ANALISI DEI DATI + LABORATORIO

prof. Claudio Barbaranelli

Dip. di Psicologia, via dei Marsi 78, 00185 Roma
tel. 06/49917623
claudio.barbaranelli@uniroma1.it

Ricevimento studenti:
Dal 25/9 al 22/12: Martedì dalle 12.30 alle 14
Dall' 1/1 al 17/9/2019: Martedì dalle 10.30 alle 12

Eventuali variazioni rispetto a quanto specificato sopra verranno comunicate per tempo sul sito di facoltà (www.psicologia1.uniroma1.it) e su elearning2.

ANALISI DEI DATI + LABORATORIO

prof. Claudio Barbaranelli

Orario Lezioni

Aula	Orario	
Aula 8	Martedi	10:00:12:00
	Mercoledi	11:00-14:00
	Giovedi	11:00-14:00

inizio 1° semestre	01/10/2018
fine 1° semestre	21/12/2018

ANALISI DEI DATI + LABORATORIO

CONTENUTO DEL CORSO

Il corso riguarderà i seguenti argomenti:
- **I trattamenti preliminari dei dati**
- **La regressione lineare multipla**
- **L'Analisi della Varianza (ANOVA)**
- **L'analisi Fattoriale Esplorativa**
- **I modelli di equazioni strutturali**

Le applicazioni informatiche verranno effettuate con i programmi SPSS e MPLUS

ANALISI DEI DATI + LABORATORIO

TESTI DI RIFERIMENTO

a) Barbaranelli, C. (2007). Analisi dei dati. II edizione. Milano: Led. (capitoli 1, 2, 3, 4, appendici 1 e 2).

b) Barbaranelli, C. (2006). Analisi dei dati con SPSS: Le analisi multivariate. Milano: Led. (capitoli 1, 2 e 3).

c) Barbaranelli, C. e D'Olimpio, F. (2007). Analisi dei dati con SPSS: Le analisi di base. Milano: Led. (capitoli 1, 2, 3, 4 e 6).

d) Lucidi e materiale integrativo presentato a lezione. Questo materiale è disponibile sul sito www.elearning2.uniroma.it.

IL TESTO "INTRODUZIONE AI MODELLI DI EQUAZIONI STRUTTURALI" NON SARA' DISPONIBILE ED E' SOSTITUITO DALLO SCRITTO "NOTE SUI SEM" SCARICABILE DAL SITO www.elearning2.uniroma.it.

ANALISI DEI DATI + LABORATORIO

METODI DIDATTICI

Gli argomenti del corso verranno presentati attraverso lezioni prevalentemente frontali sollecitando un ruolo attivo da parte degli studenti.
Le ore di laboratorio si alterneranno con le lezioni teoriche e prevedranno esercitazioni su MPLUS e su SPSS.

Gli studenti possono scaricare la *DEMO version* di MPLUS dal sito: http://www.statmodel.com
Tale versione è gratuita ed ha una licenza perpetua.

Gli studenti possono scaricare SPSS (licenza autorizzata per gli studenti e il personale della Sapienza) dal sito della Sapienza.

ANALISI DEI DATI + LABORATORIO

MODALITÀ DI FREQUENZA: La frequenza alle lezioni e ai laboratori non è obbligatoria, ma raccomandata.
MODALITÀ D'ESAME: L'esame prevede una prova scritta costituita da:
- domande a risposta chiusa e aperta relative ai testi in programma;
- esercizi sull'interpretazione di output dei programmi SPSS e MPLUS;
- esercizi sulla programmazione in linguaggio MPLUS.

Esempi di esercizi su MPLUS sono scaricabili dal sito http://elearning2.uniroma.it
Per sostenere la prova è necessario prenotarsi entro i termini definiti sul sito della Facoltà.
Le modalità d'esame NON saranno differenziate per studenti frequentanti e non frequentanti.

TRATTAMENTI PRELIMINARI DEI DATI

Trattamenti preliminari dei dati

Sommario

* Forma della distribuzione
* Valori anomali (outliers) univariati
* Normalità bivariata e multivariata
* Outlier multivariati
* Le informazioni mancanti (*missing values*)

Forma della distribuzione

Distribuzione Normale Univariata

Forma "a campana", unimodale, simmetrica rispetto alla media (quindi media e mediana coincidono, e coincidono anche con la moda), presenta due punti di flesso per $x = \mu - \sigma$, e $x = \mu + \sigma$.
Famiglia di distribuzioni normali univariate: diverse distribuzioni normali sono definite da due parametri, la media (μ) e la deviazione standard (σ) della distribuzione.
Funzione di probabilità della distribuzione normale:

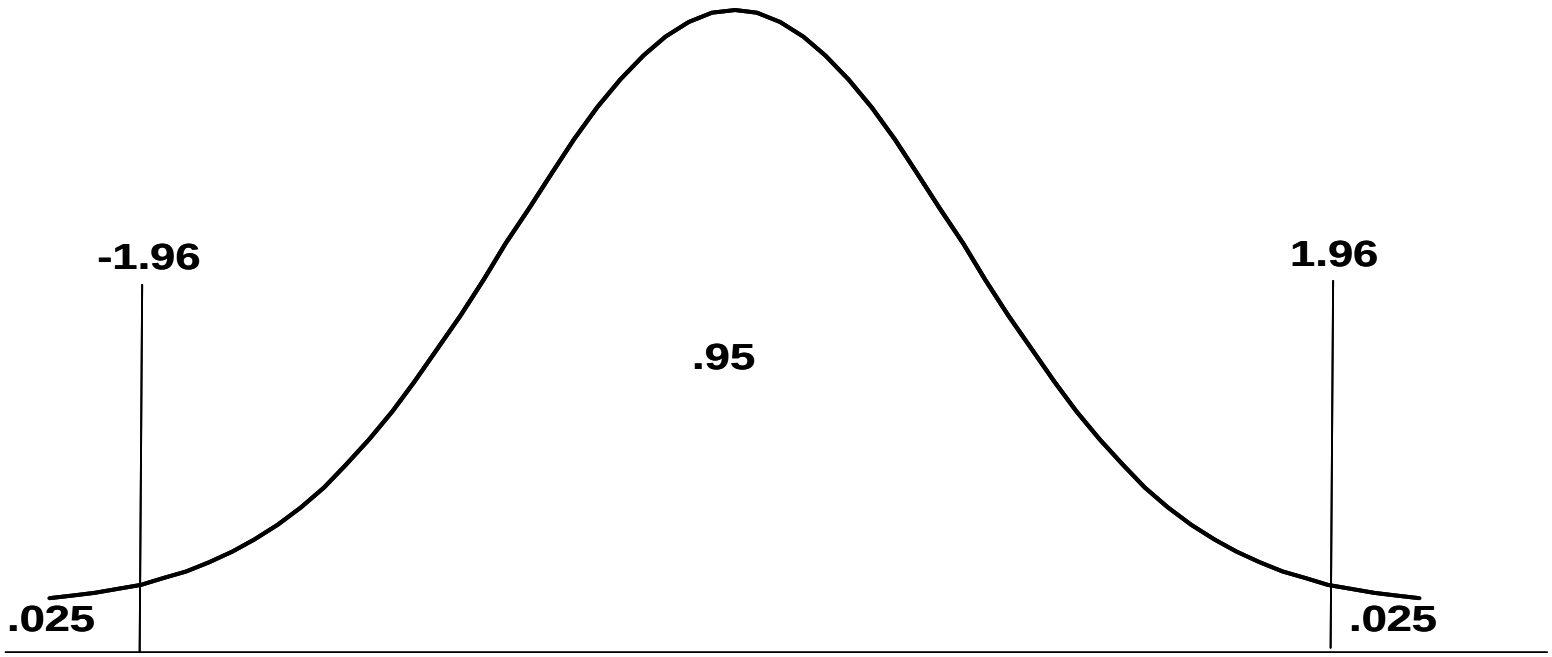
$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Forma della distribuzione

Normale standardizzata

La probabilità dei suoi valori è stata tabulata: ciò la rende particolarmente utile nella verifica delle ipotesi statistiche.

$$P([(X-V)/\sigma_E] \in [-1.96, 1.96]) = .95$$



Forma della distribuzione

Esame della normalità della distribuzione

Diversi metodi per esaminare se una variabile è normale. Le informazioni di questi diversi metodi vanno integrate.

- Indici di forma della distribuzione
- Test statistici
- Metodi grafici

Forma della distribuzione

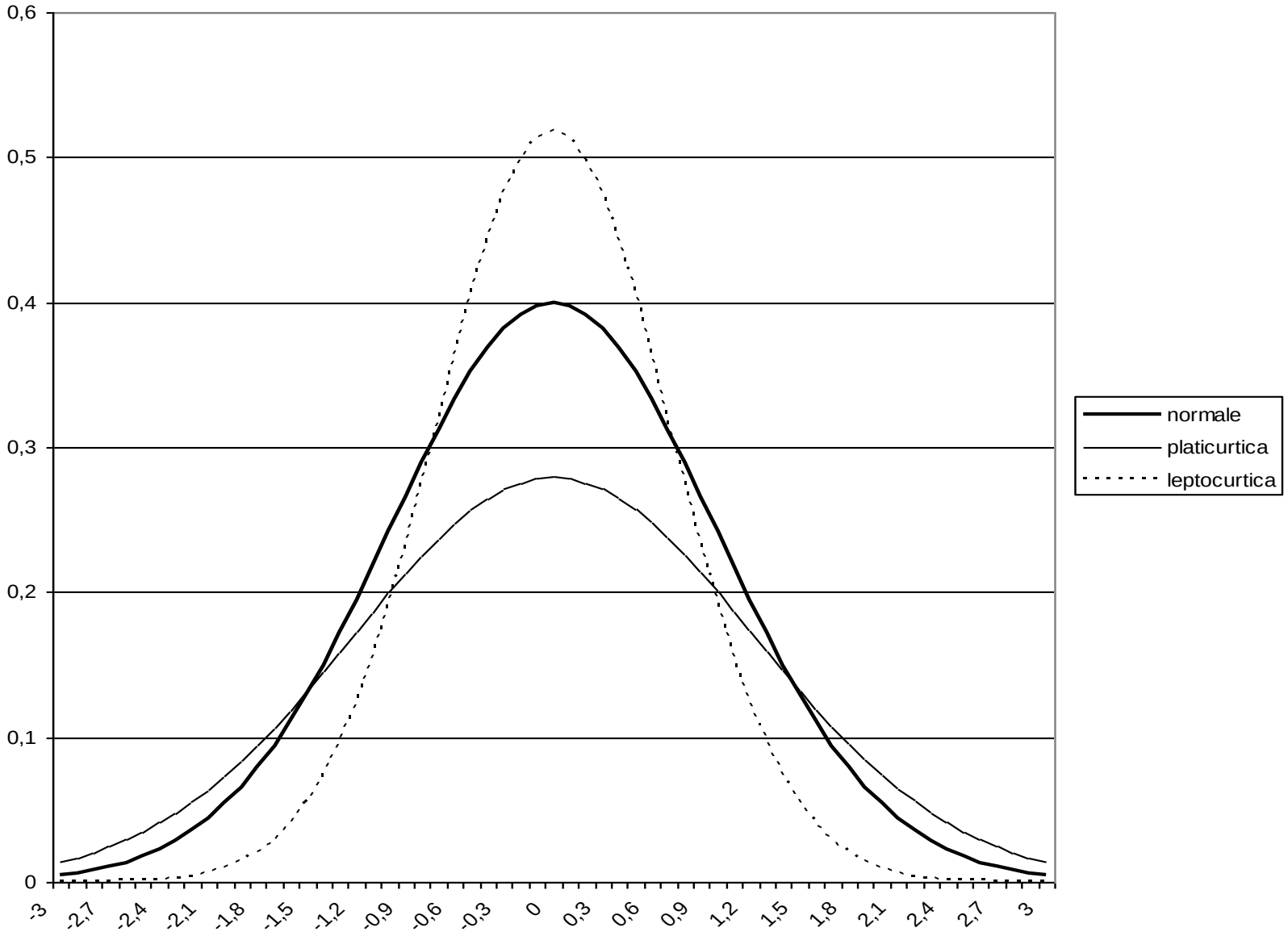
Indici che valutano la forma della distribuzione: Curtosi e Asimmetria (o *skewness*)

Curtosi: riflette il grado in cui i punteggi sono distribuiti nelle code piuttosto che nelle zone centrali della distribuzione. Uguale a 0 quando la distribuzione è perfettamente normale.

Curtosi Negativa: distribuzione platicurtica, "più schiacciata", i valori estremi sono più frequenti rispetto alla normale.

Curtosi Positiva: distribuzione leptocurtica, "più appuntita", i valori estremi sono meno frequenti.

Forma della distribuzione



Forma della distribuzione

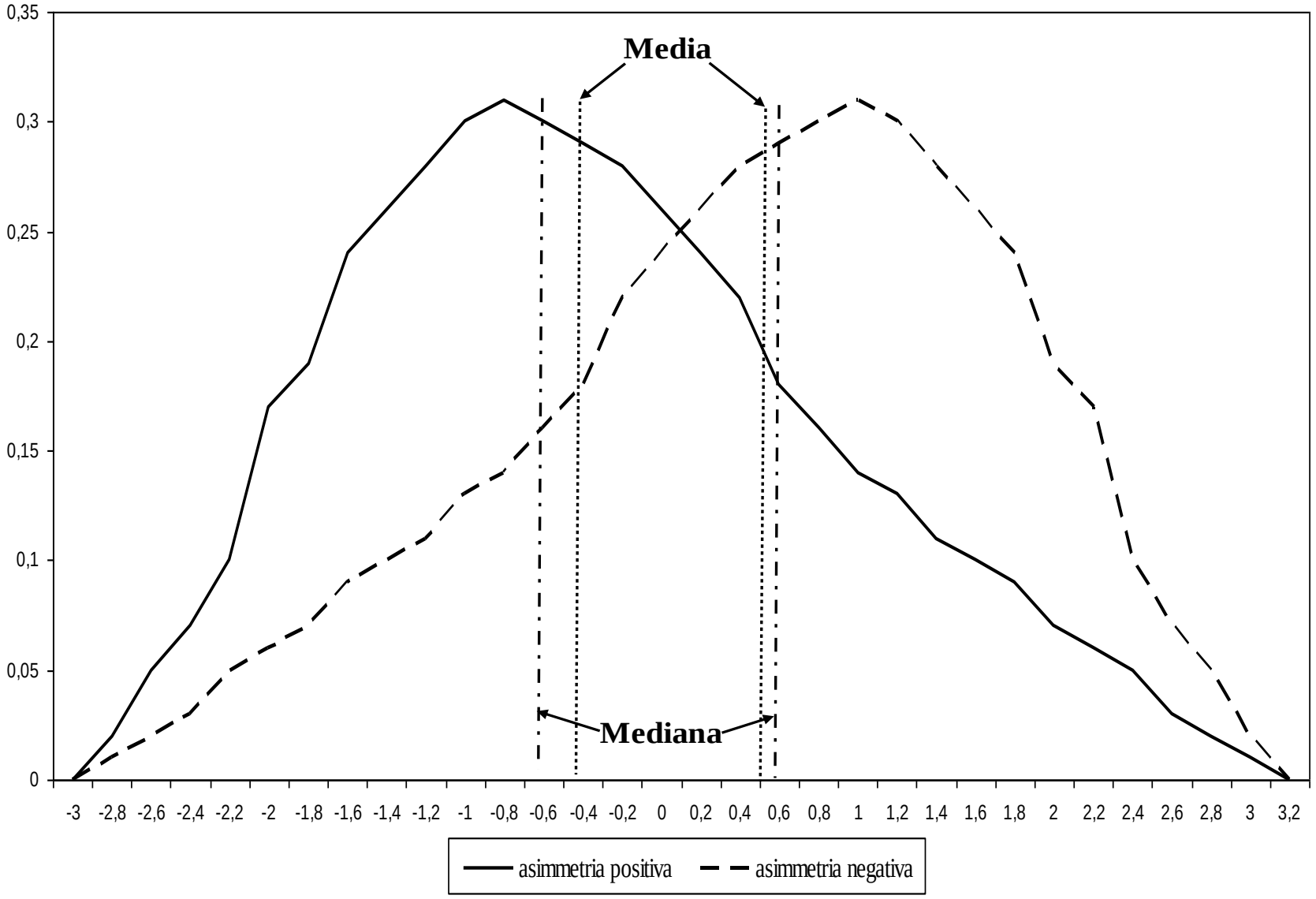
Indici che valutano la forma della distribuzione: Curtosi e Asimmetria (o *skewness*)

Asimmetria: riflette il grado in cui la distribuzione è disposta simmetricamente attorno ai valori di tendenza centrale. Uguale a 0 quando la distribuzione è perfettamente normale.

Asimmetria positiva: i valori bassi hanno frequenza maggiore, la media risulta maggiore della mediana.

Asimmetria negativa: i valori alti sono più frequenti, la media risulta inferiore alla mediana.

