniamo θ, attraverso le osservazioni raccolte nel campione C, non conoscendo nulla sulla distribuzione della v.c. X e determinare il grado di accuratezza della stima $\hat{\theta}$. La soluzione di questo problema può essere analitica, individuando - sotto certe condizioni - la distribuzione esatta o asintotica dello stimatore utilizzato, oppure empirica attraverso l'impiego di tecniche di simulazione e di ricampionamento. Di seguito è presentato un aspetto di quest'ultimo approccio, indicando gli elementi caratterizzanti della metodologia *bootstrap.*

A.2.2.2 L'idea di base

L'idea di base del *bootstrap* fu introdotta nel 1979 da Bradley Efron proponendo una metodologia basata su calcoli informatici finalizzati alla stima dello scarto quadratico medio di una stima $\hat{\theta}$. Il *bootstrap* è, quindi, una di quelle metodologie statistiche che comunemente vengono dette *computer intensive*; trova cioè la sua ragione d'essere nella possibilità di disporre di adeguati strumenti di calcolo elettronico (la metodologia è "avida" di tempo di calcolo e se non si ricorresse all'uso di calcolatori elettronici la sua implementazione sarebbe talmente laboriosa da scoraggiarne l'impiego in assoluto). Fatta eccezione per questo limite, al giorno d'oggi tutt'altro che insuperabile, il metodo, come dicono Efron e *Tibshirani* (1993) '*enjoys the advantage of being completely automatic*'. A ciò si aggiunga che la stima *bootstrap* dello scarto quadratico medio di $\hat{\theta}$, non richiedendo assunzioni teoriche particolari sulla popolazione in riferimento alla quale è calcolata, non risulta influenzata dalla complessità matematica dello stimatore impiegato. La logica su cui

poggia il *bootstrap* può essere illustrata come segue: dato un campione iniziale $\boldsymbol{x} = (x_1, \dots, x_n)$ estratto da una variabile casuale X con distribuzione $F^0(\boldsymbol{x}; \theta)$, si stima il parametro θ mediante $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$. Si devono però superare due ostacoli. Il primo è concettuale, la legge di $\hat{\theta}$ è funzione di $F^0(\cdot)$, che è ignota. Il secondo ostacolo è pratico: anche immaginando di conoscere $F^0(\cdot)$, in ben pochi casi il problema di ottenere la legge di $\hat{\theta}$ è matematicamente trattabile, pur ammettendo soluzioni approssimate. Il *bootstrap* sfrutta la simulazione per superare simultaneamente i due ostacoli e farci pervenire a una stima della distribuzione di $\hat{\theta}$ sotto $F^0(\cdot)$. Sul piano concettuale, l'idea di base è sostituire $F^0(\cdot)$ con una sua stima. Sul piano pratico, stimata $F^0(\cdot)$, si è ricondotti a una simulazione. Dal campione iniziale si estraggono con ripetizione altri campioni, $\boldsymbol{x}_b^* = (x_{b1}^*, \dots, x_{bn}^*)$, con $b = 1, \dots, B$, tratti dalle variabili casuali $\boldsymbol{X}_b^* = (X_{b1}^*, \dots, X_{bn}^*)$, dove le componenti X_{bi}^*, con $i = 1, \dots, n$, sono indipendenti e identicamente distribuite con opportuna legge marginale. Sono così ottenuti per simulazione B realizzazioni di $\boldsymbol{x}_b^*$, che sono B campioni di numerosità n. Su ciascuno di essi si fornisca una stima del parametro d'interesse $\hat{\theta}_b^* = \hat{\theta}(\boldsymbol{x}_b^*)$ e si calcoli il valore assunto dalla statistica $\hat{\theta}$. Il campione da $\hat{\theta}_b^*$ viene usato per stimare la funzione di densità di $\hat{\theta}_b^*$, con il metodo del nucleo, o la funzione di ripartizione di $\hat{\theta}_b^*$ con la funzione di ripartizione empirica.

