Il modello di misura

Attendibilità e congruenza delle misure

(cap. 12)

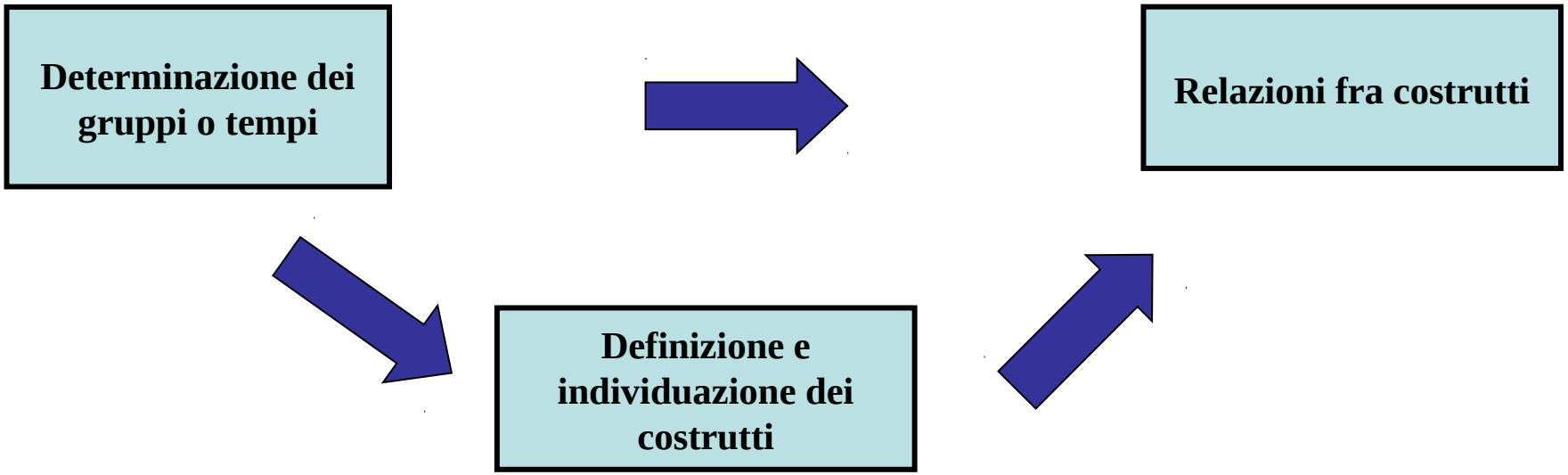
Marcello Gallucci
Univerisità Milano-Bicocca

A
M
D

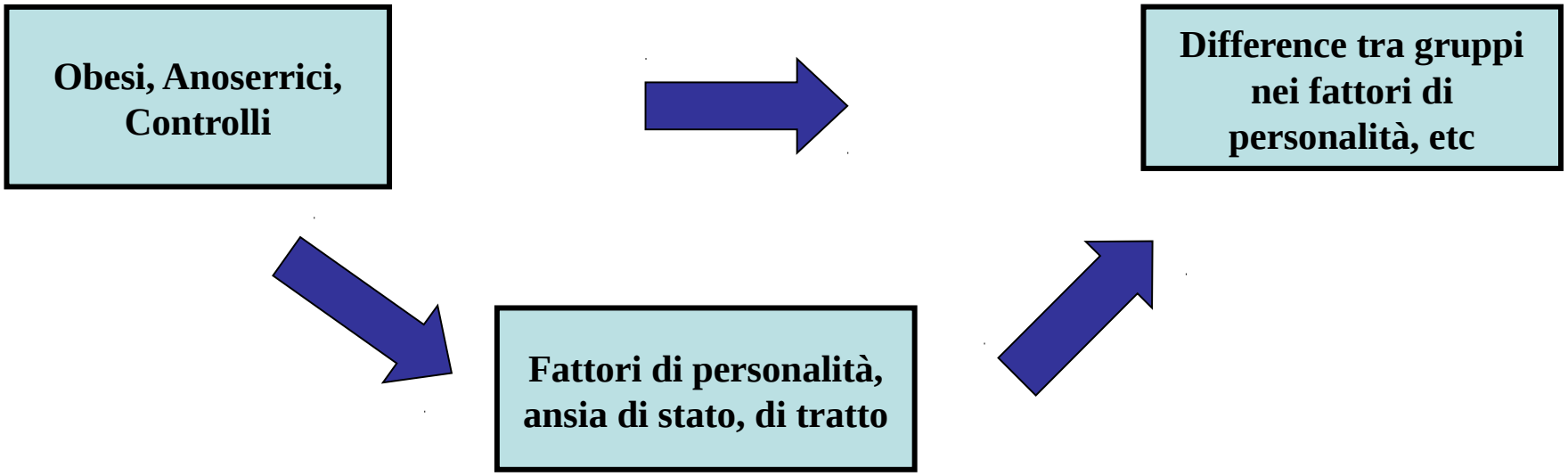
Preludio

- La maggior parte delle ricerche empiriche in ambito psicologico condividono le seguenti caratteristiche
  - Data un'ipotesi (o una serie di ipotesi) viene formulato un disegno di ricerca
  - Pianificato una set di misurazioni per i costrutti rilevanti
  - Studiate le relazioni tra costrutti

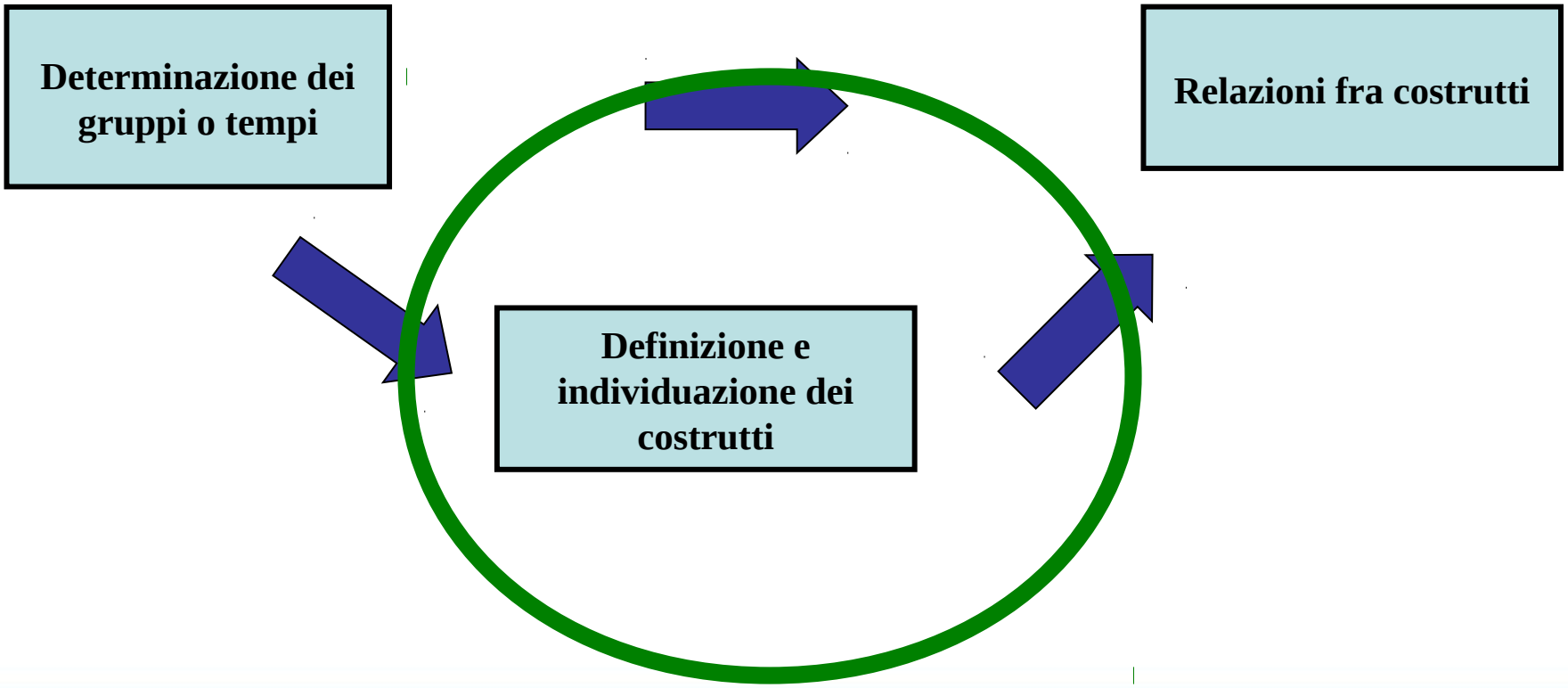
Logica delle analisi statistiche



Logica delle analisi statistiche



Modello di Misurazione



Modello di Misurazione

- Stabilire le proprietà psicometriche delle nostre misure
- Stabilire se il le misure selezionate abbiano “retto” nel nostro campione
- Nelle ricerche più sofisticate, stabilire se le nostre misure abbiano la **struttura dimensionale** e la **coerenza** attesa sulla base della letteratura

Validità vs Attendibilità

- **L'attendibilità** → la misura cattura coerentemente un costrutto

- **La validità** → la misura cattura effettivamente il costrutto atteso

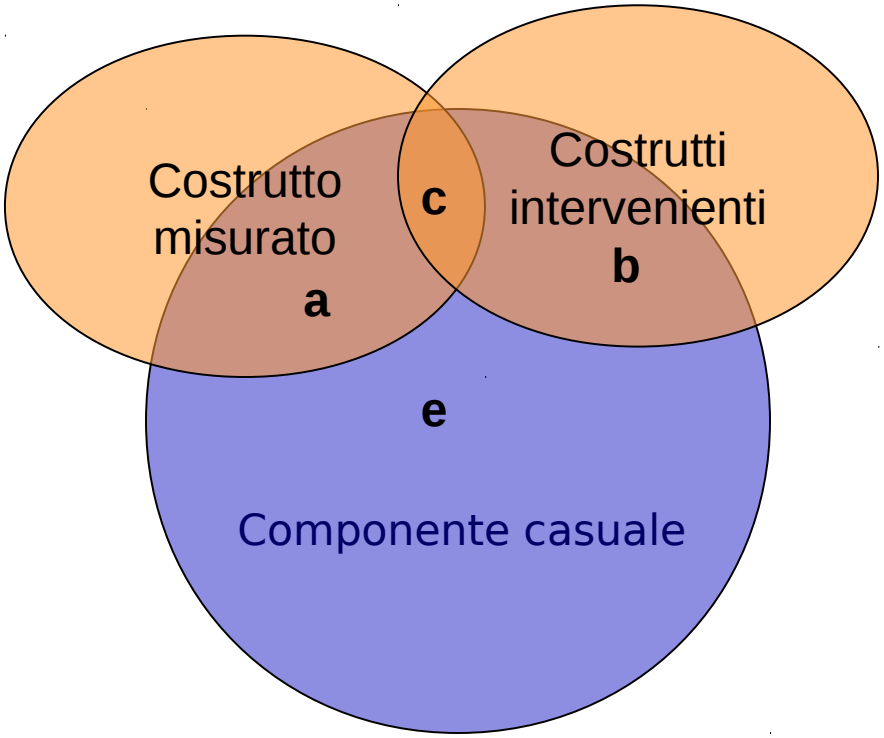
(la scala di metacognizione misura la metacognizione)

