Statistica applicata a comprendere la validità di strumenti di assessment

Prof. Giuseppe Verlato

Variabile latente

- Una variabile latente è un fenomeno non direttamente osservabile (soddisfazione, ansia, disabilità).
- E' una variabile, quindi varia da un individuo all'altro, e da un momento all'altro nello stesso individuo.
- E' latente, non è direttamente misurabile, come il peso, l'altezza, la glicemia.
- Per misurarla, occorre quindi uno strumento di assessment (un questionario o scala di misura)

Questionari

Variabile latente da indagare	**Strumento di rilevazione**
Stato di salute	SF-36 (Short Form 36 Health Survey)
Capacità di svolgere attività quotidiane per cura personale	Questionario ADL (Activities of Daily Living)
Ansietà e depressione	Hospital Anxiety and Depression Scale
Livello di coscienza	Glasgow Coma Scale

Fasi della costruzione di un questionario (De Vellis, 1991)

1. Definire chiaramente il fenomeno non osservabile (costrutto) da misurare (revisione della letteratura scientifica)
2. Generare un insieme di item (punti, elementi, domande)
3. Scegliere la struttura della scala di misura (scala di Likert, scala del differenziale semantico, scala di Stapel)
4. Far controllare l'insieme di item da un gruppo di esperti (verifica della validità del contenuto)
5. Considerare l'aggiunta di item di validazione (verifica della validità del costrutto)
6. Somministrare il questionario ad un campione di soggetti
7. **Valutare l'affidabilità della scala di misura**
8. Ottimizzare la lunghezza della scala.

+

9. **Concordanza dei risultati al test-retest, in genere su un sotto-campione.**

Scala Likert

Nella **scala di Likert** gli item vengono presentati sotto forma di affermazioni e agli intervistati viene chiesto di esprimere il proprio grado di disaccordo o accordo nei confronti di esse.

Le risposte possono essere formate da un numero dispari o pari di opzioni, di conseguenza può essere presente o meno una risposta neutrale.

In generale, direbbe che la Sua salute è:

/_1_/	/_2_/	/_3_/	/_4_/	/_5_/
Eccellente	Molto buona	Buona	Passabile	Scadente

Scala del differenziale semantico

La **scala del differenziale semantico**, invece, prevede che i soggetti intervistati possano valutare il fenomeno di interesse su una scala caratterizzata ai due estremi da due aggettivi, i quali esprimono concetti opposti (caso bipolare) o la presenza/assenza di un certo attributo (caso unipolare).

Secondo Lei, il corso è:

Inutile	___	___	___	___	___	___	___	Utile
	1	2	3	4	5	6	7	

Entrambe le scale vengono trattate come quantitative (scale di intervallo).

Scala di Stapel

La **scala di Stapel** rappresenta una variante della scala del differenziale semantico, in cui ogni attributo è descritto da un solo aggettivo (ad esempio: utile) anziché da una coppia (inutile/utile).

Si tratta quindi di una scala unipolare, generalmente a 10 punti, non verbali, con valori compresi tra +5 e –5 con esclusione dello zero. La scala è progettata per misurare simultaneamente la direzione e l'intensità degli atteggiamenti. Ad esempio:

	+5
	+4
	+3
	+2
	+1
Utilità del corso	-1
	-2
	-3
	-4
	-5

Proprietà di uno strumento di misura (assessment)

- 1.**Affidabilità:** stabilità o precisione dello strumento.
- 2.**Validità:** capacità dello strumento di rilevare l'oggetto di studio che si è interessati proprio a rilevare