A.2.2.3 Aspetti formali

Formalmente, la metodologia si articola in due fasi fondamentali; la prima è quella in cui un campione C, che d'ora in avanti chiameremo 'campione principale', di ampiezza n viene estratto *senza ripetizione* da una popolazione X; la seconda è quella in cui viene estratto da C, con *reinserimento*, un sub-campione di ampiezza n (detto 'pseudo-osservazioni' o 'campione *bootstrap*'); questa seconda fase viene ripetuta B volte. Tale ripetizione, che deve avvenire un numero 'sufficientemente grande' di volte, ha lo scopo di portare ad un'approssimazione dell'universo dei campioni di ampiezza n ottenibili con reinserimento da C. Poi si indicherà con ${}_b\boldsymbol{X}^* = \left({}_bX_1^*, \dots, {}_bX_n^* | x_1, \dots, x_n \right)$ la v.c. descritta dal b-esimo campione *bootstrap* $(b = 1, \dots, B)$ condizionato al fatto che dalla popolazione sia stato estratto il particolare 'campione principale' C; la v.c ${}_b\boldsymbol{X}^* = \left({}_bX_1^*, \dots, {}_bX_n^* | x_1, \dots, x_n \right)$ varierà nell'insieme costituito dall'universo dei campioni *bootstrap* di numerosità n estraibili da C. Al termine di questa seconda fase, l'obiettivo sarà quello di determinare lo scarto quadratico medio di $\hat{\theta}$. A tale scopo, nel *bootstrap* parametrico si assume che $F^0(\cdot) = F(\cdot, \theta^0)$, per un modello parametrico $\mathcal{F} = \{F(\cdot, \theta), \theta \in \Theta \subseteq \mathbb{R}^p\}$, con $\theta^0 \in \Theta$. Si ottiene quindi una conveniente stima $\hat{\theta}_n$ di θ^0, e si usa $F\left(\cdot, \hat{\theta}_n\right)$ come distribuzione marginale per X_{bj}^*. Nel *bootstrap* non parametrico non si assume un modello non parametrico e si utilizza come distribuzione marginale per X_{bj}^* direttamente la funzione di ripartizione empirica $\hat{F}_n(\cdot)$ del 'campione principale' anziché l'incognita distribuzione $F^0(\cdot)$ della popolazione da cui C è stato estratto. In entrambi i casi, sotto

opportune condizioni, per n sufficientemente grande, $\hat{\theta} \sim \hat{\theta}_b^*$, sotto $\hat{\theta}$. Per maggior chiarezza viene riproposto nella Tabella A.1 l'algoritmo presentato da Efron (1993).

Tabella A.1. Schema concettuale della stima *bootstrap* dello scarto quadratico medio di $\hat{\theta}$, $\sqrt{V(\hat{\theta})}$ (Efron, 1993).

Distribuzione empirica		Campioni *bootstrap* di dimensione n: $\boldsymbol{x}_b^* = (x_{b1}^*, \dots, x_{bn}^*)$, per $b = 1, \dots, B$	Stime *bootstrap* di $\hat{\theta}$: $\hat{\theta}^* = (\hat{\theta}_1^*, \dots, \hat{\theta}_b^*, \dots \hat{\theta}_B^*)$ $\hat{\theta}_b^* = \hat{\theta}(\boldsymbol{x}_b^*)$, per $b = 1, \dots, B$		Stima *bootstrap* di $\sqrt{V(\hat{\theta})}$
$\hat{F}$					
	↗	$\boldsymbol{x}_1^*$	$\hat{\theta}_1^* = \sqrt{\hat{V}(\boldsymbol{x}_1^*)}$		
		$\boldsymbol{x}_2^*$	$\hat{\theta}_2^* = \sqrt{\hat{V}(\boldsymbol{x}_2^*)}$		
		$\boldsymbol{x}_3^*$	$\hat{\theta}_3^* = \sqrt{\hat{V}(\boldsymbol{x}_3^*)}$	↘	
		⋮	⋮		$\sqrt{\hat{V}_B(\hat{\theta})} = \left[\sum_{b=1}^{B} \frac{\left[\hat{\theta}_b^* - \hat{\theta}_B\right]^2}{B-1}\right]^{1/2}$
	→	$\boldsymbol{x}_b^*$	$\hat{\theta}_b^* = \sqrt{\hat{V}(\boldsymbol{x}_b^*)}$	→	dove: $\hat{\theta}_B = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^*$
		⋮	⋮		
	↘	⋮	⋮	↗	
		⋮	⋮		
		$\boldsymbol{x}_B^*$	$\hat{\theta}_B^* = \sqrt{\hat{V}(\boldsymbol{x}_B^*)}$		

