



**Sensibilità, Specificità  
e identificare una condizione di interesse**



## Identificare una condizione di interesse

Nelle slides precedenti abbiamo visto la metodologia che si usa per identificare un deficit (o un danno) cognitivo.

Data la difficoltà metodologica/epistemologica per definire il deficit si usa un ragionamento inverso, a partire dall'assenza di deficit.

Esistono però situazioni in cui è possibile identificare la condizione di interesse (o criterio), tramite delle procedure, magari costose o invasive, che però permetterò di avere una classificazione che riteniamo il nostro riferimento “vero”.

Il test/metodo/procedura diagnostica utilizzati per ottenere la nostra classificazione di riferimento è detto **gold-standard**.



## Identificare una condizione di interesse

Nel caso in cui sia possibile distinguere due gruppi di interesse, l'obiettivo di un test diventa identificare una soglia che ci permetta di discriminare correttamente i due<sup>1</sup> gruppi, in accordo al gold-standard.

Nelle prossime slides faremo alcuni esempi delle classificazioni *una volta che è nota questa soglia*. Spiegheremo in seguito come ottenere questa soglia.

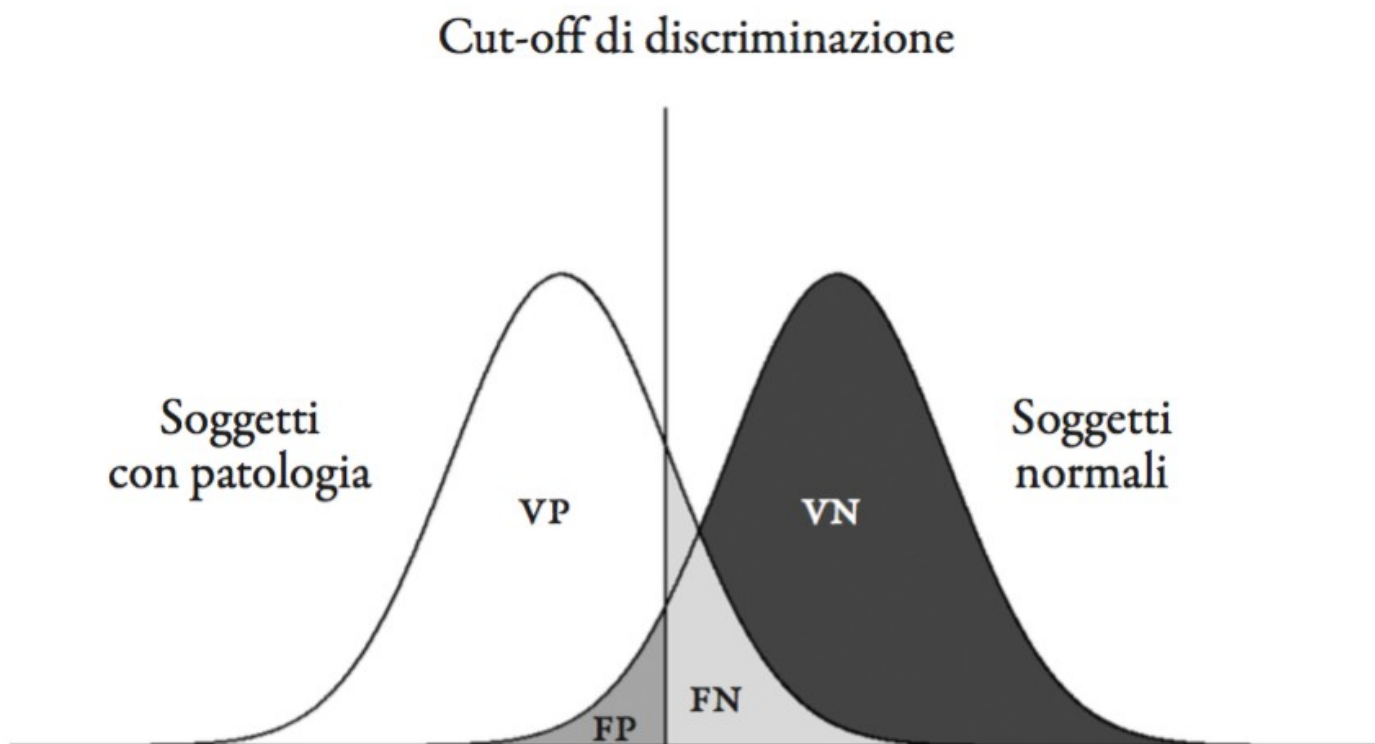
Per distinguere dal *cut-off di normalità* chiameremo questa soglia ***cut-off di discriminazione***.

1 è possibile anche pensare a casi con tre o più gruppi, ma spesso sono ricondotti a confronti fra due. Nelle slides che seguono ci concentreremo sulla più comune discriminazione tra due gruppi.

## Matrice di Confusione

I possibili risultati che si possono ottenere con un test (in riga), rispetto alla classificazione del gold-standard (in colonna), danno origine a quella che è chiamata “matrice di confusione”.

	Condizione di interesse PRESENTE (es. patologia)	Condizione di interesse ASSENTE (es. controllo)	Totali di colonna
Test positivo	Vero positivo (A)	Falso positivo (B)	A+B
Test negativo	Falso negativo (C)	Vero negativo (D)	C+D
Totali di riga	A+C	B+D	





## Sensibilità e specificità

A partire dai risultati della matrice di confusione possiamo calcolare delle importanti misure.

**Sensibilità:**  $VP / VP + FN$

È la probabilità di identificare correttamente un caso positivo (secondo il gold standard), se la persona è positiva al gold standard.