1. Affidabilità e validità

$$X_{oss} = X_{true} + X_{bias} + X_{random}$$

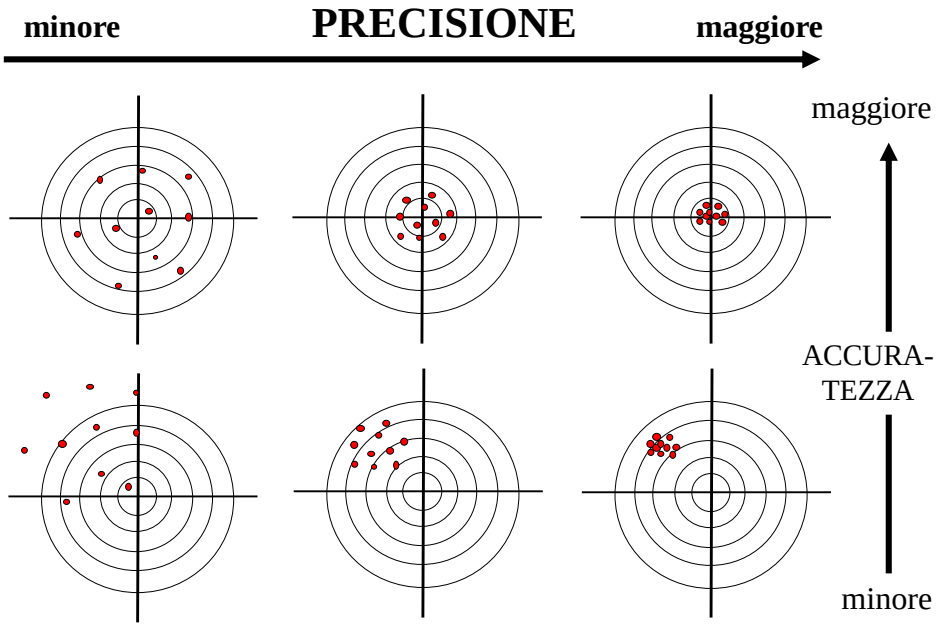
dove X_{random} è l'errore casuale
e X_{bias} è l'errore sistematico o bias

Una scala di misura è **affidabile**
quando $X_{random} = 0$

Una scala di misura è **valida**
quando sia $X_{random} = 0$ che $X_{bias} = 0$,
ovvero quando quando $X_{osservato}$ = quando X_{reale}

Errore sistematico ed errore casuale in una Scala di Misura

- ACCURATEZZA: Capacità di prevenire/evitare l'errore sistematico o distorsione (*bias*)
- PRECISIONE: Capacità di ridurre/minimizzare la variabilità casuale; è spesso valutata sulla base del coefficiente di variazione. Ripetibilità = Capacità di ottenere lo stesso valore nelle stesse condizioni (stesso operatore, stesso laboratorio, …). Riproducibilità = Capacità di ottenere lo stesso valore in condizioni differenti.


minore
PRECISIONE
maggiore
maggiore
ACCURA-
TEZZA
minore

ESEMPLIFICAZIONE
GLICEMIA VERA = 150 mg/dl
grande ERRORE CASUALE
scarsa PRECISIONE
glicemia (mg/dl)
grande ERRORE SISTEMATICO
scarsa ACCURATEZZA
buona PRECISIONE e buona ACCURATEZZA
Personale Inesperto
Ritardo eccessivo tra prelievo e lettura
GRAFICA

Verifica statistica dell'affidabilità

Aspetto da indagare		Strumento statistico
Correlazione tra gli item		Coefficiente di correlazione di Pearson
	Presi a due a due	inter-item correlation
	Un item e la scala complessiva	
	compreso l'item stesso	uncorrected item-to-total correlation
	escluso l'item stesso	corrected item-to-total correlation
Presenza di uno o più variabili latenti (dimensionalità)		Analisi fattoriale (analisi delle componenti principali)
Coerenza interna		Alpha di Cronbach Indici basati sul metodo split-half

Correlazione tra gli item

Indice di correlazione di Pearson

1) Inter-item correlation: grado di associazione tra gli item presi a 2 a 2

2) Item-to-total correlation: grado di associazione tra ciascun item e la scala nel suo complesso

   2.1) Uncorrected item-to-total correlation: la scala complessiva comprende l'item in studio
   2.1) Corrected item-to-total correlation: l'item in studio viene escluso dalla scala complessiva

Correlazione

Il coefficiente di correlazione (r) è un numero adimensionale, che varia tra -1 e +1