Tutti i possibili campioni, diversi per almeno un elemento, che si potrebbero estrarre da X, sono:

$$\binom{2n-1}{n}.$$

Nella maggior parte dei casi, però, si tratta di un numero troppo elevato, pertanto si procede estraendo da X un numero di campioni più piccolo (B) la cui grandezza dipende dalla disponibilità dei mezzi di calcolo utilizzabili nonché dal tipo di stime che si vogliono ottenere: l'ideale sarebbe un numero $B \to \infty$ ma,

fortunatamente, esistono precisi criteri per stabilire un numero $B < \infty$ che sia ugualmente soddisfacente. Tali criteri possono essere esplorati nel dettaglio in Efron ("*An introduction to the bootstrap*", 1993). Si riportano alcune regole pratiche per determinare B:

- $30 < B < 50$ per stimare i momenti;
- $B = 50$ per ottenere stime affidabili dello scarto quadratico medio di $\hat{\theta}$ sotto la distribuzione empirica $\hat{F}$: $\sqrt{\hat{V}_{\hat{F}}(\hat{\theta})}$;
- Molto raramente è richiesto ≥ 200 per stimare uno scarto quadratico medio;
- B almeno uguale a 1000 per stimare la distribuzione di campionamento dello stimatore o per costruire intervalli di confidenza.

Le stima *bootstrap* del parametro d'interesse θ è data da:

$$\hat{\theta}_B = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^*,$$

quella della varianza di $\hat{\theta}$ è:

$$\hat{V}_B(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\theta}_B\right)^2.$$

In questo modo, simulando l'estrazione di più campioni, è possibile stimare la distribuzione *bootstap* di $\hat{\theta}$ senza fare assunzioni sulla forma della popolazione per cui tale grandezza è ricercata.

A.2.3 Applicazione del *bootstrap*

L'ambito applicativo più importante del *bootstrap* è quello della costruzione di intervalli di confidenza. L'approccio classico a questa categoria di procedure inferenziali prevede l'uso di statistiche pivotali e delle loro distribuzioni esatte o asintotiche. Nella generalità dei casi, quando si ottiene con i dati campionari una stima $\hat{\theta}$ della caratteristica θ su cui si investiga, si assume che, all'aumentare dell'ampiezza del campione C, la distribuzione di $\hat{\theta}$ approssimi sempre di più quella della normale con $E(\hat{\theta}) = \theta$. Considerato ciò, l'intervallo di confidenza per θ, al livello di confidenza $1 - \alpha$, viene calcolato con la formula:

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})},$$

dove $z_{\alpha/2}$ viene ottenuto dalle tavole della distribuzione normale standardizzata e $\sqrt{\hat{V}(\hat{\theta})}$ è la stima dello scarto quadratico medio di $\hat{\theta}$. I problemi connessi con questa impostazione metodologica sono legati al fatto che il ricercatore, il più delle volte, postula l'apparteneza della popolazione oggetto di studio ad una famiglia parametrica nota (ad esempio la famiglia esponenziale): la veridicità di questa assunzione spesso non risulta controllabile. Il metodo *bootstrap* permette di superare il problema insito nella formulazione di ipotesi distributive non verificabili e consente al ricercatore di ottenere intervalli o regioni di confidenza più affidabili.

A.2.3.1 *Metodi per la determinazione degli I.C. per* θ

Esistono vari metodi *bootstrap* per la determinazione degli intervalli di confidenza per θ, i principali sono i seguenti:

a) Metodo dell'intervallo *t-bootstrap*
b) Metodo del percentile
c) Metodo del percentile corretto
d) Metodo del percentile corretto e accelerato.

a) Metodo dell'intervallo t-bootstrap

Per ciascuno dei B campioni *bootstrap* si calcola la quantità pivotale è

$$Z^* = \frac{\hat{\theta}_B - \theta}{\sqrt{\hat{V}_B(\hat{\theta})}},$$

che tende approssimativamente ad una $N(0,1)$, con $b = 1, \dots, B$, dove dove $\hat{\theta}_B = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^*$ e $\sqrt{\hat{V}_B(\hat{\theta})}$ è la stima dello scarto quadratico medio di $\hat{\theta}$, dove $\hat{V}_B(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}_B)^2$.

Pertanto, l'intervallo di confidenza per θ al livello $(1-2\alpha)$ è dato da:

$$\left[\hat{\theta}_B - \hat{t}^{(1-\alpha)}\sqrt{\hat{V}_B(\hat{\theta})}, \hat{\theta}_B - \hat{t}^{(\alpha)}\sqrt{\hat{V}_B(\hat{\theta})}\right].$$

dove $\hat{t}^{(\alpha)}$ e $\hat{t}^{(1-\alpha)}$ sono, rispettivamente, l'α-esimo e $(1-\alpha)$-esimo percentile della distribuzione *bootstrap* di Z^*, ovvero $\hat{t}^{(\alpha)}$ è tale che

$P\left[Z^* \leq \hat{t}^{(\alpha)}\right] \cong \alpha$ (o, altrimenti, $\rightarrow \frac{\left[\#Z^* \leq \hat{t}^{(\alpha)}\right]}{B} = \alpha$).