Forma della distribuzione



Forma della distribuzione

Formula per la curtosi

$$\frac{\sum_{i=1}^{N}(x_i-\bar{x})^4}{N} \Big/ \left(\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}\right)^2$$

Errore standard della curtosi = $(24/N)^{1/2}$

Di solito viene sottratto il valore 3 per rendere la curtosi uguale a 0 nel caso di perfetta distribuzione normale.

Forma della distribuzione

Formule per l'asimmetria

$$\left(\frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{N} \right)^2 \Big/ \left(\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} \right)^3$$

Errore standard della asimmetria = $(6/N)^{1/2}$

$$3\frac{(\bar{x} - \text{Mediana})}{s_x}$$

Forma della distribuzione

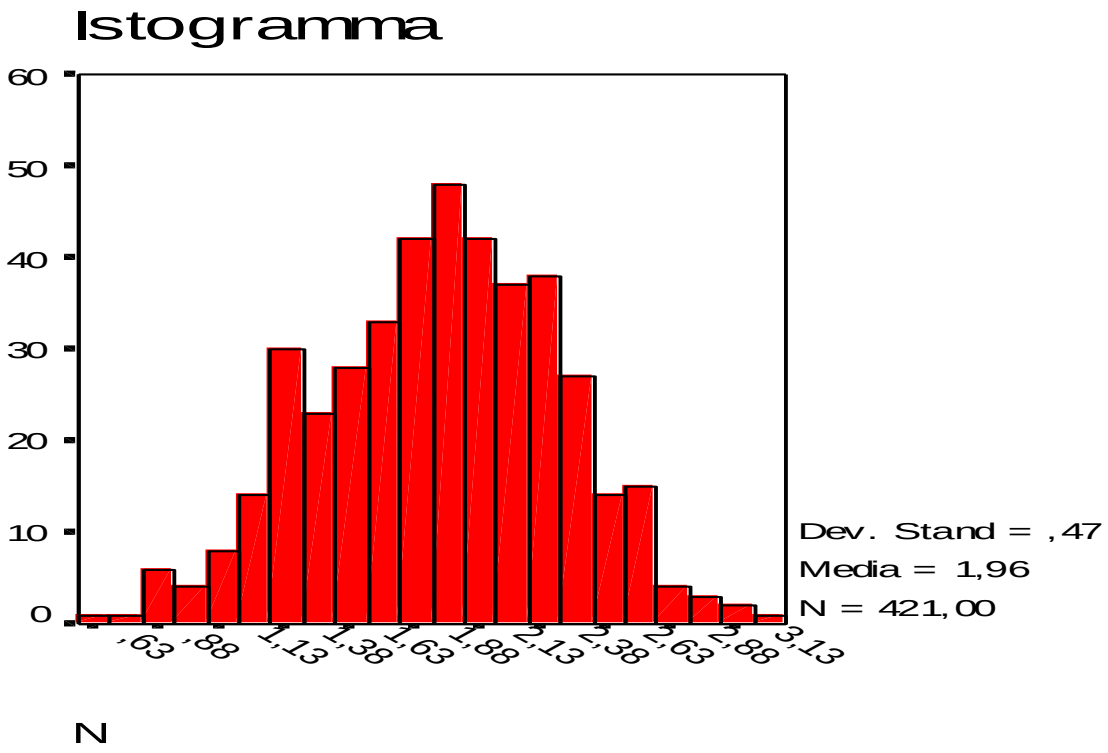
Verifica delle ipotesi per asimmetria e curtosi: dividere il singolo indice (di asimmetria o di curtosi) per il suo errore standard, ed utilizzare come distribuzione di riferimento la normale standardizzata. Test troppo potente, ovvero risulta significativo quasi sempre.
Criterio empirico: accettabili valori compresi tra –1 e 1
Nella verifica delle ipotesi su asimmetria e curtosi utilizzare un livello di alpha più basso (.01 o .001).

Test statistici di Normalità: Kolmogorov-Smirnov e Shapiro-Wilk. Se risultano significativi si deve rifiutare l'ipotesi nulla che la distribuzione sia normale. Test molto potenti che conducono troppo spesso al rifiuto dell'ipotesi nulla.

Forma della distribuzione

Grafici per l'esame della normalità

Istogramma della distribuzione di frequenze della variabile



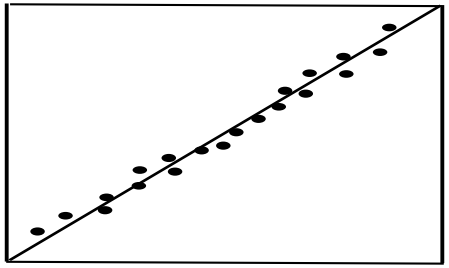
Forma della distribuzione

Grafici per l'esame della normalità

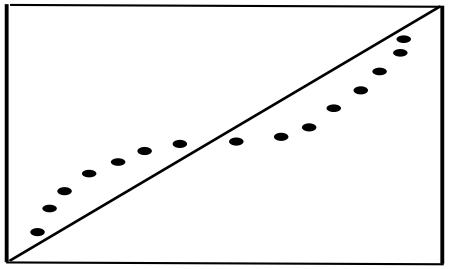
Plot dei quantili, o *Q-Q Plot* o *Cumulative Normal Plot*

Si confrontano i quantili della distribuzione della variabile, rispetto ai quantili della distribuzione normale. In ascissa sono riportati i valori osservati, in ordinata i valori attesi se la distribuzione è normale. Se la variabile si distribuisce in forma normale, i punti di tale distribuzione congiunta sono addensati sulla diagonale.

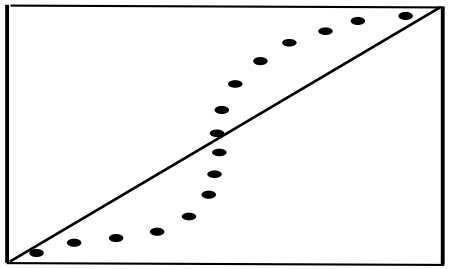
Forma della distribuzione



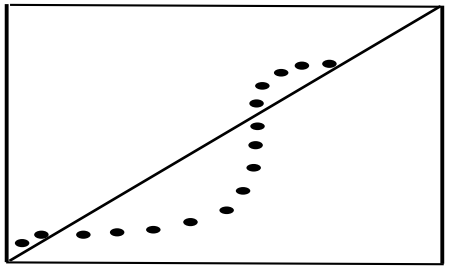
1. Normale



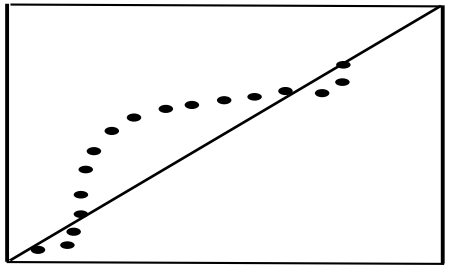
2. Platicurtica



3. Leptocurtica



4. Asimmetria negativa



5. Asimmetria Positiva

Valori anomali (outliers) univariati

Valori che si distinguono in maniera particolare rispetto agli altri valori nella distribuzione.

Outliers univariati: casi che in una variabile presentano valori estremamente elevati o estremamente bassi.

Gli outliers possono influenzare: la media, la deviazione standard, l'asimmetria e la curtosi, il coefficiente di correlazione di Pearson.

Indici che risultano meno influenzati dagli outliers: mediana e moda;

Statistiche "robuste" (es., media "*trimmed*" calcolata eliminando il 5% dei casi con punteggi più elevati e più bassi).

Valori anomali (outliers) univariati

Individuare gli outliers univariati

Standardizzare i punteggi relativi alla variabile in esame e calcolare una distribuzione delle frequenze.

Sono possibili outliers i casi che presentano un punteggio z maggiore di |3|.

Esame della distribuzione per vedere se i punteggi troppo elevati sono casi isolati dal resto dei punteggi.

In alternativa è possibile utilizzare il valore assoluto mediano (MAD) secondo una formula più complessa:

$$|X-Mdn|/(1.483*MAD)$$

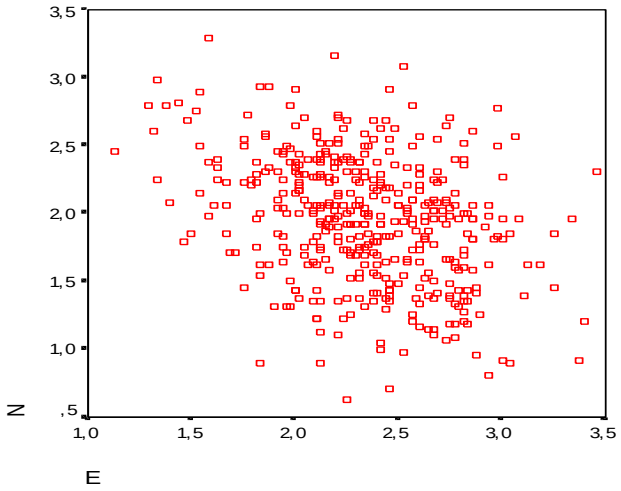
Miller (1991) propone di considerare come outlier i soggetti che presentano punteggi maggiori di |3| o di |2.5| (scelta moderatamente conservativa)

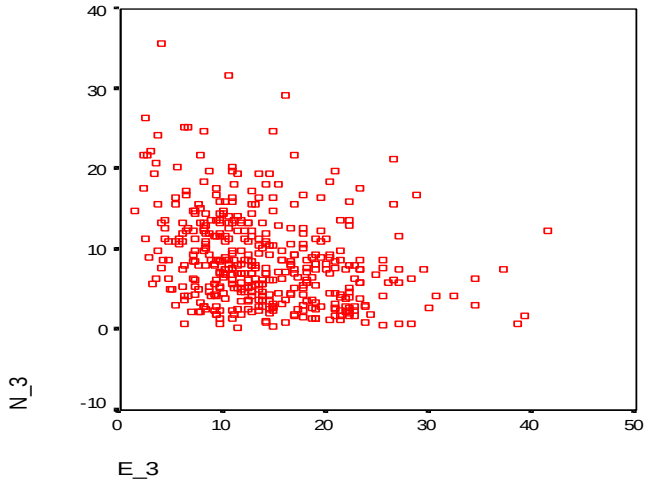
Linearità

Linearità della relazione tra due variabili

Relazione lineare tra X e Y: la variazione nei punteggi in Y attesa in concomitanza di una variazione di punteggi in X è costante per tutti i valori di X.

Diagramma di dispersione (o scatterplot)





Trasformazioni delle variabili 1

Non-linearità e non-normalità: fenomeni collegati. Tecniche per rendere "normale" la distribuzione (Tabachnick e Fidell, 1994, 2013)

Problema	Trasformazione [X*=f(X)]
Asimmetria Positiva Estrema (valori >2)	Reciproco: X* =1/X
Asimmetria Positiva Sostanziale (valori tra 1 e 2)	Logaritmo: X* =Log10(X)
Asimmetria Positiva Moderata (valori tra .5 e 1)	Radice Quadrata $X^* = \sqrt{X}$
Asimmetria Negativa Moderata (valori tra -.5 e -1)	Radice Quadrata $X^* = \sqrt{(K - X)}$
Asimmetria Negativa Sostanziale (valori tra -1 e -2)	Logaritmo = X* =Log10(K-X)
Asimmetria Negativa Estrema (valori <-2)	Reciproco = X* =1/(K-X)

Nb. K è uguale al valore più elevato della variabile X, +1

Trasformazioni delle variabili 2

Trasformazioni di Box-Cox (Box e Cox, 1964)

Si tratta di una serie di trasformazioni che servono a normalizzare una distribuzione ma sono più complicate da calcolare

$$X^{(\lambda)} = \begin{cases} \dfrac{X^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log X, & \text{if } \lambda = 0. \end{cases}$$

La constante λ serve a normalizzare i punteggi: il suo valore ottimale (che massimizza la correlazione tra i punteggi originali e quelli trasformati) può essere individuato con appositi algoritmi (possibile anche in SPSS, vedi Osborne, 2010, *http://pareonline.net/getvn.asp?v=15&n=12*)

Trasformazioni delle variabili 3

POMS (Little, 2013)

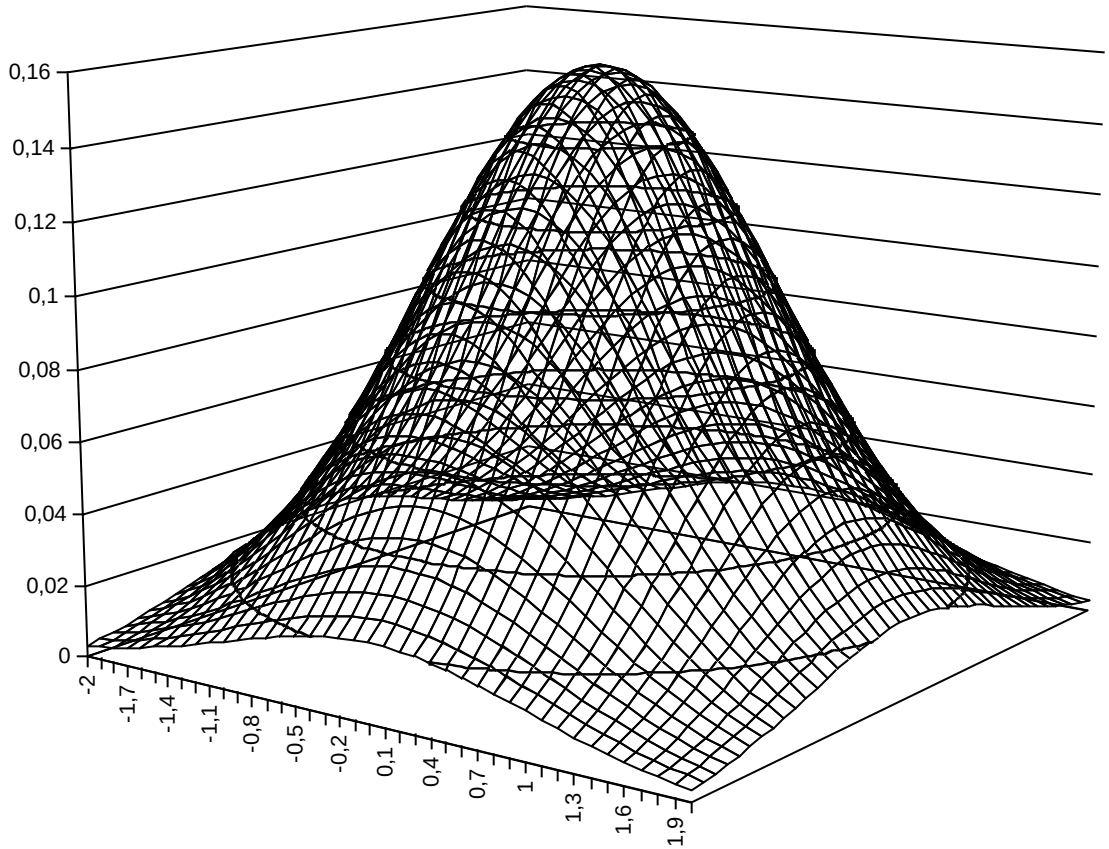
Percentage of Proportion of maximum scoring (POMS) è un tipo di trasformazione che NON serve per normalizzare i punteggi, ma per renderne omogenea l'unità di misura, senza ricorrere alla standardizzazione (che annulla le differenze nella variabilità). Dopo la trasformazione le variabili hanno tutte la stessa unità di misura.

$$\text{Punteggio trasformato} = \frac{\text{Punteggio} - \text{minimo}}{\text{Massimo - minimo}}$$

Il risultato è quello di trasformare le variabili su una scala che va da 0 (minimo) a 1 (massimo)

Normalità bivariata

Distribuzione normale bivariata: ciascuna delle 2 variabili è distribuita normalmente rispetto all'altra. La loro distribuzione congiunta ha la seguente forma:



Normalità bivariata

Distribuzione normale bivariata: ciascuna delle 2 variabili è distribuita normalmente rispetto all'altra.

Funzione di probabilità della d.n.b.:
$f(x,y; \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy}) =$

$$\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\, e^{-([(x-\mu_x)^2/2\sigma_x^2+(y-\mu_y)^2/2\sigma_y^2]-2\rho[(x-\mu_x)^2/2\sigma_x^2+(y-\mu_y)^2}$$

dove μ_x e μ_y sono le medie di x e y, σ_x^2 e σ_y^2 sono le varianze di x e y, e ρ_{xy} è la correlazione tra x e y.

Normalità multivariata

Distribuzione normale multivariata: generalizzazione della normale bivariata per k>2 variabili.
Normalità multivariata: assunzione che riguarda l'insieme delle variabili che vengono considerate in analisi.
Funzione di probabilità della normale multivariata:

$$f(y;\ \mu,\ \Sigma) = \frac{1}{(2\pi)^{p/2}\ |\ \Sigma\ |^{1/2}}\ EXP\left(\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right)$$

dove p è il numero di variabili nel vettore y, Σ è la matrice di varianze e covarianze tra le p variabili, $|\Sigma|$ è il suo determinante, μ è il centroide delle p variabili, e EXP è l'operatore della funzione esponenziale e^x.
La funzione ha in tutto p(p+3)/2 parametri.

