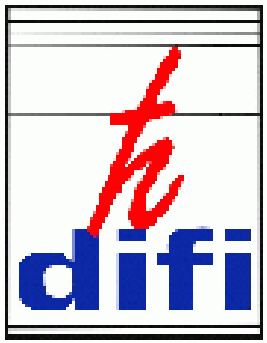


Dipartimento di Fisica

Scuola di Specializzazione in Fisica Medica

A.A. 2012/2013

Statistica nelle applicazioni sanitarie

Maria Roberta Monge:
Roberta.Monge@ge.infn.it

Test di screening e test diagnostici

- L'utilità di un test medico relativo ad una malattia è legata alla valutazione di quanto i suoi risultati siano corretti e con quale incertezza per due gruppi distinti di soggetti: il gruppo dei soggetti malati e quindi potenzialmente positivi al test ed il gruppo dei soggetti sani e quindi potenzialmente negativi al test.
- In questo contesto si differenziano due diversi tipi di test, a seconda degli scopi che si prefiggono, i test diagnostici di conferma che tendono appunto a diagnosticare la malattia ed i test di screening che invece tendono ad escludere la malattia nei soggetti sani.

Falso positivo e falso negativo

- Tutti questi tipi di test possono dare risultati corretti, ma possono anche presentare degli errori: si definisce falso positivo un soggetto sano che risulta invece positivo al test, mentre si definisce falso negativo un soggetto malato che risulta invece negativo al test.
- Questi tipi di errore si riconducono agli errori di decisione nei test statistici.

Errori di decisione nei test statistici

Nei test statistici di verifica delle ipotesi si confrontano due ipotesi, l'ipotesi nulla H_0 e l'ipotesi alternativa H_1.

Non sempre però la decisione presa su quale delle due ipotesi sia quella vera porta a risultati corretti, come si può vedere dalla tabella seguente:

Scelta corretta e scelta errata

	accetto Ho (e quindi rifiuto H_1)	**rifiuto Ho (e quindi accetto H_1)**
Ho vera	**ok**	**errore di primo tipo (probabilità α)**
Ho falsa	**errore di secondo tipo (probabilità β)**	**ok**

Sensibilità e specificità.

- L'errore di primo tipo nei test medici coincide con i falsi positivi o errore α.
- L'errore di secondo tipo nei test medici coincide con i falsi negativi o errore β.

	Positivi	**Negativi**
Malati	**ok**	**Falsi negativi (errore β)**
Sani	**Falsi positivi (errore α)**	**ok**

Questi due tipi di errore sono strettamente legati ai concetti di sensibilità e specificità che forniscono una misura del funzionamento del test medico.

La sensibilità è la capacità di un test di individuare la malattia quando essa sia presente, quindi in soggetti malati.

La specificità è la capacità di un test di escludere la malattia quando essa non sia presente, quindi in soggetti sani.

Consideriamo la seguente tabella di contingenza sul risultato del test e lo stato reale della malattia:

	Positivi al test (+)	**Negativi al test (-)**	
malati	**a**	**b**	**a+b**
sani	**c**	**d**	**c+d**
	a+c	**b+d**	**n=a+b+c+d**

Sensibilità del test

- Definiamo sensibilità del test la probabilità di essere positivi al test essendo malati:

$$P(+|m)=\frac{a}{a+b}$$

- Alla sensibilità del test è collegato l'errore falso negativo β corrispondente alla probabilità di risultare negativi al test essendo malati:

$$P(-|m)=\frac{b}{a+b}=\beta$$

- Si avrà pertanto: $\beta = 1 - sensibilità$

Specificità del test

- Definiamo specificità del test la probabilità di essere negativi al test essendo sani:

$$P(-|s) = \frac{d}{c+d}$$

- Alla specificità del test è collegato l'errore falso positivo α corrispondente alla probabilità di risultare positivi al test essendo sani:

$$P(+|s) = \frac{c}{c+d} = \alpha$$

- Si avrà pertanto: $\alpha = 1 - specificità$

Se indichiamo rispettivamente con

VP=Veri Positivi=a

FN=Falsi Negativi=b

FP=Falsi Positivi=c

VN=Veri Negativi=d

si ha che i soggetti malati sono VP+FN e quelli sani VN+FP

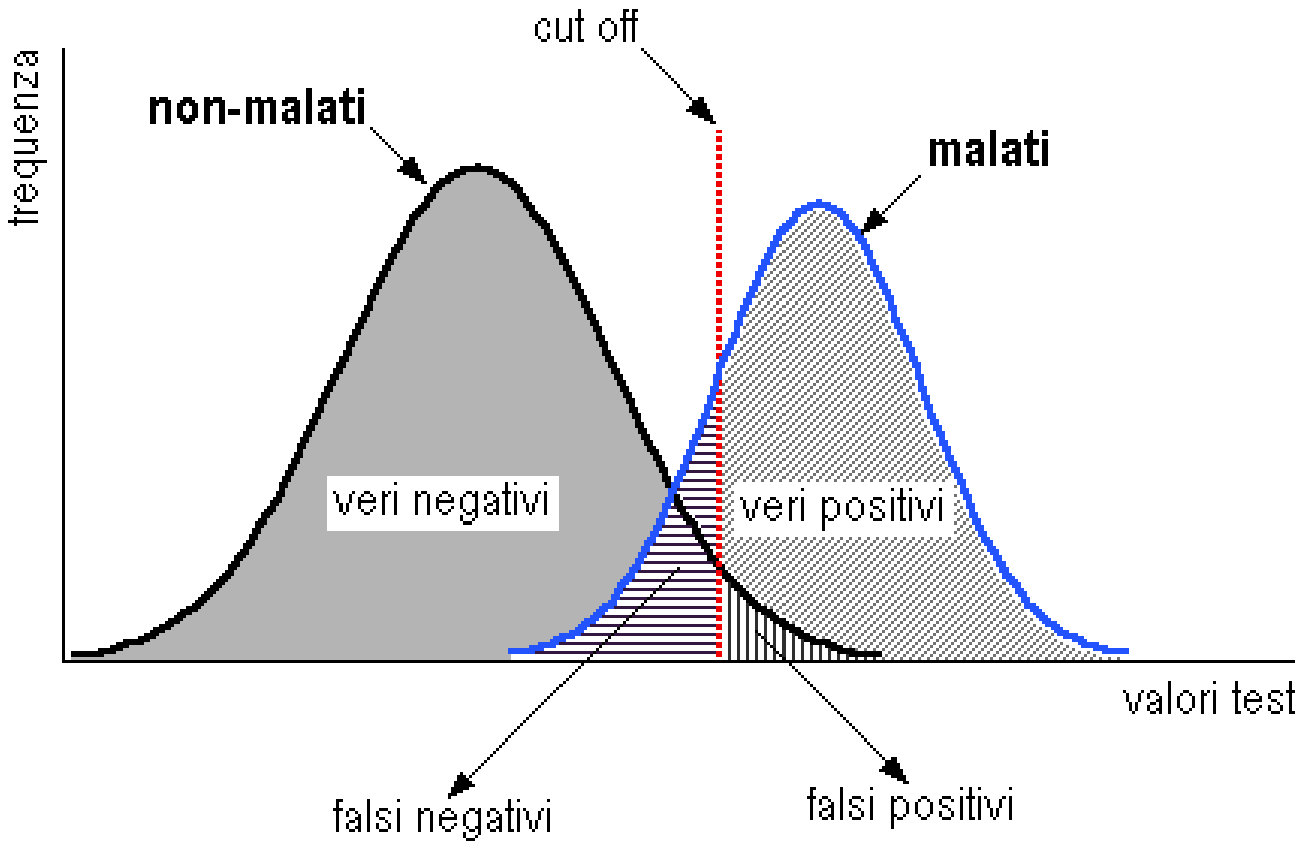
$$P(+|m)=\frac{VP}{VP+FN};P(-|m)=\frac{FN}{VP+FN}=\beta$$

