Metodi statistici per la Neuropsicologia Forense

A.A. 2023/2024

Giorgio Arcara

IRCCS San Camillo, Venezia
Università degli Studi di Padova

4b. Validità (formule)

La Validità

- **Validità di contenuto**
- **Validità convergente/divergente (validità di costrutto)**
- **Validità di criterio**
- **Validità di facciata**
- **Validità ecologica**

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

Validità di contenuto

Validità di contenuto : la proprietà degli item di essere sufficienti ed adeguati per valutare il costrutto di interesse.

Metodo di Lawshe (1975) é un metodo per calcolare validità di contenuto di item e test totale, caratterizzato da 3 step.

STEP 1) Si chiede degli esperti (panelist) di valutare ciascun item secondo queste possibilità:

- *Essenziale*
- *Utile ma non essenziale*
- *Non necessario*

Validità di contenuto

Metodo di Lawshe (1975)

STEP 2) si calcola per ogni item un Content Validity Ratio

$$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

n_e = numero di panelist (i.e. esperti) che dicono che quell'item è *essenziale*
N = numero totale dei panelist

Varia da -1 a 1, 0 se metà dei panelist dicono che è essenziale

Ogni valore pari a 1 viene aggiustato a 0.99 (per comodità e per convenzione)

Nota: CVR si riferisce a ciascun Item, non al test

Validità di contenuto

Metodo di Lawshe (1975)

Quali Item mantenere?

STEP 3) si seleziona quali item mantenere

N. di panelist	Valore minimo di CVR
5	.99
6	.99
7	.99
8	.75
9	.78
10	.62
11	.59
12	.56
13	.54
14	.51
15	.49
20	.42
25	.37
30	.33
35	.31
40	.29

Tabella da Lawshe, 1975, p.568

In alternativa mantengono gli Item in cui CVR o >=0.78 (Polit, Becl, Owen, 2007)

Validità di contenuto

Metodo di Lawshe (1975)

STEP 4)

È possibile calcolare un valore per il test (**Content Validity Index**) che non è altro che la media dei CVR di tutti gli Items.

$$CVI = \frac{\sum_{I_1}^{I_k} \left(\frac{n_e - \frac{N}{2}}{\frac{N}{2}} \right)}{k}$$

n_e = numero di panelist che dicono che quell'item è *essenziale*
N = numero totale dei panelist
k = numero degli item
I = gli items che compongono il test

Validità di contenuto

NOTE Metodo di Lawshe (1975)

CVR viene usato durante Item selection nella creazione del test

CVI viene utilizzato alla fine per riportare Ii grado di validità di contenuto del test

Validità di contenuto

Note su Metodo di Lawshe (1975)

È interessante notare come il metodo risponde solo in parte alla domanda di validità di contenuto.

Il metodo di Lawshe mi dice se un dato Item è appropriato per il test, *ma non se il test contiene tutti gli item necessari per i suoi scopi* (nell'articolo originale di Lawshe sono discussi alcuni esempi in cui gli item da cui si parte sono già stati definiti e il problema è identificare se sono approprietà)

La validità di contenuto, spesso non è considerata né riportata nei test (non è molto di moda).

Validità di costrutto

La **validità di costrutto** esprime quanto il test misura effettivamente il costrutto che intende misurare valutando corenza interna degli item, oppure correlazione con altri test.

La validità di costrutto è valutata in due maniere:

- valutando sei gli **item** misurano in maniera appropriata il costrutto.

- valutanto se **il punteggio totale del test**[1] sia il punteggio totale del test (rappresenta correttamente il costrutto?).

[1] nella maggior parte dei casi, il punteggio totale si ottiene sommando gli items (in alcuni casi facendo una media).

Validità di costrutto

Alpha di Cronbach

Uno dei classici modi per vedere su tutti gli item di un test si comportano in maniera coerente è tramite l'utilizzo di metriche come l'alpha di Cronbach. Questa però è spesso associata più all'affidabilità (vedi quindi slides su formule affidabilità).

Validità di costrutto

Analisi Fattoriale

L'analisi statistica più comunemente utilizzata per valutare gli Item è l'**analisi fattoriale.** L'analisi fattoriale (nel contesto della teoria dei test) indaga se gli item di un test si comportano in maniera coerente, dimostrando che essi misurano uno stessa variabile latente (detto **fattore**).