L'utilizzo di questo tipo di intervallo è condizionato dal fatto che esso non è invariante per trasformazione.

Metodo del percentile (BP)

Il calcolo dell'I.C. fatto con questo metodo si basa sull'utilizzo dei percentili della distribuzione *bootstrap* cumulata di $\hat{\theta}^* = \left(\hat{\theta}_1^*, \dots, \hat{\theta}_b^*, \dots \hat{\theta}_B^*\right)$. Si ricavino le seguenti quantità $\hat{\theta}^{*(\alpha)}$ e $\hat{\theta}^{*(1-\alpha)}$, così definite:

$$\hat{\theta}^{*(\alpha)} = Pr\left(\hat{\theta}^* \leq \alpha\right),$$

$$\hat{\theta}^{*(1-\alpha)} = Pr\left(\hat{\theta}^* \leq 1-\alpha\right).$$

Gli estremi dell'intervallo di confidenza per θ al livello $(1-2\alpha)$ sono:

$$\left[\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}\right] \text{ oppure } \left[\hat{\theta}_B - \frac{\hat{\theta}^{*(1-\alpha)}\sqrt{\hat{V}_B(\hat{\theta})}}{\sqrt{B}}, \hat{\theta}_B - \frac{\hat{\theta}^{*(\alpha)}\sqrt{\hat{V}_B(\hat{\theta})}}{\sqrt{B}}\right],$$

dove $\hat{\theta}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$ e $\hat{V}_B\left(\hat{\theta}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}_B\right)^2$.

Una proprietà di tale intervallo, che lo fa preferire a quello visto in precedenza, è che è invariante per trasformazioni monotone ed è anche *range-preserving*, ovvero i suoi estremi ricadono sempre entro l'intervallo di variazione del parametro d'interesse. Il metodo del percentile metodologicamente è ritenuto particolarmente affidabile in quanto offre il miglior *livello di copertura*. Un suo limite, invece, è che esso risulta poco accurato a meno che la dimensione dei sub-campioni non sia sufficientemente grande.

b) Metodo del percentile corretto (BC)

Talvolta per la costruzione di I.C. è vantaggioso ricorrere alle proprietà della distribuzione normale. Ciò può essere fatto mediante una trasformazione monotona $\varphi(\cdot)$ di $\hat{\theta}_B$ del tipo:

$$\varphi = \frac{\hat{\theta}_B - \theta}{\tau} + z_0, \qquad (A.14)$$

tale da indurre la normalità di $\hat{\theta}_B$ e dove τ è costante. Nei casi in cui è ammissibile una trasformazione come la (A.14), il metodo del percentile fornisce un intervallo di confidenza distorto. Efron ha dimostrato che ponendo:

$$\begin{cases} z_0 = G^{-1}\{F^*(\hat{\theta}_B, \hat{F})\}, \\ \hat{\theta}^{*(\alpha)} = F^{*-1}\{G(z^{(\alpha)} + 2z_0)\}, \\ z^{(\alpha)} = G^{-1}(\alpha), \end{cases}$$

l'intervallo di confidenza per θ al livello $(1-2\alpha)$ risulta uguale a:

$$\left[\hat{\theta}_B - \frac{\sqrt{\hat{V}_B(\hat{\theta})}}{\sqrt{B}}\hat{\theta}^{*(1-\alpha)}, \hat{\theta}_B - \frac{\sqrt{\hat{V}_B(\hat{\theta})}}{\sqrt{B}}\hat{\theta}^{*(\alpha)}\right].$$

Nelle applicazioni pratiche solitamente si può usare al posto di $G(\cdot)$ la distribuzione normale standardizzata $\phi(\cdot)$ per cui:

$$\begin{cases} z_0 = \phi^{-1}\{F^*(\hat{\theta}_B, \hat{F})\}, \\ z^{(\alpha)} = \phi^{-1}(\alpha). \end{cases}$$

In particolare si osserva che se $z_0 = 0$, l'intervallo BC coincide con quello BP.

c) Metodo del percentile corretto-accelerato (BCA)