Normalità multivariata

La distribuzione multivariata di *p* variabili è normale se:

- tutte le distribuzioni univariate delle variabili sono normali;

- le distribuzioni congiunte di tutte le coppie di variabili seguono la distribuzione normale bivariata;

- tutte le combinazioni lineari delle variabili sono normali.

Di solito se la distribuzione univariata di ogni singola variabile è normale, anche la distribuzione multivariata delle variabili lo è. Se c'è normalità multivariata, le relazioni tra le variabili considerate sono sicuramente lineari.

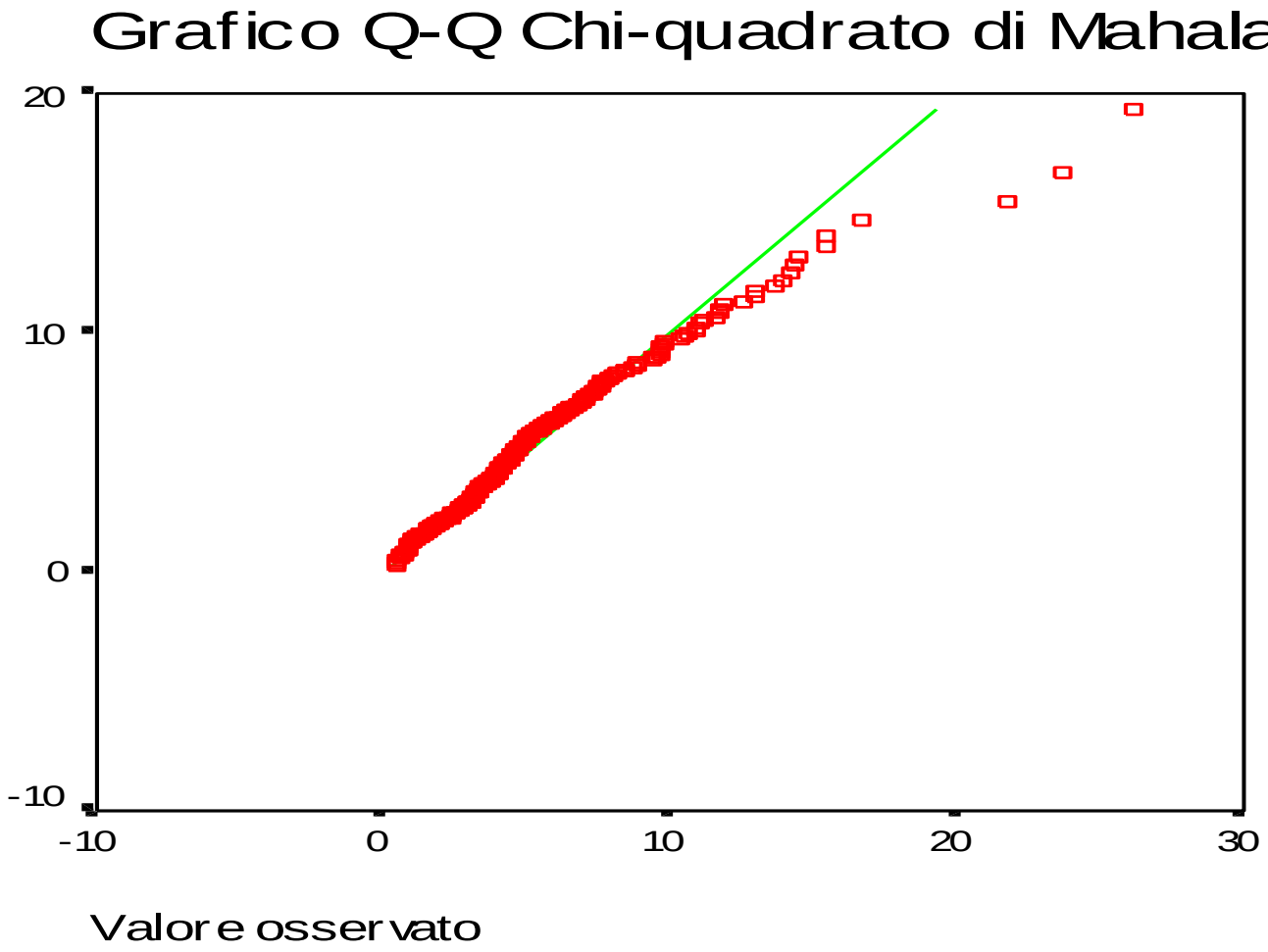
Normalità multivariata

Valutare la normalità multivariata: Test grafico basato sui quantili della distribuzione del chi quadrato.
Distanza generalizzata o distanza di Mahalanobis per ogni singolo caso:

$$D_i^2 = (\mathbf{X_i} - \overline{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X_i} - \overline{\mathbf{X}})$$

Rappresenta la distanza del vettore di punteggi di un soggetto (X_i) dal centroide del campione $\overline{X}$, pesata per le var/covarianze (S).
Se la distribuzione delle variabili è normale multivariata e il numero dei casi meno il numero di variabili è maggiore di 25, la distanza generalizzata segue la distribuzione del chi-quadrato.

Normalità multivariata



In ascissa sono riportati i valori osservati (D^2), in ordinata i valori attesi della distribuzione del chi-quadrato. Se la distribuzione è normale multivariata il grafico ha un andamento lineare.

Normalità multivariata

Coefficiente di curtosi multivariata di Mardia

$$k = \sum_{i=1}^{N} \left(D_i^2\right)^2 \Big/ N$$

Se la distribuzione delle p variabili è normale multivariata, e se n>50 soggetti) il coefficiente di curtosi multivariata di Mardia è ≤ p(p+2).

$$Z_k = \frac{k - E(k)}{\sqrt{VAR(k)}}$$

Z_k si distribuisce approssimativamente come una variabile normale standardizzata se il campione è sufficientemente ampio. Esame dell'ipotesi nulla che k < p(p+2), con un test a due code per un livello di probabilità pari a $\alpha/2$.

Outlier multivariati

Combinazioni dei punteggi delle singole variabili che risultano particolarmente "strani".

Casi che hanno una combinazione di punteggi particolarmente rara rispetto al resto del campione.

Si possono considerare outliers multivariati i casi in cui la distanza di Mahalanobis D^2 risulta significativa al livello $p<.001$ (Tabachnick e Fidell, 2007), prendendo come distribuzione di riferimento quella del chi-quadrato con *p* gradi di libertà (dove *p* =numero di variabili).

Le informazioni mancanti (*missing values*)

In fase di codifica dei dati è bene che i valori mancanti siano opportunamente codificati, in modo da distinguerli dai valori *effettivi* che possono assumere le variabili.

In fase di analisi è necessario che il ricercatore decida cosa fare dei valori mancanti.

Ci sono diverse strategie possibili.

Le informazioni mancanti (*missing values*)

a) la limitazione dell'analisi ai soli casi che presentano valori validi per tutte le variabili in esame (esclusione *listwise*);

b) la limitazione dell'analisi ai casi che di volta in volta presentano valori validi nella coppia di variabili che viene considerata (esclusione *pairwise*);

c) la sostituzione del valore mancante con la media della variabile nel campione, o con la media ottenuta dal soggetto nelle variabili considerate;

d) la sostituzione del valore mancante con una sua stima ricavata tramite procedure statistiche (regressione, EM) effettuate sui soggetti che presentano dati completi.

Statistical Package for Social Sciences

https://www.spss.it/

https://web.uniroma1.it/infosapienza/

SPSS

INTRODUZIONE A SPSS

- Le componenti fondamentali di SPSS
- Lo screening dei dati (es. valutare la normalità della distribuzione; come trattare i dati mancanti)
- L'analisi dei dati (statistiche descrittive, attendibilità, analisi degli item, analisi della varianza, correlazione e regressione, analisi fattoriale)

SPSS

LE COMPONENTI FONDAMENTALI DI SPSS

- 1. LE FINESTRE
- 2. I MENÙ
- 3. LE FINESTRE DI DIALOGO
- 4. LE BARRE DEGLI STRUMENTI
- 5. LA BARRA DI STATO

SPSS

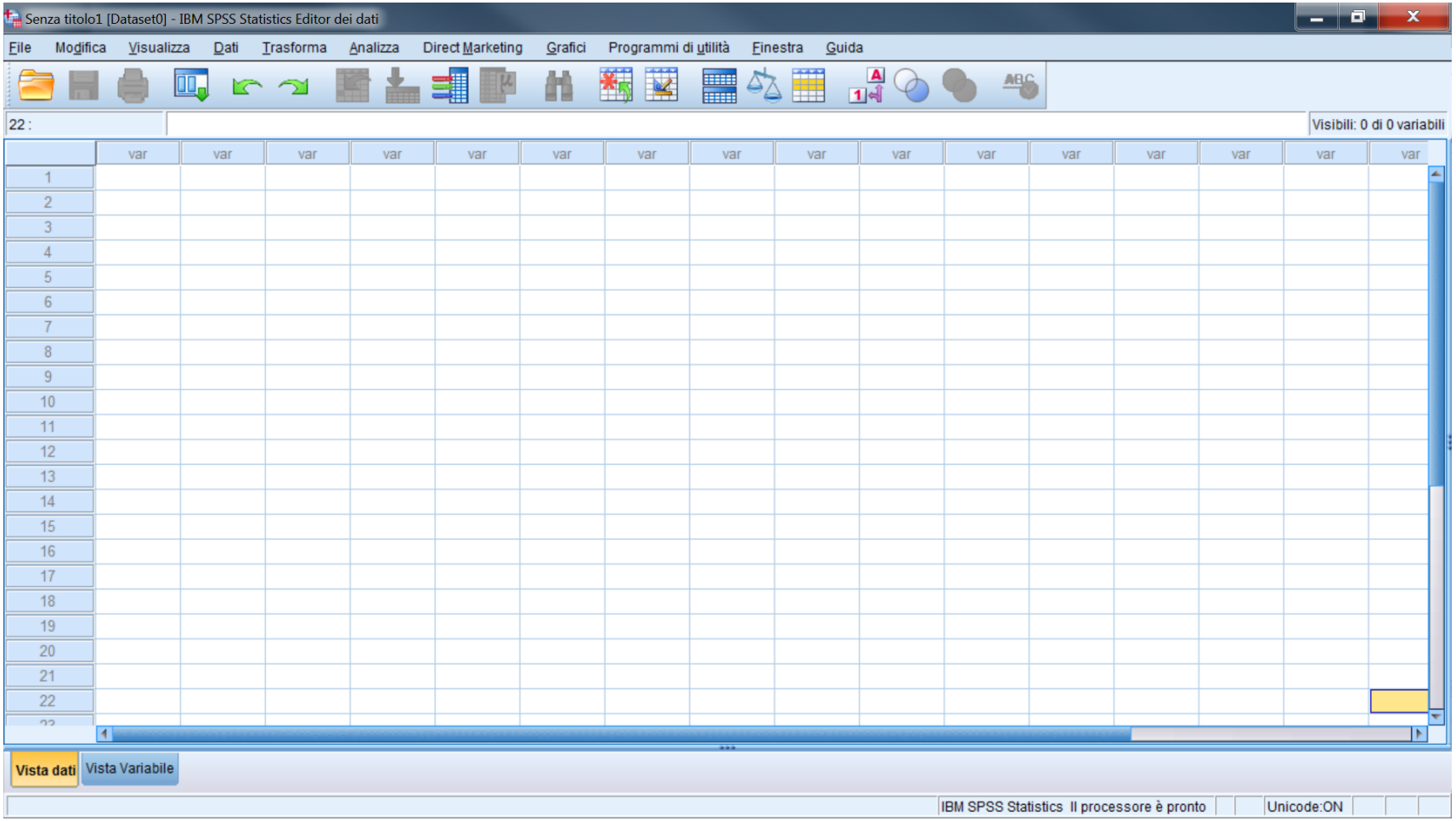
LE FINESTRE DI SPSS

- 1. LA FINESTRA EDITOR DEI DATI
- 2. LA FINESTRA VISUALIZZATORE
- 3. LA FINESTRA SINTASSI

SPSS

La finestra Editor dei dati

QUESTA FINESTRA MOSTRA I CONTENUTI DEL FILE DEI DATI



La finestra DATA EDITOR si apre automaticamente ogni volta che ha inizio una sessione SPSS. Si possono aprire più data files alla volta.

SPSS

La finestra Editor dei Dati

Molte caratteristiche della finestra data editor sono simili a quelle dei fogli elettronici (es. excel). Vi sono comunque alcune importanti differenze.

- Le righe corrispondono ai casi (unità). Ciascuna riga rappresenta un caso o un'osservazione. Ad esempio ciascun individuo che compila un questionario è un caso.

- Le colonne sono le variabili. Ciascuna colonna rappresenta una variabile o una caratteristica rilevata. Ad esempio ciascun item di un questionario è una variabile.

- Le celle contengono i valori. Ogni cella contiene un singolo valore di una variabile relativa ad un caso. La cella è l'intersezione di un caso con una variabile. Diversamente da Excel, le celle contengono solo valori, e non possono contenere formule.

SPSS
La finestra Editor dei Dati

- Il data file è rettangolare

Le dimensioni del data file sono determinate dal numero di casi e di variabili

Si possono inserire dati in ogni cella. Se si inseriscono dati in una cella che è al di fuori dei confini che delimitano il data file, il rettangolo dei dati si estende fino ad includere ogni riga e/o colonna tra questa cella e i confini del file

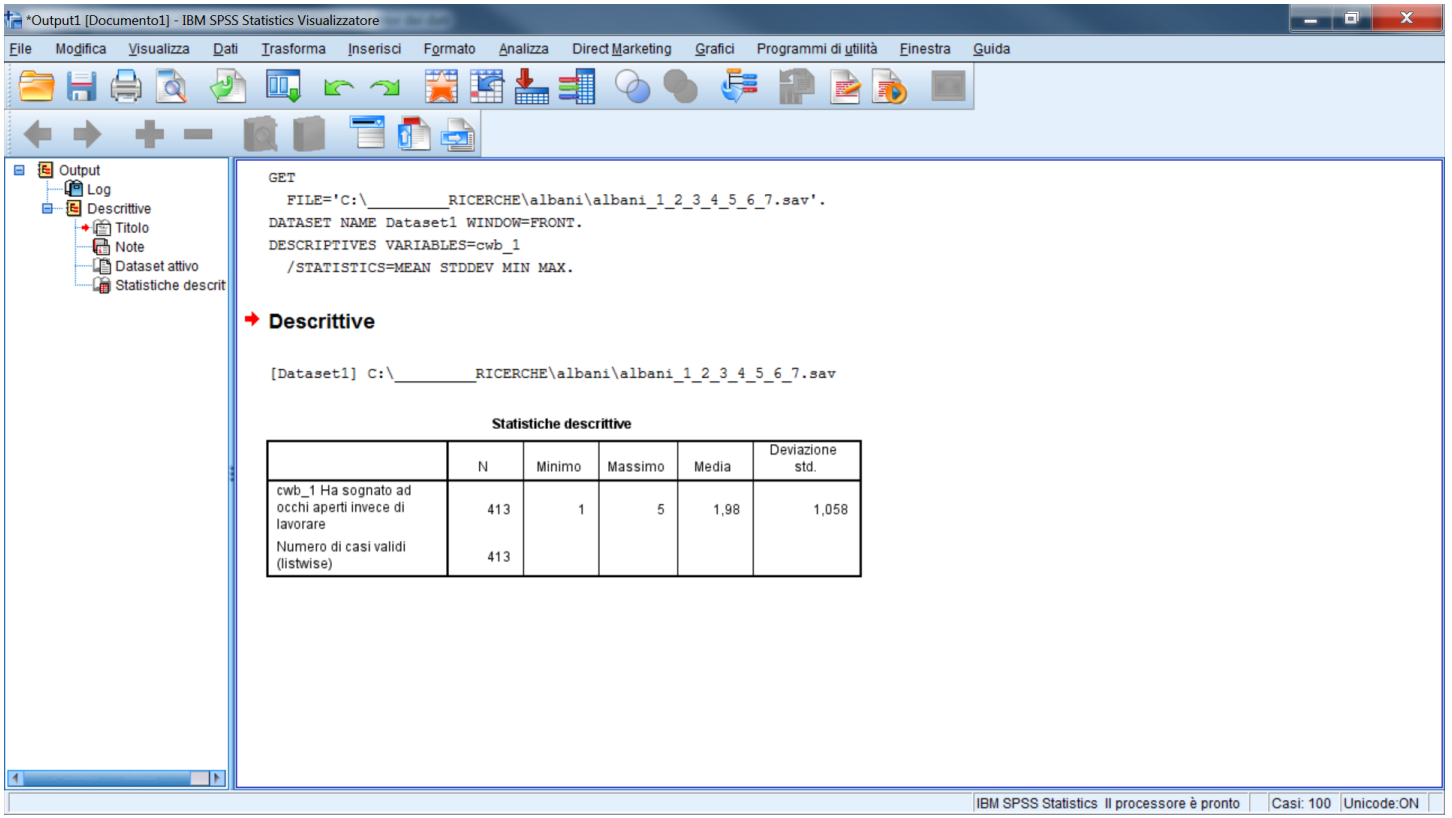
Le celle vuote all'interno dei confini del data file vengono considerate come un valore mancante, ovvero vengono convertite in valori mancanti "di sistema"

SPSS

La finestra Visualizzatore

Questa finestra si apre automaticamente la prima volta che viene eseguita una procedura che genera un output

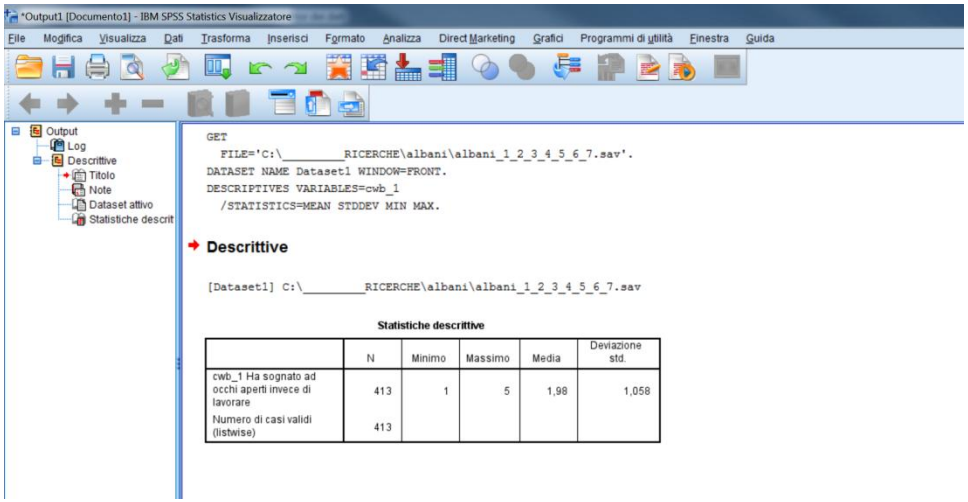
Nella finestra Visualizzatore vengono mostrati tutti i risultati statistici, le tabelle e i grafici (output)



SPSS

La finestra Visualizzatore

La finestra visualizzatore è suddivisa in due parti:



- Il quadro di sinistra fornisce una visione d'insieme dei contenuti dell'output.