Operativamente e per questi scopi l'analisi fattoriale viene condotta in una matrice *m x n*, dove *m* sono gli Item (in riga) ed *n* sono I partecipanti in colonna

	i_1	i_2	i_3	i_4	…	i_n
p_1	x_{11}	x_{12}	x_{13}	x_{14}	…	x_{1n}
p_2	x_{21}	x_{22}	x_{23}	x_{24}	…	x_{2n}
p_3	x_{31}	x_{32}	x_{33}	x_{34}	…	x_{3n}
p_4	x_{41}	x_{42}	x_{43}	x_{44}	…	x_{4n}
…	…	…	…	…	…	…
p_m	x_{m1}	x_{m2}	x_{m3}	x_{m4}	…	x_{mn}

Validità di costrutto

Analisi Fattoriale

Il risultato di un'analisi fattoriale è spesso una matrice di "loadings", cioè quanto ciascun item è correlato ad un determinato fattore.

	f_1	f_2	f_3
i_1	l_{11}	l_{12}	l_{13}
i_2	l_{21}	l_{22}	l_{23}
i_3	l_{31}	l_{32}	l_{33}
i_4	l_{41}	l_{42}	l_{43}
…	…	…	…
i_m	l_{m1}	l_{m2}	l_{m3}

Esistono delle rule-of-thumb per definire se un loading è significativo

(Es. > 0.3 o 0.5
Oppure < -0.3 o -0.5)

i = item
F = fattore
l = loading

Validità di costrutto

Analisi Fattoriale

In questo esempio gli item caricano su 2 fattori

	f_1	f_2
i_1	0.2	0.7
i_2	0.5	-0.04
i_3	0.7	0
i_4	0.9	0.8
i_5	0.7	- 0.9

Un primo fattore cattura una correlazione positiva Tra gli item 2,3,4,5, che tendono avere punteggi tutti correlati tra loro.

Un secondo fattore cattura una correlazione negativa di item 1 e 4 rispetto ad item 9. In sostanza quando sono alti I punteggi di 1 e 4 son obassi quelli di 5, e viceversa.

Analisi Fattoriale

Esistono due tipi principali di analisi fattoriale.

L'**analisi fattoriale esplorativa**, ha come scopo cercare di definire quale è la struttura fattoriale che meglio descrive (se esiste) la relazione tra items.

L'**analisi fattoriale confermativa**, ha come scopo verificare se gli item si comportano in maniera conforme ad una struttura fattoriale nota.

Validità di costrutto

Analisi Fattoriale

Nel contesto della validità di costrutto l'analisi fattoriale può essere una tecnica utilizzata per selezionare gli item che comporranno gli item del test finale (ad esempio escludendo quelli che non si comportano in maniera adeguata rispetto al costrutto o ai costrutti che si intendeva misurare.

Ad esempio, se un test intende misurare un singolo costrutto e tutti gli item eccetto uno hanno loading alti in un fattore, probabilmente quell'item andrebbe eliminato dalla versione finale del test.

Validità di costrutto

Analisi Fattoriale

Esistono molte analisi statistiche e metriche per valutare la bontà di un'analisi fattoriale

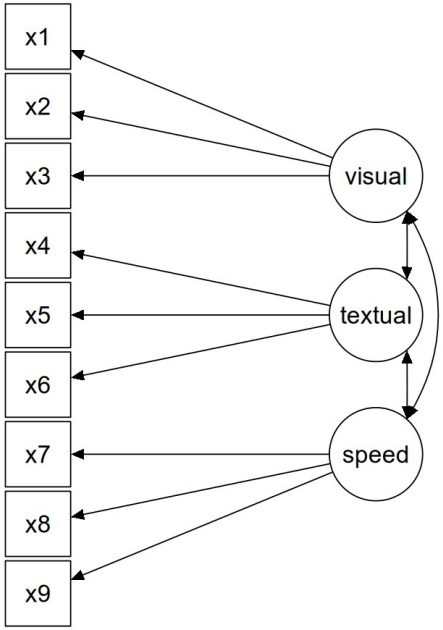
Tra I più comuni si usa il test di Goodness-of-Fit (è un test che riporta Chi-quadro) che deve essere **non significativo (p > 0.05).** Se significativo vuol dire che la struttura a fattori proposta non è sufficiente per spiegare la variabilità dei dati.