**Specificità:**  $VN / VN + FP$

È la probabilità di identificare correttamente un caso negativo (secondo il gold standard), se la persona è negativa al gold standard



## Sensibilità e specificità

Un esempio più “neuropsicologico clinico”

**Sensibilità:**  $VP / VP + FN$

È la probabilità di identificare correttamente un paziente con demenza (secondo il gold standard), se la persona ha la demenza, secondo il gold-standard.

**Specificità:**  $VN / VN + FP$

È la probabilità di identificare correttamente un sano (secondo il gold standard), se la persona non ha demenza (secondo gold standard).

## Sensibilità e specificità

Un esempio più “neuropsicologico forense”

**Sensibilità:**  $VP / VP + FN$

È la probabilità di identificare correttamente un simulatore (classificato come tale secondo il gold standard), se sappiamo che quella è un simulatore al gold standard.

**Specificità:**  $VN / VN + FP$

È la probabilità di identificare correttamente una persona non simulatrice (classificata secondo il gold standard), se sappiamo che non è un simulatore al gold-standard.





## Altre misure

Nota che esistono altre misure che possono essere calcolate a partire dai soli dati della matrice di confusione, ma sono meno comuni di Sensibilità e Specificità

$$\text{Accuratezza} = (TN + TP) / (TN + TP + FN + FP)$$



## **Decidere la soglia ottimale**

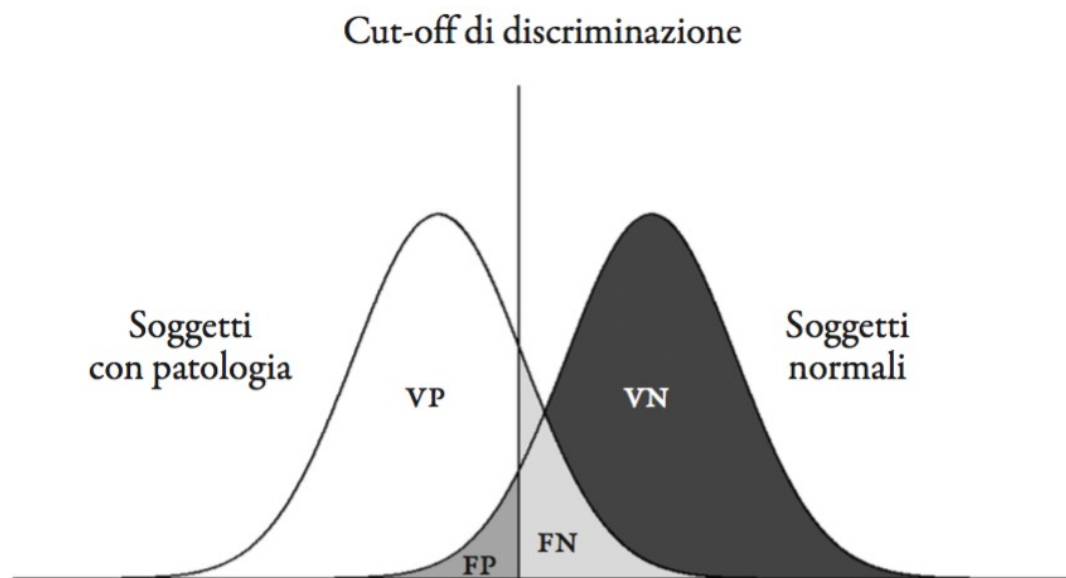
I risultati di sensibilità e specificità sono interconnessi e la soglia critica può essere definita in vari modi.

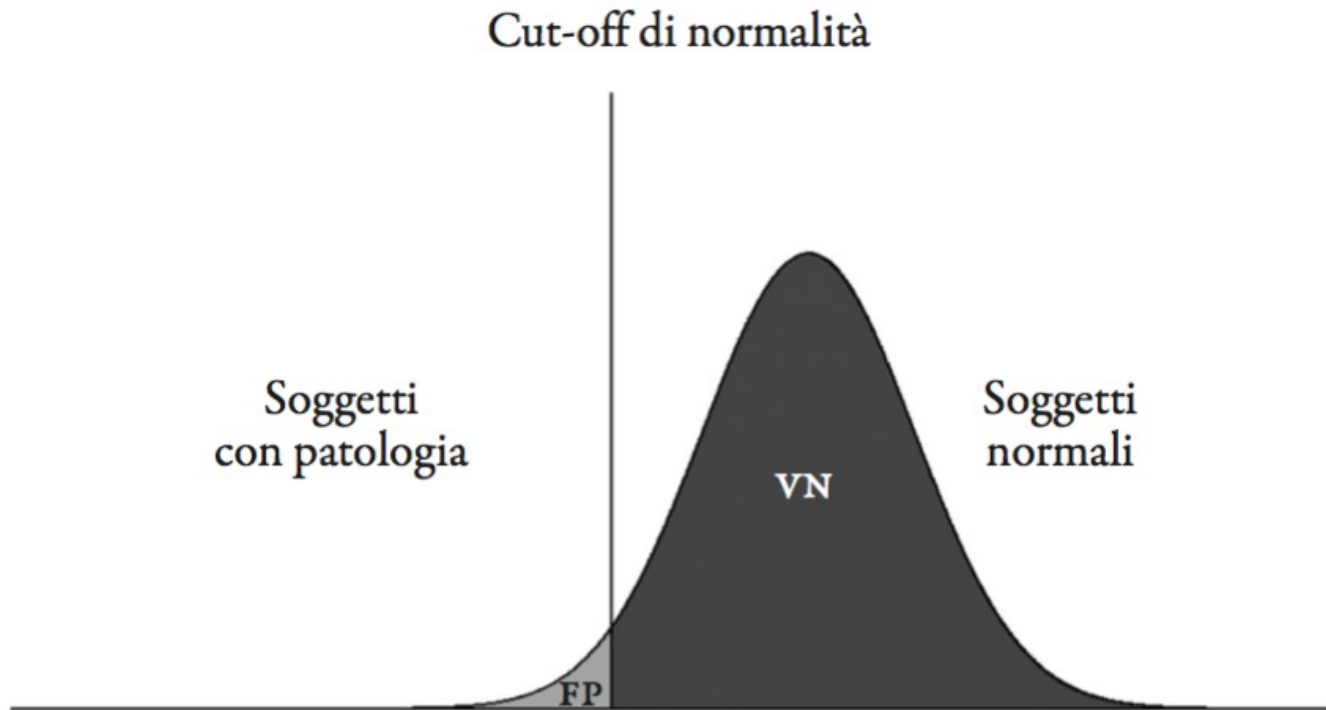
Abbassando infatti la specificità si può alzare la sensibilità e viceversa. La soglia ottimale può dipendere dagli scopi.