Questo metodo trova il suo impiego quando è possibile ipotizzare che la distribuzione *bootstrap* sia non simmetrica intorno a $\hat{\theta}_B$ e che la varianza di quest'ultima caratteristica sia funzione di θ. In questo caso occorre trasformare $\hat{\theta}_B$ in modo da stabilizzare la varianza ed eliminare l'asimmetria della distribuzione. Una trasformazione che consente di superare entrambi i problemi è quella proposta sempre da Efron:

$$\varphi = \frac{\hat{\theta}_B - \theta}{1 + a\theta} + z_0,$$

ponendo:

$$\begin{cases} \hat{\theta}^{*(\alpha)} = F_n^{*-1}\left\{G\left[z_0 + \dfrac{z^\alpha + z_0}{1 - a(z^\alpha + z_0)}\right]\right\}, \\ z_0 = G^{-1}\{F^*(\hat{\theta}_B, \hat{F})\}. \end{cases}$$

Solitamente $z_0 = \phi^{-1}\{F^*(\hat{\theta}_B, \hat{F})\}$.

L'intervallo di confidenza al livello $(1-2\alpha)$ è il seguente:

$$\left[\hat{\theta}_B - \frac{\sqrt{\hat{V}_B(\hat{\theta})}}{\sqrt{B}}\hat{\theta}^{*(1-\alpha)}, \hat{\theta}_B - \frac{\sqrt{\hat{V}_B(\hat{\theta})}}{\sqrt{B}}\hat{\theta}^{*(\alpha)}\right].$$

In casi come questo è necessario stimare sia la costante di correzione dell'asimmetria z_0 sia la costante di accelerazione a; un modo per stimare quest'ultima può essere:

$$\hat{a} = \frac{\sum_{b=1}^{B}\left[\hat{\theta}_B - \hat{\theta}_b^*\right]^3}{6\left\{\sum_{b=1}^{B}\left[\hat{\theta}_B - \hat{\theta}_b^*\right]^2\right\}^{3/2}},$$

dove $\hat{\theta}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$.

L'intervallo BCA è invariante per trasformazioni monotone ed ha un grado di accuratezza del II ordine. Naturalmente, se $a = 0$ allora BCA=BC; se $a = 0$ e $z_0 = 0$ allora BCA=BC=BP.

A.2.3.2 *Determinazione dell'errore di copertura*

Un modo per valutare la validità di un intervallo di confidenza è quello di misurare l'*errore di copertura*, dato dalla differenza tra il livello di confidenza *effettivo* e quello *nominale*. Il livello di confidenza effettivo viene determinato attraverso il calcolo del numero di intervalli *bootstrap* contenenti il parametro $\hat{\theta}$ stimato con i dati del campione principale, tenendo sempre conto del fatto che non si conosce nulla sul vero parametro θ della popolazione. È quindi necessario stimare, per ciascuno dei B campioni *bootstrap*, altrettanti intervalli di confidenza, ovvero, estrarre da ogni b-esimo campione $(b = 1, \dots, B)$ altri W sub-campioni *bootstrap* ed iterare il procedimento. In altri termini, allorchè da ogni sub-campione *bootstrap* si estraggono W sub-sub-campioni *bootstrap*, su ciascuno di essi si calcolano le statistiche $\hat{\theta}^{**}$ (nel caso dell'utilizzo del metodo del percentile si ordinano) e si determina per ciascuno dei B campioni il rispettivo intervallo di confidenza (cioè $\hat{\theta}_{INF}^{**}$ e $\hat{\theta}_{SUP}^{**}$). A questo punto, si verifica se la grandezza campionaria $\hat{\theta}$ si trova o meno entro ciascun intervallo: il livello di confidenza effettivo è dato dalla frazione di intervalli che contengono $\hat{\theta}$. Un procedimento di questo tipo è ovviamente vincolato alla possibilità di incrementare notevolmente il tempo computazionale a disposizione. Proprio per ridurre il numero di estrazioni e, conseguentemente, ottimizzare il tempo computazionale è possibile stimare solo la varianza della seconda distribuzione *bootstrap* (che richiede un numero di estrazioni inferiore), assumendo che la distribuzione di campionamento standardizzata della statistica $\hat{\theta}$ sia ben approssimata dalla distribuzione di

campionamento *bootstrap* standardizzata. Tuttavia, esistono altre tecniche di ricampionamento che riducono il numero di campioni necessario per l'implementazione del *bootstrap iterato* e che garantiscono un buon livello di accuratezza degli intervalli. Tra queste tecniche si ricordano il *bootstrap bilanciato*, il *bootstrap antitetico* e il *bootstrap bilanciato ed antitetico*.