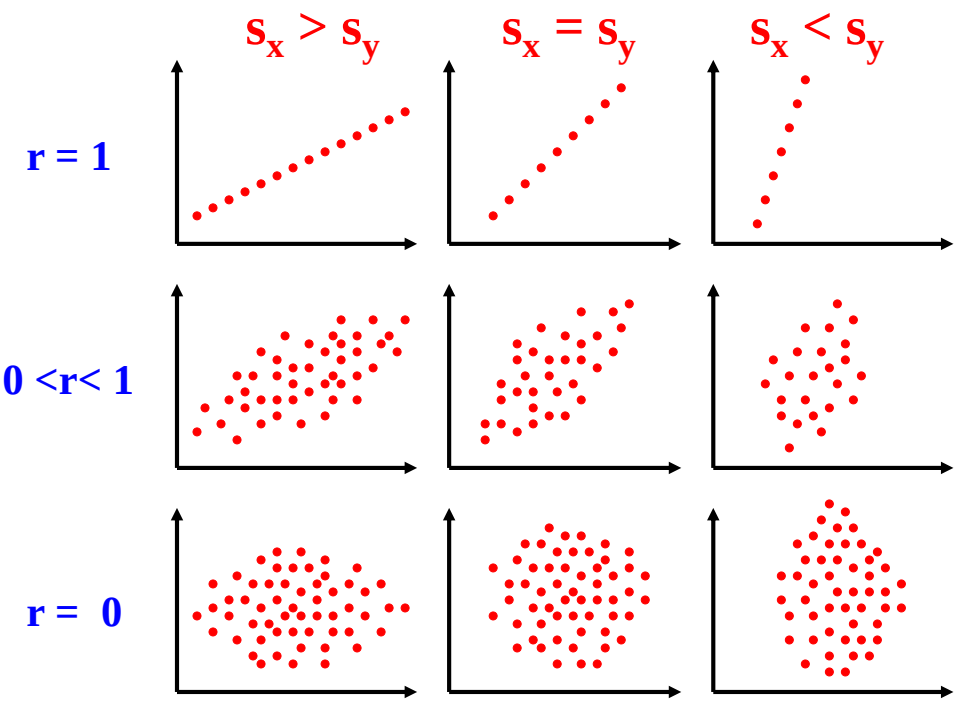
r = -1 i punti si allineano lungo una retta discendente

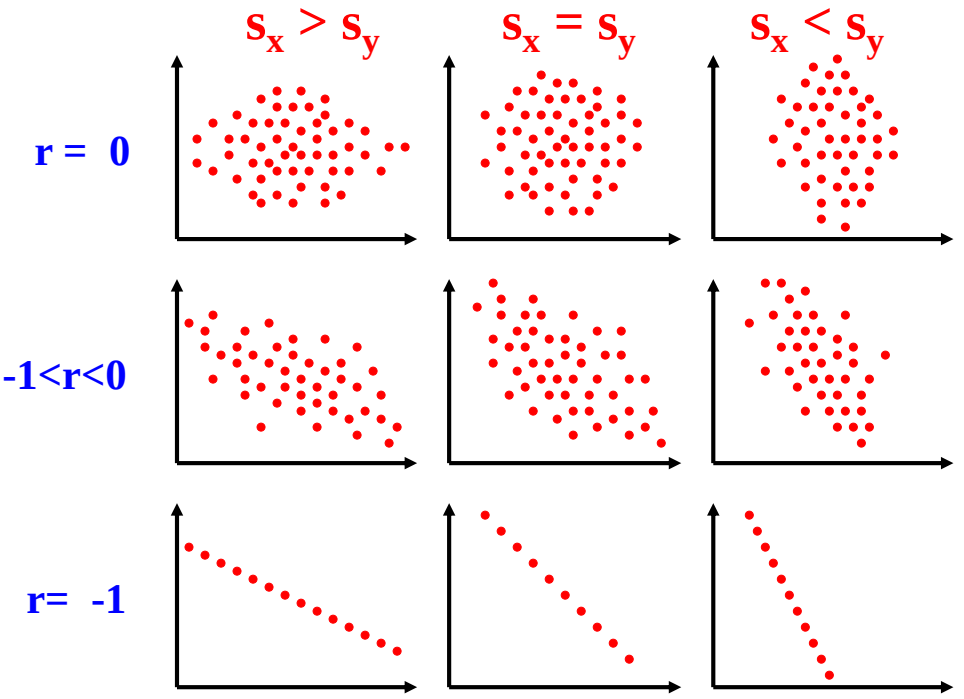
r = 0 i punti si dispongono a caso, senza mostrare un andamento crescente o decrescente

r = +1 i punti si allineano lungo una retta ascendente

$$r = \frac{\text{codevianza}_{xy}}{\sqrt{\text{devianza}_x * \text{devianza}_y}}$$

$$r^2 = \frac{\text{Devianza spiegata dalla regressione, SSR}}{\text{Devianza totale, SST}} = \frac{\Sigma (\hat{y} - \bar{y})^2}{\Sigma (y - \bar{y})^2}$$





Correlazione tra gli item

Quando [Hair et al, 2006]
1) l'inter-item correlation ha un valore <0,30
2) o l'item-to-total correlation ha un valore <0,50