Idealmente un test dovrebbe avere sia un'alta sensibilità e alta specificità.

Immaginate di spostare a sinistra o a destra la linea che indica il cut-off di discriminazione, ovviamente si modificherebbero tutte le percentuali: VP, VN, FP ed FN.

Per esempio spostando a sinistra la linea (cioè rendendo il cut-off di discriminazione più basso, si diminuiscono i FP, ma anche i VP! E aumentano i VN e FN.





Rappresentazione di un'ipotetica distribuzione dei punteggi di soggetti normali. FN = falsi negativi, valori al di sopra del cut-off di soggetti patologici; VN = veri negativi, valori al di sopra del cut-off di soggetti che sono normali. Si confrontino i possibili risultati di questo caso con quelli di un test con cut-off di discriminazione, rappresentati nella FIG. 7.I.

## Specificità e cut-off di normalità

È interessante notare come l'utilizzo di un cut-off di normalità (vedi slides valutare I deficit cognitivi), può essere immaginata come una condizione in cui non abbiamo la distribuzione dei punteggi di chi ha un deficit/danno, ma sappiamo che verosimilmente saranno più a sinistra (o in generale, peggiori) di quelli che non hanno un deficit/danno

Non sapendo dove mettere la soglia, la mettiamo nel punto che delimita 5% di campione/popolazione per I sani.

## Receiver Operating Characteristic (ROC) Curve

I possibili risultati di sensibilità e specificità sono indagati (e rappresentati graficamente) tramite la cosiddetta ROC curve.s

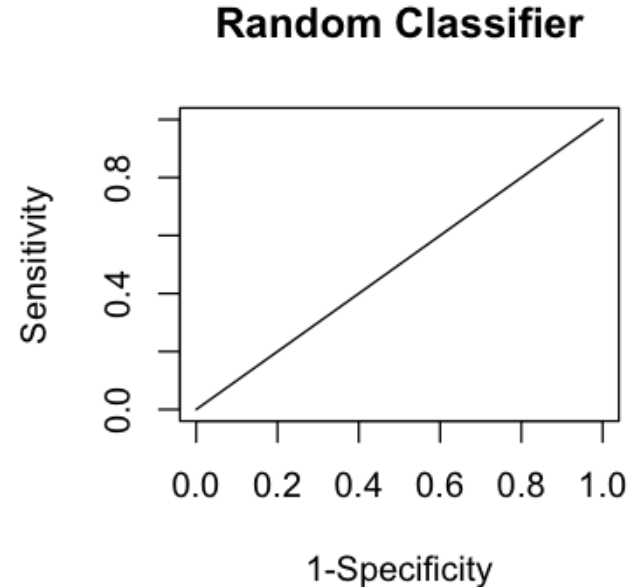
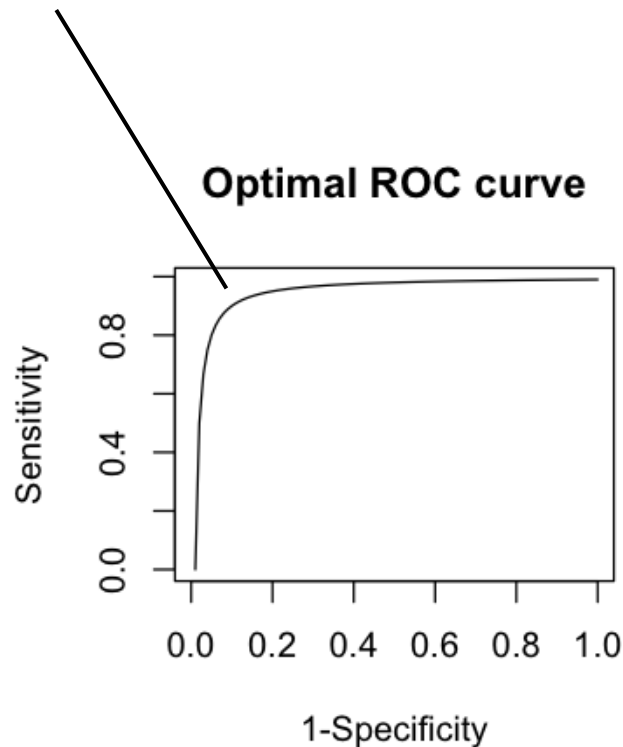
Che rappresenta sensibilità e 1-specificità in un piano, al variare delle possibili soglie.

Una ROC curve ottimale dovrebbe un angolo in alto a sinistra.