BIBLIOGRAFIA

1. Alonzo T.A., Pepe M.S. (1999), Using a combination of reference tests to assess the accuracy of a new diagnostic test, *Statistics in Medicine*; 18: 2987-3003.
2. Alvord W.G., Drummond J.E., Arthur L.O., et al. (1988), A method for predicting individual HIV infection status in the absence of clinical information, *Aids Research and Human Retroviruses*; 4: 295-304.
3. Baker S.G. (1991), Evaluation a new test using a reference test with estimated sensitivity and specificity, *Communications in Statistics A- Theory and Method*; 20: 2739-52.
4. Baker S.G., 1995, "Evaluating multiple diagnostic tests with partial verification", *Biometrics*; 51: 330-37.
5. Bamber D. (1975), The area above the ordinal dominance graph and the area below the receiver operating graph, *Journal of Mathematical Psychology*; 12: 387-415.
6. Becker M.P. (1994), Analysis of cross-classifications of counts using models for marginal distributions: an application to trends in attitudes on legalized abortion, *Sociological Methodology*; 24: 229-65.
7. Begg C.B., McNeil B.J. (1988), Assessment of radiologic tests: control of bias and other design considerations, *Radiology*; 167: 565-9.
8. Begg C.B., Metz C.E. (1990), Consensus diagnosis and gold standard, *Medical Decision Making*; 10: 29-30.
9. Beiden S.V., Campbell G., Meier K.L., Wagner R.F. (2000), On the problem of ROC analysis without truth: the EM algorithm and the information matrix, *in Proceedings of SPIE*; 3981: 126-134.
10. Biernacki C., Celeux G., Govaert G. (2003), Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models, *Computational Statistics and Data Analysis;* 41: 561-575.
11. Bottarelli E., Parodi S (2003), Un approccio per la valutazione della validità dei test diagnostici: le curve ROC (Receiver Operating Characteristic), *Ann. Fac. Medic. Vet. Di Parma* (Vol. XXIII): 49-68.
12. Branscum A.J., Johnson W.O., Hanson T.E., Gardner I.A. (2008), Bayesian semiparametric ROC curve estimation and disease diagnosis, *Statistics in Medicine*; 27: 2474-2496.
13. Celeux G., Diebolt J. (1895), The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Computational Statistics Quarterly*; 2: 73-82
14. Choi Y.K., Johnson W.O., Collins M.T., Gardner I.A. (2006), Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard, American Statistical Association and the International Biometric Society, *Journal of Agricultural, Biological, and Environmental Statistics*; Vol. 11, No. 2: 210-229.

15. Clogg C.C., Goodman L.A. (1984), Latent structure analysis of a set of multidimensional contingency tables, *Journal of the American Statistical Association*; 79: 762-71.

16. Dawid A.P., Skene A.M. (1979), Maximum likelihood estimation of observer error-rates using the EM algorithm, *Applied Statistics*; 28: 20-28.

17. DeLong E.R., DeLong D.M., Clarke-Pearson D.L. (1988), Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics*; 44: 837-845.

18. Dempster A.P.,Laird N.M.,Rubin D.B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society* B; Vol. 39, No. 1: 1-38.

19. Dendukuri N., Joseph L. (2001), Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests, *Biometrics*; 57: 158-167.

20. Diebolt J., Ip E.H. (1996), Stochastic EM: Method and application. In Markov Chain Monte Carlo in Practice, W.R.Gilks, S. Richardson, D.J.Spiegelhalter, Chapter 15, 259-273. Chapman & Hall/CRC, London.

21. Efron B., Tibishrani R.J. (1993), An introduction to the bootstrap, Chapman & Hall, London;

22. Enϕe C., Georgiadis M.P., Johnson W.O. (2000), Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown, Preventive *Veterinary Medicine*; 45: 61-81.

23. Espeland M.A., Handelman S.L. (1989), Using latent class models to characterize and assess relative error in discrete measurements, *Biometrics*; 45: 587-99.

24. Espeland M.A., Murphy W.C., Leverett D.H. (1988), Assessing diagnostic reliability and estimating incidence rates associated with a strictly progressive disease: dental caries, *Statistics in Medicine*; 7: 403-16.

25. Espeland M.A., Platt O.S., Gallagher D. (1989), Joint estimation of incidence and diagnostic error rates from irregular longitudinal data, *Journal of the American Statistical Association*; 84: 972-45.