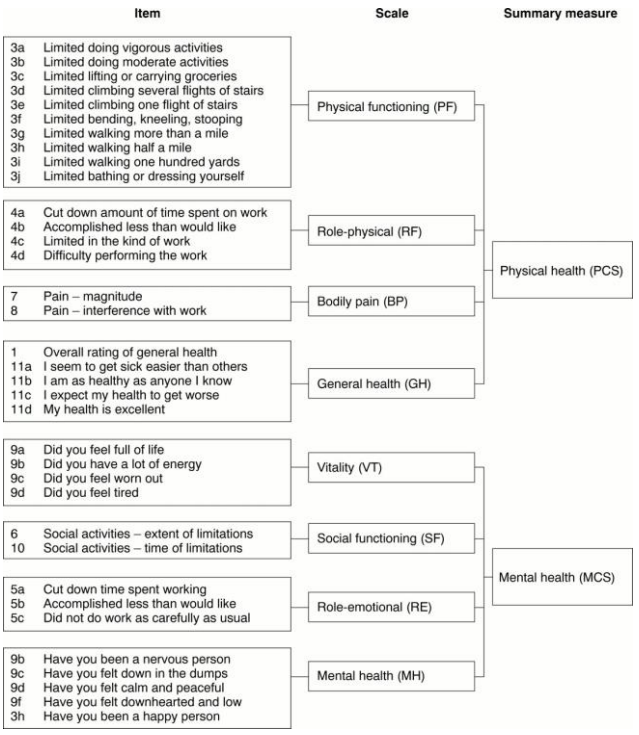
Gli item con bassa correlazione devono essere studiati con ulteriori analisi statistiche:
Analisi fattoriale: gli item in questione appartengono ad una dimensione diversa dagli altri ?

Alpha di Cronbach: gli item in questione abbassano la coerenza interna del questionario ?

Verifica della dimensionalità

L'SF-36 contiene 2 dimensioni principali (salute fisica e salute mentale), ciascuna delle quali contiene a sua volta 4 sottodimensioni.

L'analisi delle componenti principali mira a rilevare le diverse dimensioni (variabili latenti) di una scala di misura.



Analisi delle componenti principali

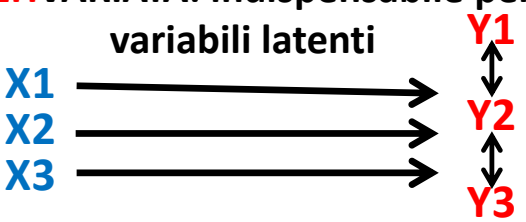
E' un' analisi MULTIVARIATA

Analisi UNIVARIABILE : sufficiente per una tesi

una X → **una Y**

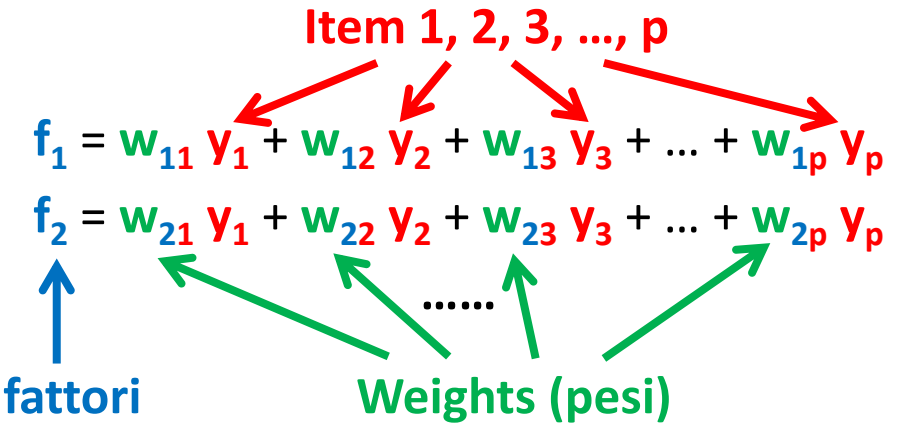
Analisi MULTIVARIABILE: richiesta dai reviewer delle riviste internazionali

Analisi MULTIVARIATA: indispensabile per studiare variabili latenti



Analisi delle componenti principali

I fattori (**variabili latenti**) sono una combinazione lineare ponderata delle variabili originali.