- Escludendo gli studi di validità, la maggior parte degli studi

empirici di cui trattiamo si preoccupa dell'attendibilità, ed assume

(sulla base della letteratura) la validità

Fonti di variazione dei punteggi

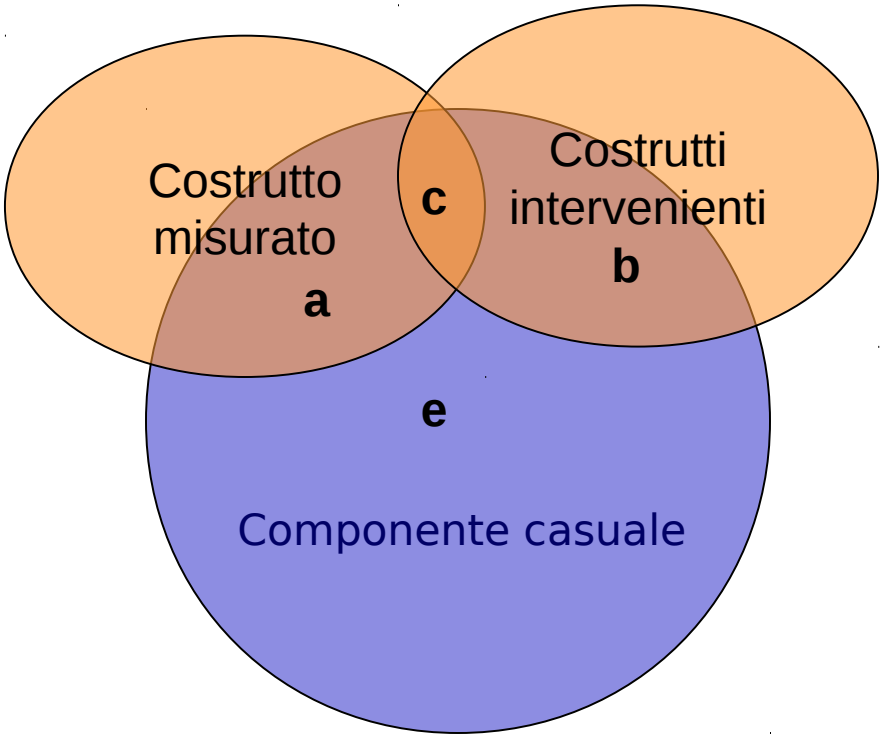


a+c+b
Variabilità Sistematica

a+c+b+e
Variabilità osservata

Attendibilità vs Validità

- La validità attiene alla corrispondenza tra misura e costrutto misurato, **a** grande rispetto a **b**

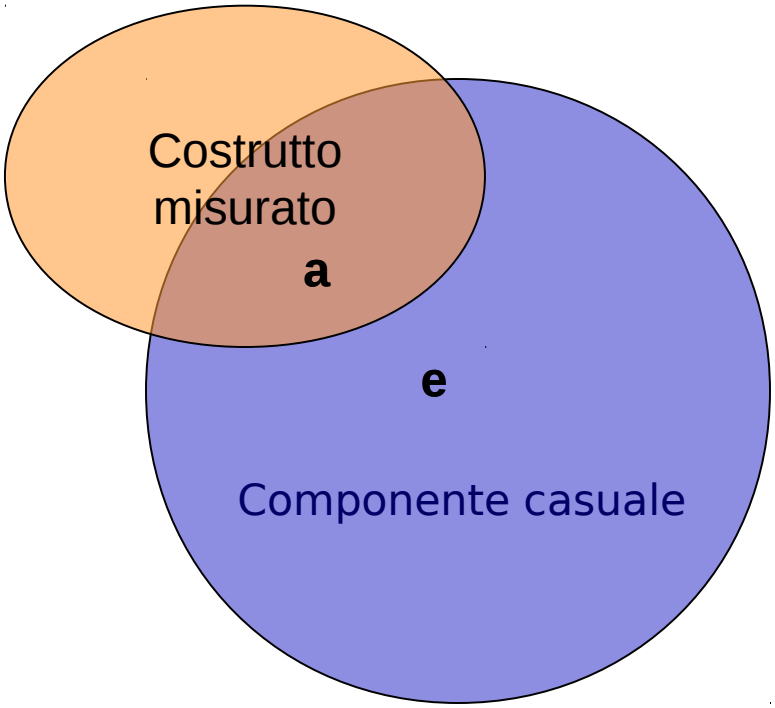


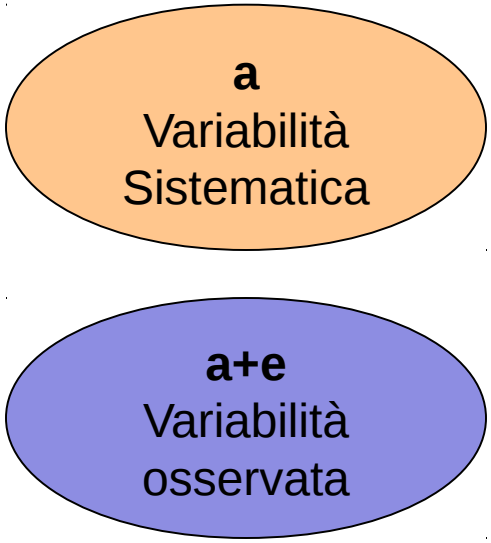
a+c+b
Variabilità Sistematica

a+c+b+e
Variabilità osservata

Attendibilità

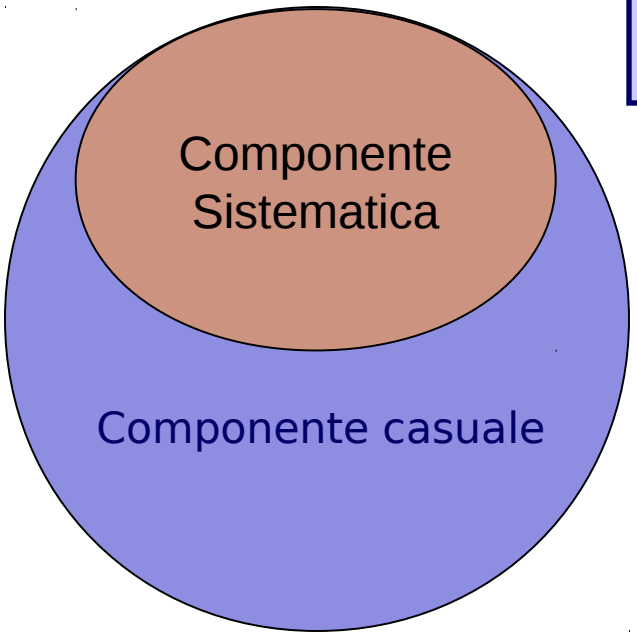
- L’attendibilità attiene alla reproducibilità e coerenza dei punteggi di una misura (**a** grande rispetto a **e**)





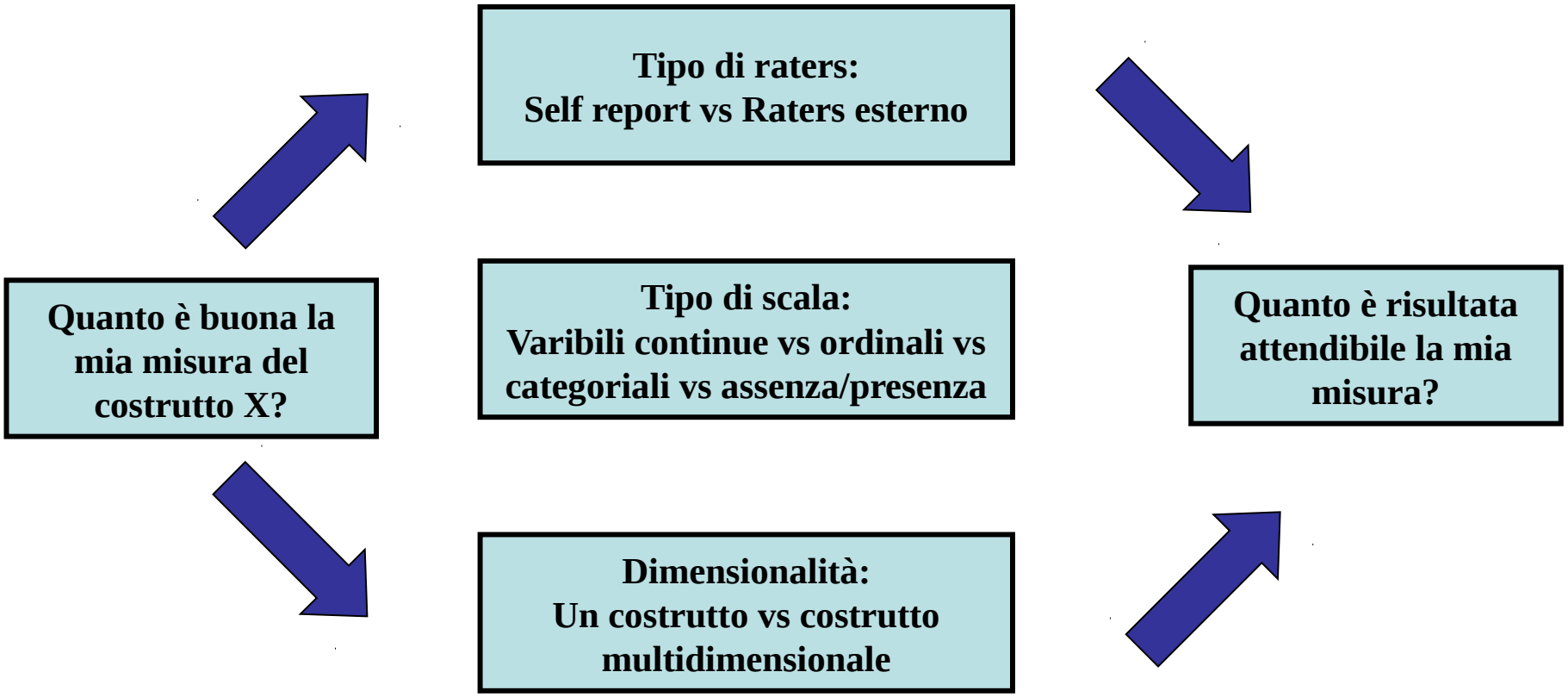
Il concetto di buona misura

- (A parte i dettagli) Tutte le forme di studio della bontà della misura insistono sul concetto di componente “sistematica” e componente “casuale” del dato osservato