26. Formann A.K. (1994), Measurement errors in caries diagnosis: some further latent class models, *Biometrics*; 50: 865-71.

27. Garret J. A., Stephenson B. (2001), Some Issues in Resolution of Diagnostic Tests using an Imperfect Gold Standard, , School of Statistics, University of Minnesota, published in: *Statistics in Medicine*, 20, 1987-2001. Technical Report 628

28. Gart J.J., Buck A.A. (1966), Comparison of a screening test and a reference test in epidemiologic studies. II: a probabilistic model for the comparison of diagnostic tests, *American Journal of Epidemiology*; 83: 593-602.

29. Georgiadis M.P., Johnson W.O., Singh R., Gardner I.A. (2003), Correlation-adjusted estimation of sensitivity and specificity of two diagnostic test, *Applied Statistics*; 52: 63-76.

30. Gilks W.R., Clayton D.G., Spiegelhalter D.J., Best N.G, McNeil A.J., Sharples L.D., Kirby A.J. (1993), Modeling complexity: applications of Gibbs sampling in medicine, *Journal of the Royal Statistical Society B*; 55: 39-52.

31. Goldberg J.D., Wittes J.T. (1978), The estimation of false negatives in medical screening, *Biometrics*; 34: 77-86.

32. Goodman L.A. (1974), Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*; 61: 215-31.

33. Hadgu A., Miller W. (2001). Comment on: Using a combination of reference tests to assess the accuracy of a diagnostic test, *Statistics in Medicine*; 20: 656-658.

34. Hadgu A., Qu Y. (1998), A biomedical application of latent class models with random effects, *Applied Statistics*; 47: 603-616.

35. Hall P., Zhou X.H. (2003), Nonparametric estimation of component distributions in a multivariate mixture, *Annals of Statistics*; 31: 201-224.

36. Hanley J.A., McNeil B.J. (1983), The meaning and use of the area under the Receiver Operating Characteristic (ROC) curve, *Radiology*; 148: 839-43.

37. Hanson T., Wesley O.J., Gardner I.A. (2003), Hierarchical Models for Estimating Herd Prevalence and Test Accuracy in the Absence of a Gold Standard, *American Statistica Association and the International Biometric Society, Journal of Agritural, Biologica, and Environmental Statistics*; Vol. 8, No. 2: 223-230.

38. Hartley H. (1958), Maximum likelihood estimation from incomplete data", *Biometrics*; 14: 174-194.

39. Henkelman R.M., Kay I., Bronskill M.J. (1990), Receiver operator characteristic (ROC) analysis without truth, *Medical Decision Making*; 10: 24-29.

40. Hiu S.L., Zhou X.H. (1998), Evaluation of diagnostic tests without gold standards, *Statistical Methods in Medical Research*; 7: 354-370, Division of Biostatistica and the Regenstrief institute for Health Care, Indiana University School of Medicine, Indianapolis, Indiana, USA, published by: SAGEhttp://smm.sagepub.com/content/7/4/354.

41. Hoppin J.W., Kupinski M.A., Kastis G.A., Clarkson E., Barrett H.H. (2002), Objective Comparison of Quantitative Imaging Modalities Without the Use of a Gold Standard, *IEEE transactions on medical imaging*; Vol. 21, No. 5.

42. Hsieh H.N., Su H.Y., Zhou X.H. (2009), Inteval Estimation for tha Difference in Paired Area under the ROC Curves in the Absence of a Gold Standard Test, *UW Biostatistics Working Paper Series*, University of Washington.

43. Hui S.L., Walter S.D. (1980), Estimating the error rates of diagnostic tests, *Biometrics*; 36: 167-71.

44. Joseph L., Gyorkos T.W., Coupal L. (1995), Bayesian estimation of disease prevalence and the parameters of diagnostics tests in the absence of a gold standard, *American Journal of Epidemiology*; 141: 263-72.

45. Lang J.B., Agresti A. (1994), Simultaneous modeling joint and marginal distributions of multivariate categorical responses, *Journal of the American Statistical Association*; 89: 625-32.

46. Li C.R., Liao C.T., Liu J.P. (2008), On the exact interval estimation for the difference in paired areas under the ROC curves, *Statistics in Medicine*; 27: 224-242.

47. Lindsay J. (2007), ROC Curves & Wilcoxon and Mann-Whitney Tests, *Tutorial Presentation*, CHL 5210 Categorical Data Analysis.