$$P(-|s)=\frac{VN}{VN+FP};P(+|s))=\frac{FP}{VN+FP}=\alpha$$

- Se un test non è sensibile, mancherà di individuare la malattia in qualche soggetto effettivamente malato e quindi aumenterà il numero dei falsi negativi aumentando β.
- Se un test non è specifico, indicherà in maniera errata la malattia in soggetti sani e quindi aumenterà il numero dei falsi positivi e di conseguenza α.

Cut off o cut point o valore di soglia

Come si può vedere dalla figura, il valore di cut off o valore di soglia, scelto per discriminare fra positivi e negativi al test, fa variare sensibilità, specificità, α e β.

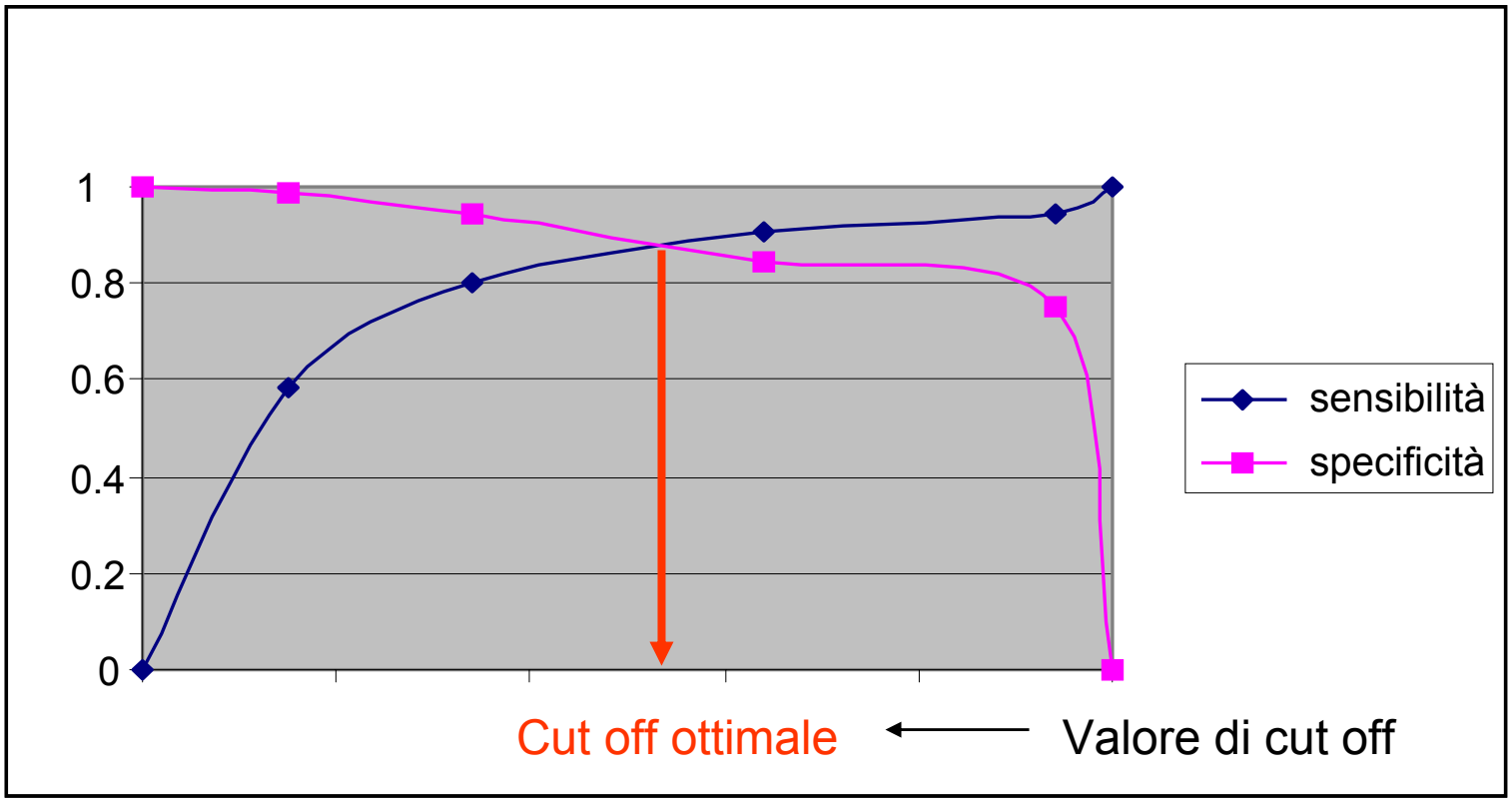


- Un cut off più spostato verso destra permetterà di identificare correttamente la maggior parte dei sani, conferendo al test un'elevata specificità e quindi pochi falsi positivi, ma sottostimerà la proporzione di malati, conferendo al test una bassa sensibilità.
- Un cut off più spostato verso sinistra permetterà di identificare correttamente la maggior parte dei malati, conferendo al test un'elevata sensibilità e quindi pochi falsi negativi, ma sottostimerà la proporzione dei sani, conferendo al test una bassa specificità.