Analisi delle componenti principali

Molte variabili raccolte

Analisi delle componenti principali

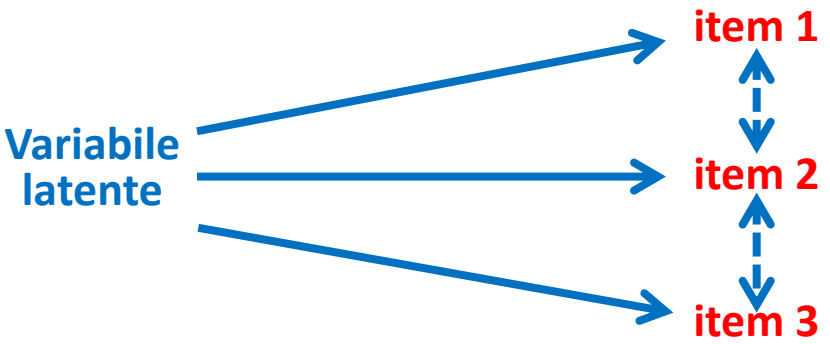
Alcuni fattori (variabili latenti)

L'analisi fattoriale consente di:

1) determinare quali variabili latenti sono alla base della scala di misura
2) ridurre il numero di variabili da prendere in considerazione (non più tutte le domande del questionario ma alcune dimensioni o variabili latenti)
3) attribuire un significato ad ogni fattore, identificando gruppi di item collegati alla stessa variabile latente

Alpha di Cronbach

La variazione di ciascun item dovrebbe riflettere in primo luogo la variazione della variabile latente sottostante, ma anche ad **un termine d'errore (variabilità individuale**). Di conseguenza i singoli item dovrebbero essere fortemente correlati tra loro.



Matrice di varianza-covarianza = varianza del vettore degli item. La **variabilità condivisa tra gli item** è rappresentata in blu, e la **variabilità non condivisa** è rappresentata in rosso

	item 1	item 2	item 3	item 4	item 5
item 1	σ^2	COV_{12}	COV_{13}	COV_{14}	COV_{15}
item 2	COV_{21}	σ^2	COV_{23}	COV_{24}	COV_{25}
item 3	COV_{31}	COV_{32}	σ^2	COV_{34}	COV_{35}
item 4	COV_{41}	COV_{42}	COV_{43}	σ^2	COV_{45}
item 5	COV_{51}	COV_{52}	COV_{53}	COV_{54}	σ^2

Alpha di Cronbach = $\frac{\text{Variabilità condivisa tra gli item}}{\text{Variabilità totale}}$

$$\alpha = \frac{p}{1-p} \cdot \left(1 - \frac{\sum \sigma^2_{item}}{\sigma^2_{totale}}\right)$$

Fattore di correzione

% var non-condivisa

% var condivisa

Alpha di Cronbach

Alpha di Cronbach	Interpretazione	Interpretazione
< 0,70	Bassa correlazione tra gli item	Scartare gli item poco correlati con il punteggio totale (correlazione prossima a zero)
	Costrutto eterogeneo	Rivalutare le dimensioni del questionario e applicare l'alpha alle singole dimension
	Poche domande	
0,70 - 0,90	Valori accettabili	Proseguire con la validazione del questionario
> 0,90	Item ridondanti	Ottimizzare (ridurre) la lunghezza della scala di misura

Split-half items (item spezzati a metà)

- La scala viene suddivisa in due gruppi omogenei di item (ad esempio, item pari e dispari)
- Si calcola la correlazione tra i punteggi totali dei due sottogruppi di item (split-half R)

Validità

- di contenuto (concettuale)
- di criterio (empirica)
- di costrutto (concettuale ed empirica)

- di facciata

Validità

VALIDITA'	Definizione	Come misurarla
di contenuto	Capacità dello strumento di misurare effettivamente ciò che si intende rilevare. Lo strumento deve essere coerente con quanto si intende rilevare e rappresentativo degli obiettivi dell'indagine. L'insieme degli item prescelti per un concetto deve coprire l'intero dominio di significato del concetto stesso.	Viene controllata su un piano **puramente teorico**, scomponendo il concetto studiato in tutte le sue dimensioni e assicurandosi che vengano tutte indagate dallo strumento. Ad esempio, un questionario sul quoziente d'intelligenza copre tutti gli aspetti dell'intelligenza stessa? Analisi indipendente dello strumento da parte di un **gruppo di esperti.**
di criterio	Non viene verificata sul piano teorico ma su quello empirico, ovvero all'atto di rilevare i dati. Viene valutata la corrispondenza tra i risultati ottenuti con il nuovo strumento e un criterio esterno correlato o un altro strumento già validato (**gold standard).**	Si confrontano i risultati ottenuti con il test nuovo da validare e un test vecchio già validato. Se i due strumenti misurano la stessa cosa, si valuta la **validità convergente**. Se invece misurano aspetti diversi, si valuta la **validità divergente.** Se i due strumenti vengono somministrati allo stesso tempo, si valuta la **validità concorrente**. Lo strumento ha **validità predittiva**, se è in grado di predire eventi successivi (ad es., il test d'ammissione all'Università predice la media dei voti).
di costrutto	Riunisce e riassume le validazioni effettuate sul piano teorico (di contenuto) e sul piano empirico (di criterio). Corrispondenza tra i risultati empirici ottenuti e la costruzione teorica proposta.	Si ha validità di costrutto se la riflessione teorica e le rilevazioni empiriche collimano nell'indicare che lo strumento riesce a rappresentare il costrutto (fenomeno non osservabile). Si possono utilizzare anche degli item di validazione (**validation items**).