48. Little R.J.A. e Rubin D.B. (2002), Statistical analysis with missing data, seconda edizione, *Wiley Interscience.*

49. Liu H., Li G., Cumberland W. G., Wu T. (2005), Testing Statistical Significance of the Area under a Receiving Operating Characteristics Curve for Repeated Measures Design with Bootstarpping, University of California at Los Angeles, *Journal of Data Sciences*;3: 257-278.

50. Liu J.P., Ma M.C., Wu C.Y., Tai J.Y. (2006), Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves, *Statistics in Medicine*; 25: 1219-1238.

51. Mantel N. (1951), Evaluation of a class of diagnostic tests, *Biometrics*; 7: 240-46.

52. McClish D.K. (1989), Analyzing a portion of the ROC curve, *Medical Decision Making*; 9: 190-195.

53. Metz C.E. (1978), Basic principles of ROC analysis, *Seminars Nuclear Medicine*; Vol VIII, No. 4: 283-298.

54. Nagelkerke N.J.D., Fidler V., Buwalda M. (1988), Instrumental variables in the evaluation of diagnostic test procedures when the true disease state in unknown, *Statistics in Medicine*; 7: 739-44.

55. Neyman J. (1947), Outline of statistical treatment of the problem of diagnosis, *Public Heath Reports*; 62: 1449-56.

56. Norris M., Johnson W. O, Gardner I. A. (2009), Modeling bivariate longitudinal diagnostic outcome data in the absebce of a gold standard, *Statistics and its interface*, Vol. 2: 171-185.

57. Obuchowsky N., Lieber M.L. (1998), Confidence intervals for the receiver operating characteristics area in studies with small samples, *Acad. Radiology*.; 5: 561-571.

58. Pepe M.S. (2003), The statistical evaluation of medical tests for classification and prediction, *Oxford University Press*: New York; 28.

59. Qu Y., Tan M., Kutner M.H. (1996), Random effects models in latent class analysis for evaluating accuracy of diagnostic tests, *Biometrics*; 52: 797-810.

60. Qu Y., Hadgu A. (1998), A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test, *Journal of the American Statistical Association*; 93: 920-928.

61. Raykar V. C., Yu S., Zhao L.H., Valandez G.H., Florin C., Bogoni L., Moy L. (2010), Learning From Crowds, *Journal of Machine Learning Research*, 11: 1297-1322.

62. Rindskopf D., Rindskopf W. (1986), The value of latent class analysis in medical diagnosis, *Statistics in Medicine*; 5: 21-28.

63. Staquet M. Rozencweig M., Lee Y.J. et al. (1981), Methodology for the assessment of new dichotomous diagnostic tests, *Journal of Chronic Diseases*; 34: 599-610.

64. Thibodeau L. A. (1981), Evaluating diagnostic tests, *Biometrics*; 37: 801-804.

65. Vacek P.M. (1985), The effect of conditional dependence on the evaluation of diagnostic tests, *Biometric*; 41: 959-68.

66. Valenstein P.N. (1998), Evaluating diagnostic tests with imperfect standards, *American Journal of Clinical Pathology*; 93: 252-58.

67. Walter S.D., Irwig L.M. (1988), Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review, *Clinical Epidemiology*; 41: 923-37.

68. Wang C., Turnbull B. W., Gröhn, Yrjö, S*ϕ*ren S. Nielsen (2006), Nonparametric Estimation of ROC Curves Based on Bayesian Models When the True Disease State is Unknown, *American Statistical Association and the International Biometric Society Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 12, No. 1: 128-146.

69. Wu C.F.J. (1983), On the Convegence Properties of the EM Algorithm, *Ann. Stat*.; Vol. 11, No. 1: 95-103.

70. Yanagawa T., Gladen B.C. (1984), Estimating disease rating from a diagnostic test, *American Journal of Epidemiology*; 119: 1015-23.

71. Yang I., Becker M.P. (1997), Latent variable modeling of diagnostics accuracy, *Biometrics*; 53: 948-58.

72. Young M.A. (1983), Evaluating diagnostic criteria: a latent class paradigm*, Journal of Psychiatric Research*; 17: 285-96.

73. Zhou X.H. (1998), Correcting for verification bias in studies of a diagnostic tests accuracy, *Statistical Methods in Medical Research*; 7: 337-53.

74. Zhou X.H., Castelluccio P., Zhou C. (2005), Nonparametric estimation of ROC curves in the absence of a gold standard, *Biometrics*; 61: 600-609.

75. Zhou X.H., Obuchowsky N.A., McClish D.K. (2002), Statistical methods in diagnostic medicine, Wiley: New York.