Altre metriche sono il **RMSEA (**root Mean squared error of Approximation) che deve essere il più basso possibnile (di solito <0.06 è considerata una soglia minima).

Validità di costrutto

Analisi Fattoriale

L'analisi fattoriale è spesso rappresentata in maniera grafica distinguendo I fattori latenti (in questo caso cerchi) Dalle variabili osservate (in questo caso quadrati)



https://lavaan.ugent.be/tutorial/cfa.html

Validità di costrutto

Analisi Fattoriale

Per interpretare se un'analisi fattoriale supporta effettivamente la validità di costrutto. Andrebbe visto se gli item si comportano effettivamente come il testa assume si dovrebbero comportare.

Es. se un test sostiene di misurare un costrutto specifico (es. Attenzione visuospaziale) tutti gli item dovrebbero caricare su un solo fattore (almeno principalmente).

È innanzitutto utile vedere se gli autori di un test hanno effettuato analisi fattoriale per indagare la relazione tra gli item e li hanno selezionati (escludendone alcuni) o in generale se c'è stata una qualche fase di selezione degli items (es. Item selection tramite correlazione).

Validità di costrutto

Analisi Fattoriale

Stabilire l'adeguatezza di un'analisi fattoriale non è comunque semplice perché spesso è possibile fare giustificazioni a posteriori dei risultati finali, senza che I risultati abbiano influenzato la scelta se tenere o meno degli items.

Ad esempio, potrei avere un test che assume di misurare uno specifico costrutto, trovare con un'analisi fattoriale esplorativa che in realtà la variabilità degli item è spiegata da due fattori e mantenere comunque tutti gli item dicendo che probabilmente sono influenzati da due fattori separati, coerenti con aspettative da teoria (Es. Attenzione visuospaziale e funzioni esecutive).

Validità di costrutto

Analisi Fattoriale

Nota che a volte l'analisi fattoriale è usata anche per I punteggi di sottoscale
Che compongono una batteria (es. Per APACS Arcara & Bambini 2016)

I principi sono gli stessi, ma in questo caso non si tratta di Items, ma di gruppi di items che compongono sottoscale (in questo caso es. Il task di Linguaggio Figurato 1, il task di Linguaggio Figurato 2 etc.)

Principal Component Analysis (PCA)

Un'analisi che va citata insieme all'analisi fattoriale è la **Principal Component Analysis (PCA)**

PCA e analisi fattoriale sono concettualmente simili e I risultati (una matrice di loadigns) sono anch'essi simili. La differenza è che nella PCA non c'è stima di "variabili latenti", ma lo scopo è ottenere nuove variabili che catturano porzioni di varianza comune tra più punteggi (es. Items, o punteggi totali di test). Tali variabili sono dette **componenti principali** (da cui il nome dell'analisi).
Questa differenza si rifletta in una diversità a livello della formula e nei risultati che si otterranno.

Senza approfondire è qui importante solo ricordare che potreste trovare in un test una PCA invece di un'analisi fattoriale. Potete interpretarle in maniera analoga, anche se sarebbe più corretto utilizzare l'analisi fattoriale perchè assume l'esistenza di variabili latenti, cosa che è spesso implicita quando si tratta di item o punteggi totali di test psicologici.

Validità di costrutto

Nota Analisi fattoriale e PCA

È importante ricordarsi che analisi fattoriale e PCA ci aiutano a capire come si raggruppano gli items e se lo fanno in maniera coerente con il costrutto (o I costrutti) che dovrebbero misurare.

Non ci garantiscono però che il costrutto *sia quello giusto* e pertanto sono solo un'informazione parziale sulla validità.

Ad esempio, immaginiamo che voglia valutare validità di costrutto di un test di memoria. Gli autori potrebbero riportare un'analisi fattoriale che mostra che gli item correlano bene fra di loro e sono riconducibili da un'unica variabile latente. Questo è ottimo ma non mi garantisce che questa variabile latente sia effettivamente la memoria. Il test potrebbe essere stato costruito in maniera sbagliata e gli item essere consistenti per un altro motivo (es. magari il test cattura perlopiù velocità psicomotoria).