Una buona misura avrà una componente sistematica grande rispetto a quella casuale

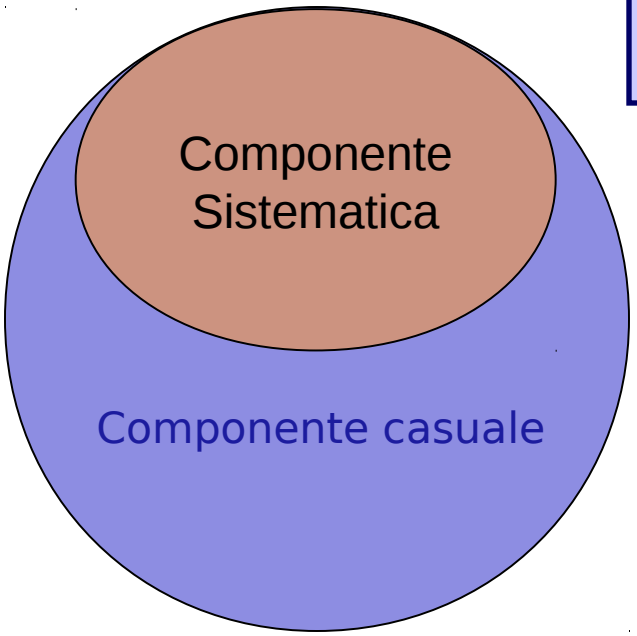
Bontà delle misure



Il concetto di buona misura

- Le differenze tra le varie tecniche di assessment della bontà di una misura risiede nel tipo di informazione da trattare (e.g. un valore numerice vs una categoria) e dalla definizione di componente sistematica

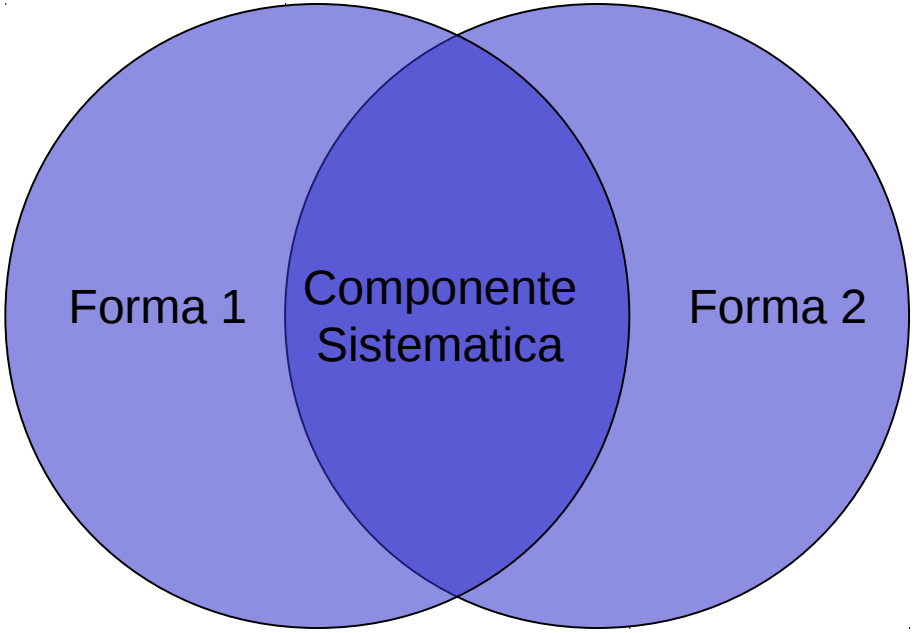
Una buona misura avrà una componente sistematica grande rispetto a quella casuale



La componente sistematica

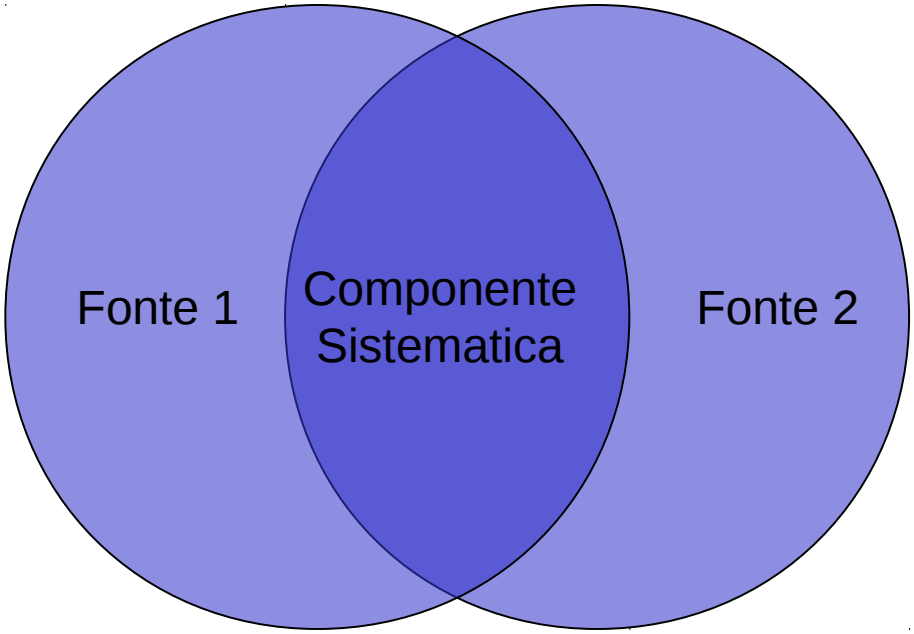
- A volte detta la componente “vera” della misura

- Se la misura è attendibile, differenti forme di misura convergeranno nella componente sistematica e non in quella casuale



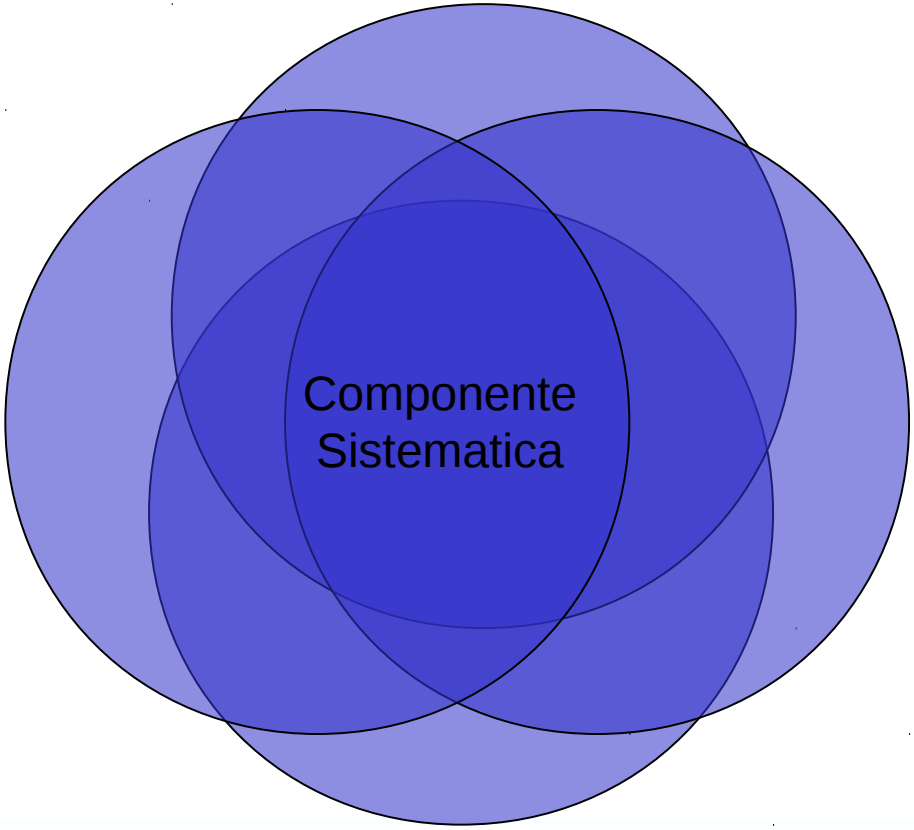
La componente sistematica

- In generale, differenti fonti di informazione relative al costrutto misurato dovrebbero convergere nell'indicare la quantità/qualità del costrutto per il caso sotto osservazione



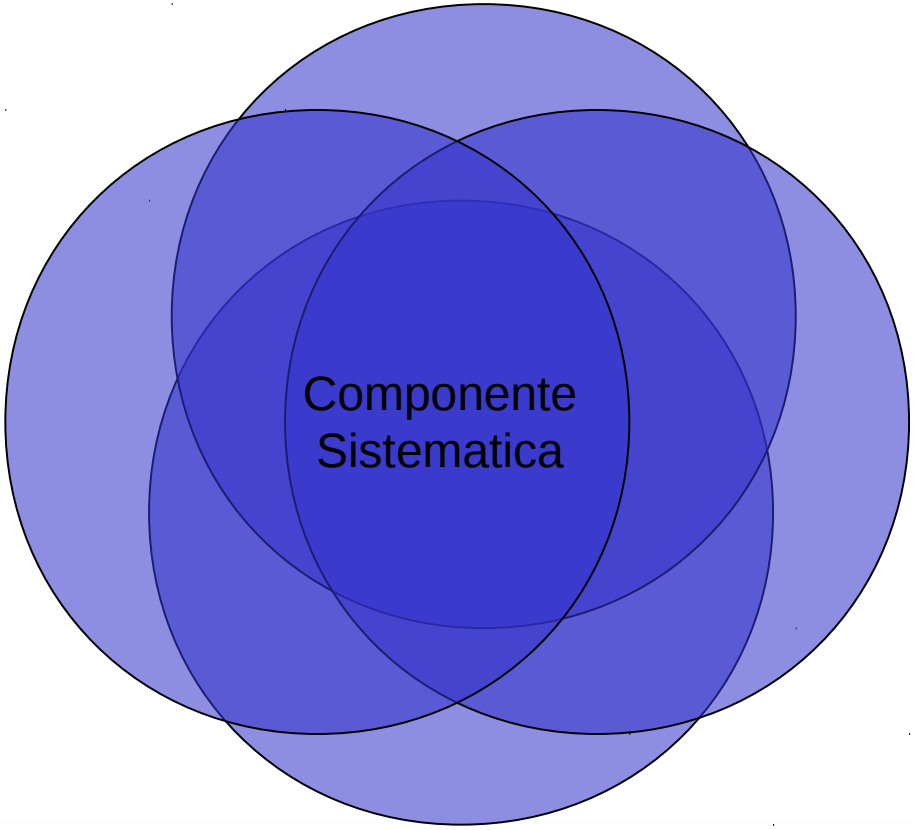
Repetita iuvant

- In statistica, se una cosa si può ripetere due volte, si può ripetere N volte!