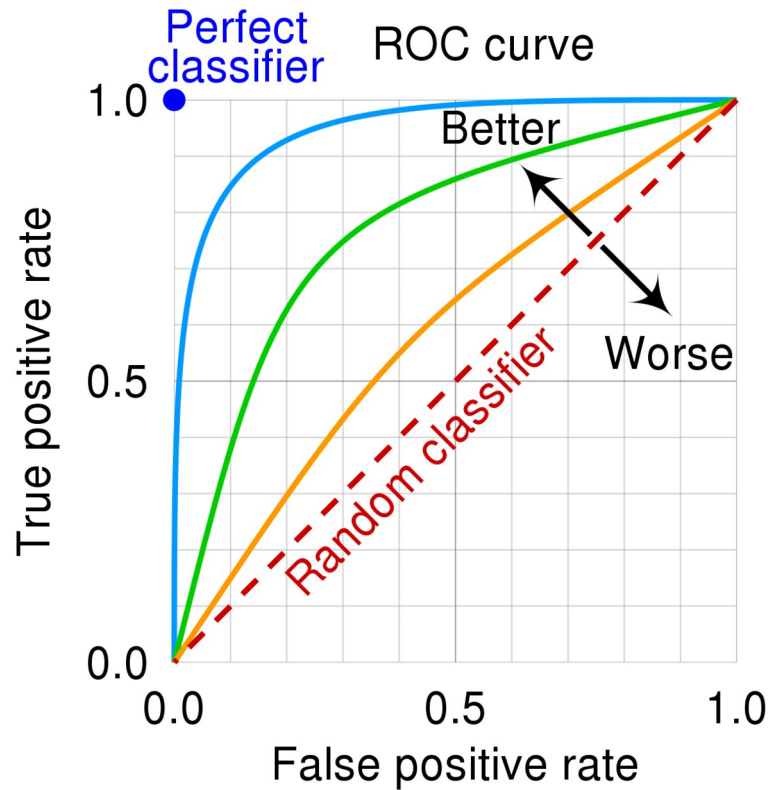
Quanto più la curva ROC si avvicina ad una diagonale, quanto più si comporta come un “random classifier”, indipendentemente dal valore di soglia, la classificazione corrisponde al lanciare una moneta e decidere l'appartenenza ad uno dei due gruppi.

## Receiver Operating Characteristic (ROC) Curve

Ogni punto della curva corrisponde ad un possibile valore nel test in questione (es. 15 nel MoCA)



## Receiver Operating Characteristic (ROC) Curve



[https://en.wikipedia.org/wiki/Receiver\\_operating\\_characteristic](https://en.wikipedia.org/wiki/Receiver_operating_characteristic)



## Receiver Operating Characteristic (ROC) Curve

In genere una soglia ottimale è il punto **più in alto a sinistra** e cioè con più alta sensibilità e specificità, oppure il valore in cui la media di sensibilità e specificità è massimo.

Ci sono casi (non rari) in cui le soglie ottimali potrebbero essere due.

Es.

- Una soglia con 0.92 sensibilità e 0.93 specificità
- E una soglia con 0.93 sensibilità e 0.92 sensibilità

In tal caso la soglia maggiore viene scelta rispetto alle necessità del test (privilegiando sensibilità o specificità).



## Receiver Operating Characteristic (ROC) Curve

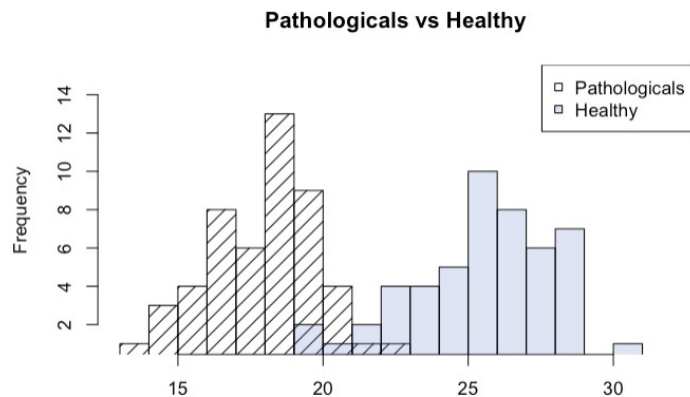
Per avere una misura generale (al di là della specifica soglia) di quanto bene funziona il nostro test si usa come misura l'area totale sotto la curva. (Area Under Curve, **AUC**), dove 1 rappresenta l'intera superficie delimitata dagli assi del grafico della curva ROC.

Un'AUC di 0 indica dunque un random classifier (perché la diagonale divide a metà lo spazio delimitato dagli assi).

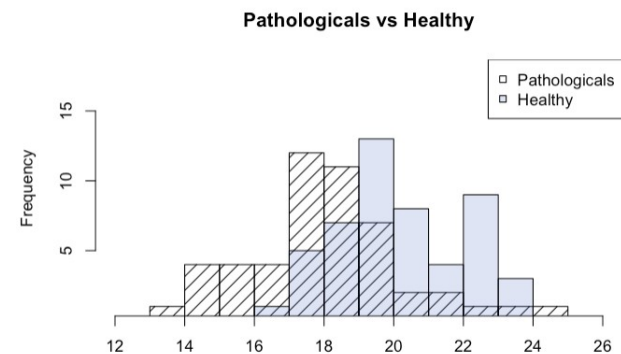
È da notare che quanto più le due distribuzioni di dati saranno separate. Tanto migliore saranno sensibilità e specificità del mio test (perché saranno meno sia i FP sia i FN).

In sostanza sarà possibile calcolare una soglia che mi discrimina bene (cut-off di discriminazione)

Se invece le distribuzioni dei due gruppi sono troppo sovrapposte è possibile che invece che Sensibilità e Specificità siano molto basse (random classifier).



