

Metodi statistici per la Neuropsicologia Forense

A.A. 2023/2024

Giorgio Arcara

IRCCS San Camillo, Venezia
Università degli Studi di Padova





4c. Affidabilità (formule)

Affidabilità

- Affidabilità test-retest
- Affidabilità inter-rater
- Consistenza Interna

Esistono diverse classificazioni di affidabilità Questa che sto utilizzando è principalmente basata su Urbina, 2004, Essentials of Psychological Testing.

Affidabilità test-retest

Affidabilità test-retest: è un indice che rappresenta la consistenza nel tempo di due misurazioni assumendo che non è avvenuto nessun cambiamento sistematico.

La formula standard e più diffusa quella della correlazione di Pearson (vedi anche Appendice di queste slides)

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

\bar{x} = mean of the values of the x-variable

y_i = values of the y-variable in a sample

\bar{y} = mean of the values of the y-variable

Affidabilità test-retest

Il più grande problema dell'affidabilità test retest è che essa non cattura eventuali effetti *sistematici* di differenza tra i punteggi. Il valore infatti ci dice quanto due valori sono *correlati* o consistenti, ma non se sono più o meno gli stessi

Immaginiamo un caso in cui ad ogni osservazione si aggiunga una quantità fissa k. Conseguirà che anche la media sarà aumentata di k

$$r = \frac{\sum (x_i - \bar{x})(y_i + k - (\bar{y} + k))}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i + k - (\bar{y} + k))^2}}$$

Segue da semplice algebra che le “k” si cancellano e quindi si ritorna alla formula iniziale: un aumento sistematico non cambia il valore della correlazione.

Vedi anche slides su Correlazione nell'Appendice concetti base di statistica

Affidabilità test-retest

Nei test cognitivi l'effetto sistematico più comune è riconducibile all'*effetto pratica*, l'etichetta utilizzata per indigare quei cambiamenti (Spesso sistematici) che si osservano nella somministrazione del test e che sono associati ad una migliore performance.

Attenzione dunque all'utilizzo e all'interpretazione dei valori di test-retest reliability che sono spesso (nei compiti cognitivi) solo parziali per indicare la *stabilità* di una performance.

Sarebbero infatti da accompagnare da analisi (es. *t-tests*) che permettano di vedere se c'è una differenza sistematica tra prima e seconda valutazione.

Considerazioni su affidabilità test-retest

Spesso l'affidabilità test-retest viene (erroneamente) considerata utile solo nel caso in cui avvengano misurazioni ripetute.

Dobbiamo invece immaginare una valutazione come una *fotografia* di una scena statica. La affidabilità test-retest stima quanto è precisa la vostra macchina fotografica.

In un test ad elevata affidabilità, non ci aspettiamo tante possibili variazioni, invece in un test con bassa affidabilità ci aspettiamo molte variazioni.

La nostra fotografia sarà una stima “puntuale”, ma potrebbe essere lontana da quella che consideriamo il punteggio vero (che ricordiamo è un'entità teorica).