Validità di costrutto

Correlazione

Lo strumento più comune per verificare la validità di costrutto di un test è indagare la correlazione del punteggio totale al test con correlazione dei punteggi ad altri test.

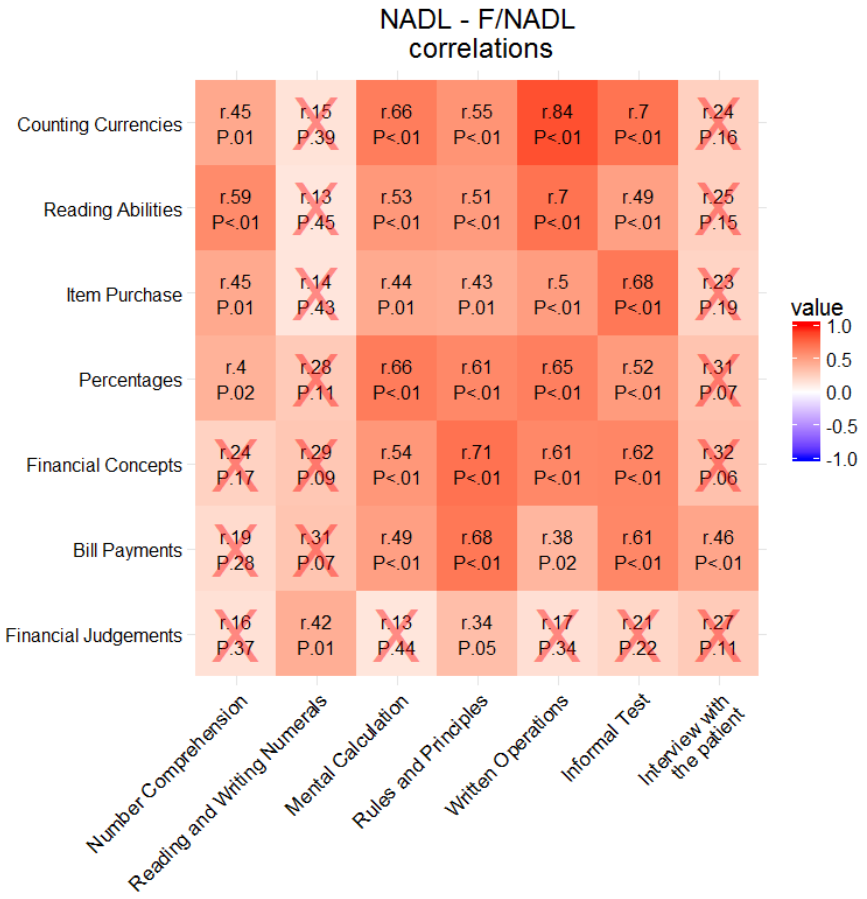
Non esiste una regola di quanto dovrebbe essere almeno questa correlazione, ma è ragionevole pensare che debba essere almeno *moderata* (> 0.3), ma verosimilmente anche di più.

I test neuropsicologici (Specie nei sani) tendono sempre ad essere molto correlati fra di loro e correlazioni più alte sono attese (di fatto queste correlazioni hanno portato ad ipotizzare al cosiddetto fattore g dell'intelligenza)

https://en.wikipedia.org/wiki/G_factor_(psychometrics)

Validità di costrutto

Correlazione



Questa figura riporta le correlazioni tra alcuni test del NADL (Semenza et al., 2014) e il NADL-F (Arcara et al., 2019).

Validità di costrutto

Correlazione

L’aspetto critico dell’interpretare valori “soglia” di correlazioni tra due diversi test è che I valori sono anche legati ai rispettivi errori dei test (quindi ad aspetti legati a numerosità campionarie e affidabilità, vedi future slides).

Inoltre è sempre possibile “giustificare” a posteriori un’eventuale differenza o valore basso (Es. La correlazione è bassa presumibilmente perchè questo test implica di più le funzioni esecutive rispetto all’altro).