Scelta del cut off

- Abbiamo visto che sensibilità e specificità sono inversamente correlate tra loro a seconda della scelta del valore di cut off: se si aumenta la specificità spostando il valore di soglia, diminuisce la sensibilità e viceversa.
- E' possibile dimostrare che, se si assumono normali le due popolazioni malati e sani, il valore di soglia ottimale che minimizza gli errori di classificazione (test più potente) è pari al valore corrispondente al punto di intersezione delle due curve di sensibilità e specificità in funzione del valore di soglia medesimo.

Sensibilità e specificità vs cut off



- Per il valore di soglia per cui si ha specificità 1, tutti i sani risultano tali (non si hanno falsi positivi), ma si ha anche che tutti i malati risulteranno sani (sensibilità 0)
- Viceversa per il valore di soglia per cui si avrà sensibilità 1, tutti i malati risultano tali, ma anche tutti i sani risulteranno malati (specificità 0).

- La soglia ottimale massimizza contemporaneamente sensibilità e specificità, minimizzando gli errori di classificazione.
- Non è però detto che questa scelta, dettata da considerazioni puramente probabilistiche, sia davvero la migliore, ma dipende dal tipo di test che si vuole fare e dal suo impatto clinico, economico e sociale.

Test diagnostici di screening e di conferma

- Solitamente nello studio di una patologia, si procede dapprima con test diagnostici molto sensibili e poco specifici che servono per identificare le persone a rischio di malattia (Test diagnostici di screening).
- Quindi i pazienti risultati positivi a questi primi test vengono sottoposti a test più specifici e meno sensibili per discriminare fra reali malati e falsi positivi al test precedente (Test diagnostici di conferma).

Valori predittivi

- Sensibilità e specificità sono delle misure di un test diagnostico ma non rispondono a due importanti problemi clinici: *se un paziente risulta positivo al test, qual è la probabilità che abbia la malattia diagnosticata*? E *se invece il risultato del test è negativo qual è la probabilità che sia davvero sano*?
- A queste domande si può rispondere mediante i valori predittivi.

Valore Predittivo Positivo.

- Si definisce Valore Predittivo Positivo VPP la probabilità di essere malati essendo risultati positivi al test diagnostico. Esso è dato da:

$$VPP = P(m|+) = \frac{VP}{VP + FP} = \frac{a}{a + c}$$

- È maggiormente influenzato dalla specificità del test in quanto più questa è alta, più è piccola la probabilità α di avere falsi positivi (c è piccolo) e di conseguenza il valore predittivo del test aumenta.

Valore Predittivo Negativo

- Si definisce Valore Predittivo Negativo VPN la probabilità di essere sani essendo risultati negativi al test diagnostico:

$$VPN = P(s|-) = \frac{VN}{VN + FN} = \frac{d}{b + d}$$

- È maggiormente influenzato dalla sensibilità del test in quanto più questa è alta, più è piccola la probabilità β di avere falsi negativi (b è piccolo) e di conseguenza il valore predittivo del test aumenta.

CONCLUSIONI

Un test diagnostico di screening, come abbiamo visto, è un test che tende soprattutto ad escludere la malattia a chi non l'ha. Si attua su persone potenzialmente sane. Deve perciò avere un Valore Predittivo Negativo alto e quindi un'alta sensibilità e pochi falsi negativi.

Un test diagnostico di conferma è invece un test che tende a diagnosticare la malattia nei soggetti realmente malati e quindi deve avere soprattutto pochi falsi positivi. Deve pertanto avere un Valore Predittivo Positivo alto e quindi un'alta specificità.

Come si può osservare dalle definizioni precedenti, le definizioni di Valore Predittivo Positivo VPP e di Valore Predittivo Negativo VPN non sono altro che un'applicazione del Teorema di Bayes:

$$VPP = P(m|+) = \frac{P(+|m)\cdot P(m)}{P(+|m)\cdot P(m) + P(+|s)\cdot P(s)}$$

$$VPN = P(s|-) = \frac{P(-|s)\cdot P(s)}{P(-|s)\cdot P(s) + P(-|m)\cdot P(m)}$$

e rappresentano delle probabiltà a posteriori, mentre P(m) rappresenta la probabilità a priori di malattia cioè la sua prevalenza. I denominatori sono rispettivamente P(+) e P(-), probabilità di risultare positivi o negativi al test.

- I valori predittivi di un test, come visto, dipendono dalla prevalenza della malattia P(m).
- Se si definisce l'odds di prevalenza OP della malattia come

$$OP = \frac{P(m)}{P(s)}$$

allora si ha

$$VPP = \frac{sensibilità}{sensibilità + \frac{\alpha}{OP}} = \frac{1-\beta}{(1-\beta) + \frac{\alpha}{OP}}$$

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

Rapporti di probabilità

- Per decidere la bontà di un test diagnostico è possibile usare i rapporti di probabilità o verosimiglianza.
- Si definisce rapporto di probabilità positivo LR^+ la quantità

$$LR^+ = \frac{P(+|m)}{P(+|s)} = \frac{sensibilità}{1-specificità} = \frac{1-\beta}{\alpha}$$