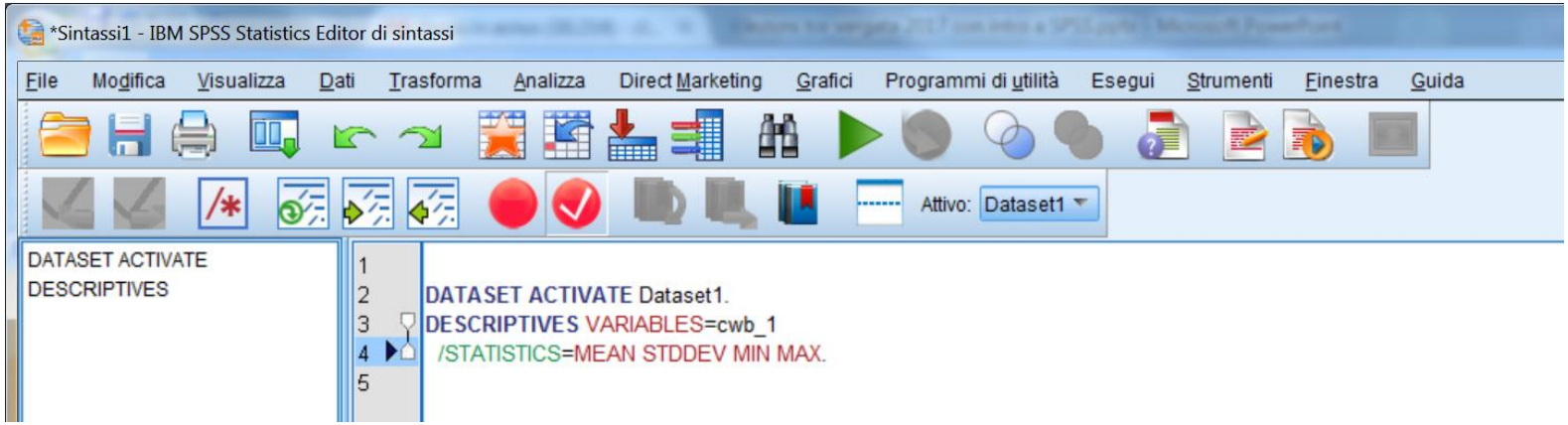
-Il quadro di destra contiene gli elementi veri e propri dell'output (tabelle, grafici e testo).

La maggior parte delle caratteristiche dell'output (es. il colore o l'ampiezza del testo) sono modificabili

SPSS

La finestra Sintassi

I comandi SPSS possono essere eseguiti utilizzando il linguaggio di programmazione di SPSS in un file di sintassi



Un file "sintassi" è un file di testo che contiene dei comandi.

I comandi scritti nel linguaggio di programmazione di SPSS possono essere salvati in modo tale da rendere possibile la ripetizione delle analisi in un momento successivo.

Una interessante risorsa per file di sintassi è: http://www.spsstools.net/en/

SPSS

I menu di SPSS

Ciascuna finestra in SPSS ha la propria barra dei menù, che consente la selezione dei menu appropriati per quel tipo di finestra.

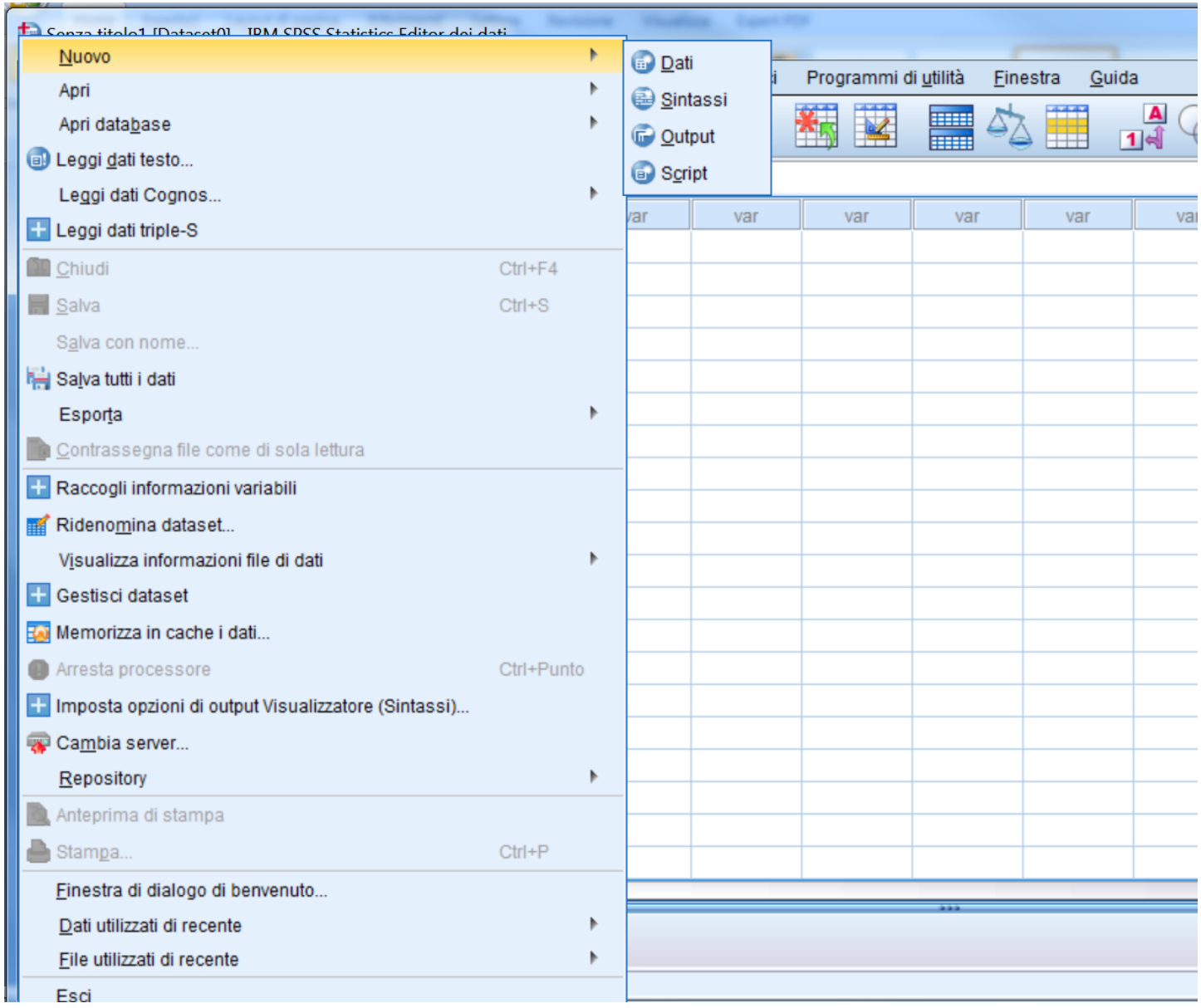
I menu ANALIZZA e GRAFICI sono disponibili su tutte le finestre, rendendo più semplice la creazione di nuovi output senza dover passare ad altre finestre.

SPSS

Il menu file

Si tratta di un menù particolarmente importante poiché mette in comunicazione il programma SPSS con l'esterno.
Il menù File può essere utilizzato per creare un nuovo file scegliendo File/Nuovo. A seconda del tipo di file desiderato è possibile scegliere tra file di dati (Dati), file testo per i comandi nel linguaggio di programmazione (Sintassi), file che contengono risultati sia in formato SPSS (Output), file che consentono di automatizzare alcune operazioni tramite appositi programmi (Script).

SPSS
Il menu file



SPSS
Il menu file

Tra i diversi tipi di file di dati che è possibile aprire segnaliamo i seguenti, rimandando il lettore al manuale di SPSS e alle funzioni di aiuto in linea per le ulteriori opzioni relative al menù:

- File SPSS (*.sav), SPSS/PC+ (*.sys) e Portabile SPSS (*.por);

- File testo "Tab delimitati", ovvero con i valori separati da tabulazioni, o fissi (*.dat);

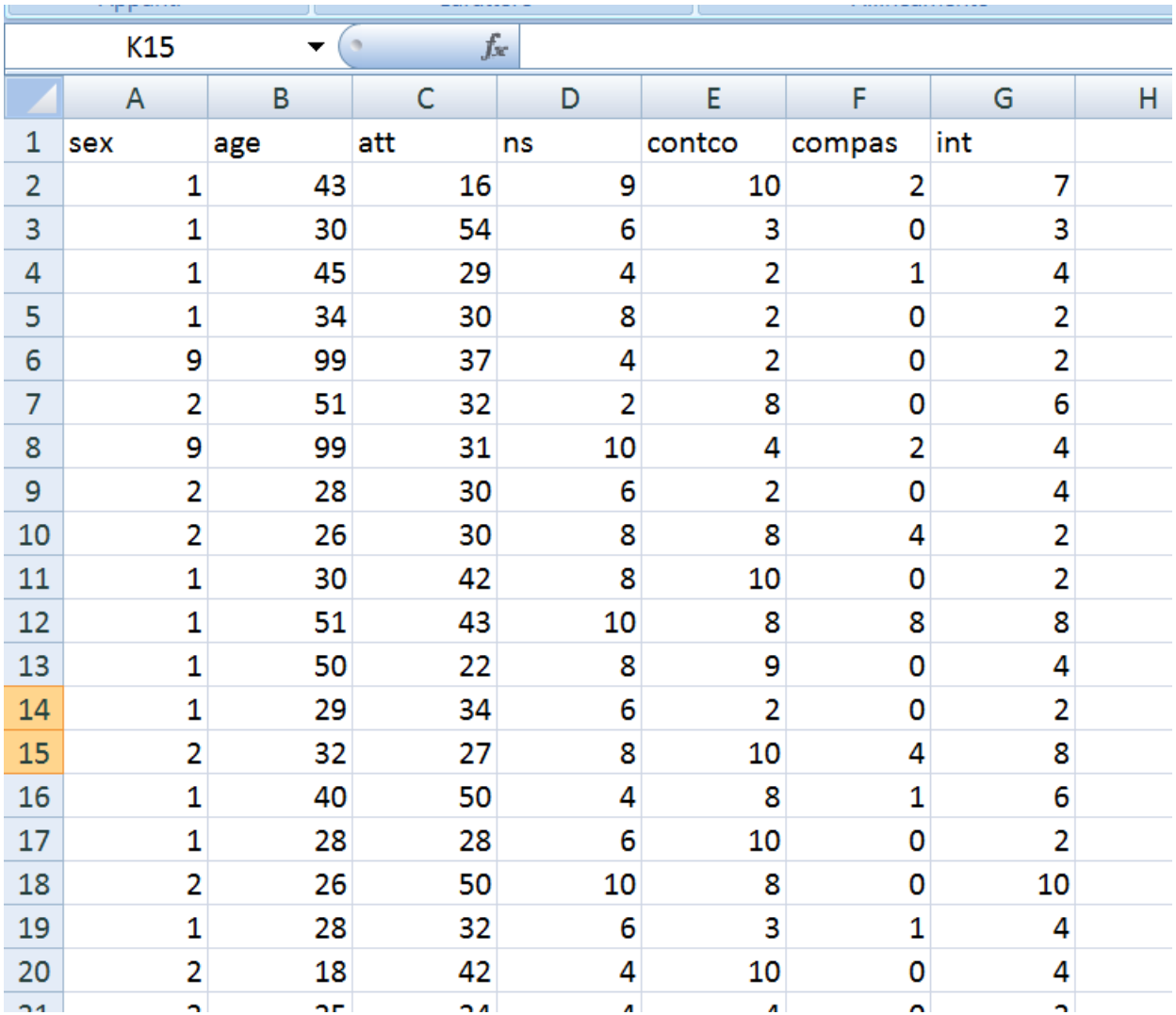
- File di fogli elettronici come Excel (*.xls, xlsx), o Lotus (*.wk3, *.wk1, *.wks).

- File SYLK - Symbolic Link per fogli elettronici di Microsoft Excel e Multiplan (*.slk).

- File dBASE IV, III o II (*.dbf)

- File SAS (*.sd2, *.ssd01, *.ssd04, *.sd7, *.sas7bdat, *.ssd01, *.xpt).