Convergenza di informazioni

- In buona sostanza, tutte le tecniche di rilevazione della bontà di una misura condividono la stessa logica: Il tratto misurato emerge dalla convergenza tra “fonti” diverse di misurazione dello stesso costrutto



Caso I

- Scala di valutazione di caratteristiche categoriche (*nominali: aggressivo vs non aggressivo, attivo vs passivo, etc*)

- Griglia di valutazione di raters esterni (*due terapeuti che classificano un trascritto di una serie di sedute*, *tre docenti che classificano atti comportamentali degli alunni, etc)*

Congruenza tra raters

- Assumiamo che due raters valutino N pazienti relativamente alla presenza di un comportamento (aggressivo)

	Rater 1	Rater 2
Paziente 1	Si	No
Paziente 2	Si	Si
Paziente 3	No	No
...		
Paziente N	No	Si

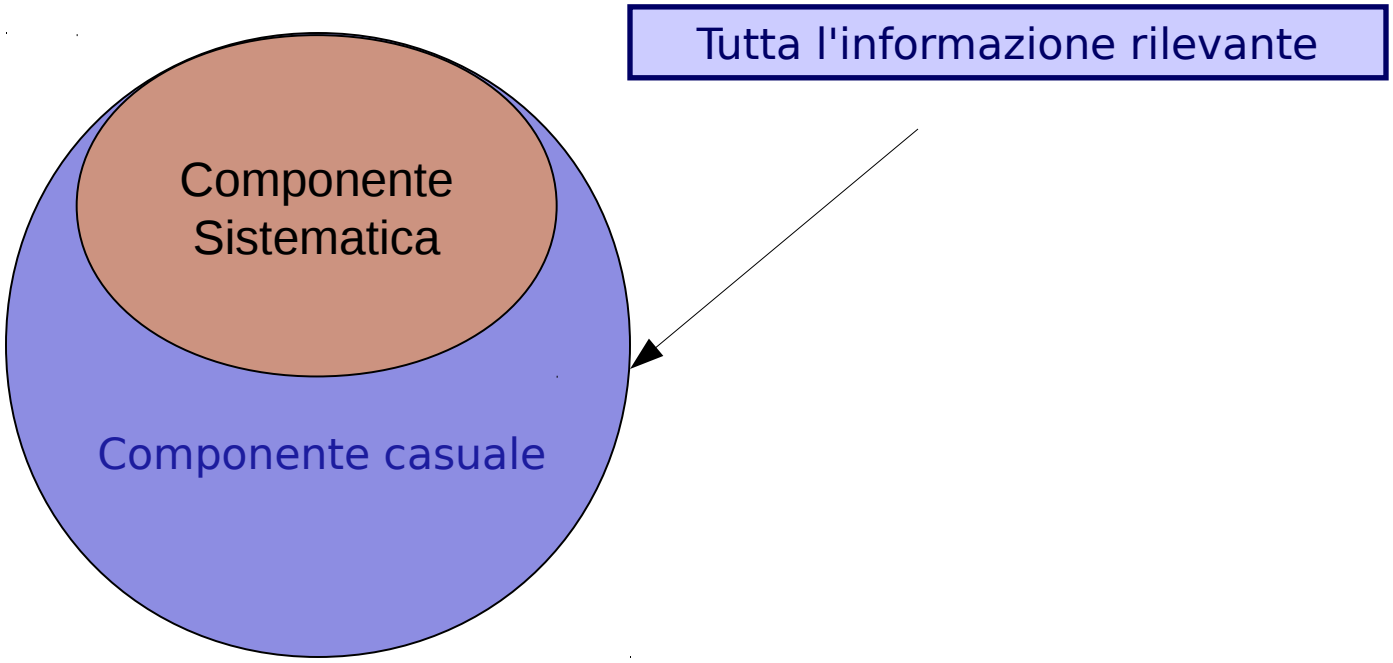
Congruenza tra raters

- Incrociando i giudizi otteniamo una tabella di valutazione

	Rater 1		
	No	Si	Tot
No	30	10	**40**
Si	30	30	**60**
Tot	**60**	**40**	100

Quale sarà l'informazione a disposizione?

- L'informazione rilevante sarà data data dalla combinazione di giudizi uguali (aggressivo – aggressivo, non aggressivo-non aggressivo)



Probabilità di convergenza

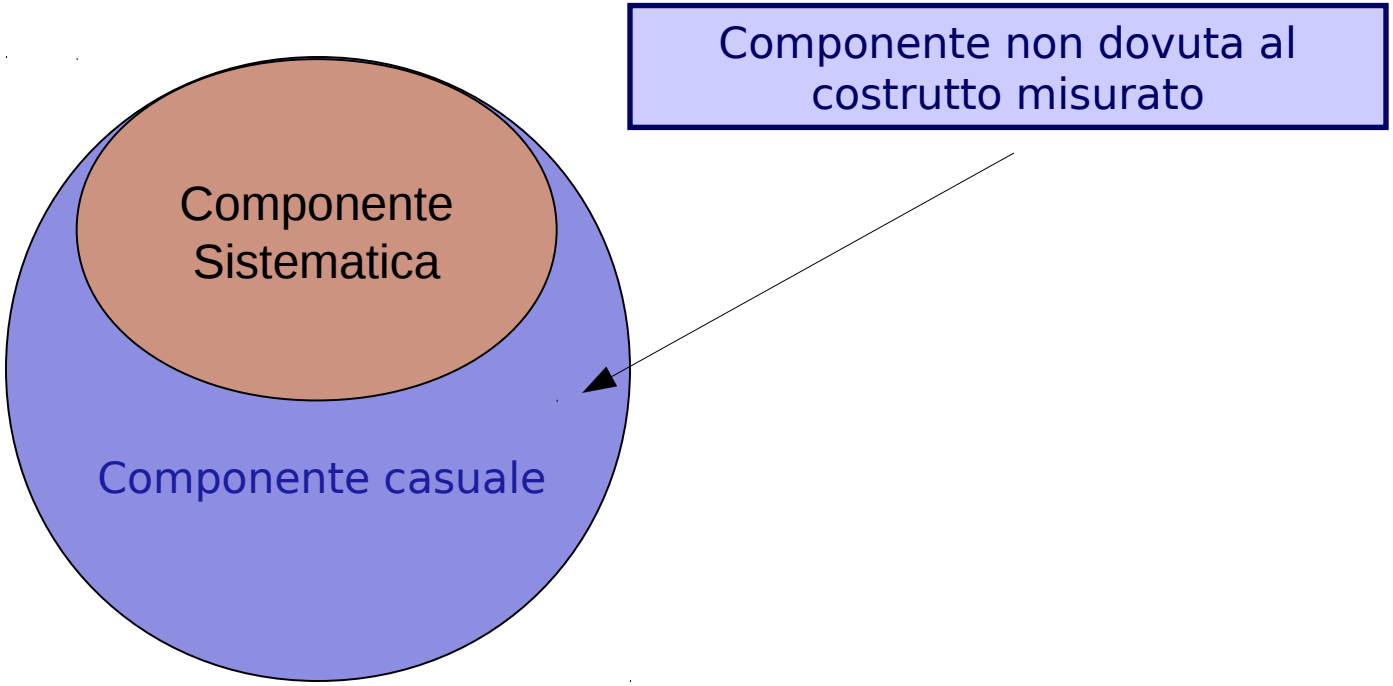
- Sarà data dalla proporzione di giudizi congruenti

	Rater 1		
	No	Si	Tot
No	**<u>30</u>**	10	**40**
Si	30	**<u>30</u>**	**60**
Tot	**60**	**40**	100

$$30/100 + 30/100 = .60$$

Quale sarà la componente casuale?

- La componente casuale sarà data dalla combinazione di giudizi uguali (aggressivo – aggressivo, non aggressivo-non aggressivo) dovuta a puro caso



Congruenza Causale

- **Tabella frequenze attese**: Se i raters fossero completamenti indipendenti (guidati dal caso) le frequenze nelle celle dipenderebbero solo dalle frequenze marginali

	Rater 1		
	No	Si	Tot
No	24	16	**40**
Si	36	24	**60**
Tot	**60**	**40**	100

$$Freq(cella) = Freq(Rater1) * Freq(Rater2) / Freq(Tot)$$

Congruenza Causale

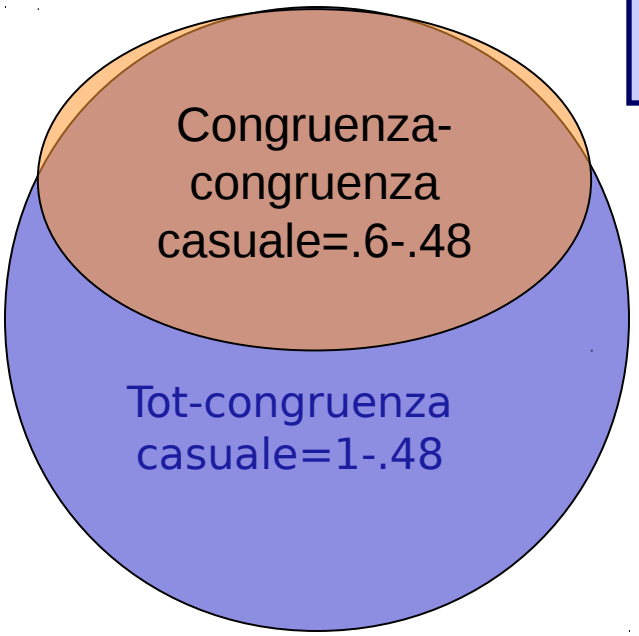
- La congruenza dovuta al caso è data dalla somma delle proprozioni delle celle congruenti nella tabella delle frequenze attese

	Rater 1		
	No	Si	Tot
No	24	16	**40**
Si	36	24	**60**
Tot	**60**	**40**	100

$$24/100 + 24/100 = .48$$

Convergenza

- La congruenza dipendenderà da quanto i rater convergono rispetto a quello che farebbe il caso