- Si definisce rapporto di probabilità negativo LR^- la quantità

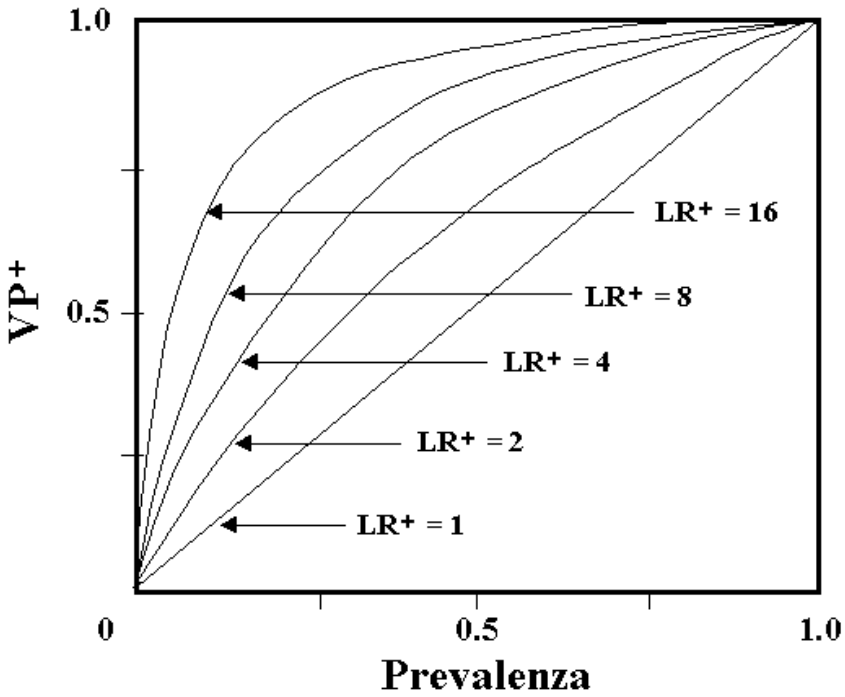
$$LR^- = \frac{P(-|m)}{P(-|s))} = \frac{1-sensibilità}{specificità} = \frac{\beta}{1-\alpha}$$

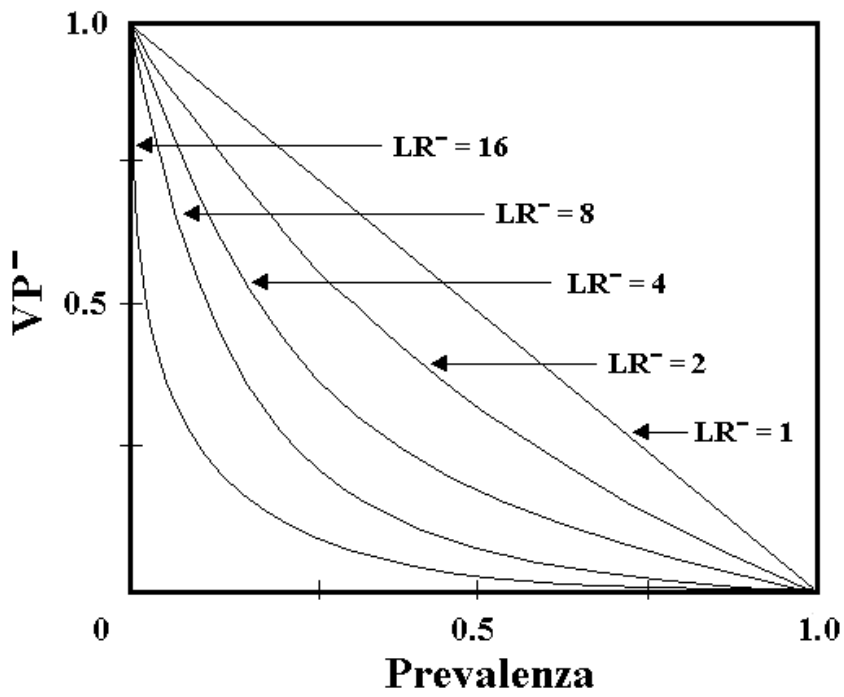
- Più alto è LR^+ migliore è il test diagnostico. Infatti deve risultare alto il rapporto fra la sensibilità del test (che si vuole grande) e la probabilità di avere falsi positivi (che si vuole piccola).
- Più basso è LR^- migliore è il test diagnostico. Infatti deve risultare basso il rapporto tra la probabilità di avere falsi negativi (che si vuole piccola) e la specificità del test (che si vuole grande).

- Si possono esprimere i valori predittivi positivo e negativo in funzione dei rapporti di probabilità e dell'odds di prevalenza come

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

- Ne segue che VPP aumenta in modo non lineare con la prevalenza e con maggior rapidità all'aumentare di LR^{+}, mentre VPN diminuisce con la prevalenza tanto più rapidamente quanto più è elevato LR^{-}.


1.0
0.5
VP+
LR+ = 16
LR+ = 8
LR+ = 4
LR+ = 2
LR+ = 1
0
0.5
1.0
Prevalenza


1.0
0.5
VP-
LR- = 16
LR- = 8
LR- = 4
LR- = 2
LR- = 1
0
0.5
1.0
Prevalenza

Esercizio 1

Tra i 40 e 50 anni la probabilità che una donna abbia il cancro al seno è 0.8%. Se una donna ha un cancro al seno la probabilità che il mammogramma sia positivo è del 90%. Se non ha il cancro al seno c'è comunque una probabilità del 7% che il mammogramma risulti positivo.

a) **Una donna fa una mammografia e il risultato è positivo. Calcolare il valore predittivo positivo.**

b) **Una donna fa una mammografia e il risultato è negativo. Calcolare il valore predittivo negativo.**

c) **Calcolare LR^+ e LR^-.**

a)Valore predittivo positivo

$P(+m)=P(m)\,P(+|m) = 0.008 \cdot 0.9 = 0.0072$

$P(+s)=P(s)\,P(+|s)= (1-0.008) \cdot 0.07 = 0.992 \cdot 0.07 = =0.06944$

$VPP=P(m|+) = 0.0072/(0.0072+0.06944)=0.0939$ **(9.4%)**

Il valore predittivo positivo vale 0.094. Possiamo affermare che la donna risultata positiva ha una probabilità di circa il 9% di essere ammalata. Occorre fare ulteriori indagini diagnostiche.

b)Valore predittivo negativo

$P(-s)=P(s)\ P(-|s)=(1-0.008)\cdot(1-0.07)=0.992\cdot0.93=0.9226$

$P(-m)=P(m)\ P(-|m)=0.008\cdot(1-0.9)=0.008\cdot0.1=0.0008$

$VPN=P(s|-)=0.9226/(0.9226+0.008)=0.9991$ (99.9%)

Il valore predittivo negativo vale 0.9991. Possiamo affermare che la donna risultata negativa ha una probabilità di circa il 99.91% di essere sana.