Validità di costrutto

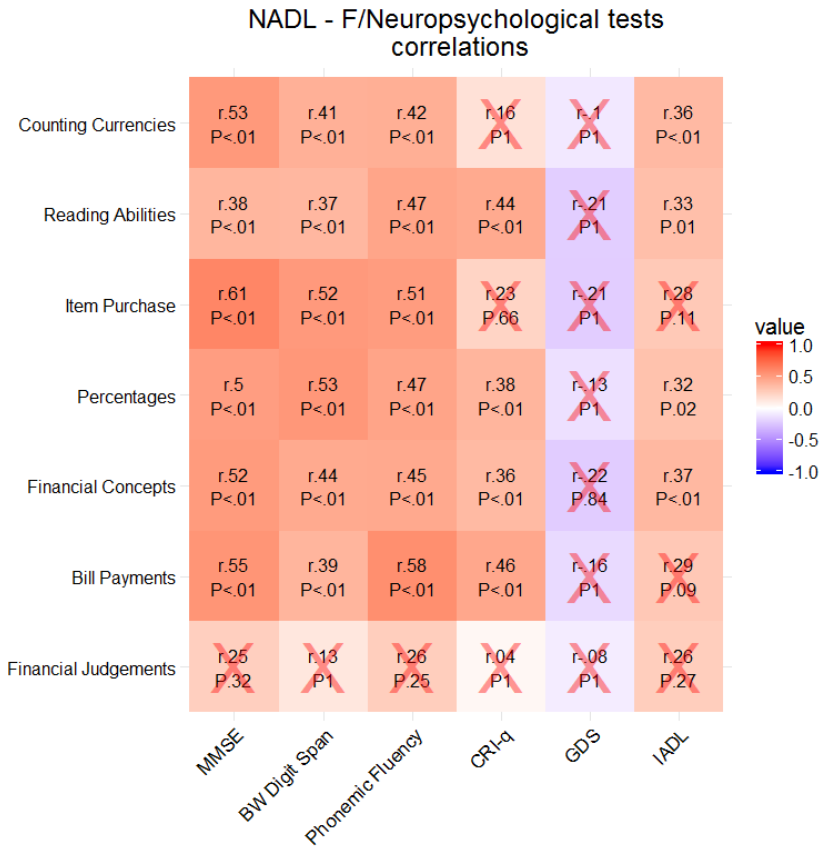
Correlazione

Anche l’assenza di correlazione o correllazione molto bassa può essere a sostegno di validità di costrutto.

In generale la validità di costrutto indica quanto bene stiamo misurando lo specifico costrutto. Dimostrare che stiamo isolando bene la specifica funzione cognitiva è a sostegno della validità di costrutto.

Validità di costrutto

Correlazione



Questa figura riporta le correlazioni tra alcuni test di test di vario tipo e il NADL-F (Arcara et al., 2019).

Notate che non correla con GDS (geriatric Depression Scales), a sostegno di validità di costrutto.

Validità di costrutto

Altre analisi

In generale la validità di costrutto non è vincolata a specifiche analisi statistica e andrebbe valutata rispetto alla specifica domanda a cui cerca di rispondere. In altre parole, piuttosto che cercare nei manuali / articoli specifiche analisi (es. È stata fatta analisi fattoriale? Ci sono correlazioni?) Dovreste cercare di rispondere a questa domanda

Esistono evidenze sperimentali che suggeriscono che il test stia misurando quello che intende misurare?

Ad esempio un test che ha come obiettivo valutare la capacità di lavorare potrebbe non avere nessuna correlazione, ma essere buono nel discriminare persone che sono state licenziate, da quelle che hanno mantenuto il lavoro (in questo caso una regressione logistica o altre analisi più legate alla discriminazione).

Validità di costrutto

Altre analisi

Esempio da Kershaw & Webber, 2008

Differences on the FCAI Between People With and People Without an Administrator

As predicted, adults without a legally appointed administrator performed significantly better than the adults who had an administrator on all dimensions of the FCAI overall, Wilks' λ $F(7,167) = 22.15$, $p < .01$, partial $\eta^2 = .486$. Univariate analyses showed that the group without administrators scored higher than the group with administrators on all six FCAI subscales, and overall. Independent samples t tests, using a Bonferroni-type adjustment for significance level ($p = .05/7 = .007$) showed that adults without an administrator performed significantly better than adults with an administrator on all FCAI subscales, and on the FCAI total score (Table 4).

