VALIDAZIONE STATISTICA DI UN TEST DIAGNOSTICO
Pilloline di Statistica Biomedica

Colonna S.
Spine Center - Bologna
Scuola di Osteopatia OSCE (Osteopatic Spine Center Education) - Bologna

La performance di un'indagine diagnostica corrisponde complessivamente al suo grado di accuratezza, ovvero alla capacità di identificare come positivi all'indagine i soggetti affetti da una data malattia e come negativi all'indagine i soggetti che, invece, non ne sono affetti. Sono classicamente definiti come indici di performance diagnostica quelli che misurano, in modi diversi, tale performance e studi di performance diagnostica gli studi che hanno come finalità la misura della performance di un'indagine o, spesso, il confronto tra le performance di più indagini.

Gli indici fondamentali di performance diagnostica per valutare la validità dei test manuali più utilizzati nella letteratura medica sono:

- Specificità e Sensibilità
- Valore predittivo (positivo e negativo)
- Ripetibilità e Riproducibilità

• Specificità e Sensibilità

La sensibilità e la specificità sono due criteri che vengono impiegati per valutare la capacità di un test di individuare, fra le unità di una popolazione, quelle provviste del «carattere» ricercato e quelle che invece ne sono prive. Quasi sempre il carattere ricercato è rappresentato dalla presenza di una patologia; quindi possiamo dire che la sensibilità e la specificità servono a valutare la capacità di un test di individuare i soggetti sani da quelli malati.
La sensibilità di un esame è la probabilità che un soggetto malato presenti un risultato positivo. Un esame è cioè sensibile al 100% quando tutti i malati risultano positivi.
La specificità di un esame è la probabilità che un soggetto sano presenti un risultato negativo. Più bassa è la specificità più è alta la proporzione dei dei falsi positivi, cioè soggetti sani che risultano positivi al test. Un esame è specifico al 100% quando tutti i sani risultano negativi.
È chiaro che un test sensibile e specifico al 100% non lascerebbe dubbi. Purtroppo molti test possono presentare falsi positivi (soggetti sani che risultano positivi) e falsi negativi (cioè soggetti malati che invece risultano negativi).
Per fare un esempio pratico valuteremo la sensibilità di un nuovo ipotetico test che ci indica se il tendine del sovraspinato è rotto oppure no. Per avere la certezza della lesione lo paragoneremo ai risultati ottenuti da un'artroscopia che ci farà una fotografia reale della situazione.
Ad esempio valutiamo 100 soggetti con male e deficit funzionale alla spalla e di questi il nostro test individua 76 soggetti con rottura del sovraspinato (patologici) e 24 senza rottura (sani). In realtà noi

non sappiamo in quei 76 ci possono essere dei falsi positivi (soggetti in realtà sani) e nei restanti 24 ci possono essere dei falsi negativi (soggetti in realtà patologici)
Eseguendo l'artroscopia scopriamo che di soggetti con lesione del tendine ne sono 80 e 20 senza lesione; ma nel campione dei sani (20) sono inclusi 4 soggetti che il test dava per positivi (falso positivo), mentre sono inclusi nel campione dei patologici (80) 8 soggetti che il test li dava negativi (falsi negativi).

	Positivi all'artroscopia	Negativi all'artroscopia	Totale
Soggetti positivi al test	Veri positivi (72)	Falsi positivi (4)	76
Soggetti negativi al test	Falsi negativi (8)	Veri negativi (16)	24
Totale	80	20	

Esempio di tabella di contingenza applicata al nostro esempio

Per valutare la Sensibilità del nuovo test la formula sarà: veri positivi/ (veri positivi + falsi negativi).
Quindi la Sensibilità sarà 72/(72+8) = 0.90
Per valutare la Specificità del nuovo test la formula sarà: veri negativi/ veri negativi+ falsi positivi.
Quindi la specificità sarà 16/(16+4) = 0.80

I dati della sensibilità e specificità possono essere riportati come variabile era 0 a 1 o come più spesso avviene in percentuale. Nel nostro caso, quindi, avremo che il nuovo test ha una sensibilità del 90% e una specificità del 80%. Il che vuol dire che questo ipotetico test riesce ad individuare meglio i casi patologici rispetto ai sani.

- **Valore Predittivo**

I valori predittivi si dividono in: valore predittivo positivo e valore predittivo negativo.
Il valore predittivo positivo di un test è la probabilità di essere malati dei soggetti risultati positivi al test studiato. La formula per calcolarlo è: veri positivi/ (veri positivi + falsi postivi). Utilizzando sempre il caso dell'ipotetico test della spalla avremo che il test ha identificato 76 soggetti positivi ma non tutti sono postivi. Applicando la formula sopra riportata si avrà che il valore predittivo positivo 72/(72+4) = 0.95 o 95%

Il valore predittivo negativo di un test è la probabilità di essere sani dei soggetti risultati negativi al test. La formula per calcolarlo sarà: veri negativi/ (veri negativi+falsi negativi). Nel nostro caso avremo che il valore predittivo negativo sarà 16/(8 +16) = 0,67 o 67%
Come avevamo già visto per la sensibilità e specificità, il nuovo ipotetico test ha più difficoltà ad individuare i soggetti veri negativi, ciò non affetti dalla patologia ricercata.