Buona discriminabilità tra gruppi,  
Avremo una buona curva ROC e alta AUC



Scarsa discriminabilità tra gruppi,  
Avremo una scarsa curva ROC e bassa AUC



È da notare anche che a differenza dei test che si basano su cut-off di normalità, ***per I test che mirano ad avere un cut-off di discriminazione potrebbe non esistere una soglia adeguata alla discriminazione.***

Questo problema non c'è per I test con cut-off di normalità perché lì potremo sempre delimitare il 5% delle prestazioni peggiori (la verità è che non sappiamo dove si trova la distribuzione con deficit o danno).




## **Sensibilità/Specificità sono proprietà del test non della valutazione**

Non proprietà della valutazione. Quella infatti dipenda dalle realtà (metafisica) della situazione.

Potrebbe avere senso se si parlasse del futuro (es. Probabilità di sviluppare l'alzheimer entro 5 anni), ma spesso la realtà c'è già (solo non sappiamo quale è)

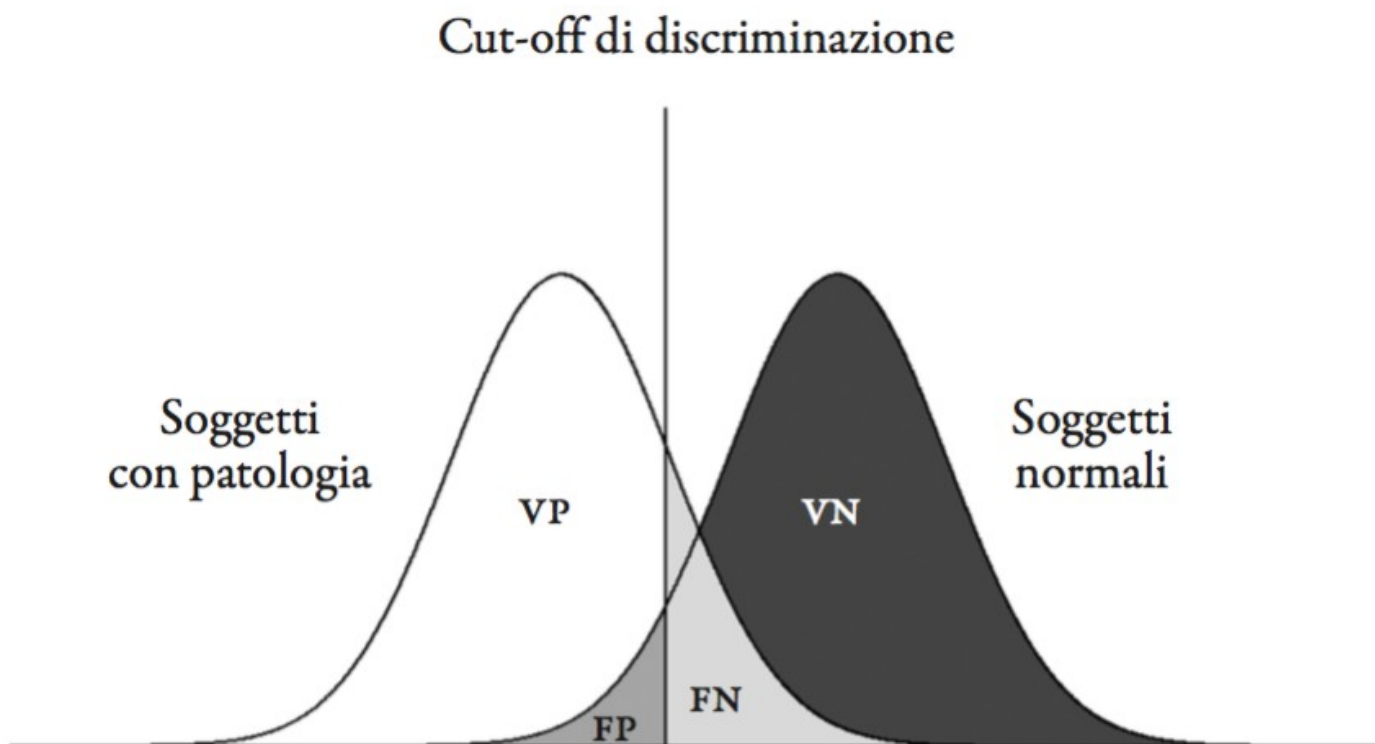
Es. Non ha senso dire ho 10% di Probabilità di avere Alzheimer (o ce l'ho o non ce l'ho). Occorre sempre non “girare” le probabilità. Il test mi dice il 10% delle persone con questa performance sviluppano Alzheimer. L'esempio è più semplice se si usano condizioni tipo “tumore” (o c'è oppure no)