Una buona misura avrà una componente sistemativa grande rispetto a quella casuale

$$\frac{C_{osservata} - C_{casuale}}{1 - C_{casuale}}$$

Kappa di Cohen

- Abbiamo inventato la K di Cohen: indice di congruenza tra raters

$$K = \frac{C_{osservata} - C_{casuale}}{1 - C_{casuale}}$$

Congruenza perfetta

$$K = \frac{1 - C_{casuale}}{1 - C_{casuale}} = 1$$

I giudici vanno a caso

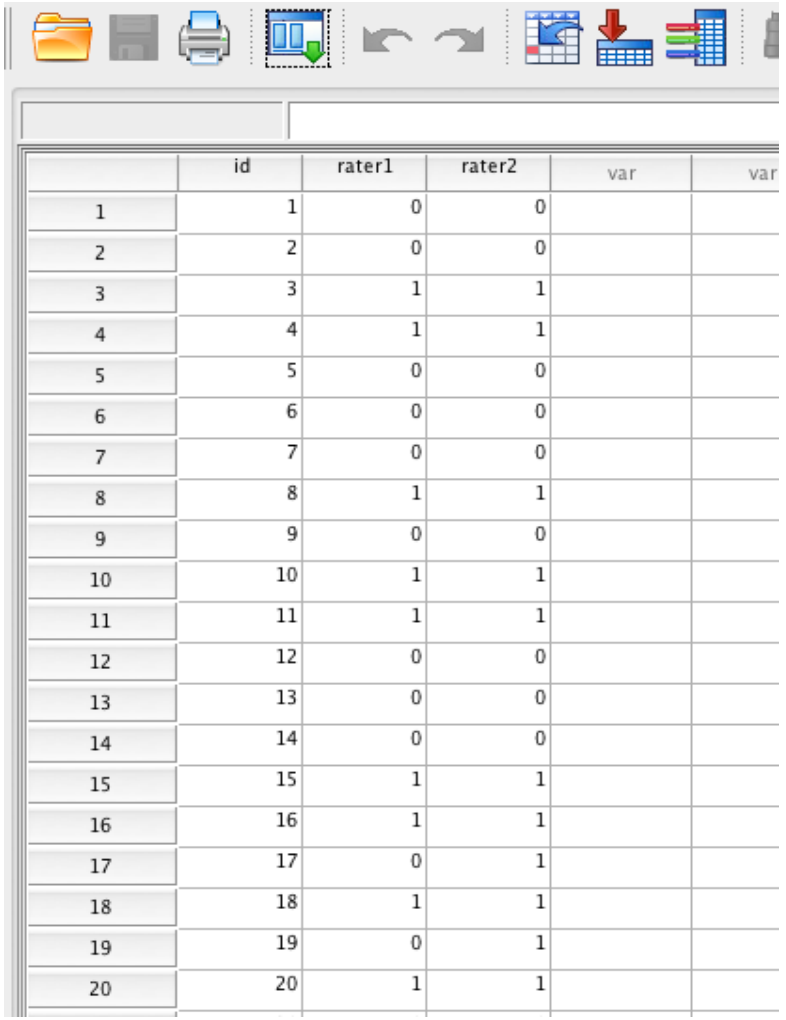
$$K = \frac{C_{casuale} - C_{casuale}}{1 - C_{casuale}} = 0$$

Caratteristiche

- Si adatta alle valutazioni nominali
- Richiede punteggi >.80 (generalmente accettato come buono)
- Dipende dal numero di categorie. Maggiore è il numero di categorie, più basso può essere il punteggio
- Non cattura congruenze sfalsate (tutte le volte che Rater 1 dice A, rater 2 dice B)
- E' indifferente all'ordine

Esempio

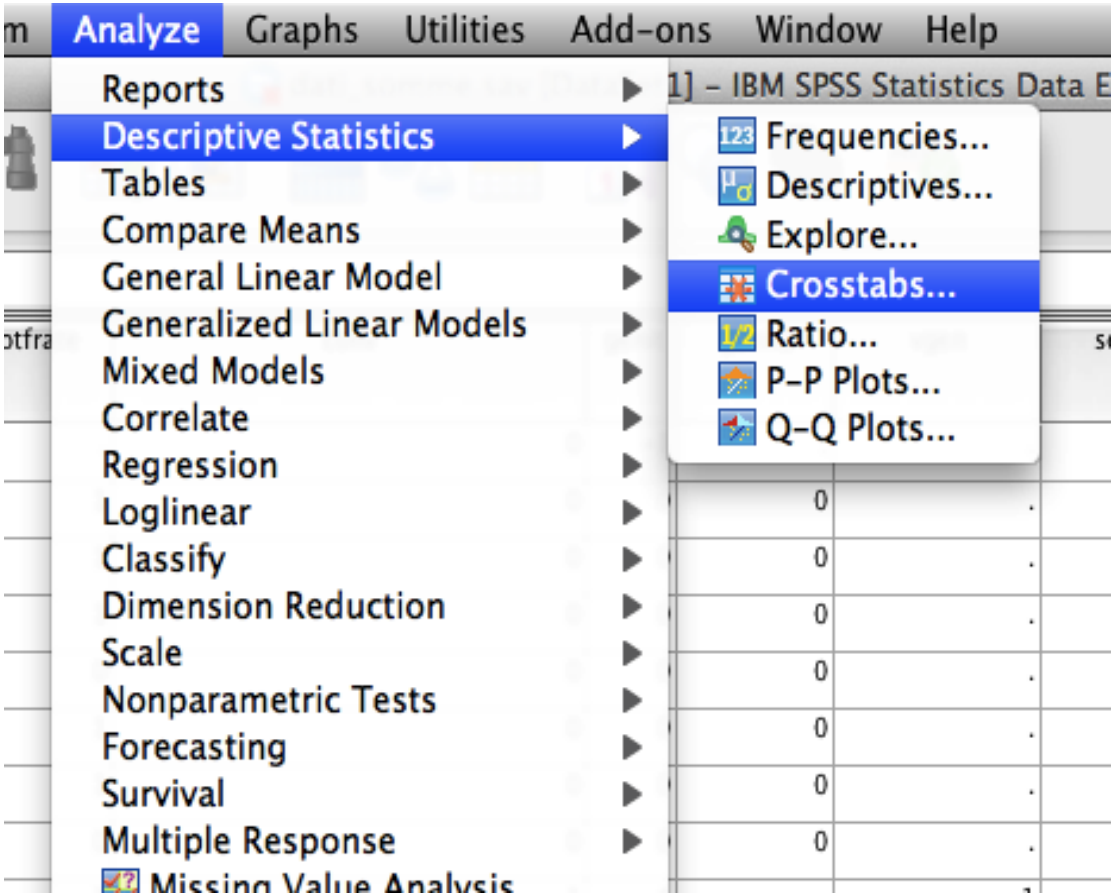
30 pazienti giudicati da due raters



	id	rater1	rater2	var	var
1	1	0	0		
2	2	0	0		
3	3	1	1		
4	4	1	1		
5	5	0	0		
6	6	0	0		
7	7	0	0		
8	8	1	1		
9	9	0	0		
10	10	1	1		
11	11	1	1		
12	12	0	0		
13	13	0	0		
14	14	0	0		
15	15	1	1		
16	16	1	1		
17	17	0	1		
18	18	1	1		
19	19	0	1		
20	20	1	1		