- **Ripetibilità e Riproducibilità**

La Ripetibilità è la bontà dell'accordo tra i risultati di misurazioni successive dello stesso misurando condotte nelle stesse condizioni di misurazione.
La Riproducibilità è la bontà dell'accordo tra i risultati di misurazioni successive dello stesso misurando condotte in condizioni di misurazione non omogenee".
Ciò vuol dire che se io esamino la stessa articolazione ad esempio per due volte, subito successive o il giorno dopo, devo riscontrare lo stesso risultato. Questa viene indicata come ripetibilità intra esaminatore. E' auspicabile che ciò che apprezzo io sia più o meno uguale a ciò che apprezza un'altro operatore utilizzando lo stesso esame. Questa viene indicata come ripetibilità inter esaminatore.
Le variabili utilizzate (cosa stiamo misurando) nell'ambito di un qualsiasi studio possono essere di tre tipi: continue, ordinali, o nominali.

Le variabili continue possono assumere un numero infinito di valori all'interno di un certo ambito. Inoltre, la distanza che c'è, ad esempio, fra 3 e 4, è la stessa esistente fra, ad esempio, 20 e 21. Questo vuol dire che se consideriamo il peso, tipica variabile continua, un soggetto che pesa 80 kg avrà un peso che è effettivamente doppio rispetto a un soggetto che pesi 40 kg. Età, pressione arteriosa, glicemia, sono tutti esempi di variabili continue. Le variabili continue, soprattutto nel caso di misure biologiche, assumono spesso una distribuzione caratteristica, graficamente simile a una campana rovesciata (curva gaussiana). In altre parole, molte osservazioni cadono in un range di valori vicini alla media, mentre man mano che ci si allontana dalla media il numero di osservazioni diminuisce. Se pensiamo ad esempio all'altezza dei soggetti di una popolazione, molti soggetti avranno un'altezza vicina a quella media di quella popolazione, mentre man mano che ci si sposta verso valori di statura più elevati o più bassi si riduce il numero di soggetti
Le variabili ordinali si differenziano da quelle continue poiché possono assumere solo un numero finito di valori all'interno di uno specifico intervallo. Inoltre, pur essendo i valori posti secondo un ordine predeterminato (ad es. un scompenso cardiaco di classe IV è più grave di uno di classe III, che a sua volta è più grave di uno scompenso di classe II), non c'è equidistanza fra i valori (non possiamo cioè affermare che uno scompenso di classe IV sia il doppio grave di uno di classe II o quattro volte più grave di uno scompenso di classe I). Gli stadi di malattia, o le misure di qualità di vita sono tipicamente variabili ordinali. Infine, le variabili nominali esprimono una qualità del tipo "tutto o nulla", senza nessun ordine prestabilito. Ne sono un esempio il sesso, la razza, la presenza/assenza di una complicanza ecc.).
Saremo autorizzati a utilizzare un test parametrico, cioè basato sui parametri media e deviazione standard, solo nel caso in cui la variabile di interesse sia continua e normalmente distribuita.
In tutti gli altri casi sono da preferire i test non parametrici; tali test si basano sui ranghi delle osservazioni, non sul loro reale valore. In altre parole, le osservazioni vengono messe in ordine crescente, e a ognuna si attribuisce un numero corrispondente alla posizione che quell'osservazione occupa nella graduatoria (rango). I test statistici non parametrici vengono quindi basati sul confronto fra le somme dei ranghi.
Se si vuole studiare la correlazione tra due variabili continue si utilizza il il test di correlazione di Pearson se invece si utilizzano dati non parametrici è più appropriato utilizzare il test di correlazione di Spearman.

Nel nostro caso, della valutazione di un test diagnostico, dobbiamo fare riferimento ai test di correlazione non parametrica, in letteratura riportati anche con termini di correlazione tra ranghi o di concordanza.

Il coefficiente di correlazione di Spearman è sovente indicato con il simbolo greco ρ (rho) e può variare - tra +1 e –1 quando la correlazione è massima, con valore positivo oppure negativo; quando non esiste correlazione è vicino a zero.

Ancora più appropriata del test di correlazione di Spearman è la statistica Kappa di Cohen che costituisce uno degli strumenti più utilizzati per saggiare l'accordo fra vari esaminatori nel valutare x numero di soggetti. I valori K possono variare da 0 a 1.

Per interpretare i risultati della kappa di Cohen si può fare riferimento ai seguenti intervalli:

- 0.01 – 0.20 slight agreement
- 0.21 – 0.40 fair agreement
- 0.41 – 0.60 moderate agreement
- 0.61 – 0.80 substantial agreement
- 0.81 – 1.00 almost perfect or perfect agreement.

Nella valutazione di test diagnostici manuali, come quelli che stiamo portando come esempio, ancora più specifico del K di Cohen e un derivato che è il PABAK (Prevalente Adjusted- Bias Adjusted- K), che sarebbe il test K adattato a queste valutazioni specifiche.

Ci sono, ovviamente, degli ulteriori test di statistica per poter ancora di più approfondire alcuni tipi di valutazioni manuale, ma quelli esposti sono sicuramente i più utilizzati.