Aprire un file di dati in formato excel

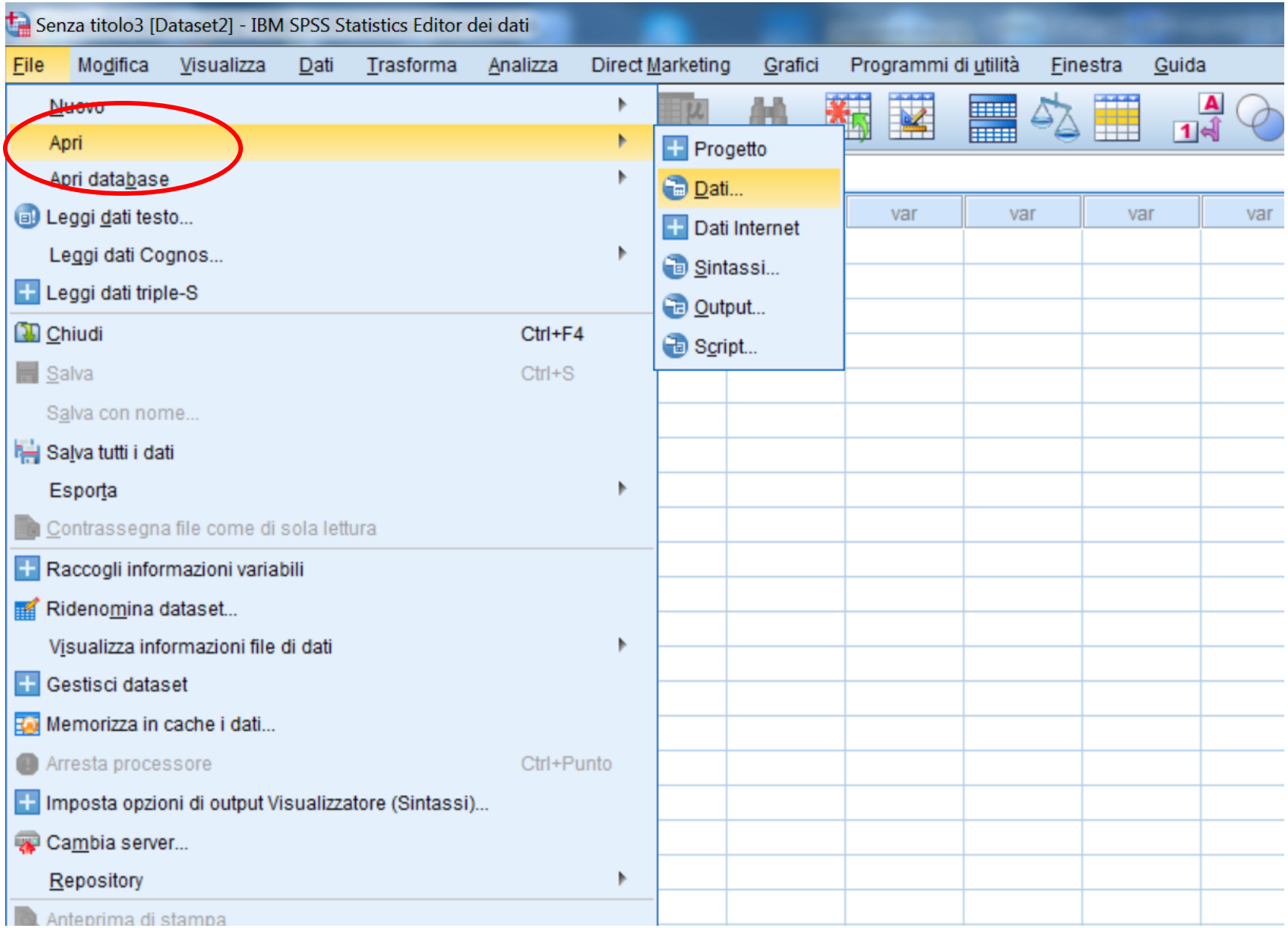


K15

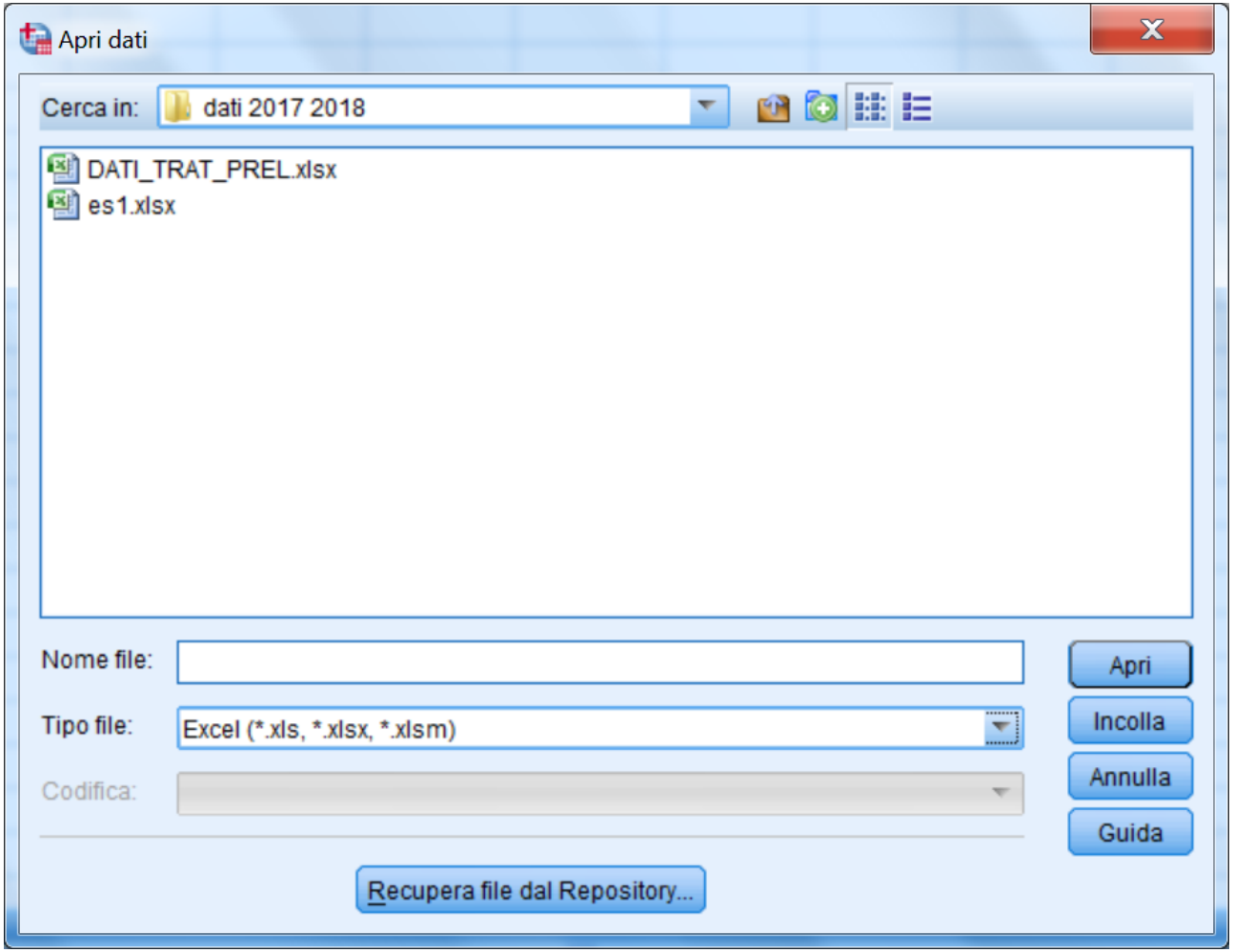
	A	B	C	D	E	F	G	H
1	sex	age	att	ns	contco	compas	int	
2	1	43	16	9	10	2	7	
3	1	30	54	6	3	0	3	
4	1	45	29	4	2	1	4	
5	1	34	30	8	2	0	2	
6	9	99	37	4	2	0	2	
7	2	51	32	2	8	0	6	
8	9	99	31	10	4	2	4	
9	2	28	30	6	2	0	4	
10	2	26	30	8	8	4	2	
11	1	30	42	8	10	0	2	
12	1	51	43	10	8	8	8	
13	1	50	22	8	9	0	4	
14	1	29	34	6	2	0	2	
15	2	32	27	8	10	4	8	
16	1	40	50	4	8	1	6	
17	1	28	28	6	10	0	2	
18	2	26	50	10	8	0	10	
19	1	28	32	6	3	1	4	
20	2	18	42	4	10	0	4	

DATI_TRAT_PREL.xlsx

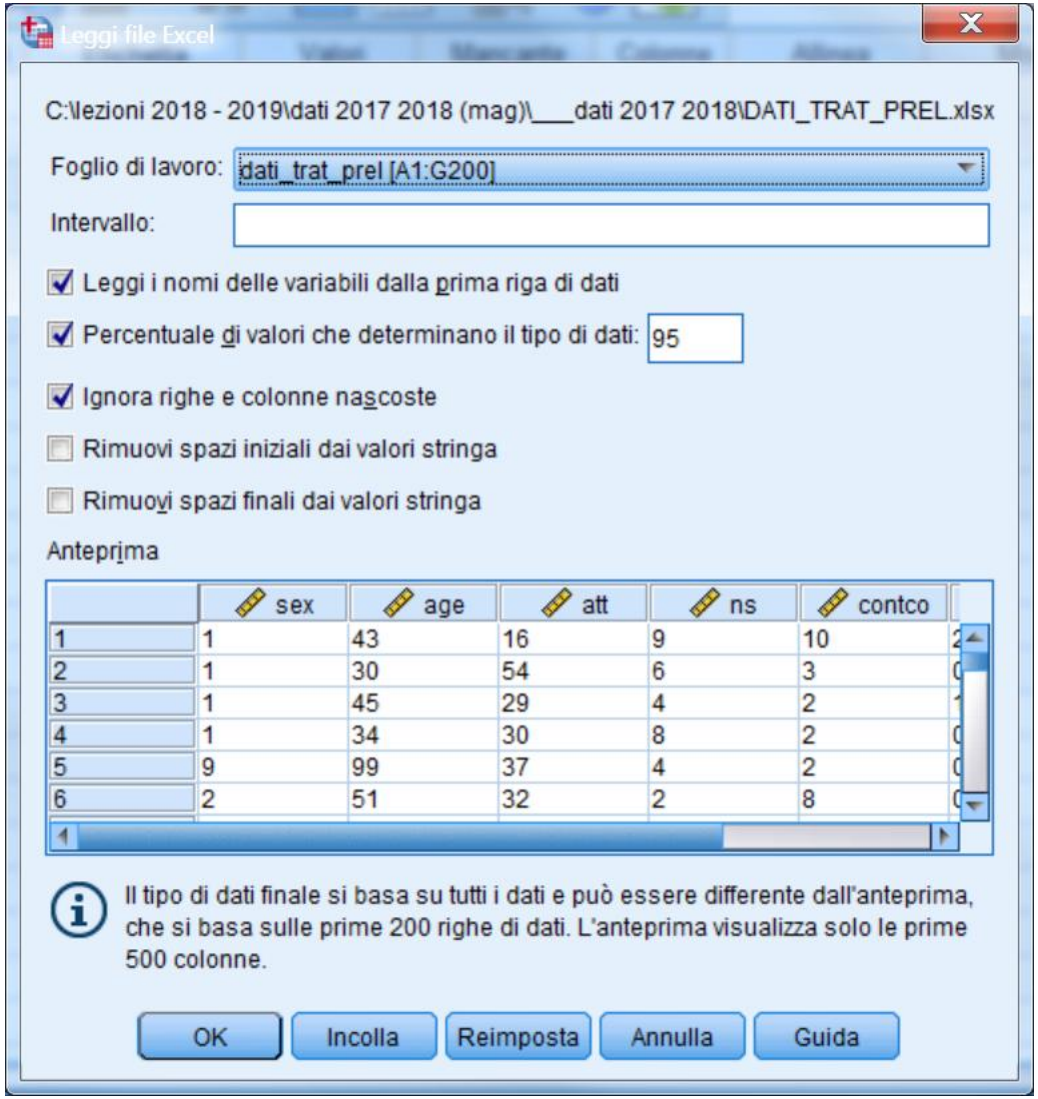
Aprire i dati



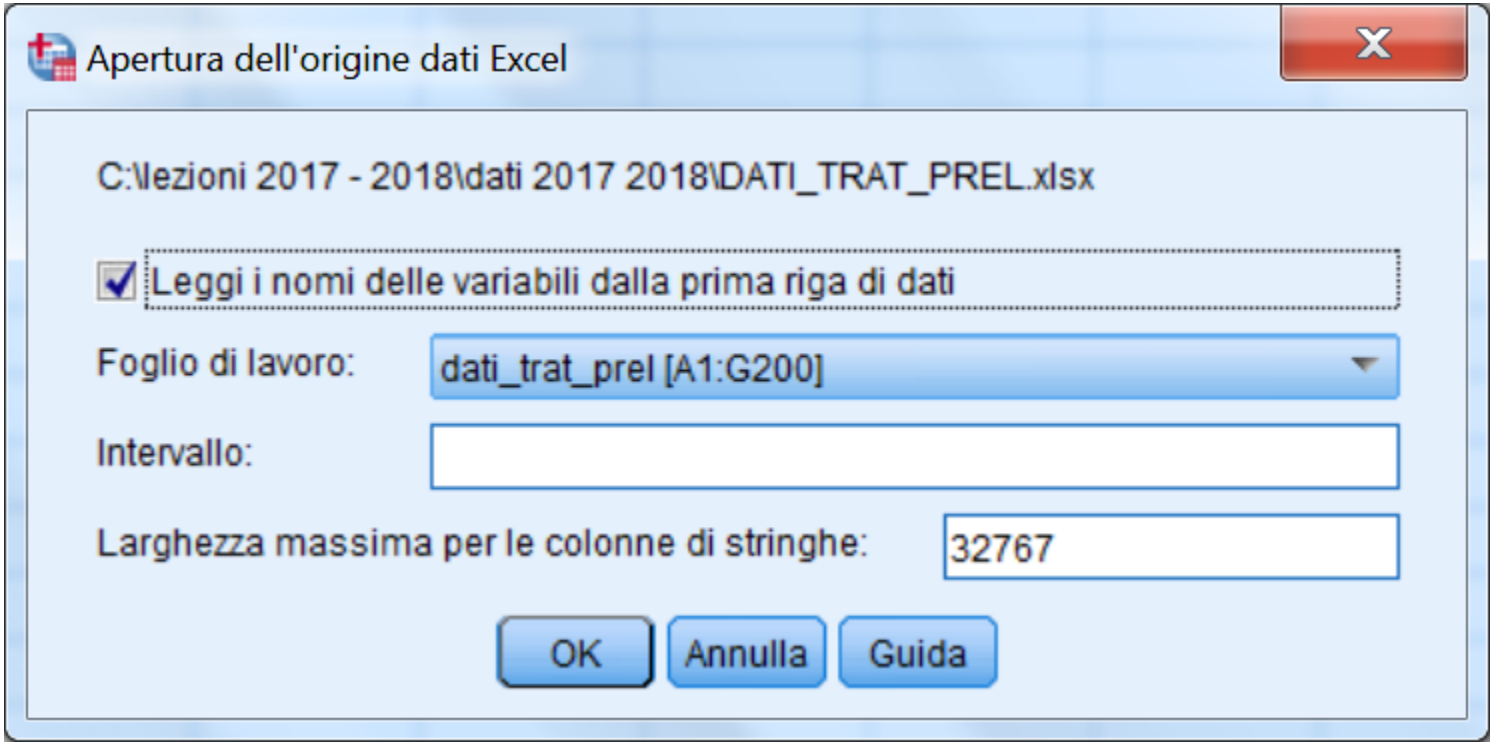
Aprire i dati



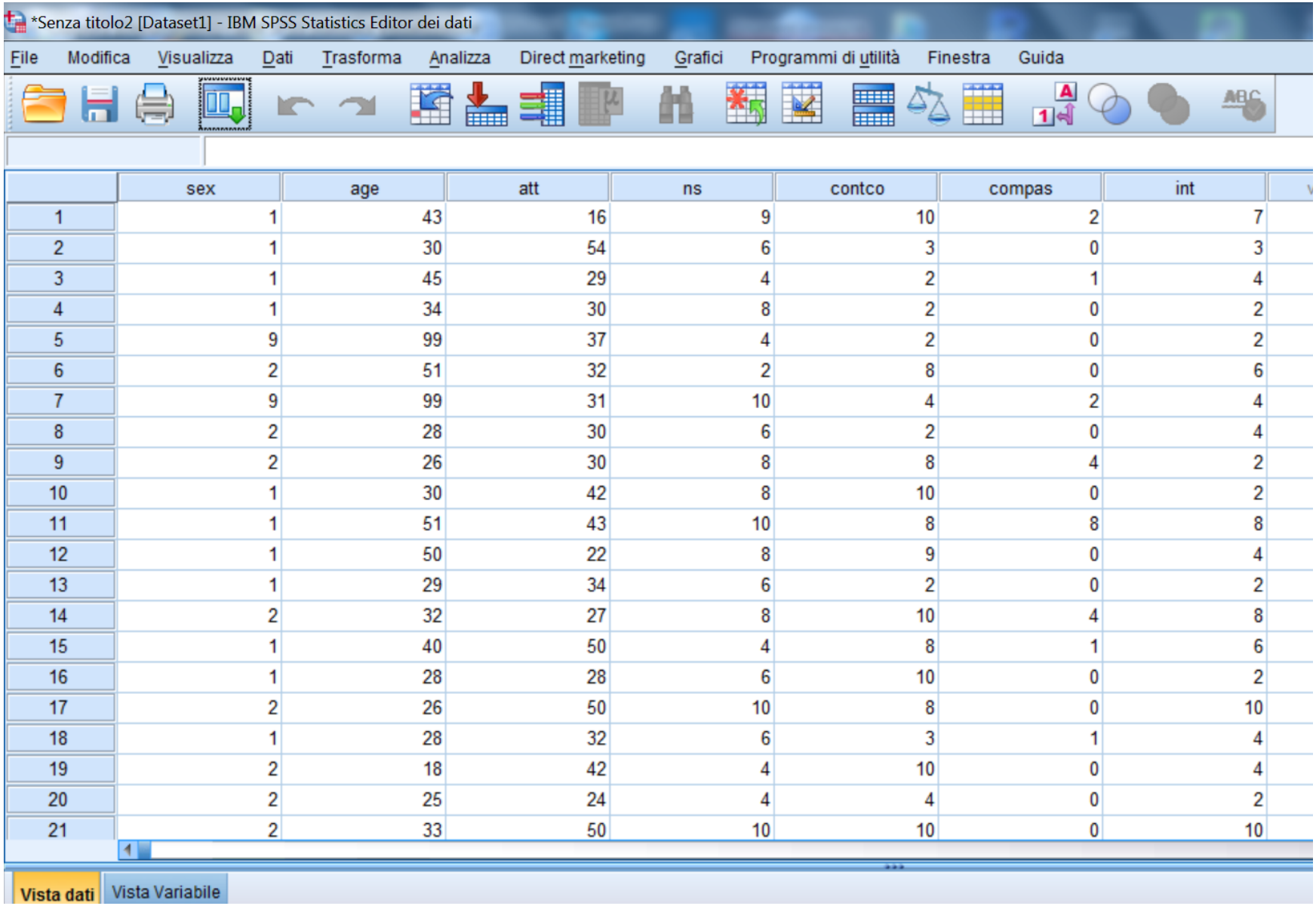
Leggere i dati (spss 25)



Leggere i dati (versioni precedenti di spss)