VALIDITA' DI CRITERIO

Lo strumento di misura viene confrontato con un "**gold standard**".

Ad esempio le risposte ad un questionario sul fumo possono essere confrontate con la concentrazione di CO o di cotinina (metabolita della nicotina) nel sangue (Olivieri et al, 2002).

Analogamente le risposte a un questionario sull'asma possono essere confrontate con le prove allergologiche, la spirometria o la diagnosi clinica di asma (de Marco et al, 1998).

Bibliografia

de Marco R, Cerveri I, Bugiani M, Ferrari M, Verlato G (1998) An undetected burden of asthma in Italy: the relationship between clinical and epidemiological diagnosis of asthma. *Eur Respir J*, 11: 599-605

Olivieri M, Poli A, Zuccaro P, Ferrari M, Lampronti G, de Marco R, Lo Cascio V, Pacifici R (2002) Tobacco smoke exposure and serum cotinine in a random sample of adults living in Verona, Italy. Arch Environ Health, 57: 355-359

VALIDITA' DI FACCIATA

VALIDITA'	Definizione	Come misurarla
di facciata / accettabilità (validità apparente)	Analisi dell'adeguatezza del materiale-stimolo ai soggetti da esaminare	Attraverso la richiesta di un giudizio qualitativo intuitivo ai partecipanti (**tempo, risposte multiple, refusi di compilazione**)

Una volta che il questionario è stato sviluppato …

Test-retest:

1. nelle stesse condizioni (stabilità, ripetibilità) con gli stessi osservatori (**intra-rater agreement**)
2. o in condizione diverse (concordanza, riproducibilità) con osservatori diversi (**inter-rater agreement**)

In genere il test-retest viene effettuato su un sottocampione.

Concordanza fra osservatori

- Per dati nominali: kappa di Cohen
- Per dati ordinali: kappa di Cohen pesata (weighted kappa)
- Per dati quantitativi: r di Pearson, coefficiente di correlazione intraclasse

Esiste un *gold standard* (riferimento aureo)?

SI'

Si calcola la SENSIBILITÀ e la SPECIFICITÀ del test contro il gold standard

NO: si calcola la misura KAPPA di CONCORDANZA, introdotta da COHEN

Esempio:

Nella diagnosi del cancro colo-rettale il gold standard è la colonscopia + biopsia

***Inter-rater* o *intra-rater agreement*: concordanza fra i giudizi espressi da operatori diversi o dallo stesso operatore**

Studio sulla discinesia tardiva

giudice a		giudice b: *presente*	giudice b: *assente*	*totale*
	pres.	**123**	**10**	**133**
	ass.	**6**	**29**	**35**
	totale	**129**	**39**	**168**

$\chi^2 = 88.23 \longrightarrow P < 0.0000001$

Bergen et al (1992) Studio sulla discinesia tardiva.

Il chi-quadrato ci dà scarse informazioni.

Abbiamo semplicemente falsificato l'ipotesi nulla che le diagnosi del primo clinico fossero del tutto indipendenti dalle diagnosi del secondo clinico.

Tuttavia questo risultato può essere conseguito anche da due clinici molto MEDIOCRI.

Misura di concordanza osservata (Proportion of agreement)

giudice b

giudice a	*presente*	*assente*	*totale*
pres.	**a**	**b**	**a+b**
ass.	**c**	**d**	**c+d**
totale	**a+c**	**b+d**	**N**

$$P_0 = \frac{a + d}{N}$$

Misura di concordanza osservata (Proportion of agreement)

giudice b

giudice a	*presente*	*assente*	*totale*
pres.	**123**	**10**	**133**
ass.	**6**	**29**	**35**
totale	**129**	**39**	**168**

$$P_0 = \frac{123 + 29}{168} = \frac{152}{168} = 0.905$$

Bergen et al (1992) Studio sulla discinesia tardiva.