Il FCAI è un test per valutazione competenze finanziarie. Per dimostrare validità di costrutto sono stati confrontati I pazienti con e senza tutore legale (tramite MANOVA).

Validità di costrutto

Altre analisi

È possibile anche combinare analisi precedentemente citate per supportare validità di costrutto

Ad esempio Montemurro et al. (2023) per dimostrare che il test Tele-GEMS era utile per valutare cognizione generale e hanno usato il seguente approccio.

1) raccolto dati su pazienti (Sclerosi Multipla in quel caso), su vari test.
2) fatto PCA su questi test e ottenuto un punteggio globale (prima componente) che catturava 60% della varianza.
3) visto se punteggio Tele-GEMS correlava con questo punteggio. Correlazione era 0.50 (moderata/alta) e interepretata a supporto.

Validità di costrutto

Altre analisi

Un altro ipotetico esempio su come combinare analisi per validità di costrutto:

Se l'analisi fattoriale mostra che gli item si comportano come se catturassero due variabili latenti, una associata ad attenzione ed una a memoria, sarebbe utile e a supporto di validità di costrutto se I punteggi ottenuti sommando queste due parti correlassero rispettivamente con un altro test di attenzione ed un altro test di memoria.

Validità di criterio

La **validità di criterio** valuta quanto un test è associato o predice un altra variabile di outcome (osservabile), detta **gold-standard.**

Nella validità di criterio lo scopo è vedere come il nostro test predice un altra variabile di outcome, detta **gold-standard.**

Le analisi che si possono utilizzare (e come valutare) dipendono se il criterio è una *variabile continua* (es. Il punteggio ad un altro test) oppure una *variabile categoriale* (es. L'appartenza ad un gruppo, es MCI vs Controlli).

Validità di criterio

Variabile continua

Se il gold standard è una variabile continua, allora si può vedere tramite una correlazione. In tal caso la correlazione dovrebbe essere molto alta (> 0.9).

A partire dalla correlazione possiamo calcolare la varianza spiegata nota anche come r^2 che non è altro che il coefficiente di correlazione al quadrato.

A volte per calcolare la validità di criterio si usa la regressione lineare. Nel caso di due variabili la regressione lineare è strettamente imparentata con il coefficiente di correlazione (e infatti I risultati di alcuni aspetti sono matematicamente identici)

Validità di criterio

Variabile categoriale

Nel caso di variabili categoriali (Es. Sopra/sotto cut-off, oppure MCI vs Controlli), esistono una famiglie di analisi diverse che vengono utilizzate. La categoria più rilevante (spesso patologica) è detta anche **condizione di interesse.**

In questa fase diciamo solo che esistono 4 tipi di possibili risultati quando si confronta un test con il suo criterio (categoriale)

	Condition of interest		
	Present (e.g. examinee with a pathology)	Absent (e.g. healthy examinee)	Row Total
Positive test (defective performance)	True Positive (A)	False Positive (B)	A+B
Negative test (adequate performance)	False Negative (C)	True Negative (D)	C+D
Column Total	A+C	B+D	

Tutte le metriche legate a questo fanno parte dei concetti di Sensibilità/Specificità
Che saranno trattate in apposite lezioni (dopo I cut-off)

Validità di facciata

La **validità di facciata** valuta quanto un test appare nella sua forma di misurare ciò che intende misurare

Spesso è valutata tramite domande ad hoc ad esperti o tramite confronto con test gold standard esistenti, ma solo tramite confronto qualitativo.

Talvolta la sola informazione che si ha a disposizione di un test è proprio la validità di facciata. Non ci sono altre prove che il test misuri ciò che vuole misurare se non il fatto che sia costruito da item che *plausibilimente* misurano proprio quel costrutto.

Validità ecologica

La **validità ecologica** valuta quanto un test è in grado di prevedere comportamenti al di fuori del setting di valutazione (quindi nella vita quotidiana)

Spesso è studiata tramite *correlazioni* con altre scale riferite a comportamenti con la vita quotidiana (diversamente da molti test l'interesse non è in costrutti ma in altri comportamenti)

Può essere anche essere valutata tramite altre analisi, purché mostrino che I punteggi al test siano associati al fenomeno di vita quotidiana di interesse.