Aprire i dati



	sex	age	att	ns	contco	compas	int
1	1	43	16	9	10	2	7
2	1	30	54	6	3	0	3
3	1	45	29	4	2	1	4
4	1	34	30	8	2	0	2
5	9	99	37	4	2	0	2
6	2	51	32	2	8	0	6
7	9	99	31	10	4	2	4
8	2	28	30	6	2	0	4
9	2	26	30	8	8	4	2
10	1	30	42	8	10	0	2
11	1	51	43	10	8	8	8
12	1	50	22	8	9	0	4
13	1	29	34	6	2	0	2
14	2	32	27	8	10	4	8
15	1	40	50	4	8	1	6
16	1	28	28	6	10	0	2
17	2	26	50	10	8	0	10
18	1	28	32	6	3	1	4
19	2	18	42	4	10	0	4
20	2	25	24	4	4	0	2
21	2	33	50	10	10	0	10

SPSS
Il menu file

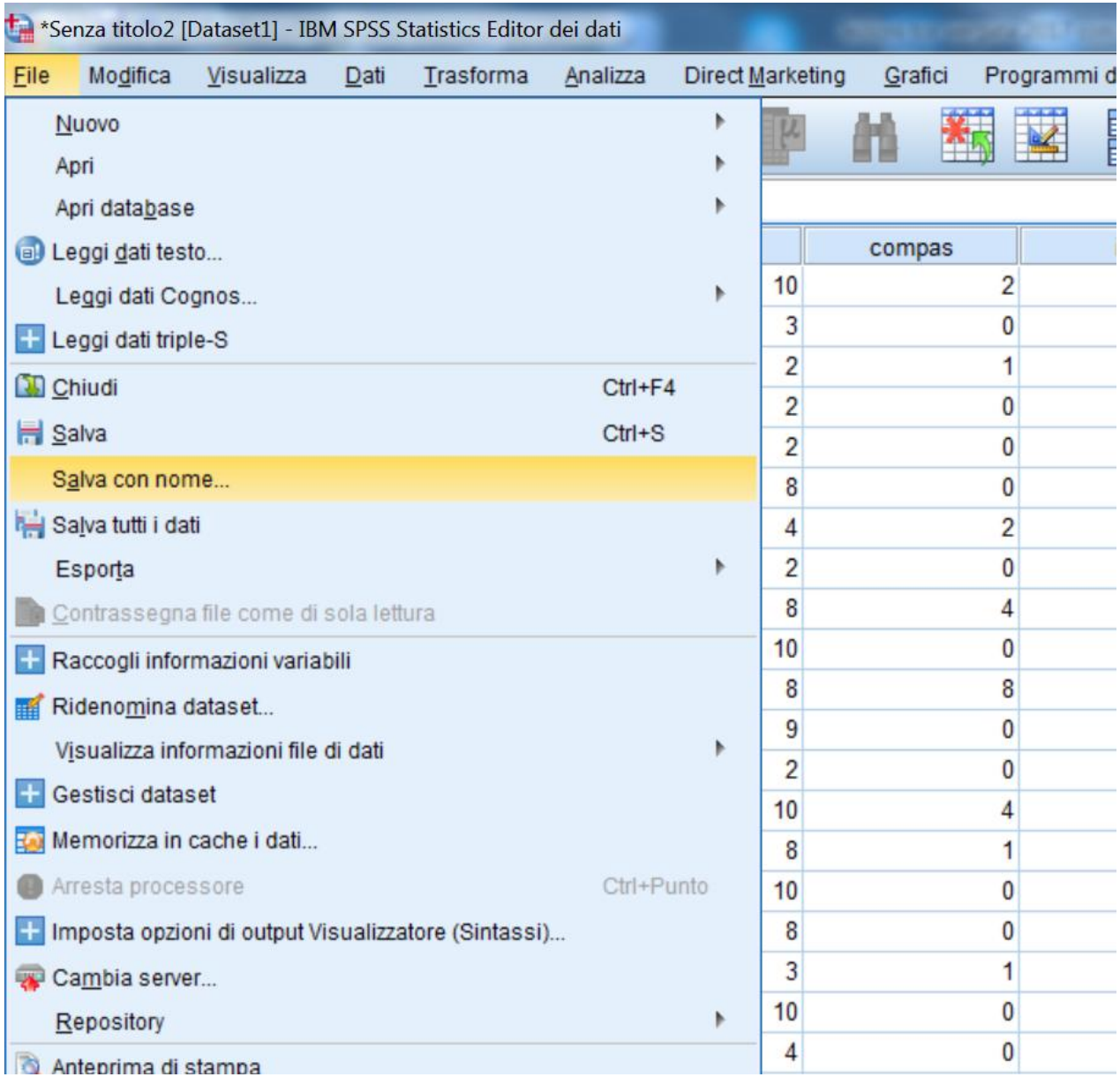
Salvare ed esportare un file dati

Per salvare un file di dati scegliere dal menu: File ⇒Salva oppure File ⇒Salva con nome.

Nel secondo caso si aprirà una finestra di dialogo analoga a quella relativa all'apertura dei file che consente di specificare il percorso per il file da salvare, e di definire il tipo di file che viene salvato.

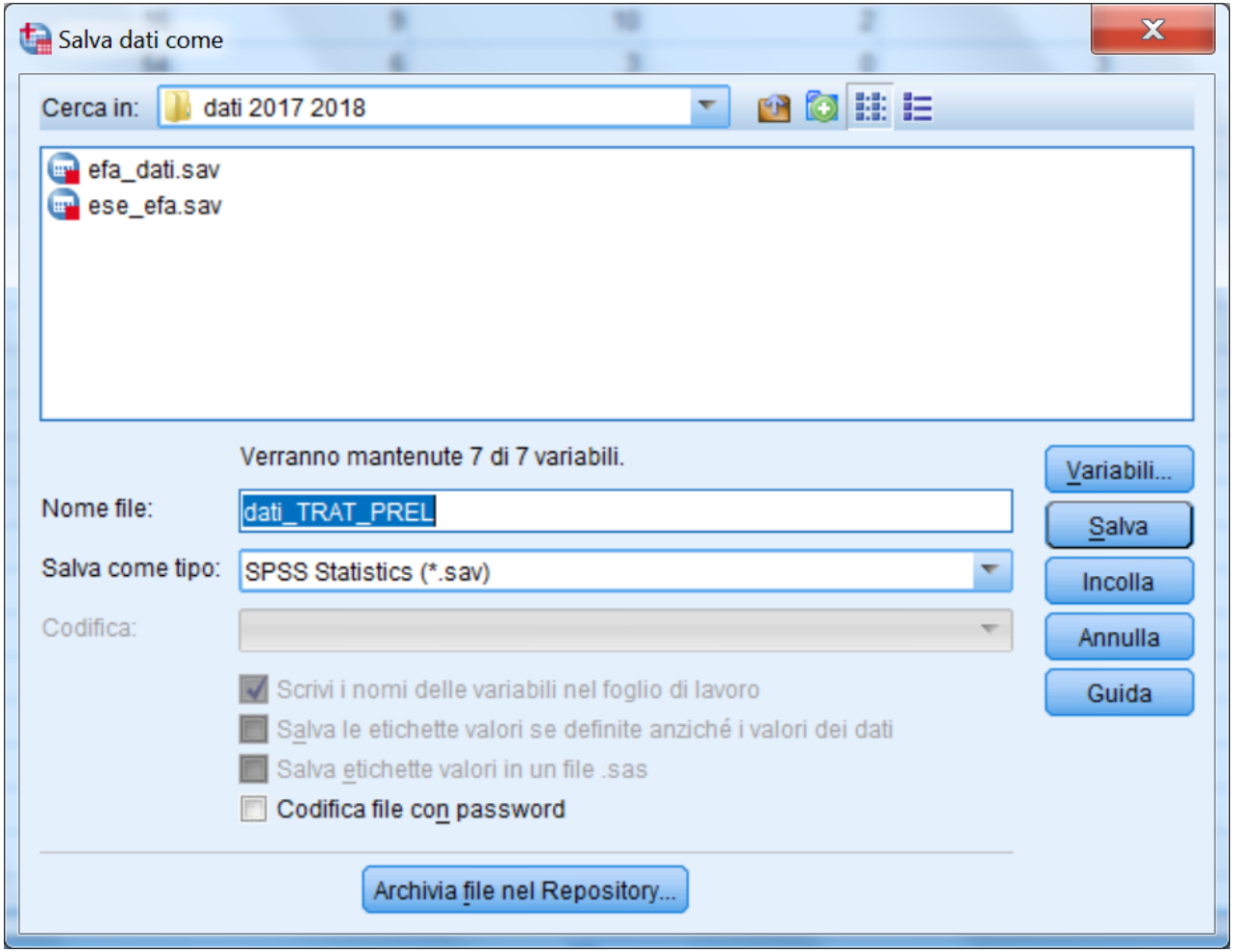
I formati di file definibili sono quelli esaminati nella slide relativa all'apertura dei file.

Salvare i dati



Salvare i dati

Salviamo il nostro file importato da excel per usarlo come file .sav nei prossimi esempi (altrimenti alla chiusra del programma andrebbe perso).



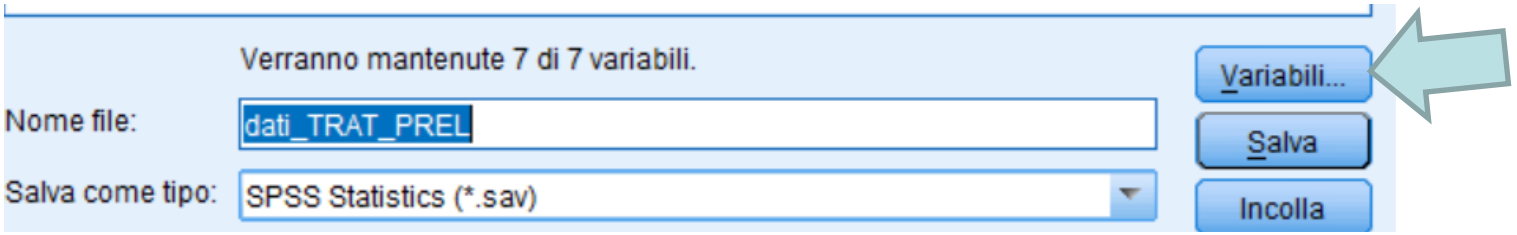
SPSS
Il menu file

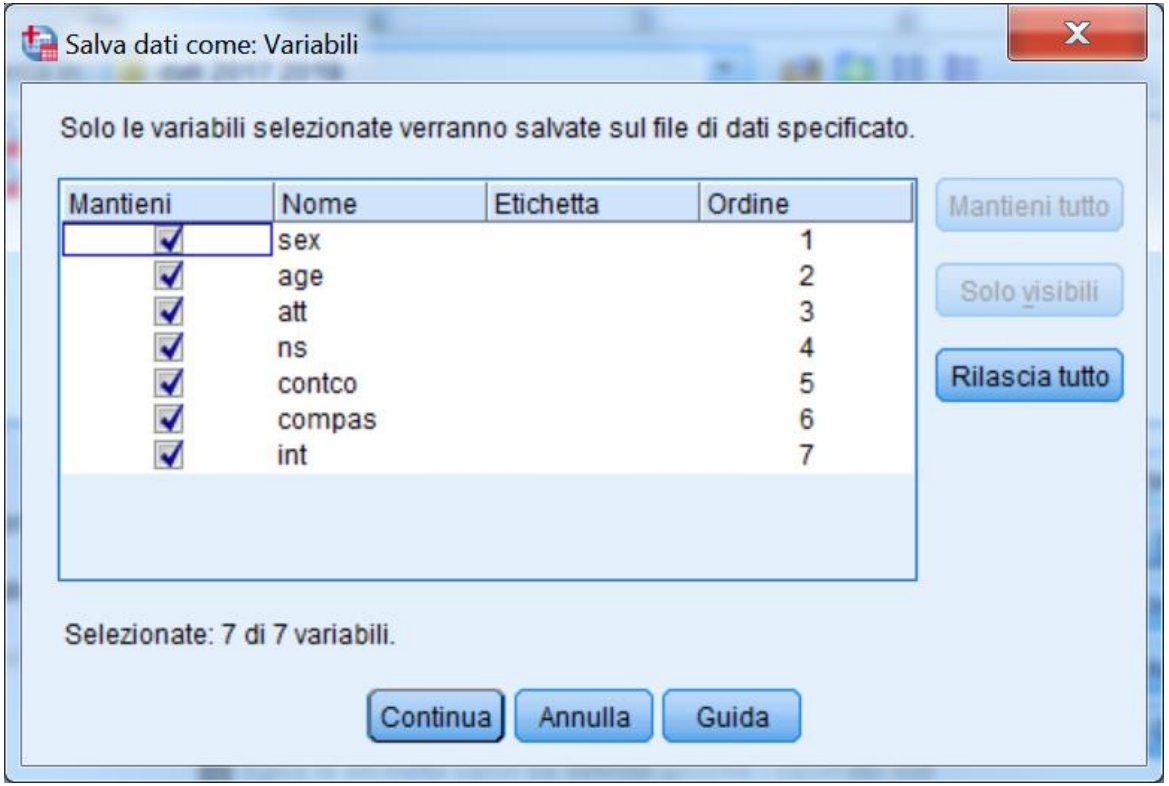
E' possibile escludere variabili dal file che viene salvato cliccando sul pulsante "Variabili" e scegliendo quali variabili eliminare.

Nella figura successiva viene mostrata la finestra di dialogo che consente di filtrare le variabili: se viene lasciata l'opzione di default tutte le variabili vengono mantenute nel file che viene salvato.

Per eliminare una variabile dal nuovo file è sufficiente effettuare un clic del mouse sul quadrato corrispondente alla variabile nella colonna "Mantieni".

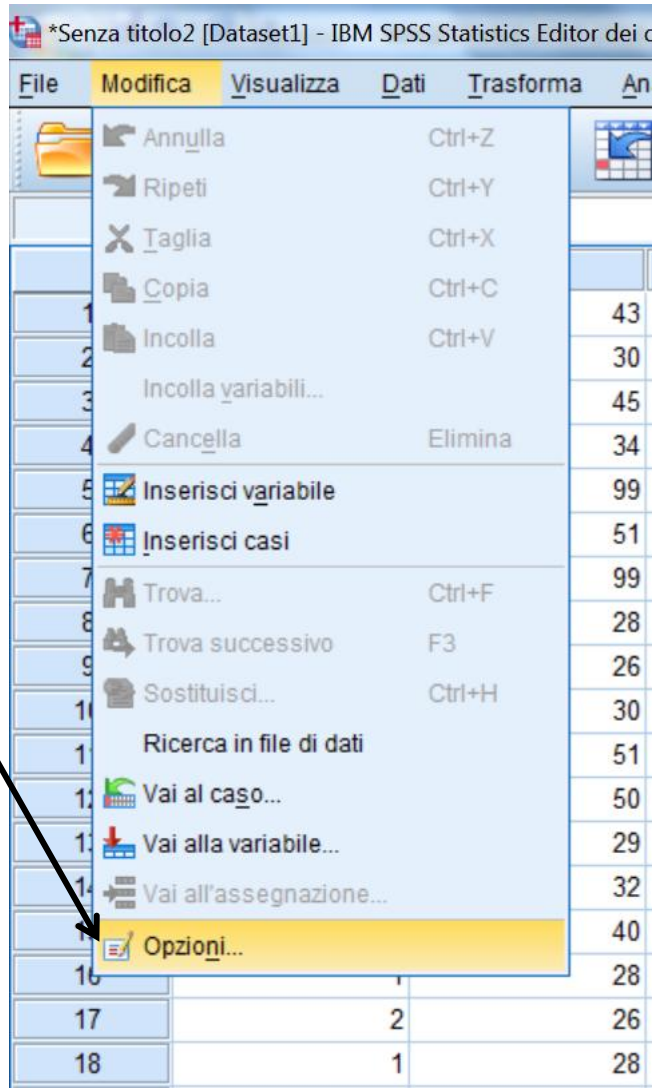
SPSS
Il menu file





SPSS
Il menu modifica

Questo menu consente di copiare, tagliare, incollare e cancellare righe e colonne nell'Editor dei dati, trovare dei valori specifici per una data variabile e definire le opzioni di base per il programma (es., definire il tipo di visualizzazione delle variabili negli elenchi e negli output) per le quali si rimanda ai manuali specifici e alle funzioni di aiuto in linea.