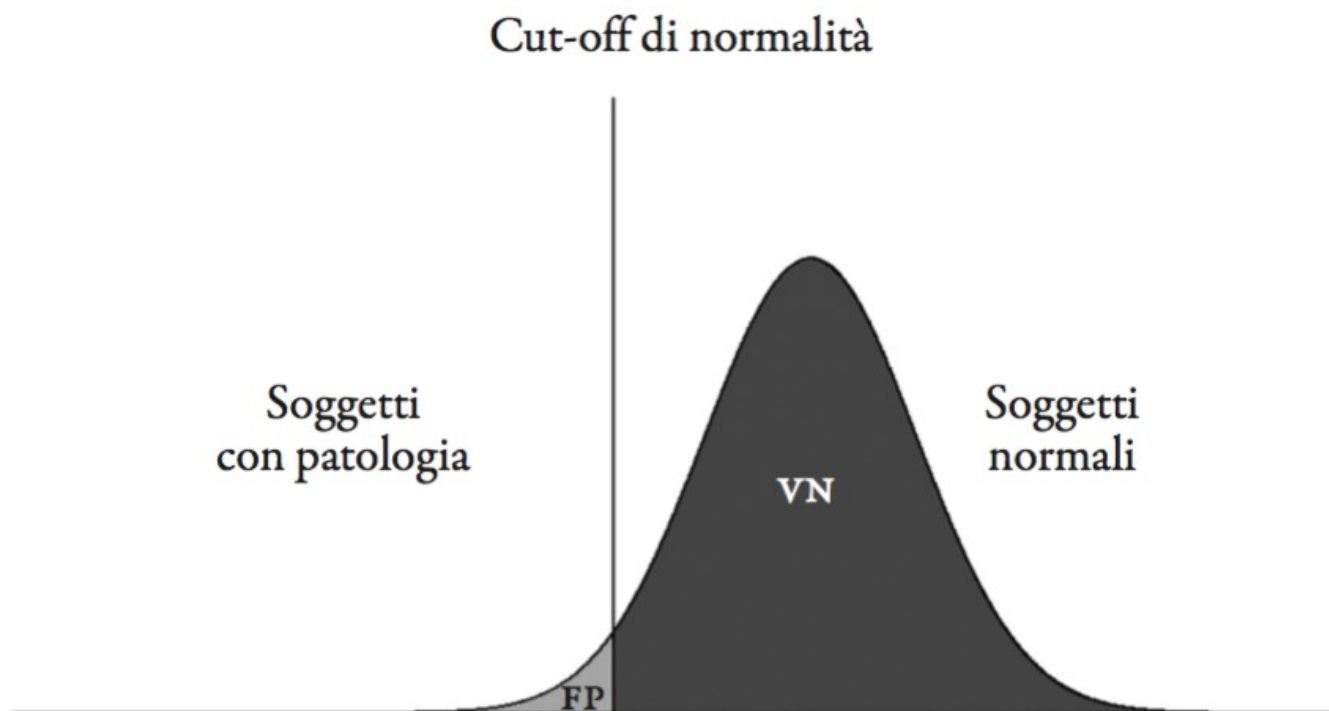


È da notare che i test in cui è possibile valutare sensibilità/specificità (cioè c'è un gold-standard), spesso sono validati secondo **validità di criterio** (vedi anche slides su validità).

In questo caso, spesso il concetto di validità di costrutto o su altre validità decade un po'. L'importante è infatti non tanto cosa si sta misurando, ma l'efficacia che risulta nella classificazione

Rimangono invece valide e importanti le considerazioni su affidabilità del test.





Rappresentazione di un'ipotetica distribuzione dei punteggi di soggetti normali. FN = falsi negativi, valori al di sopra del cut-off di soggetti patologici; VN = veri negativi, valori al di sopra del cut-off di soggetti che sono normali. Si confrontino i possibili risultati di questo caso con quelli di un test con cut-off di discriminazione, rappresentati nella FIG. 7.1.



## Limiti Sensibilità e Specificità

Sensibilità e Specificità ci dicono cose importanti:

- Assumendo che sono veramente positivo quale è la probabilità di essere sotto soglia al mio test?
- Assumendo che sono veramente negativo quale è la probabilità di essere sopra soglia?

Questo è utile, ma è molto simile al limite che avevamo per il cut-off di normalità, Facciamo un ragionamento “al contrario”, ***dobbiamo assumere cosa siamo e vedere se siamo sotto soglia.***

- $P(\text{test sotto cut-off} \mid \text{Patologico})$
- $P(\text{test sopra cut-off} \mid \text{Sano})$

Ricordiamo che il cut-off di normalità era stato definito

$$P(\text{test sotto cut-off} \mid \text{Sano}) = 0.05$$

Idealmente sarebbe meglio avere una situazione al contrario:

$$P(\text{Patologico} \mid \text{test sotto-cut-off}) = ?$$

Questa però è molto difficile da calcolare (vedi slide dopo su PPV, NPV e Teorema di Bayes).

## Positive Predictive Value e Negative Predictive Value

**Positive Predictive Value (PPV)** =  $TP / (TP + FP)$

**Negative Predictive Value (NPV)** =  $TN / (TN + FN)$

il **PPV** indica la proporzione di persone realmente positive (al gold-standard), tra tutte quelle che risultano positive al test.

Es. il numero di reali persone con Alzheimer (secondo gold-standard), che risultano positive al test, cioè che vengono classificate come Alzheimer nel test che dovrebbe distinguere Alzheimer da Sani,

Il **NPV** indica invece la proporzione di persone realmente negative (al gold-standard), tra tutte quelle che risultano negative al test.

Es. il numero di reali persone Sane (secondo gold-standard), che risultano negative al test, cioè che vengono classificate come Sane nel test che dovrebbe distinguere Alzheimer da Sani,

[https://epidemiology.sruc.ac.uk/shiny/apps/predictive\\_values/](https://epidemiology.sruc.ac.uk/shiny/apps/predictive_values/)



## Positive Predictive Value e Negative Predictive Value

**RICORDA:** Come per Sensibilità e specificità, PPV e NPV sono concetti che si riferiscono al test e che si riferiscono ad una specifica soglia.

Cambiando soglia, noi cambiamo (potenzialmente) sia Sensibilità, sia Specificità, sia PPV, sia NPV.



## **Alcuni app utili**

[https://epidemiology.sruc.ac.uk/shiny/apps/predictive\\_values/](https://epidemiology.sruc.ac.uk/shiny/apps/predictive_values/)

<https://neurotroph.shinyapps.io/Sensitivity-Specificity/>

<https://micncltools.shinyapps.io/TestAccuracy/>

<https://micncltools.shinyapps.io/ClinicalAccuracyAndUtility/>

## Positive Predictive Value e Negative Predictive Value

Le formule di PPV e PPN possono (tramite semplice algebra rispetto a formule precedenti, riscritte nella seguente maniera

**Positive Predictive Value** =  $\text{Sensitivity} \times \text{prevalence} / (\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence}))$

**Negative Predictive Value** =  $\text{Specificity} \times (1 - \text{prevalence}) / \text{Specificity} \times (1 - \text{prevalence}) + (1 - \text{sensitivity}) \times \text{prevalence}$ .