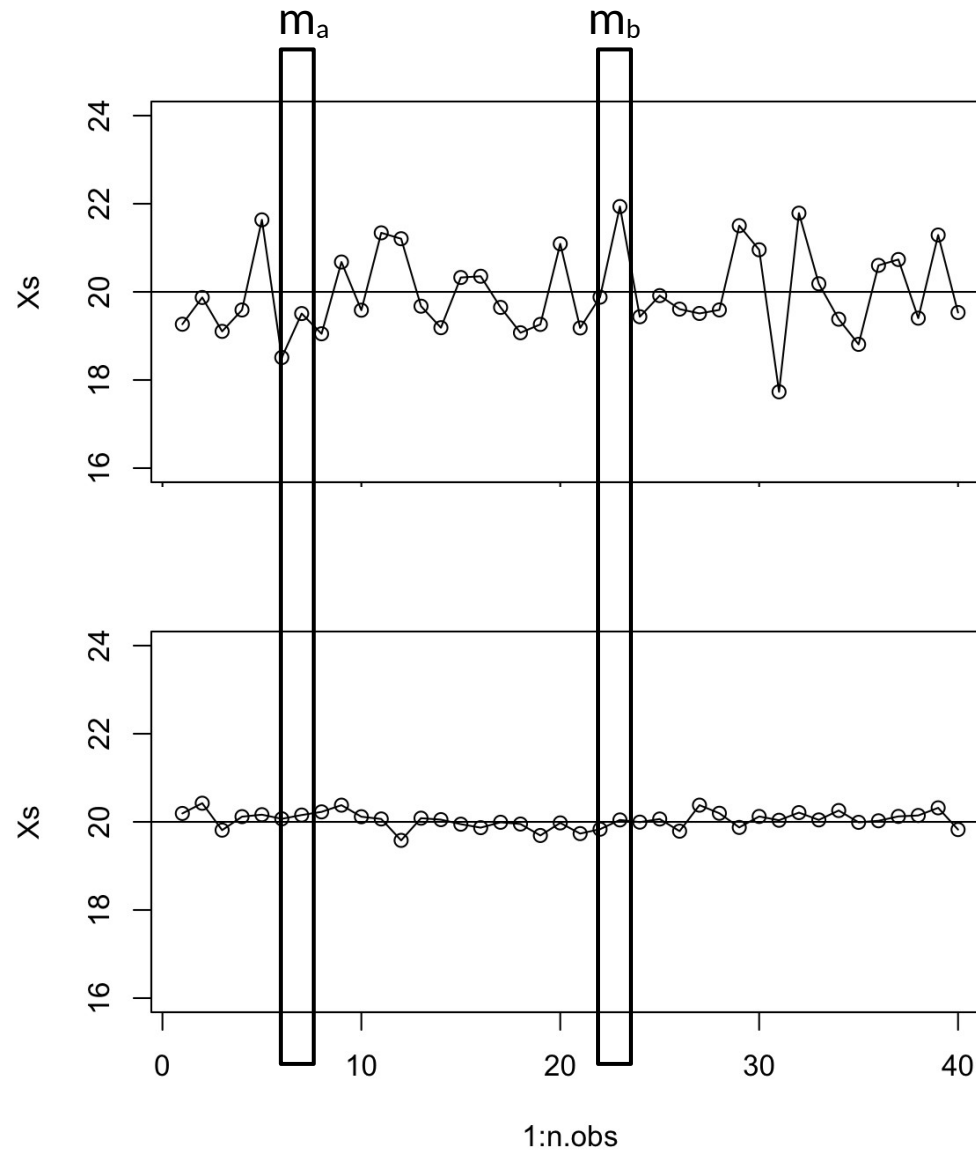
Considerazioni su affidabilità test-retest

Nella slides successiva si mostra questo concetto, ipotizzando delle singole misurazioni m_a ed m_b in due test (uno a bassa affidabilità test-retest e uno ad alta affidabilità test-retest).

Per ogni test verrà fatta una sola valutazione e non c'è interesse nella rivalutazione nel tempo. Nel test con alta affidabilità test-retest, avremo un punteggio osservato che è meno dipendente dalla specifica istanza di misurazione.

Affidabilità test-retest

Test bassa affidabilità
test-retest



Test alta affidabilità
test-retest

Affidabilità test-retest

Perché viene usata il coefficiente di correlazione di Pearson (r) se non è adeguato?

Probabilmente per semplice convenzione e per semplicità (il coefficiente r di Pearson può esprimere facilmente la varianza condivisa tra le variabili, semplicemente calcolando r^2)

Non è completamente sbagliato e può fornire un quadro abbastanza completo se accompagnato da t-tests (per indagare differenze sistematiche).

Affidabilità inter-rater

Affidabilità inter-rater: è un'indice che rappresenta la coerenza dei punteggi di più rater.

La formula più utilizzata per l'affidabilità inter-rater è l'**Intra Class Correlation (ICC)**. A differenza della correlazione (che può essere usata solo per 2 variabili), l'ICC invece può essere usata per k variabili (dove k è il numero di raters).

Esistono diverse formule di Intraclass correlation, che sostanzialmente rispondono a diverse domande e che assumono una diversa relazione tra gruppi e osservazioni (nel nostro caso raters e pazienti osservati).

Le formule più recenti per intraclass correlation si basano su scomposizione di varianza fatta tramite ANOVA o tramite mixed effect models (utile da sapere ma non approfondiamo).

https://en.wikipedia.org/wiki/Intraclass_correlation

Affidabilità inter-rater

La versione di ICC più utilizzata per calcolare l'affidabilità inter-rater è la **ICC(2,1) agreement**

Nel contesto di test con questa versione di ICC si assume che più rater vedano la stessa prestazione dello stesso soggetto e fornisce una stima che tiene conto dell'agreement in termini assoluti (se il punteggio è lo stesso, non se il punteggio è correlato)

$$ICC(2,1) = \frac{(MSB - MSE)}{\left(MSB + (n_j - 1)MSE + \frac{n_j(MSJ - MSE)}{n_c} \right)}$$

MSB = Variance between subjects

MSJ = Variance between judges (raters)