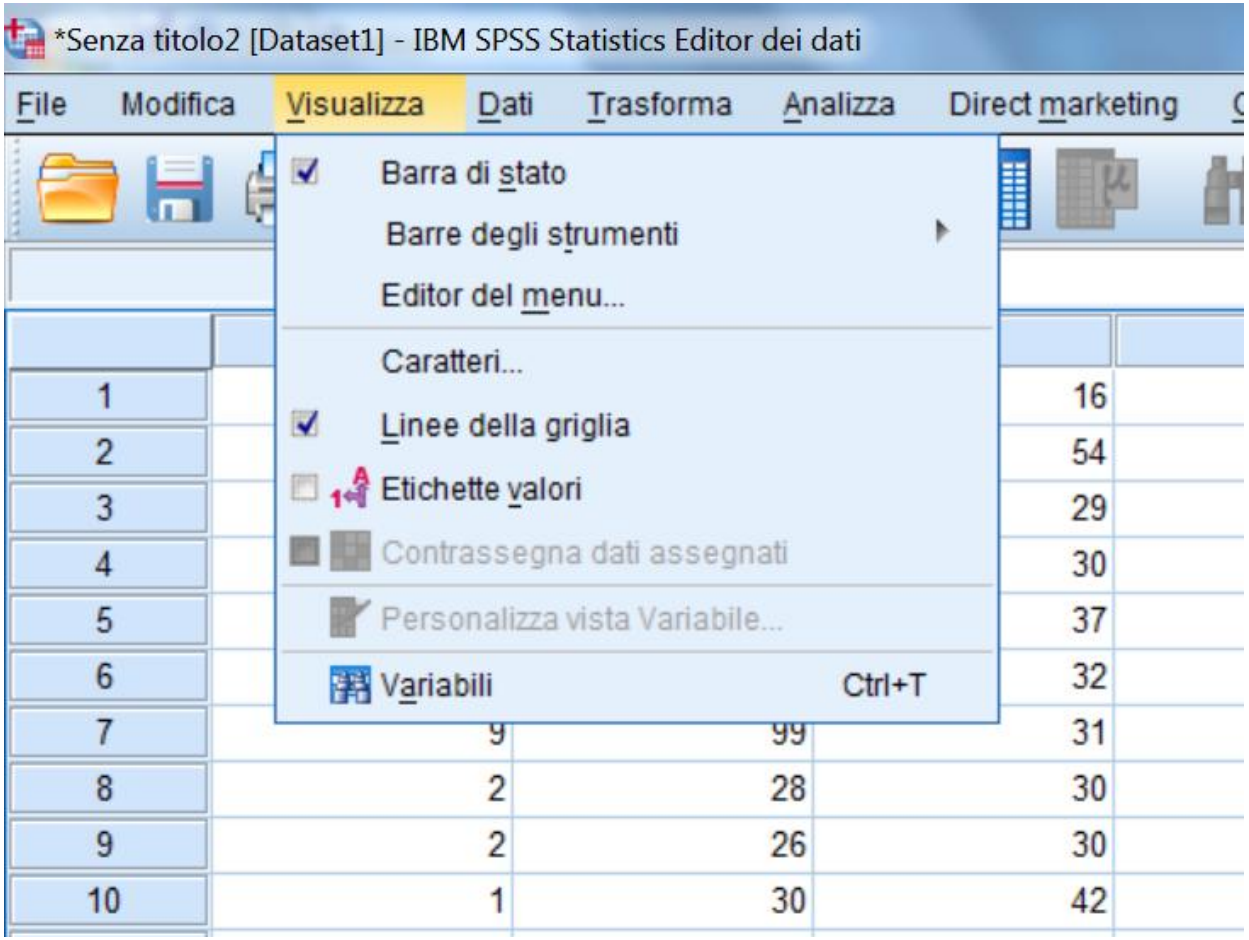


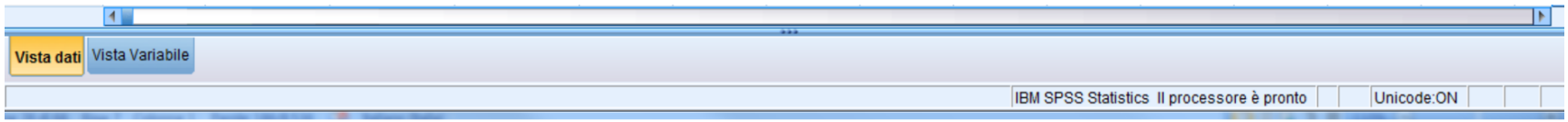
SPSS
Il menu visualizza

Questo menù definisce il modo in cui vengono visualizzate la barra di stato, le barre degli strumenti, le variabili, le griglie della tabella dei dati, e definisce i caratteri utilizzati per visualizzare le diverse informazioni. In particolare:

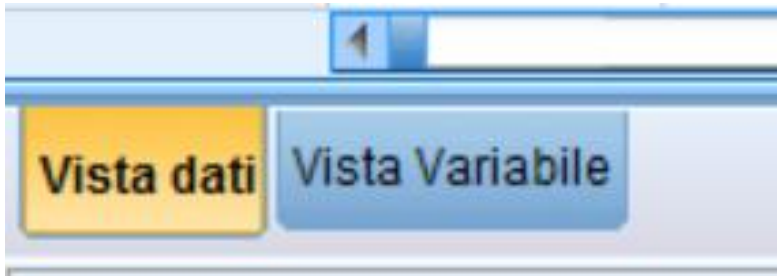
- L'opzione *Barra di stato* consente di mostrare o nascondere la barra di stato, ovvero quella zona della parte inferiore di una finestra SPSS nella quale sono visualizzate le informazioni sullo stato di esecuzione dei programmi, sullo stato del filtro e della ponderazione dei casi

SPSS
Il menu visualizza





SPSS
Vista dati



*Senza titolo2 [Dataset1] - IBM SPSS Statistics Editor dei dati

File Modifica Visualizza Dati Trasforma Analizza Direct marketing Grafici Programmi

5 :

	sex	age	att	ns	c
1	1	43	16	9	
2	1	30	54	6	
3	1	45	29	4	
4	1	34	30	8	
5	9	99	37	4	
6	2	51	32	2	
7	9	99	31	10	
8	2	28	30	6	
9	2	26	30	8	
10	1	30	42	8	

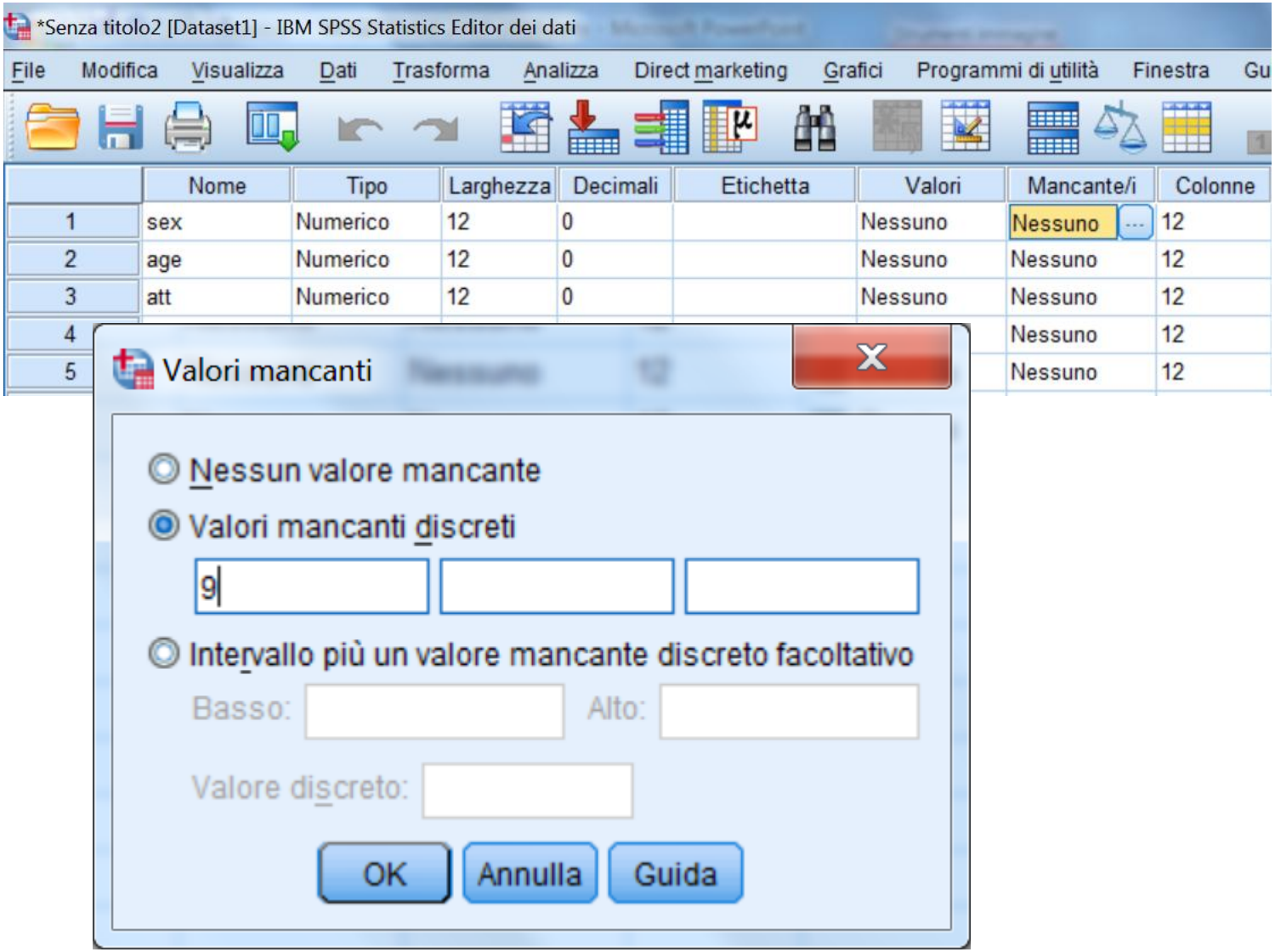
SPSS
Vista Variabile

*Senza titolo2 [Dataset1] - IBM SPSS Statistics Editor dei dati

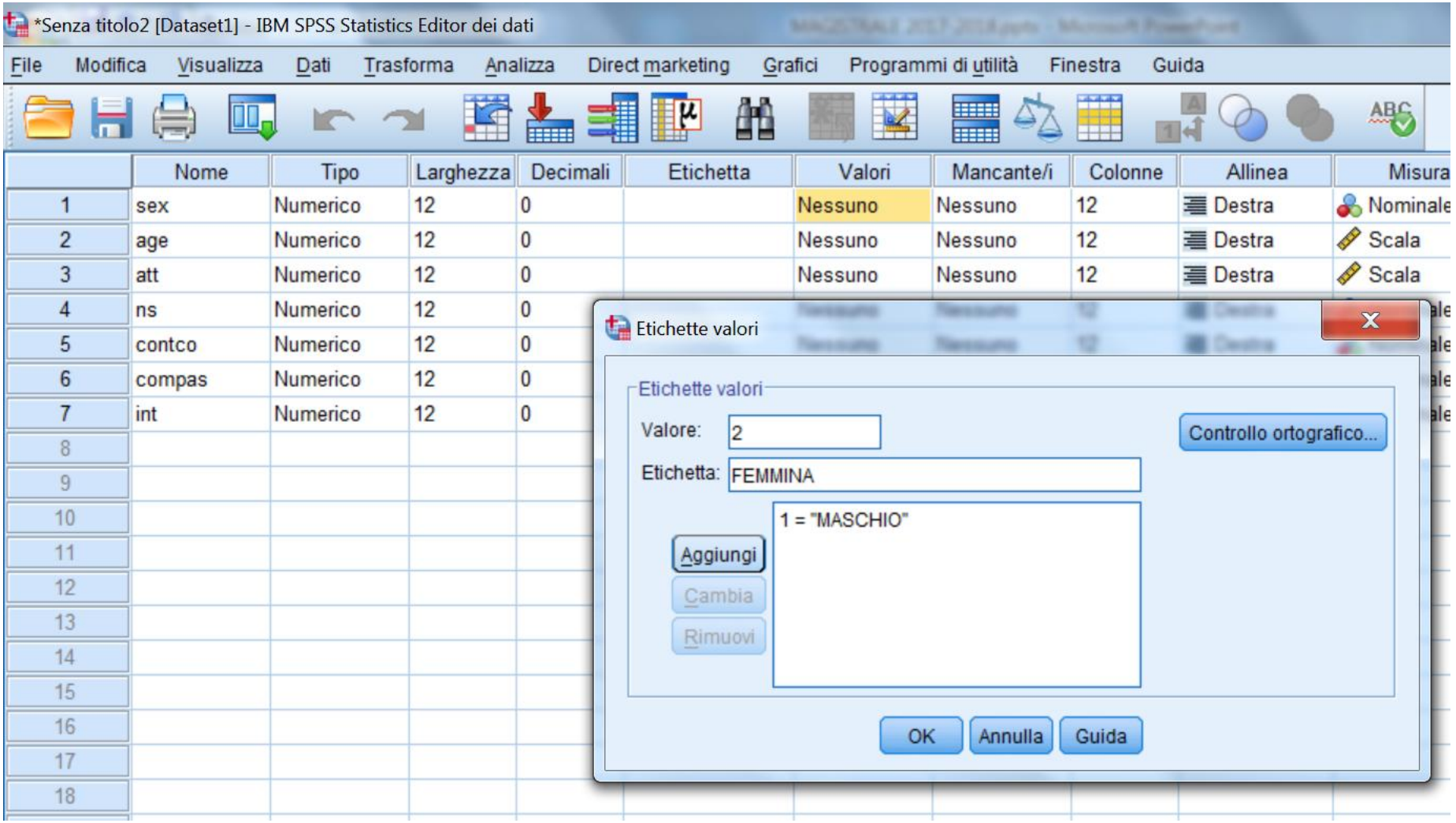
File Modifica Visualizza Dati Trasforma Analizza Direct marketing Grafici Programmi di utilità Finestra Guida

	Nome	Tipo	Larghezza	Decimali	Etichetta	Valori	Mancante/i	Colonne	Allinea	Misura	Ruolo
1	sex	Numerico	12	0		Nessuno	Nessuno	12	Destra	Nominale	Input
2	age	Numerico	12	0		Nessuno	Nessuno	12	Destra	Scala	Input
3	att	Numerico	12	0		Nessuno	Nessuno	12	Destra	Scala	Input
4	ns	Numerico	12	0		Nessuno	Nessuno	12	Destra	Nominale	Input
5	contco	Numerico	12	0		Nessuno	Nessuno	12	Destra	Nominale	Input
6	compas	Numerico	12	0		Nessuno	Nessuno	12	Destra	Nominale	Input
7	int	Numerico	12	0		Nessuno	Nessuno	12	Destra	Nominale	Input
8											
9											

SPSS
Vista Variabile

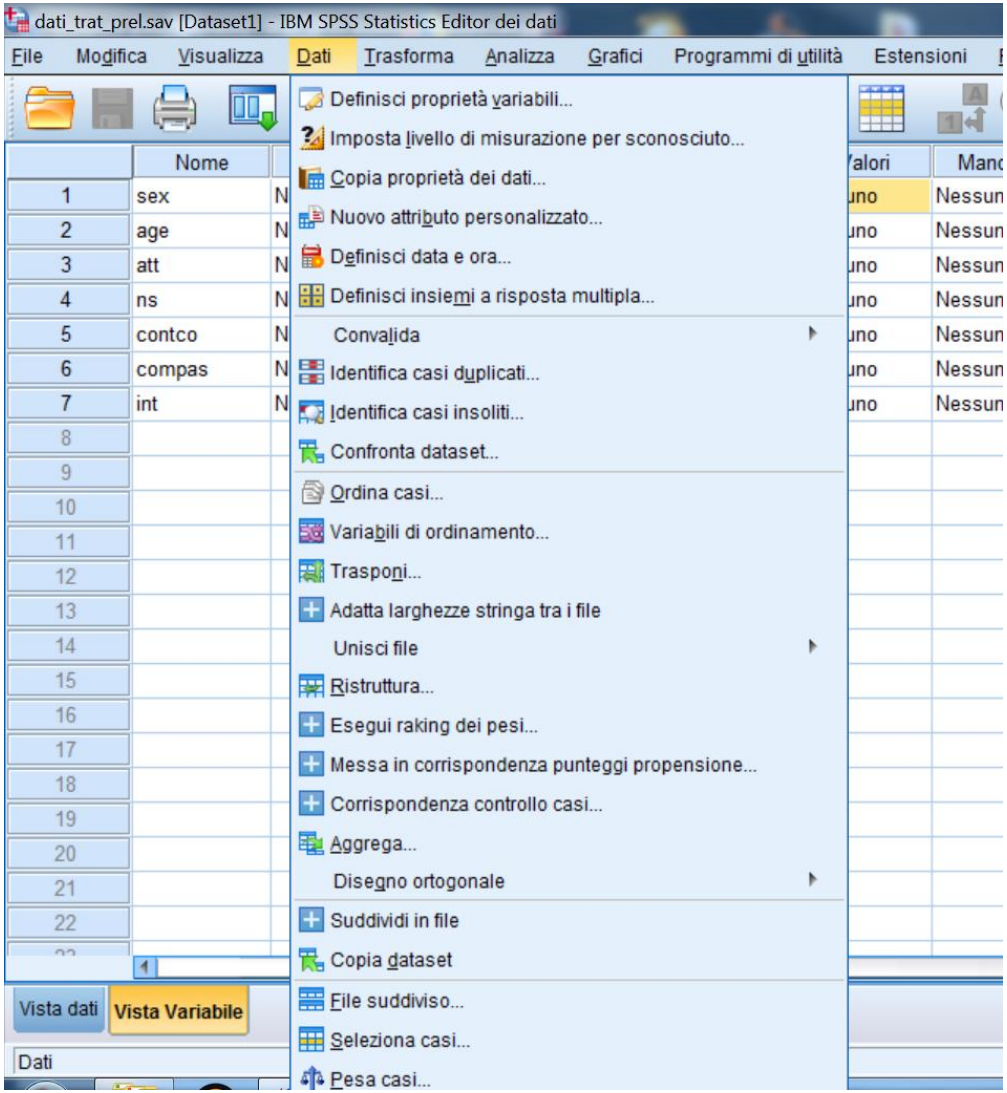


SPSS
Vista Variabile



SPSS
Il menu dati

Effettua operazioni sulle variabili e sui casi.