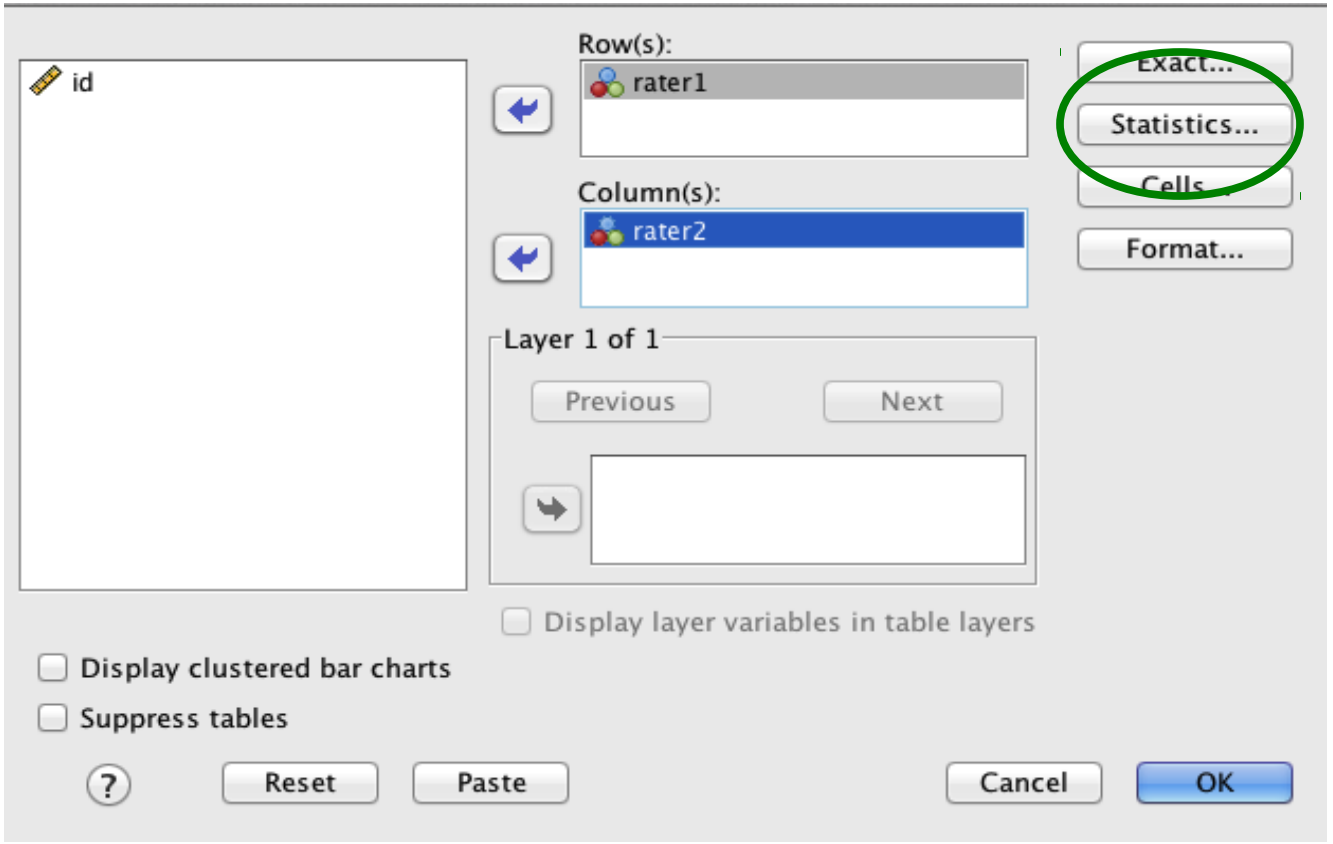
Esempio

30 pazienti giudicati da due raters



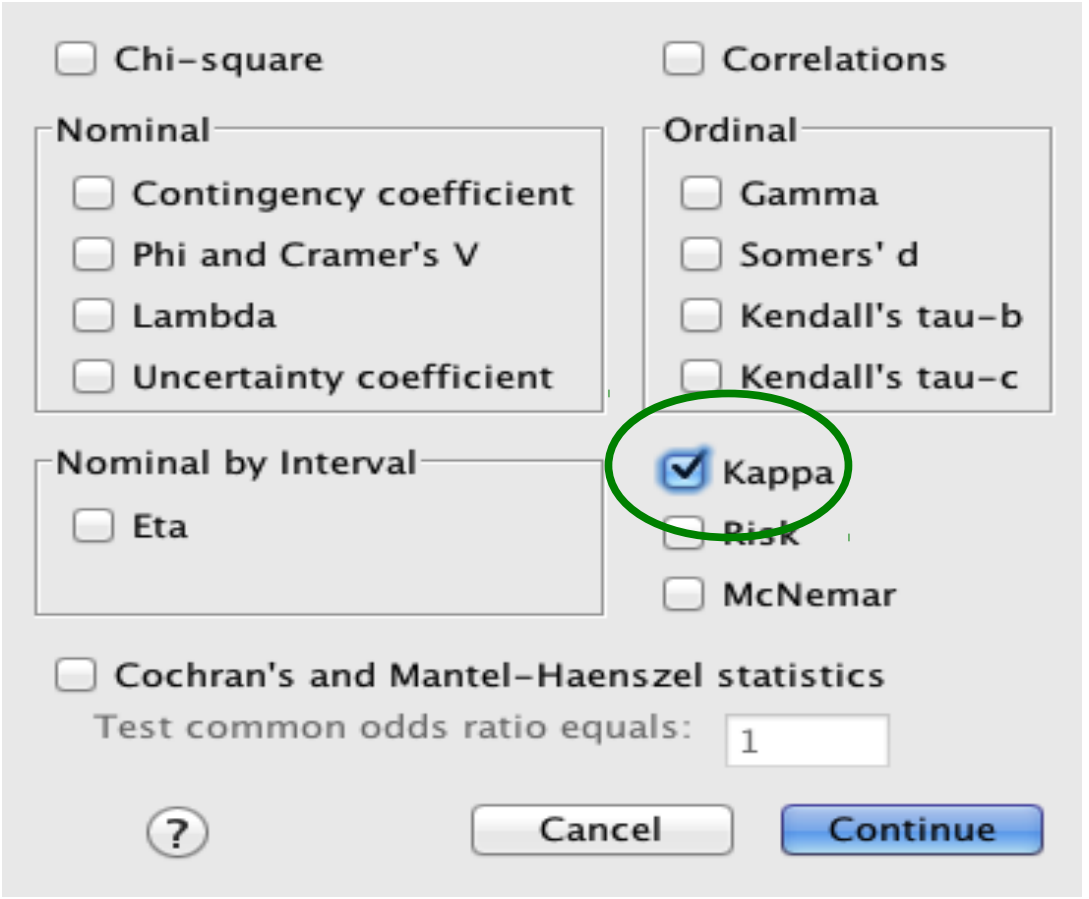
Esempio

Inserisco le variabili dove sono contenuti le risposte dei raters



Esempio

Seleziono K



Esempio

Risultati

rater1 * rater2 Crosstabulation

Count

		rater2		Total
		0	1	
rater1	0	12	2	14
	1	1	15	16
Total		13	17	30

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.798	.110	4.382	.000
N of Valid Cases		30			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

Repetita iuvant!

- Se abbiamo più di due raters, possiamo usare il K di Fleiss

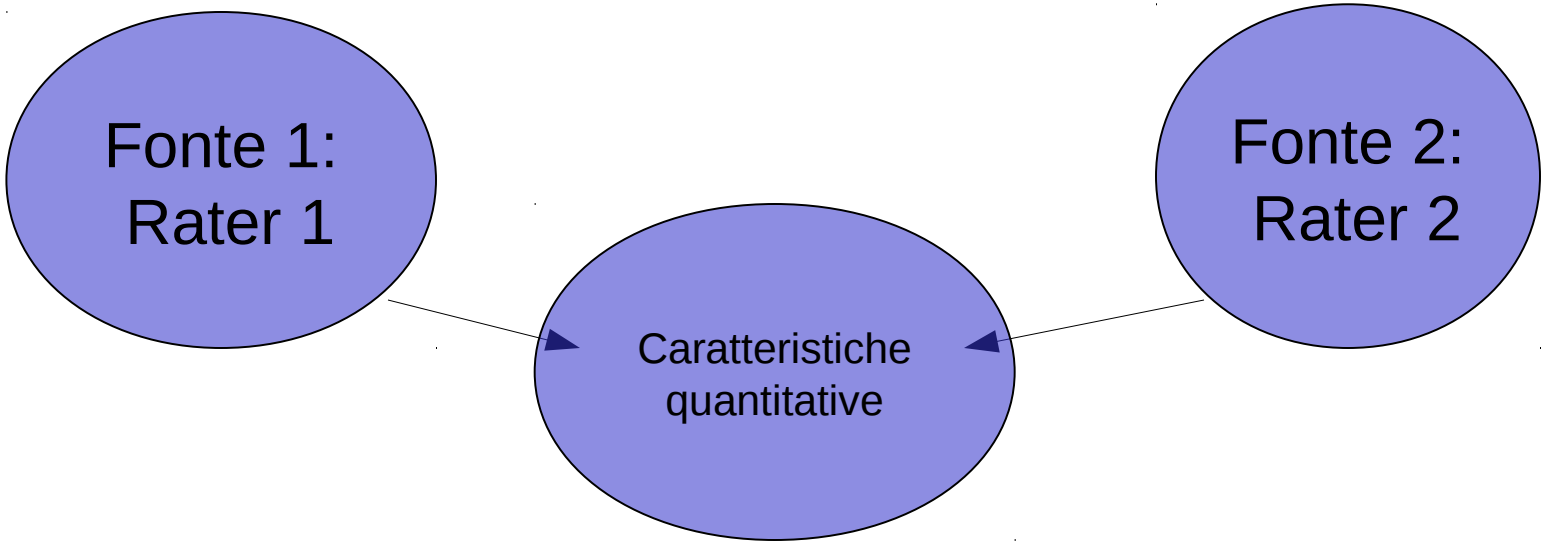
$$K = \frac{\bar{C}_{osservata} - \bar{C}_{casuale}}{1 - \bar{C}_{casuale}}$$

Dove la congruenza (probabilità di uguale risposta) viene sostituita dalla congruenza media fra raters

Caso II

- Scala di valutazione di caratteristiche continue (intensità di un tratto, item su scala Likert)

- Griglia di valutazione di raters esterni (*due terapeuti che compilano un questionario basato su scala Likert per vari pazienti)*



Congruenza tra raters

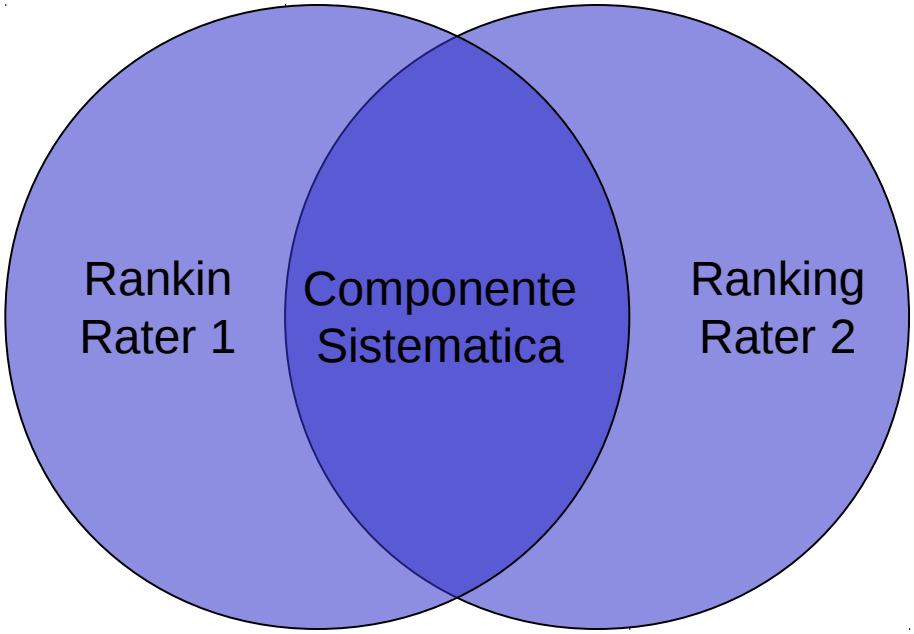
- Assumiamo che due raters valutino N pazienti relativamente alla frequenza di un comportamento (da per nulla a spesso in 7 passi)

	Rater 1	Rater 2
Paziente 1	5	7
Paziente 2	7	7
Paziente 3	0	1
...		
Paziente N	6	4

La componente sistematica

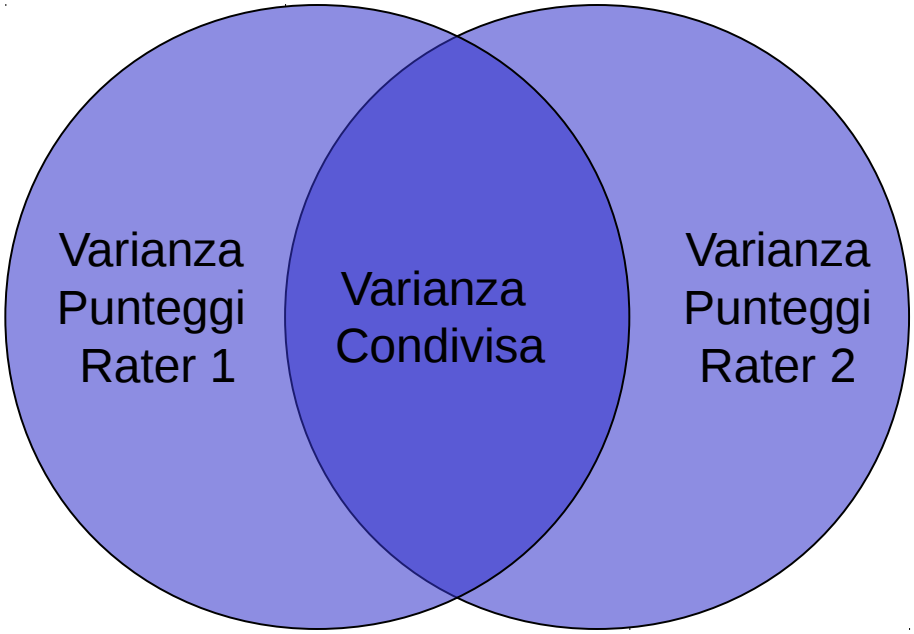
- La componente sistematica sarà data dalla correlazione (o varianza condivisa) dei punteggi dati due raters

- La componente di errore è data dalla varianza non condivisa



Coefficiente di correlazione-intraclasse

- Il coefficiente indica quanta variabilità dei punteggi è dovuta a differenze effettive tra soggetti (cioè riscontrate da tutti i raters) rispetto alle differenze tra i raters