MSE = variance of interaction judges by subjects

n_r = number of raters / judges

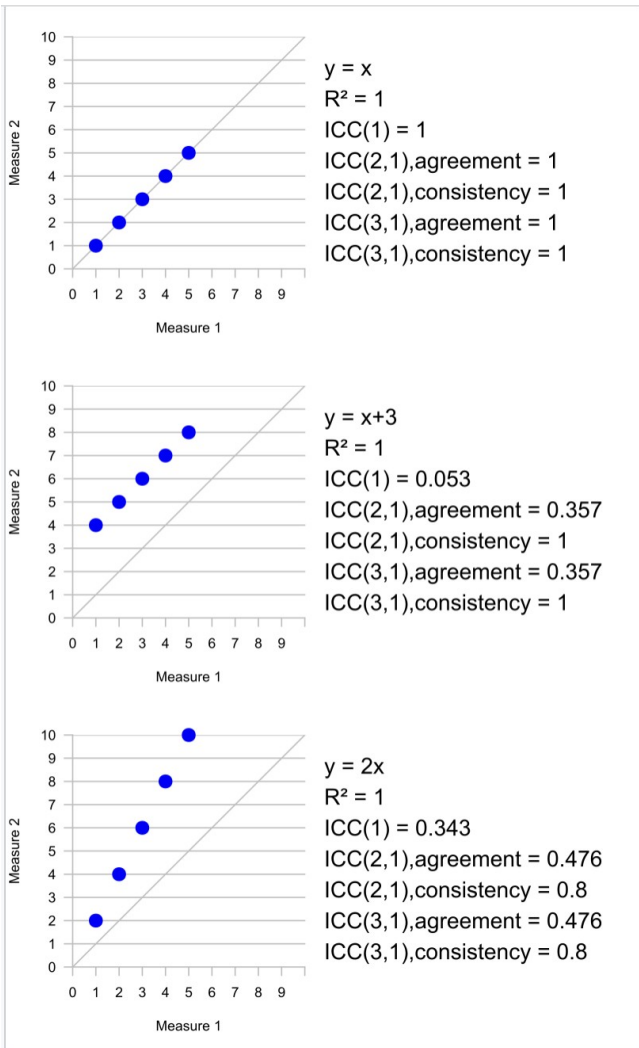
n_j = number of cases observations

Affidabilità inter-rater

Da pagina Wikipedia Intraclass Correlation

La figura mostra come diverse relazioni sono catturate da diversi coefficienti.

Nota che R^2 è il coefficiente di correlazione r elevato al quadrato, ma è anche usato nelle regressioni (in breve altre analisi statistiche dove si indaga la relazione tra variabili usando equazione del tipo $y = bx + e$).



Different intraclass correlation coefficient definitions applied to three scenarios of inter-observer concordance. □

Alcune considerazioni su ICC

Perché non è usata ICC anche per test-retest visto che tiene conto anche di differenze sistematiche?

Probabilmente più per prassi e convenzione. In letteratura è possibile trovare alcuni studi che la utilizzano anche per test-retest. In caso di effetto pratica sarebbe infatti più adeguata.

Consistenza interna

Consistenza interna: rappresenta la consistenza tra gli item di un test

Un approccio tradizionale per calcolare la consistenza di item di un test, è quella di dividere in due il test e indagare misure di somiglianza (es. correlazione) tra i punteggi delle due metà nei vari partecipanti. Questo metodo è detto **split-half reliability (o affidabilità split-half)**

Il grosso limite di questo approccio è che il risultato dipende da come dividiamo a metà il test (potremmo sovrastimare o sottostimare).

Da un punto di vista concettuale le misure di consistenza interna vanno ad indagare invece quale sarebbe la correlazione usando tutte le possibili divisioni split-half

Formula di Kuder-Richardson (K-R 20)

È una formula che viene usata quando gli item sono dicotomici (0/1)

$$r_{k-R 20} = \left(\frac{n}{n-1} \right) \frac{s_t^2 - \sum pq}{s_t^2}$$

N = numero di items nel test

S_t^2 = varianza del punteggio totale nel test

$\sum pq$ = somma del prodotto p x q in ogni item di ogni test

P = proporzione delle persone che passano l'item (o risposta 1)

q = proporzione delle persone che non passano in in ciascun item (o risposta 0)

Cronbach's alpha

È una generalizzazione della formula KR-20, nel caso di items non dicotomici

$$\alpha = \left(\frac{n}{n-1} \right) \frac{s_t^2 - \sum s_i^2}{s_t^2}$$

n = numero di items nel test

S_t^2 = varianza del punteggio totale nel test

S_i^2 = varianza di ogni item del test

Altri metodi per consistenza interna

Anche se i metodi KR-20 e alpha sono i più comuni essi hanno dei limiti intrinseci (Es. Assumono unidimensionalità)

In alcuni quindi da preferire altri metodi. Esempio se i punteggi degli item non sono normali meglio il metodo **GLB (Greatest Lower Bound)**.