Poiché il valore predittivo negativo è molto alto, possiamo affermare che l'esame mammografico è un buon test di screening.

c) Calcolare LR+ e LR-.

$LR^{+}=0.9/0.07=12.86$

$LR^{-}=0.1/0.93=0.1075$

In definitiva per valutare la bontà di un test si può osservare il rapporto:

$$\mathbf{LR^{+}/LR^{-}}$$

Se il valore è maggiore di 50 il test è buono.

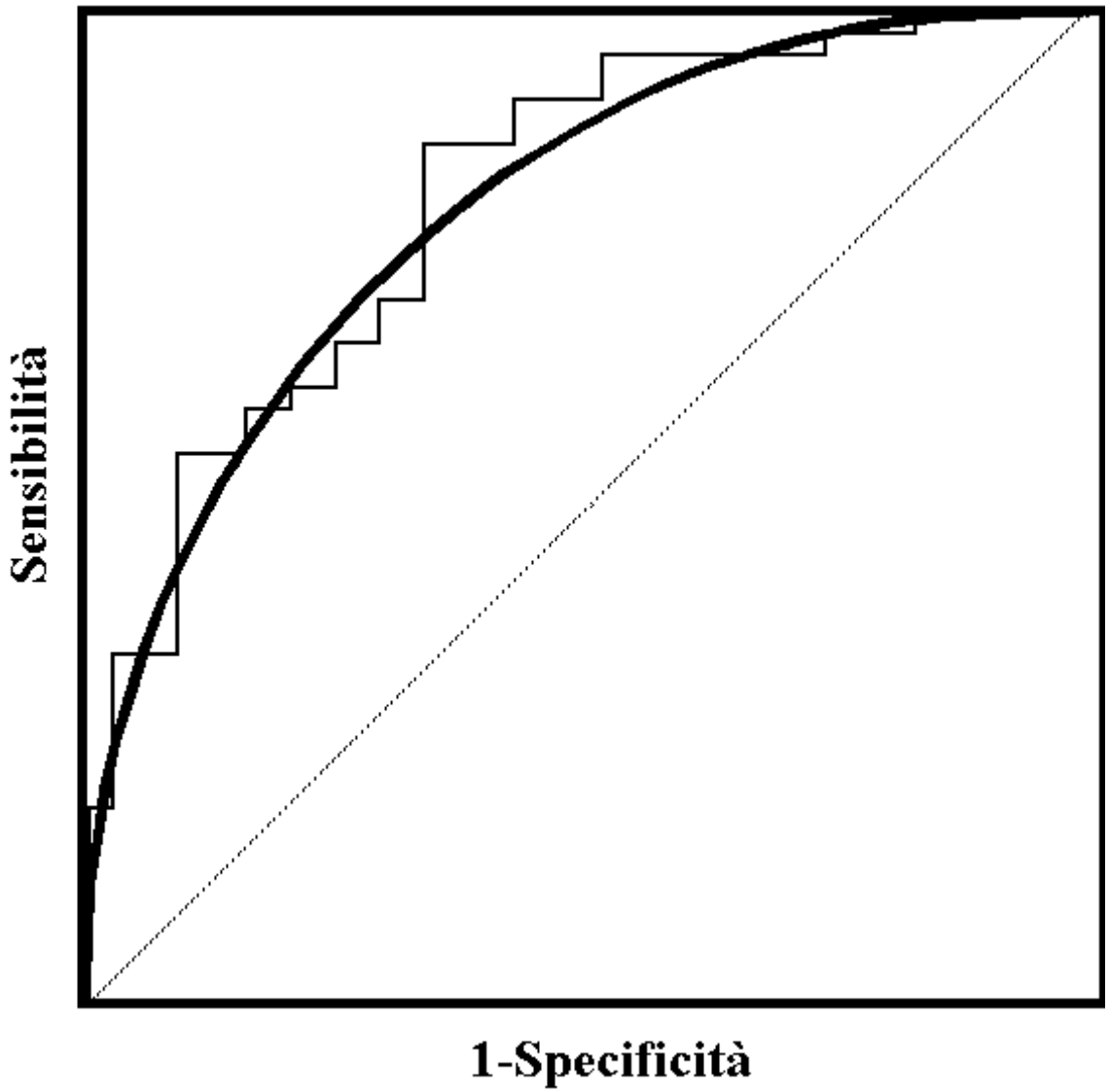
Nel nostro caso :

$LR^{+}/LR^{-}=119.6$ → Il test risulta buono.

Curva ROC

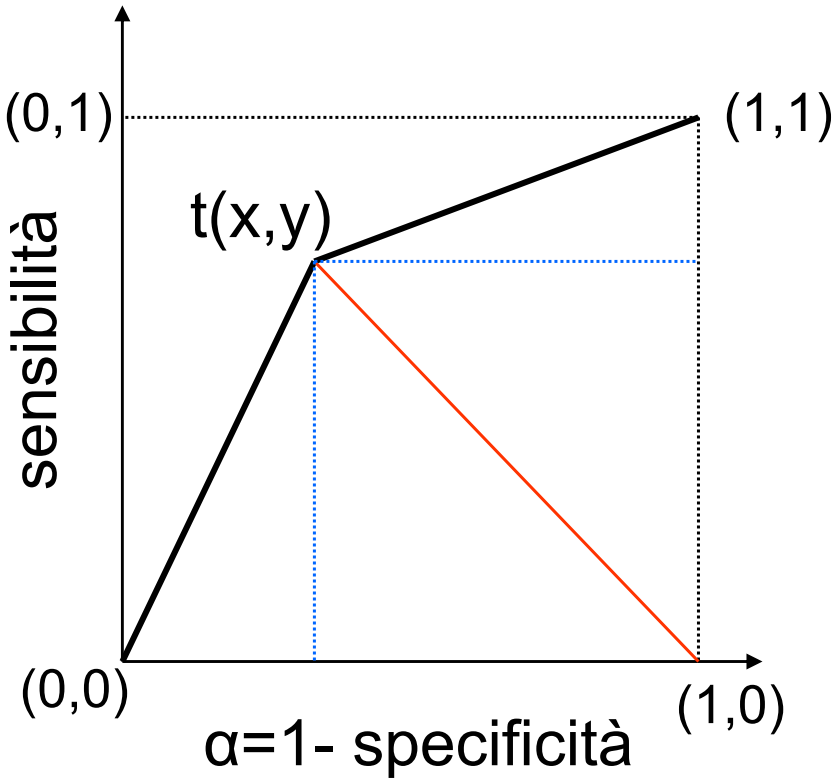
- La curva ROC (Receiver Operating Characteristic) è una curva utilizzata per la valutazione di un classificatore binario e consiste nella rappresentazione dell'andamento della sensibilità rispetto a (1-specificità)=α al variare del valore del cut off del test diagnostico.
- Deve il suo nome al fatto che è stata sviluppata per la prima volta durante la seconda guerra mondiale per riconoscere il segnale reale (oggetti nemici) captato dai radar (receiver) dal rumore di fondo.