Notare che spuntano Sensitivity e Specificity (che conosciamo), ma anche la “**prevalence**” (Cioè il tasso di base o base rate)

**Prevalence:**  $\text{Real Positive} / \text{Total population} = (TP + FN) / (TP + FP + TN + FN)$

*La prevalence è il numero totale di Positivi (rispetto al gold standard), sul totale del campione*



## Esperimento mentale

AT è un professionista di 35 anni che coordina uno studio di commercialisti. Da circa un anno i colleghi si lamentano di alcuni errori a lavoro. AT si dimentica a volte di appuntamenti importanti o di clienti. Il CdA cerca di togliergli il suo ruolo sostenendo di non avere le capacità cognitive per potere ricoprire questo ruolo, ma AT non vuole cederlo. In questa disputa il CdA richiede una valutazione e che somministra il MoCA.

Supponiamo che un cut-off di 18 al Moca ha 95% di Sensibilità e 92% Specificità per identificare Alzheimer da Controlli.

AT ottiene 16.

*Cosa concludereste?*



## Esperimento mentale

Siete periti di parte per AT, un professionista di 60 anni che coordina uno studio di commercialisti. In seguito ad un incidente stradale con trauma cranico, la sua capacità di concentrarsi si riduce drammaticamente e non è più in grado di seguire lo studio. Dopo circa due mesi dall'incidente l'assicurazione manda un perito che somministra il MoCA.

Supponiamo che un cut-off di 18 al Moca ha 95% di Sensibilità e 92% Specificità per identificare Alzheimer da Controlli.

TA ottiene 16.

*Cosa concludereste?*

## Esperimento mentale

L'esempio di prima mostra come validità e affidabilità non considerino alcuni aspetti di base, che sono relativi alla possibilità in partenza che l'individuo avesse l'alzheimer.

non è possibile calcolare questa probabilità per un caso specifico (e avrebbe anche poco senso), ma è possibile calcolarla *per un certo contesto*.

In tali casi tutte le valutazioni condotte in un certo contesto potrebbero essere caratterizzate da una diversa prevalenza, con implicazioni sulla valutazione.

Es. la probabilità di avere Alzheimer di persone che sono visitate in un Centro Valutazione Alzheimer è diversa dalla probabilità di avere l'Alzheimer in un contesto che valuta disturbi cognitivi in seguito ad un incidente stradale a fini assicurativi.

la probabilità di simulazione in un contesto di valutare a fini assicurativi, è diverso dalla probabilità di simulazione in un contesto di valutazione in un centro Valutazione Alzheimer, perchè verosimilmente, cambia la *prevalenza*.





## Importanza PPV e PPN

### Esempio COVID

- tamponi routinari in struttura non a rischio speciale (es. Ospedale di riabilitazione)
- tamponi in farmacia

Anche con test con altissima sensibilità e specificità in caso di test massivi c'è un certo rischio di veri positivi (con conseguenti problemi), che varia a seconda del contesto.

È probabilmente più facile che un tampone sia un falso positivo in ospedale rispetto che in farmacia. Questo perché in ospedale ci sono numerosi test fatti senza un quesito diagnostico, mentre chi arriva a fare valutazione in farmacia verosimilmente ha sintomi o ha avuto contatto con persone positive al COVID.



## Importanza PPV e NPV in pratica

In ambito forense molti test riportano i valori di PPV per diversi valori di prevalenza

Questo permette di esplorare vari scenari (realistici) e vedere come si classifica il nostro paziente.

Questo approccio è sensato, ma assume che i valori di prevalenza esplorati siano adeguati. Potrebbe però dare informazioni importanti: es. Che qualsiasi sia il valore di prevalenza esplorato non ci sono grosse differenze rispetto alle conclusioni da trarre (es. Che il paziente è verosimilmente negativo o verosimilmente positivo).

## Importanza PPV e NPV in pratica

**TABLE 5**  
**Cutting Scores, Sensitivity, and Specificity for the Control Versus Probable Malingering Groups**

<i>Reliable Digit Span Cutoff</i>	<i>Sensitivity</i>	<i>Specificity</i>
5	21	100
6	38	97
7	67	93
8	88	80

**TABLE 6**  
**Predictive Power of the Reliable Digit Span at Selected Base Rates**

<i>Base Rate</i>	<i>Cutoff</i>			
	5	6	7	8
Positive Predictive Power				
.5	1.0	.95	.92	.81
.4	1.0	.88	.87	.74
.3	1.0	.85	.80	.65
.2	1.0	.80	.68	.53
.1	1.0	.57	.54	.33
Negative Predictive Power				
.5	.56	.61	.75	.87
.4	.65	.70	.81	.91
.3	.74	.78	.87	.93
.2	.83	.87	.91	.97
.1	.92	.94	.97	.99

Es. da Mathias et al., 2002 , *Assessment*

Notare come un cut-off di 5 per una serie di scenari di prevalence (base rate) realistici su base di letteratura ha un PPV di 1. questo significa che persone che sono positive al test sono verosimilmente positive.



## Limiti Sensibilità, Specificità PPV e PPN

È importante notare che queste considerazioni (così come per Sensibilità e Specificità) si applicano al gruppo e al test, ma mai alla specifica misurazione

Sono probabilità riferite al test (Sensibilità e Specificità) o al test in combinazione al contesto (PPV e PPN).