Però anche due persone completamente ignoranti di medicina riescono a raggiungere una certa concordanza, per semplice effetto del caso.

A noi interessa valutare di quanto la concordanza fra gli esperti eccede la concordanza attesa per semplice effetto del caso.

Concordanza ATTESA
per effetto del caso
(agreement expected just by chance)

$$P_{exp} = \frac{a_{exp} + d_{exp}}{N}$$

Sotto l'ipotesi di indipendenza statistica

atteso = (totale riga) * (totale colonna) / (totale generale)

giudice a \ giudice b	*presente*	*assente*	*totale*
pres.	**a**	**b**	**a+b**
ass.	**c**	**d**	**c+d**
totale	**a+c**	**b+d**	**N**

$$a_{exp} = \frac{(a+b) * (a+c)}{N}$$

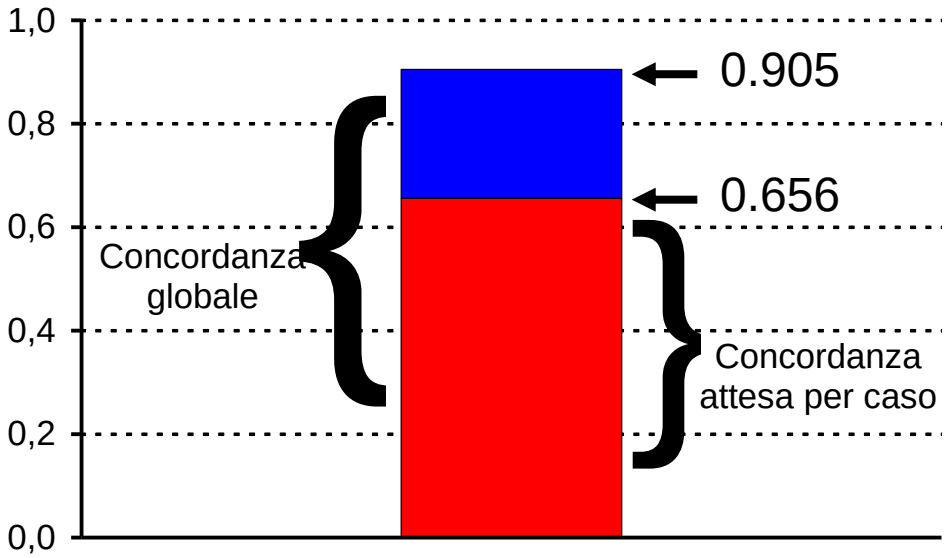
Concordanza attesa

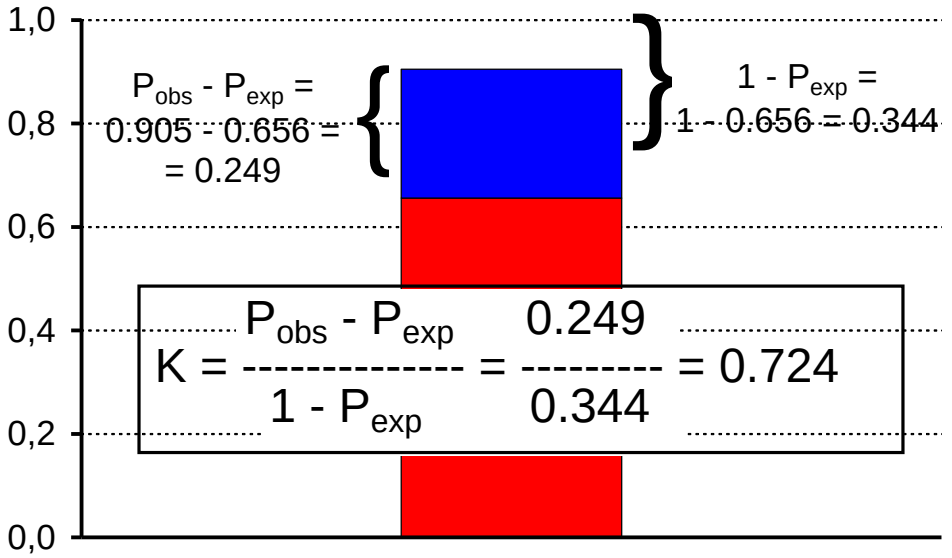
giudice a \ giudice b	*presente*	*assente*	*totale*
pres.	**123**	**10**	**133**
ass.	**6**	**29**	**35**
totale	**129**	**39**	**168**