- La curva ROC lega quindi la probabilità di ottenere un vero positivo tra le persone effettivamente malate (sensibilità) alla probabilità α di ottenere un falso positivo tra le persone sane al variare della soglia del test scelta (valore di cut off).
- Si può parlare di curva ROC se sensibilità e α variano con continuità.
- Solitamente però i dati disponibili sono discreti, per cui si ottiene sperimentalmente una curva spezzata (ROC plot) da cui la curva ROC è ottenuta per interpolazione (smoothing).


Sensibilità
1-Specificità

Capacità discriminante del test

- La capacità discriminante del test, cioè quanto bene il test eseguito riesce a distinguere tra sani e malati, è proporzionale all'area sottesa dalla curva ROC (AUC=Area Under Curve) ed equivale alla probabilità che un soggetto estratto a caso sia classificato correttamente come malato se positivo al test e come sano se negativo.



- Consideriamo infatti il generico punto t che si ottiene per un particolare valore di soglia del test che porta ad avere un ben determinato valore x di α e y di sensibilità.
- L'area del quadrilatero (AUC) si ottiene dalla scomposizione in due triangoli, entrambi di base 1 e altezza rispettivamente pari alla sensibilità P(+|m) e alla specificità P(-|s).

- Si considerano infatti eventi mutuamente escludentesi e quindi disgiunti (essere malati e classificati positivi ed essere sani e classificati negativi) e quindi la probabilità di essere classificati correttamente è data dalla somma delle due probabilità.

- Nel caso di un test perfetto che distingue completamente fra malati e sani (capacità discriminante del 100%), non si hanno né falsi positivi né falsi negativi, quindi $\alpha=\beta=0$.

  Il punto t si riduce in questo caso al punto di coordinate (0,1) e l'area sotto la curva corrisponde all'area del quadrato di coordinate (0,0),(0,1),(1,0),(1,1) vale AUC=1.
- Risulta evidente che il test sarà tanto migliore quanto più t è vicino al punto (0,1).
- Il ROC plot si ottiene dai diversi punti t in funzione della soglia del test.

- La curva ROC corrispondente alla diagonale tra i punti (0,0) e (1,1) del quadrato con AUC=1 rappresenta un test di valore informativo nullo in quanto $1-\beta=\alpha \Rightarrow P(+|m)=P(+|s)$

  Essa è detta *chance line* e si ha AUC=0.5.
- Negli altri casi, il punto della curva più vicino all'angolo superiore sinistro del quadrato rappresenta il miglior compromesso fra sensibilità e specificità e si ha $1-\beta>\alpha \Rightarrow P(+|m)>P(+|s)$ con AUC>0.5.
- Se $1-\beta<\alpha \Rightarrow P(+|m)<P(+|s)$ i falsi positivi sono più dei veri positivi e quindi il test non ha senso.

Classificazione della capacità discriminante di un test

La classificazione della capacità discriminante di un test proposta da Swets (1998) è la seguente:

AUC<0.5	Il test non ha senso
AUC=0.5	Test non informativo
0.5<AUC≤0.7	Test poco accurato
0.7<AUC≤0.9	Test moderatamente accurato
0.9<AUC<1.0	Test altamente accurato
AUC=1.0	Test perfetto

- Sotto l'ipotesi di curve ROC proprie in cui AUC può essere considerata distribuita in modo approssimativamente normale, è possibile testare la significatività della capacità discriminante del test mediante la formulazione delle seguenti ipotesi statistiche:

$$H_0 : E(AUC) = 0.5$$

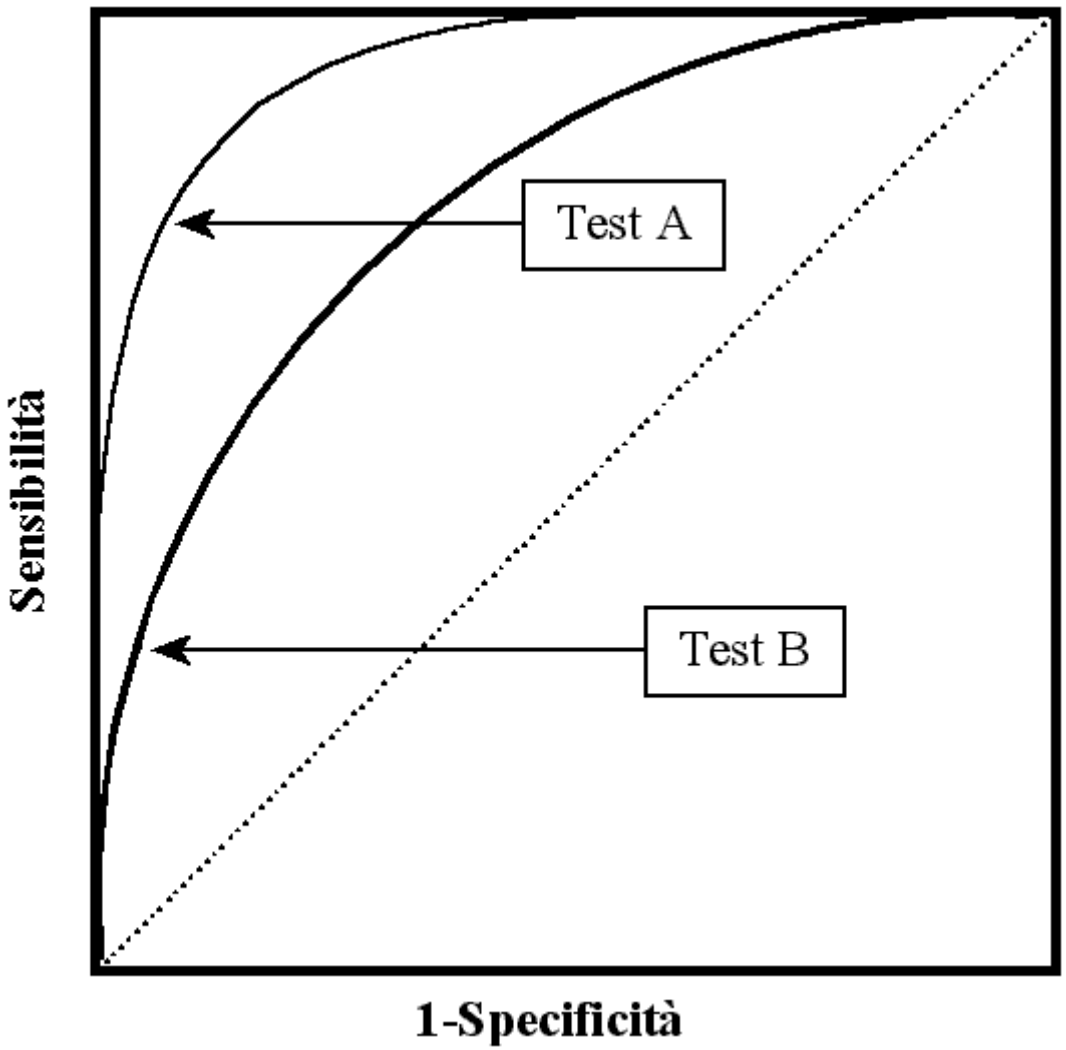
$$H_1 : E(AUC) > 0.5$$

- E' inoltre possibile confrontare due test mediante un test di significatività con le ipotesi:

$$H_0 : E(AUC_A - AUC_B) = 0$$

$$H_1 : E(AUC_A - AUC_B) \neq 0$$

Confronto di due test diagnostici tramite ROC



In questo caso risulta evidente la superiorità del test A la cui curva ROC si trova interamente al di sopra di quella corrispondente al test B. Con un test statistico si può verificare se la differenza risulta significativa.

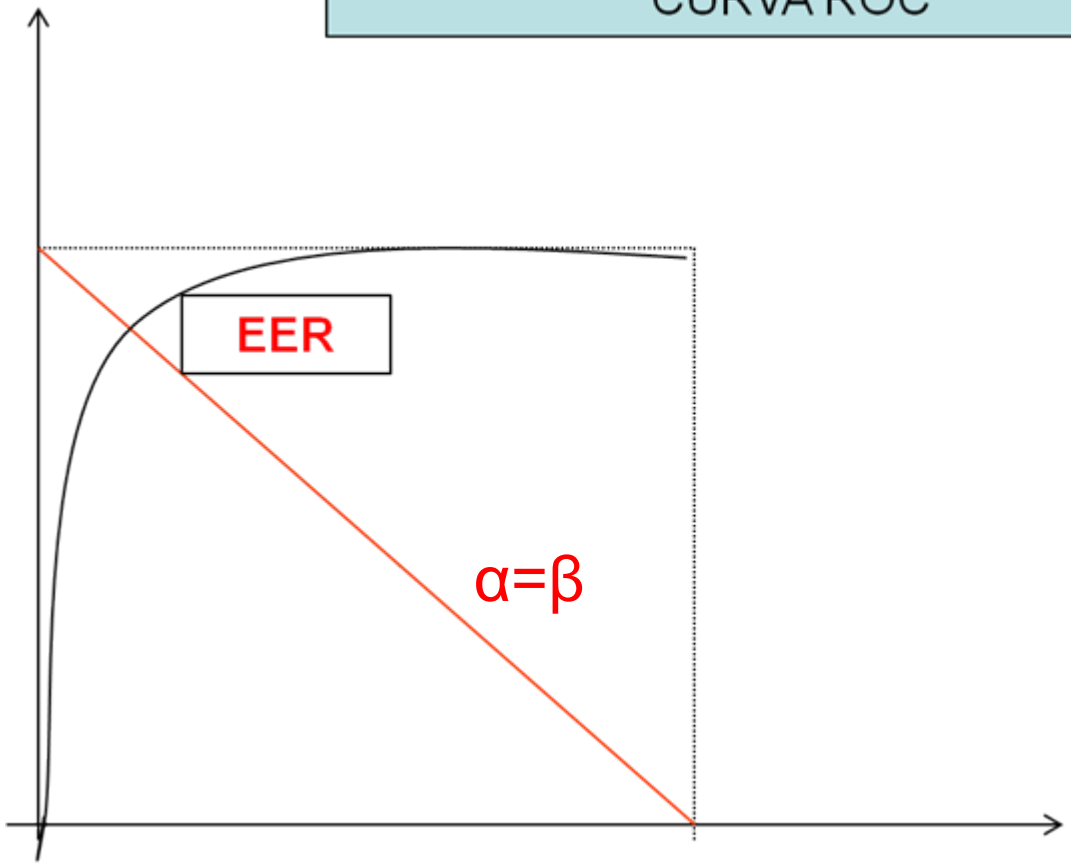
Equal Error Rate EER

E' detto anche Crossover Error Rate (CER) e rappresenta il tasso di errore che caratterizza un test la cui soglia di decisione viene fissata in modo che la proporzione di falsi positivi sia approssimativamente uguale a quella dei falsi negativi: $\alpha = \beta \Rightarrow 1-\alpha = 1-\beta \Rightarrow specificità = sensibilità$

Corrisponde al taglio ottimale che si ottiene dal punto di incontro delle curve della specificità e della sensibilità.

Poichè le curve ROC solitamente vengono costruite per punti, l'EER solitamente non può essere calcolato in modo preciso, ma lo si stima usando il punto più vicino a quello in cui la retta che unisce i punti (0,1) e (1,0) incontra la curva ROC.