Infatti per una specifica misurazione, a meno che non si voglia indagare qualcosa che avverrà nel futuro (es. Probabilità di sviluppare entro 1 anno l'Alzheimer) , per una specifica persona c'è una verità (es. per l'Alzheimer quella persona o ce l'ha, o non ce l'ha).

Le probabilità sono nel test di essere corretto quando sviluppato e immaginando un suo utilizzo ripetuto, ma *non* nel darci un'informazione corretta nella specifica misurazione.

## Una considerazione su probabilità

Se volessimo esprimere Sensibilità e specificità in termini di probabilità (come fatto per il cut-off di normalità). Potremmo usare una denotazione come quella che segue.

$$y: P(o < y | C) = k \quad \text{Sensibilità}$$

$$y: P(o \geq y | N) = j \quad \text{Specificità}$$

Dove  $y$  è il cut-off di discriminazione,  $o$  è il punteggio osservato.  $C$  è l'appartenenza al gruppo di persone con Condizione di Interesse (es. Patologici) e  $N$  appartenenza al gruppo di Normali (cioè sani).

$k$  = Sensibilità,  $j$  = specificità



## Un possibile errore nell'interpretare queste probabilità

I PPV e NPV (si vedano anche simulazioni del codice allegato), mostrano come l'intuizione su queste probabilità possano portare a degli errori. In particolare potrebbe essere spontaneo cercare di interpretare come se esse ci indicino questo<sup>\*</sup>

$$z : P(C | o < z) = k$$

“z è il valore tale che la probabilità di avere la condizione di controllo, se si ottiene un punteggio sotto cut-off è uguale a k” (NOTA: è diversa dalla definizione di Sensibilità! Vedi slide precedente)

Nelle prossime slides vedremo quale è il problema nell'ottenere questa probabilità.

<sup>\*</sup> nota che riflessioni simili sono state fatte su cut-off di normalità



# Teorema di Bayes

Il teorema di Bayes è un teorema molto importante che mette in relazione diverse probabilità. È importante perché esprime in maniera chiara la relazione tra probabilità che (intuitivamente) potrebbero essere confuse per identiche.



## Probabilità condizionali e teorema di Bayes

Il teorema di Bayes è un teorema molto importante che mette in relazione diverse probabilità. È importante perchè esprime in maniera chiara la relazione tra probabilità che (intuitivamente) potrebbero essere confuse per identiche.

$$P(A|B) \neq P(B|A)$$

$$P(\textit{patologia} | \textit{punteggio sotto cut - off}) \neq P(\textit{punteggio sotto cut - off} | \textit{patologia})$$

Quello che sarebbe ideale avere è la probabilità in arancione, ma quella che abbiamo è quella in blu



## Teorema di Bayes

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

$$P(patologia|test\ sotto\ cut-off) = \frac{P(patologia) \cdot P(test\ sotto\ cut-off|patologia)}{P(test\ sotto\ cut-off)}$$

## Teorema di Bayes

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

$$P(patologia|test\ sotto\ cut-off) = \frac{P(patologia) \cdot P(test\ sotto\ cut-off|patologia)}{P(test\ sotto\ cut-off)}$$

$$P(patologia|test\ sotto\ cut-off) = \frac{3\% \cdot 95\%}{8\%} = 35\%$$



## Teorema di Bayes

Al di là dell'aspetto matematico, il teorema di Bayes ci dice in sostanza una cosa:  
Per calcolare le probabilità che a noi veramente interessa dovremmo conoscere altre Probabilità che di fatto non potremo sapere mai (cioè Probabilità a Priori, che è anche associata a Prevalenza, se si fa riferimento ai concetti di PPV e PPN

Per questa ragione i test ci danno informazioni “indicative” sulle probabilità di avere una malattia. Ma sta al neuropsicologo clinico giungere alla sua diagnosi, anche ignorando il risultato di un test, se ci sono elementi che suggeriscono che probabilità a priori (o base rate) possano alterare i valori. Questo verrà approfondito nella sezione della interpretazione.



## Il gruppo di riferimento per Sensibilità e specificità

Idealmente vorremmo che il gruppo di sani e di pazienti sia composto da persone *il più simili possibili* all'individuo che vogliamo valutare, perchè solo in quella maniera il punteggio atteso (sia nello scenario che abbia la condizione di interesse, sia nel caso non l'abbia) sia possibile.

Esistono anche metodi che permettono di calcolare curve ROC includendo covariate (cioè altre variabili che possono influenzare la soglia e quindi Sensibilità e Specificità. Questo è una cosa del tutto analoga ai metodi di regressione per calcolare il cut-off di normalità. Questi metodi però non sono state praticamente mai usate in neuropsicologia clinica e forense e non saranno approfonditi (Sarebbero utili, però).



## Considerazioni conclusive (1/2)

In quanto riportato sull'identificazione di condizioni di interesse valgono molte delle considerazioni già fatte nel contesto di deficit/danno.

Anche per Sensibilità, Specificità PPV, etc. vale il fatto che sono tanto più precise quanto più alta è la numerosità campionaria (perchè più rappresentativa della popolazione).

I mio cut-off di discriminazione potrebbe essere diverso da quello della popolazione (che è quello che mi interessa), e tanto più grande è il mio campione tanto più sicuro sono che il mio cut-off sarà simile a quello della popolazione.



## Considerazioni conclusive (2/2)

È da considerare che spesso nel caso di test con condizione di interesse non sono disponibili cut-off che tengono conto di variabili (es. Età, scolarità, sesso, etc.) perchè i metodi più comunemente usati per le analisi ROC non le considerano.

Per tali ragione valgono tutta una serie di limiti legati alla rappresentatività del campione considerato, che va considerato quando si *interpretano* i risultati.

Ogni altra considerazione rimane pertinente (es. Meglio norme paese-specifiche, norme recenti, etc.)