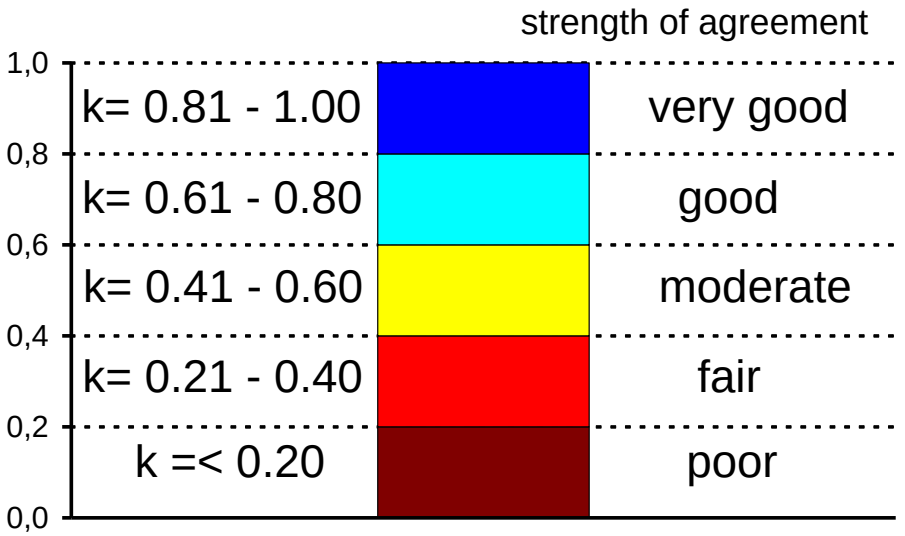
a_{atteso} = (133 * 129) / 168 = 102.125

d_{atteso} = (35*39) / 168 = 8.125

$$P_e = \frac{102.1+8.1}{168} = \frac{110.25}{168} = 0.656$$


1,0
0,8
0,6
0,4
0,2
0,0
0.905
0.656
Concordanza globale
Concordanza attesa per caso


1,0
0,8
0,6
0,4
0,2
0,0
$P_{obs} - P_{exp}$ = 0.905 - 0.656 = = 0.249
1 - P_{exp} = 1 - 0.656 = 0.344
$K = \frac{P_{obs} - P_{exp}}{1 - P_{exp}} = \frac{0.249}{0.344} = 0.724$



Altman DG (1991) Practical statistics for medical research. Chapman & Hall: London.

RIPETIBILITA' / RIPRODUCIBILITA' con dati quantitativi

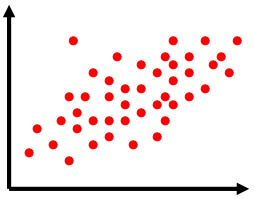
Coefficiente di correlazione di Pearson

$$r = \frac{\text{codevianza}_{xy}}{\sqrt{\text{devianza}_x * \text{devianza}_y}}$$

Coefficiente di correlazione intraclasse

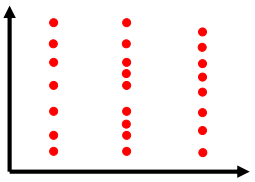
$$ICC = \frac{\text{varianza}_{\text{tra gruppi}}}{\text{varianza}_{\text{tra gruppi}} + \text{varianza}_{\text{entro gruppi}}}$$

Coefficiente di correlazione di Pearson
Dati appaiati

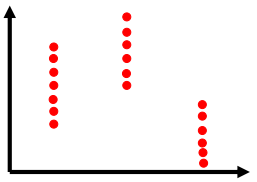


Coefficiente di correlazione intraclasse
Dati suddivisi in gruppi

ICC prossimo a zero



ICC prossimo a uno



References

De Vellis RF. Scale Development. Theory and Application. Newbury Park CA: Sage Publications Inc., 1991.

Item-to-item or item-to-total correlation

Hair JF, Black WC, Babin B, Anderson RE, Tathan RE. Multivariate data analysis (6° edition). Prentice Hall: New Jersey, 2006

ICC

Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. Tutor Quant Methods Psychol 2012; 8:23-34

Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005; 19:231-40

Alpha di Cronbach

Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Ed 2011; 2:53-55

Esempio

Galeoto G, Sili A, Tamburlani M, Farina M, Mannocci A, Mollica R, Servadio A. Costruzione e validazione di uno strumento di valutazione dei rischi ambientali e delle limitazioni alla movimentazione manuale dei carichi. Clin Ter 2017; 168(6):e349-356