Correlazione intraclasse

- Data la varianza tra casi Vc e la varianza tra i raters Vr

$$ICC = \frac{V_c}{(V_c + V_R)}$$

- Che equivale a dire: componente sistematica diviso informazione totale

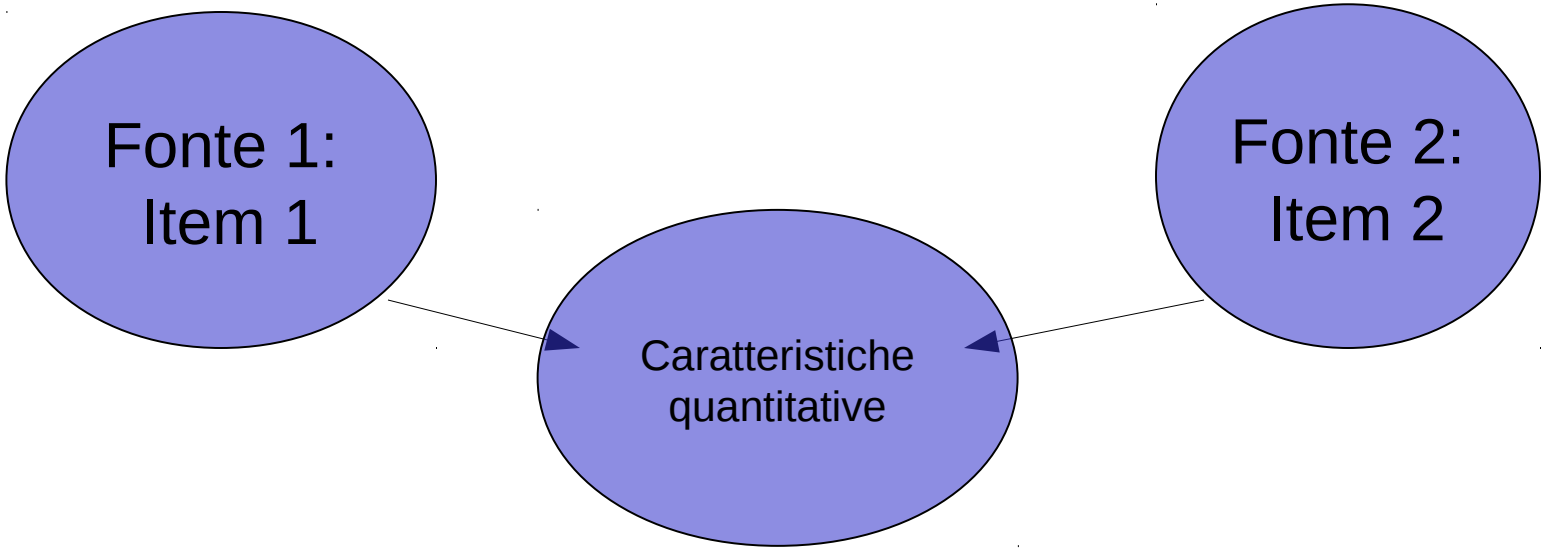
Caratteristiche

- Si adatta alle valutazioni quantitative
- Cattura congruenze sfalsate (se rater 1 attribuisce sistematicamente un punteggio minore del rater 2, i raters risulteranno convergenti)
- Si applica anche quando si hanno più di due raters

Caso III

- Scala di valutazione di caratteristiche continue (intensità di un tratto, items su scala Likert)

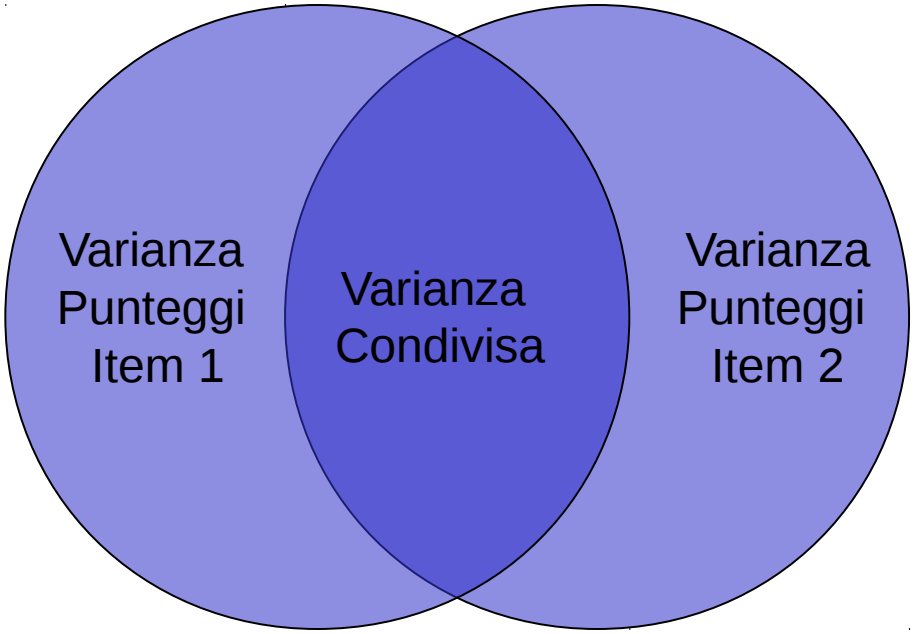
- Self-report (questionari classici)



Cogruenza nelle risposte tra item

- Se due item misurano lo stesso costrutto, i due item saranno correlati

- Dunque condivideranno varianza



Correlazione Pearson

- Data lE varianze V_{item} degli items

$$r = \frac{Covarianza_{items}}{\sqrt{(V_{item1} * V_{item2})}}$$

- Che equivale a dire: componente sistematica (comune) diviso informazione massima ottenibile

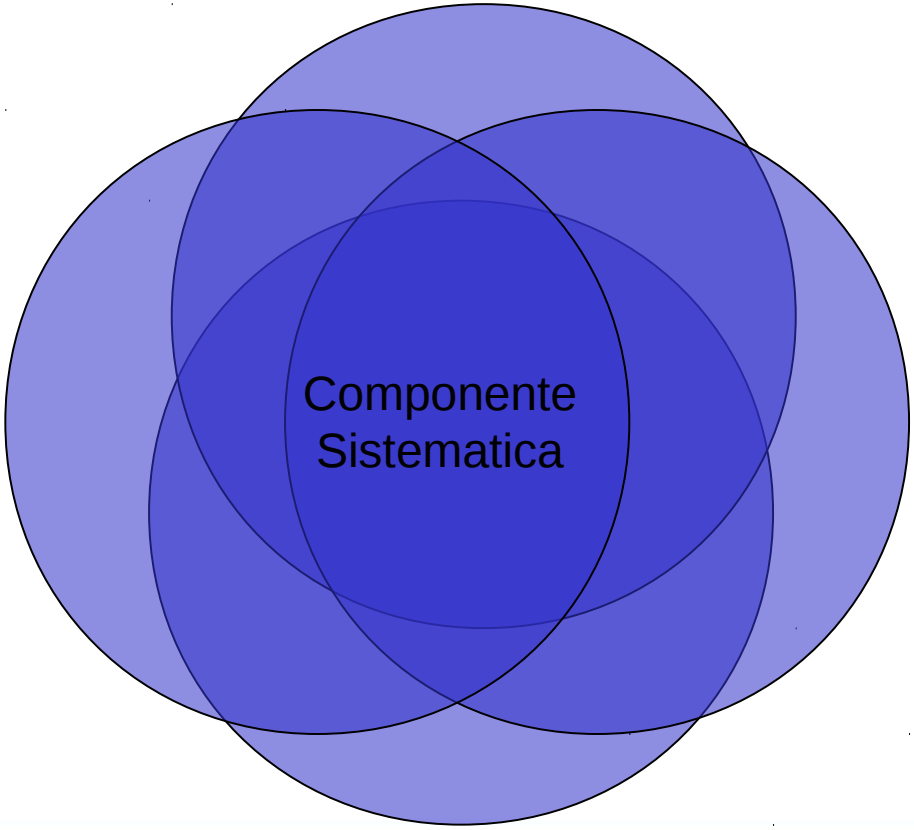
Caratteristiche

- Sebbene scale da due items siano rare, la correlazione di Pearson viene usata come indice di attendibilità in varie situazioni:

  - Test-retest di una scala
  - Split-half reliability (divido il test in due parti e le correlo)
  - Attendibilità di misure non parcellizzate in items (ad esempio tempi di reazione, etc)

Repetita iuvant

- Se abbiamo una serie di items (che misurano lo stesso costrutto), considereremo la varianza condivisa dagli items



Alpha di Cronbach

- Considero la media delle correlazioni fraK items come indice di congruenza (assumo le correlazioni siano tutte positive)

$$\bar{r} = \sum r_i$$

- La componente sistematica la metto a confronto con l'informazione totale

$$var_{tot} = 1 + K\,\bar{r}$$

$$\alpha = K\,\frac{\bar{r}}{(1 + K\,\bar{r})}$$

- Nella versione non standardizzata sostituisco la covarianze alle correlazioni

Caratteristiche

- E' l'indice di coerenza interna più usato
- Assume che le variabili siano tutte orientate nella stessa direzione
  - Se gli item non hanno tutti correlazioni positive, gli item con correlazioni negative con gli altri vanno moltiplicati per -1 prima di calcolare l'alpha di Cronbach
- Può essere calcolato anche su variabili dicotomiche

Esempio

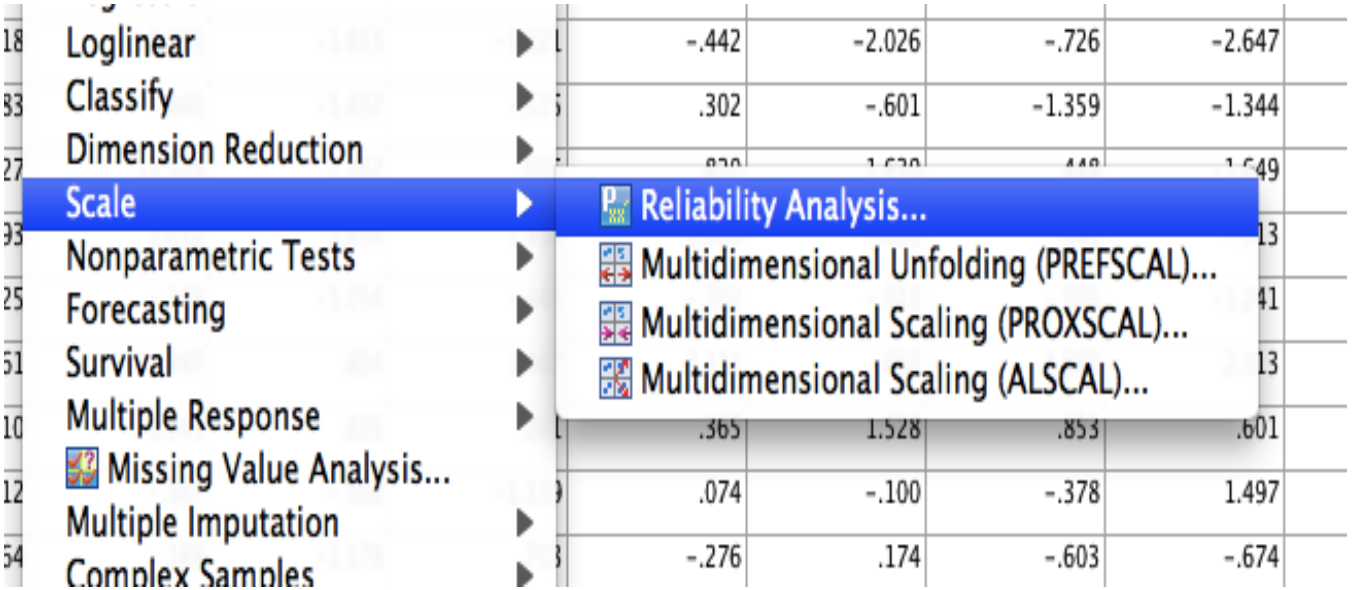
4 Item self-report

(in questo caso standardizzati, ma non serve in genereale)

	ss1	ss2	ss3	ss4
1	.013	.452	1.046	.490
0	-.185	-.286	.550	.273
6	-1.635	.318	-1.255	-1.813
9	-1.087	.183	-.940	-1.437
6	.880	-.327	-1.524	-.362
8	1.672	.493	1.012	1.856
1	.724	-1.225	-.148	-1.754
4	-.097	1.061	.447	.854
8	.387	.510	1.041	.835
8	-.198	-.912	-.367	-.562
5	-1.487	-.764	.188	-1.178
5	-1.204	-.767	.184	-.667
3	-.584	.316	.277	1.286