In caso di multidimensionalità può essere meglio usare il metodo **omega (ω)**.

Nota che GLB e omega sono legati ad analisi fattoriali (vedi slides su validità) e sono computazionalmente più complessi .

Item selection

Un concetto associato a KR-20, Cronbach's alpha, e altre misure di consistenza in terna è quello dell'*item selection* e cioè quella procedura nella creazione di un test di identificazione degli item da mantenere nel test finale.

Alcune misure sono *specifiche* per items (un po' come i loadings nell'analisi fattoriale)

Esempi che vengono dall'alpha di cronbach

r-drop: indica la correlazione tra il test, e il test tolto quell'item (si ha dunque un valore r-drop per item)

r-cor: indica la correlazione tra l'item e il test totale controllando per overlap con altri item e per affidabilità globale.

Interpretare affidabilità a livello del test

Ogni sorgente di affidabilità può essere usata per stimare l'errore di misura del test

Il coefficiente di affidabilità (indipendente dalla formula) può infatti essere interpretato come la percentuale di variabili riconducibile al punteggio vero, trasformando semplicemente il coefficiente in percentuale. Questo segue proprio la definizione che viene fatta a priori di cosa è l'affidabilità

Ad esempio, una correlazione test-retest 0.8, vuol dire che 80% di variabilità test-retest è riconducibile alla variabilità del punteggio vero

Ad esempio, una consistenza interna di 0.9, vuol dire che 90% di variabilità nella consistenza interna è riconducibile alla variabilità del punteggio vero, etc.

Interpretare affidabilità a livello del test

L'errore associato ad ogni sorgente di affidabilità è pertanto calcolabile come

$$\text{Errore} = (1 - C_a) * 100$$

Dove C_a è il coefficiente di affidabilità in questione (es. Affidabilità test retest, affidabilità inter rater, etc).

Combinare diversi tipi di errore

Calcolando diverse sorgenti di errore è possibile sommarle (assumendo la loro indipendenza). Per ottenere un errore totale e calcolare quindi la variabilità del punteggio che si presume sia riconducibile al punteggio vero

Es. supponiamo di avere un test che ha le seguenti caratteristiche:

Affidabilità test-retest = 0.85 (r di pearson)

Affidabilità inter-rater = 0.70 (ICC(21))

Consistenza interna = 0.95 (alpha di Crobach)

Errore test-retest = $(1-0.85) * 100 = 15\%$

Errore inter-rater = $(1-0.7) * 100 = 30\%$

Errore consistenza interna = $(1-0.97) = 3\%$

Errore totale = $30 + 15 + 5 = 48\%$

A livello di test 48% della varianza è riconducibile ad errore,

E 52% $(100 - 48)$ della varianza è riconducibile al punteggio vero ed è anche l'affidabilità complessiva del test

Stimare il punteggio vero

Le formule di affidabilità ci permettono inoltre di stimare il punteggio vero

$$T' = r_{xx}(X - M) + M$$

T' = stima del punteggio vero

r_{xx} = affidabilità complessiva (vedi slide precedente)

X = punteggio osservato

M = media del campione di riferimento

Stimare il punteggio vero

$$T' = r_{xx}(X - M) + M$$

1) Per stimare il punteggio vero entra due volte in gioco la Media del campione di riferimento (denotata con M). Questo è per tenere in considerazione un fenomeno detto *regressione verso la media*. In ogni misura con errore è sempre più probabile che una misura successiva sia più vicina alla media del campione di riferimento. Questo perché (specie se l'errore dello strumento è grande) ogni deviazione della media ha probabilità di essere dovuta al caso.

Correggere per la media ha però un grosso limite intrinseco: con che media correggere? Ad esempio, se ho un paziente con TBI è corretto usare (per M), La media del gruppo di sani con cui è stato sviluppato il test?

Questo tipo di limiti rende complessa l'applicazione di formule per stimare il punteggio vero. Vedremo in futuro comunque in utilizzi clinici e pratici (es. cut-off, valori di cambiamento), **quasi sempre non sono usati i punteggi veri ma i punteggi osservati.**

Stimare il punteggio vero

Alcune note:

2) La stima del punteggio vero è spesso accompagnata anche dalla creazione di *intervalli di confidenza* attorno alla stima (riprenderemo questo quando parleremo di cut-offs, vedi anche Appendice)

Come considerare i coefficienti di affidabilità

Ogni coefficiente di affidabilità, in generale stimare la variabilità riconducibile
Al punteggio vero, rispetto alla variabilità totale (cioè quella osservata dai punteggi),
utilizzando diverse formule

$$r_{xx} = \frac{\sigma_{true}}{\sigma_{tot}}$$