CURVA ROC



Esempio 2

Si vuole scegliere il punto di cut off in un test sulla valutazione della ferritina serica nell'organismo. In un gruppo di malati di Anemia per carenza di ferro (IDA: iron deficiency anemia) e di sani si misura la ferritina ottenendo la seguente tabella:

ferritina serica (mmol/l)	con IDA (% of total)	senza IDA (% of total)
< 15	474	20
15-34	175	79
35-64	82	171
65-94	30	168
> 94	48	1332

Il test migliore è quello che riesce a discriminare fra sani e malati e quindi con alta sensibilità e specificità. Scegliendo ≤34 come valore anomalo di ferritina per separare i malati dai sani otteniamo la seguente tabella:

ferritina serica (mmol/l)	con IDA (% of total)	senza IDA (% of total)
<= 34	474 + 175	20 + 79
> 34	82 + 30 + 48	171 + 168 + 1332

Rispetto al cutpoint <15 aumentano i VP di 175 unità ma diminuiscono i VN di 79 unità e quindi diminuisce la specificità.

Sensibilità e specificità per cutpoint 34

ferritina serica (mmol/l)	con IDA (% of total)	senza IDA (% of total)
<= 34	649	99
> 34	160	1671

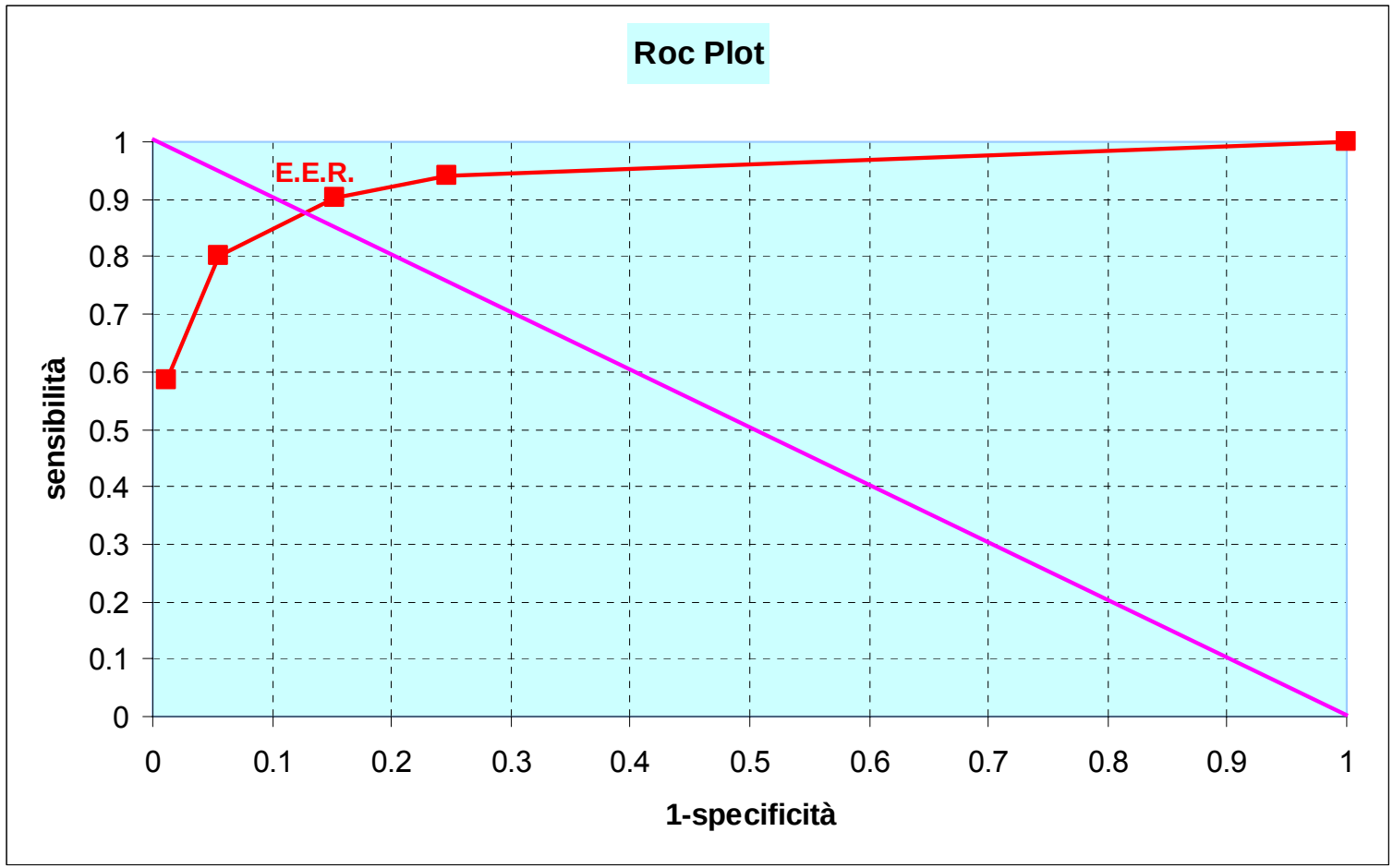
La sensibilità e la specificità per questo cutpoint o punto di cutoff 34 sono:

Sensibilità:P(+|m) = 649 / (649 + 160) = 80.2%

Specificità: P(-|s)= = 1671 / (1671 + 99) =94.4%

Ad ogni punto differente corrisponde una diversa sensibilità e specificità. Calcoliamo la sensibilità e la specificità considerando come punti di cut off i 4 valori di ferritina:

Cutpoint che indica un valore anomalo di ferritina (mmol/l)	Sensibilità	Specificità
< 15	58.6%	98.9%
<= 34	80.2%	94.4%
<= 64	90.4%	84.7%
<= 94	94.1%	75.3%



L’area sotto la curva ROC risulta pari a 0.9344 (93.4%) e quindi il test risulta altamente accurato

L’incontro tra la diagonale del quadrato e la curva ROC indica il punto in cui la frequenza dei falsi negativi è uguale alla frequenza dei falsi positivi: nel nostro caso 64 è il valore del cut off che più si avvicina a quel punto. Esso corrisponde al cut off ottimale che minimizza gli errori di classificazione.