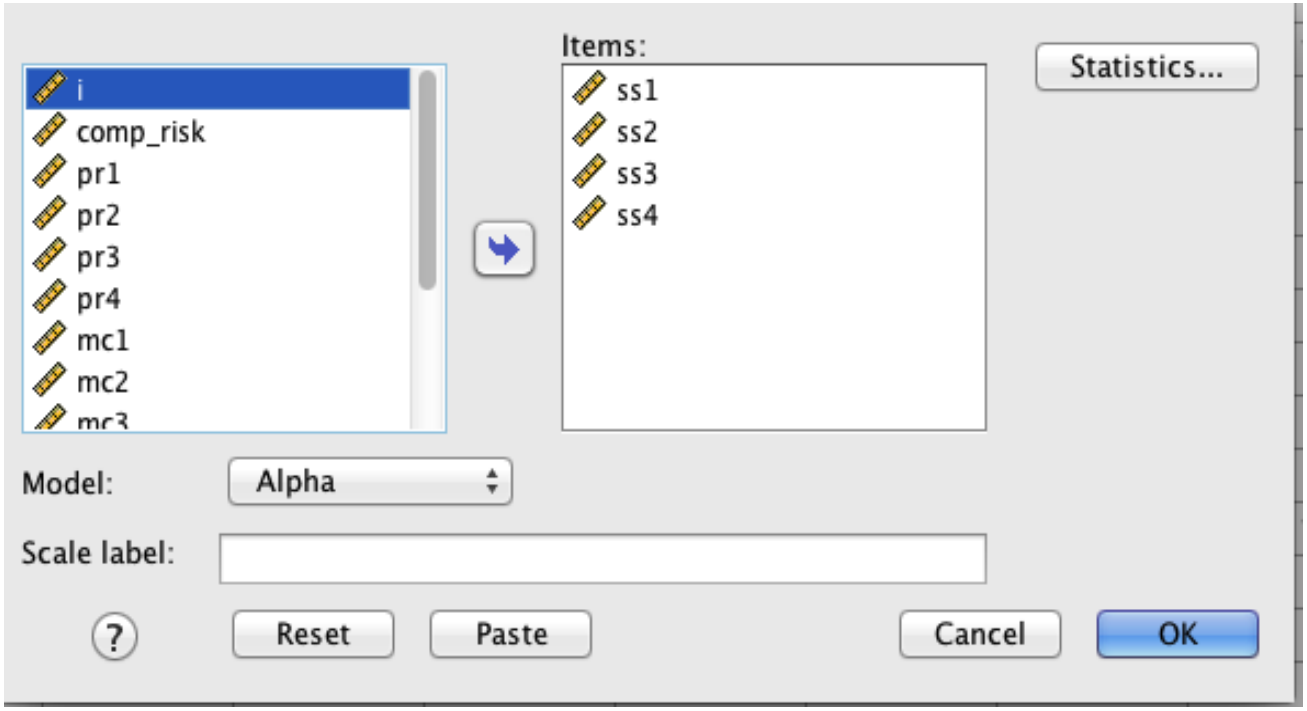
Esempio

Scale



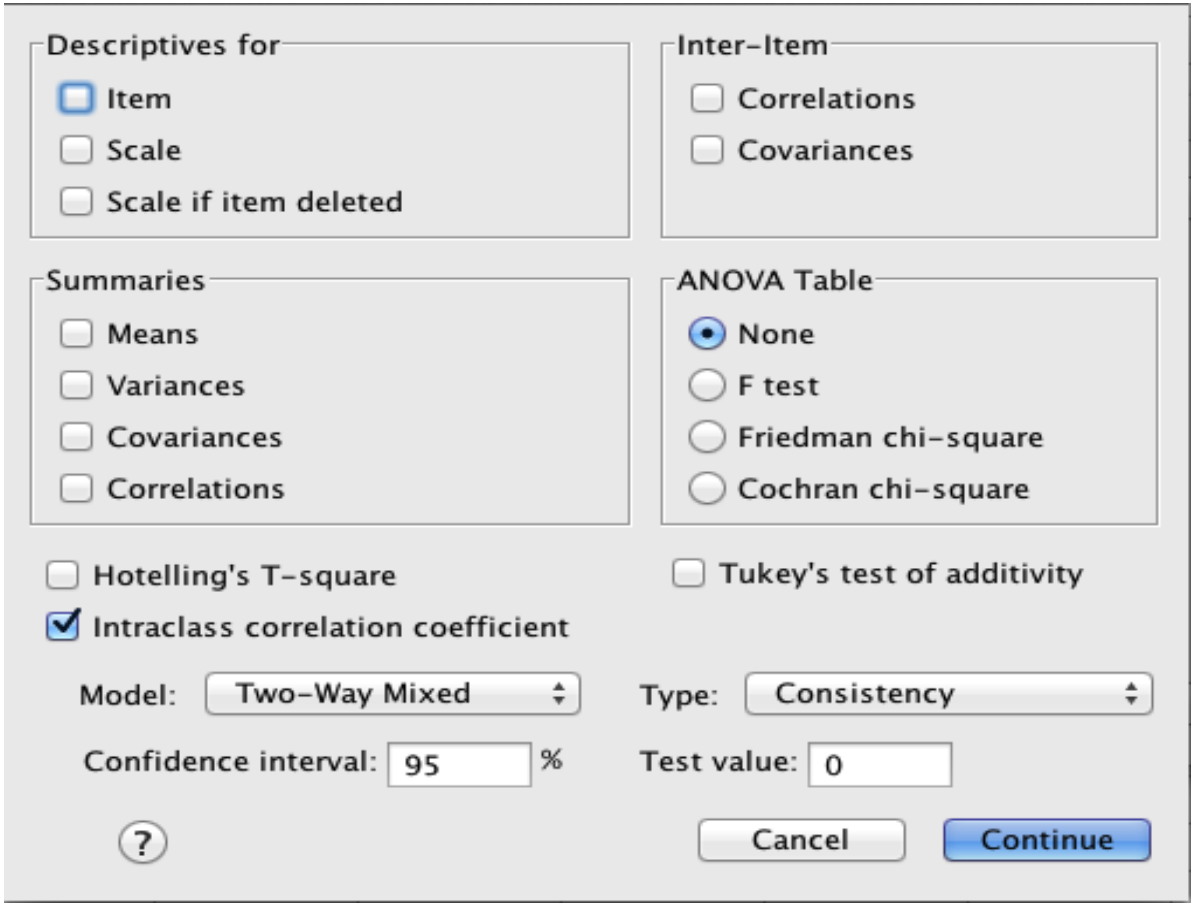
Esempio

Inserisco gli item



Esempio

Posso chiedere ICC



Esempio

Risultati

Reliability Statistics

Cronbach's Alpha	N of Items
.823	4

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.538[b]	.445	.630	5.664	109	327	.000
Average Measures	.823[c]	.763	.872	5.664	109	327	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.
b. The estimator is the same, whether the interaction effect is present or not.
c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Validità

- Per quanto riguarda la validità di una misura, esistono vari metodi per stabilirla. Questi metodi definiscono diversi tipologie di validità

- Contenuto (≈ interna)
- Criterio (≈ esterna)
- Costrutto (≈ nomologica)

Validità di contenuto

- Riguarda il disegno della misura
- Qualita’ del paradigma di misura (operazioni di misurazione)
- Punto cruciale: massimizare la variazione nei punteggi dovuta al costrutto rispetto a quella dovuta a fattori teoricamente irrelevanti (costrutti intervenienti)

Validità di contenuto

- **Qualita' teorica/psicometrica:** Qualita' della definizione teorica, sistematicita' nella generazione degli items, procedure di selezione degli items e di validazione del test
- **Qualita' degli items:** Un solo concetto in una frase, linguaggio chiaro, affermazioni specifiche, no domande "dirette" in qualche modo
- **Condizioni di somministrazione** Non troppi items, atmosfera collaborativa, concentrazione
- **Riduzione bias valutativi** ordine randomizzato ma fisso degli items, istruzioni adeguate, direzione items bilanciata, evitare effetti "sequenza"

Validità di criterio

- Si focalizza sulla capacità della misura di predire comportamenti o processi (criteri) associati al costrutto

- **Concorrente**: Misura e criterio sono misurati allo stesso tempo

- **Predittiva**: La misura e' ottenuta prima del criterio

- I criteri sono stabiliti teoricamente come conseguenze rilevanti che devono essere predette dalla misura (ad es., scala di depressione e benessere soggettivo)

Validità di criterio

- Predittori e criteri NON sono proprieta' intrinsiche delle variabili
- Si verifica tramite modelli di regressione
- Correlazione tra VI (predittore) e VD (criterio) o regressione lineare semplice
- Quando abbiamo piu' predittori simultaneamente usiamo la regressione multipla
- **Validita' incrementale**: La capacita' predittiva unica di una VI (al netto delle altre VI, coefficienti parziali)

Validità di costrutto

- Attiene alla dimostrazione che il costrutto misurato è effettivamente quello inteso e non un altro costrutti

- Si esamina mediante la correlazione con altre misure

- **Validità convergente**: la nostra misura dovrà mostrare correlazioni alte con misure alternative del costrutto

- **Validità discriminante**: la nostra misura dovrà mostrare correlazioni basse con misure di altri costrutti

Validità e Attendibilità

- Una misura deve essere contemporaneamente sia valida che attendibile

		Attendibilità	
		Insufficiente	Buona
Validità	Insufficiente	**Punteggi casuali**	**Misura qualcos'altro**
	Buona	**Impossibile**	**Misura buona**

Fine della Lezione