SPSS
Il menu dati

I sotto-menù più utili sono:

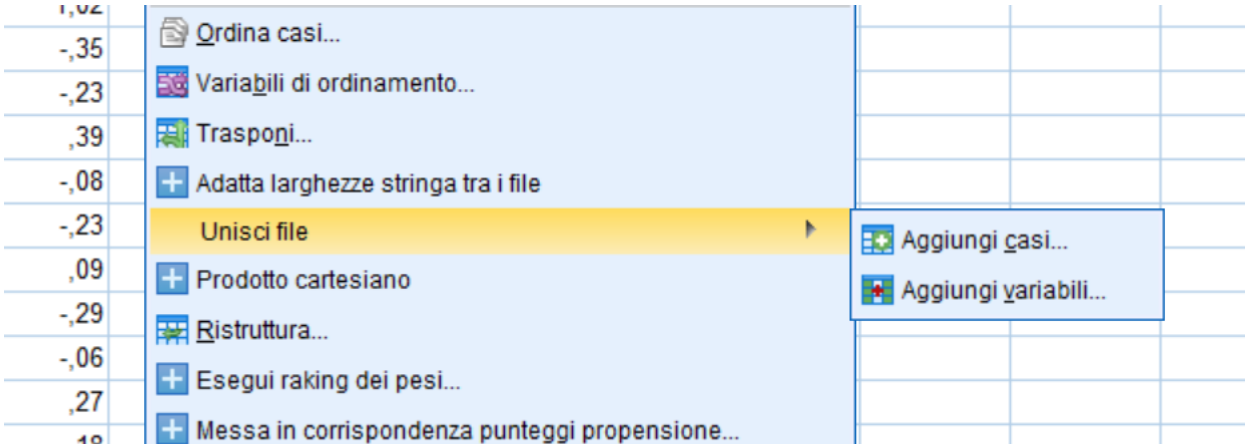
"*Copia proprietà dei dati*" consente all'utente di prendere un file dati SPSS esterno ed utilizzarlo come modello per la definizione del file dati corrente. In particolare, sia le proprietà del file (es., etichetta del file, insiemi a risposta multipla, ecc.), sia quelle delle variabili (es., etichette dei valori, valori mancanti, etichette delle variabili, ecc.) del file "modello" possono essere utilizzate per definire quelle del file corrente

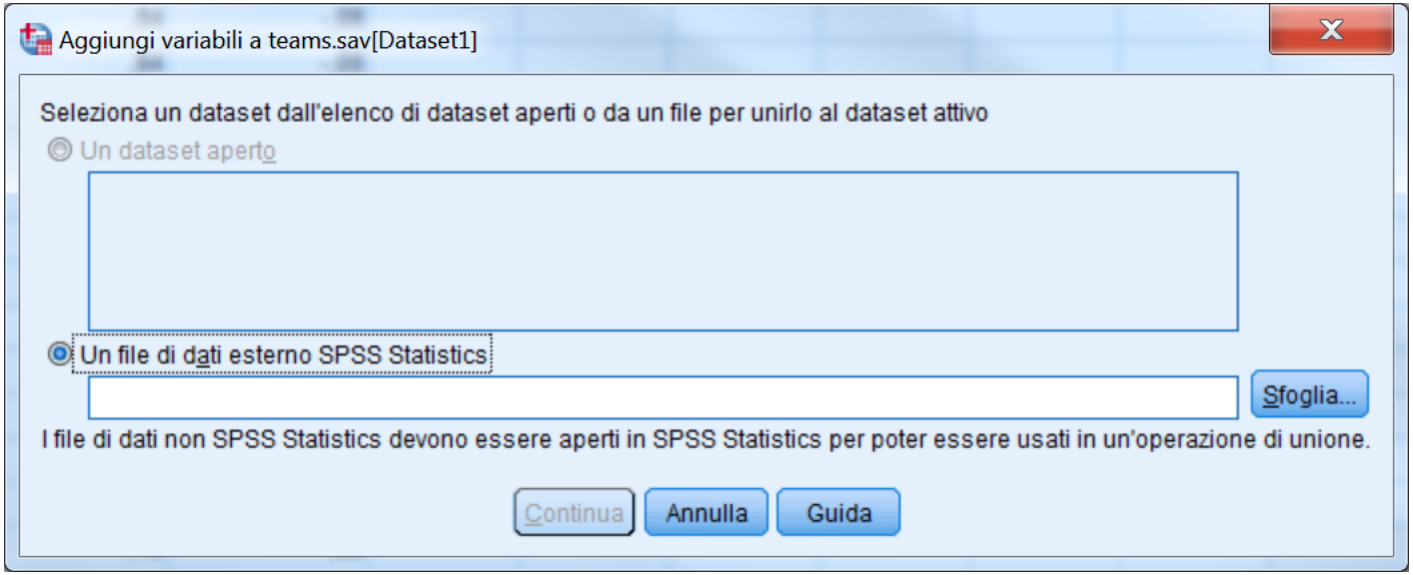
SPSS
Il menu dati

I sotto-menù più utili sono:

"*Unisci file*" consente di unire due file in un unico file e presenta due diverse modalità fondamentali: *Aggiungi casi* e *Aggiungi variabili.*
***Aggiungi casi* unisce il file attivo con un secondo file che contiene le stesse variabili ma casi differenti.**
***Aggiungi variabili* unisce il file attivo con un file dati esterno che contiene gli stessi casi ma variabili differenti da quelle nel file attivo. I casi devono avere *lo stesso ordine* in entrambi i file. Se si utilizza una "variabile chiave" per appaiare i casi, i due file devono *essere ordinati* in modo crescente rispetto alla variabile chiave.**

SPSS
Il menu dati



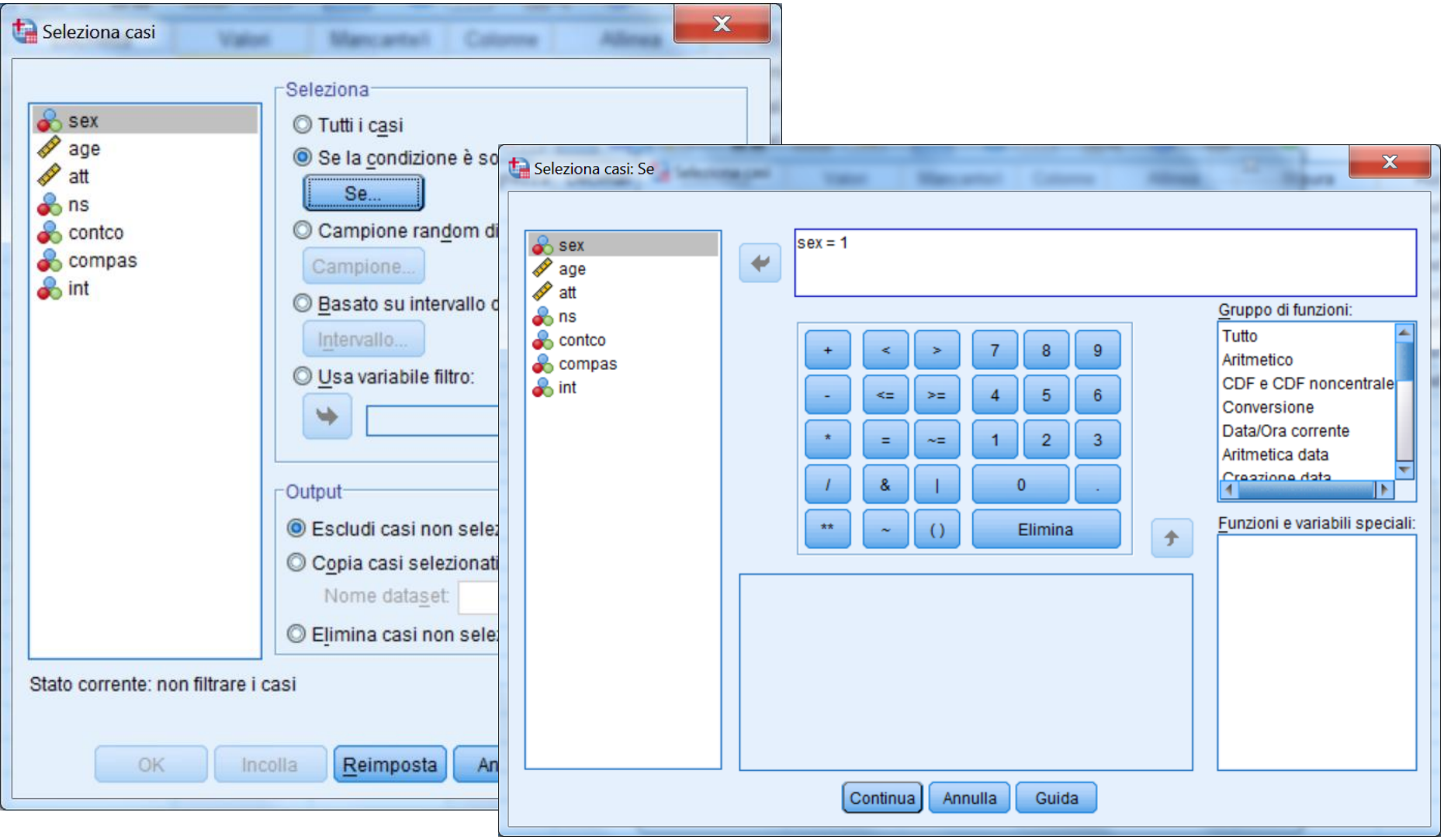


SPSS
Il menu dati

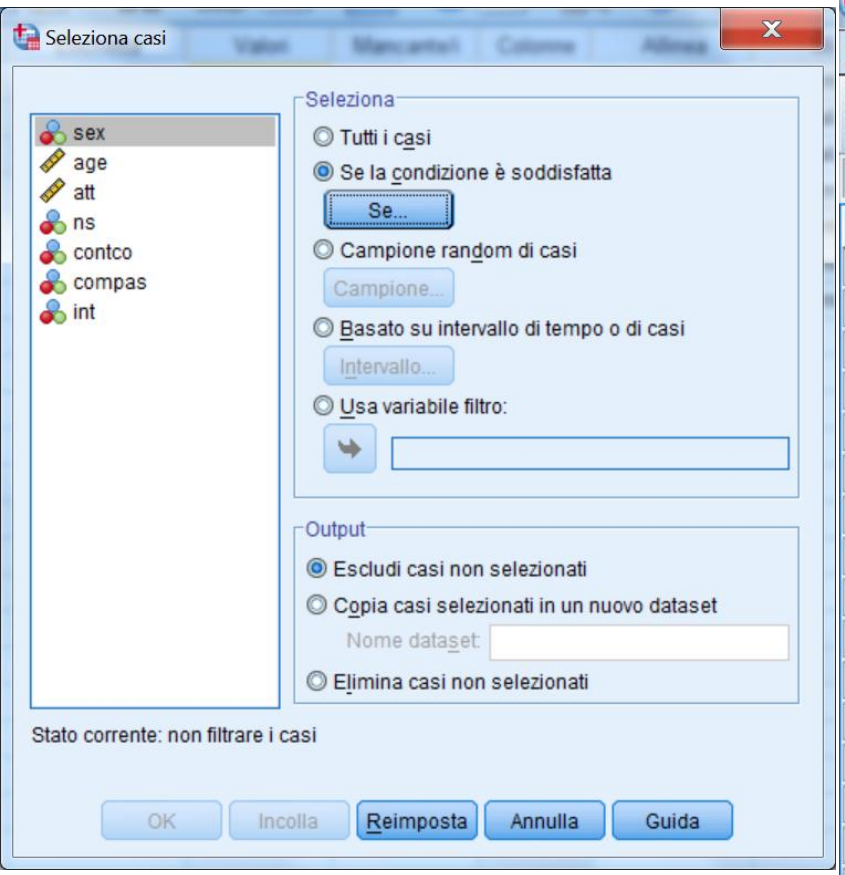
I sotto-menù più utili sono:

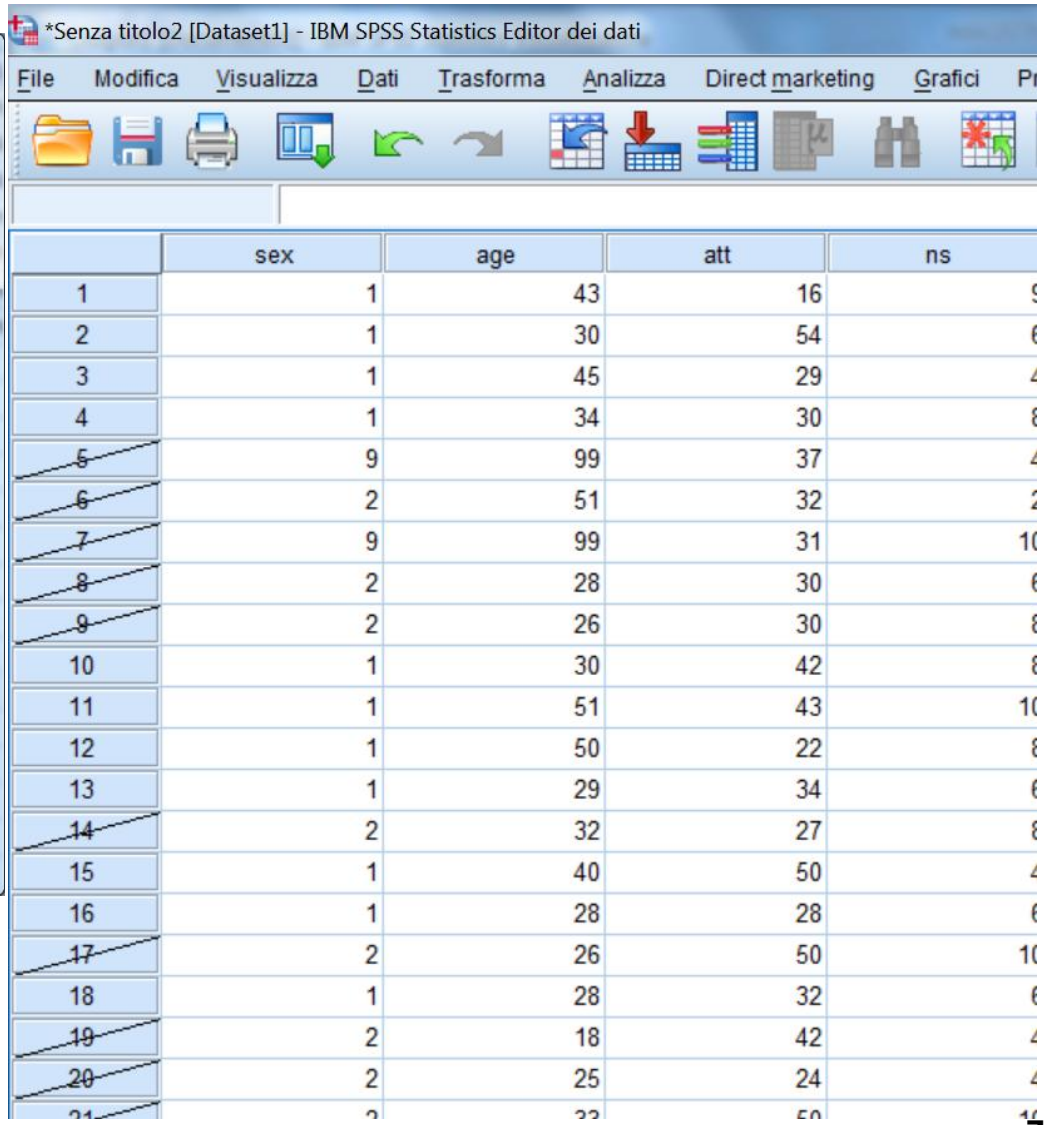
"*Seleziona Casi*" consente di definire sottoinsiemi di casi che vengono selezionati tramite un criterio specificato dall'utente stesso. Per la selezione dei casi l'utente può specificare un'operazione di natura più o meno complessa, oppure avvalersi del generatore di numeri casuali di SPSS. I casi non selezionati possono essere *filtrati* o *cancellati* del tutto dal file. La modalità che prevede che i casi siano filtrati crea una nuova variabile, "filter_$", che serve per indicare lo stato attuale del filtro. Il valore di tale variabile è uguale a 1 per i casi che soddisfano la condizione di selezione, mentre è uguale a 0 per i casi che non soddisfano tale condizione e che quindi vengono esclusi dall'analisi.

SPSS
Il menu dati



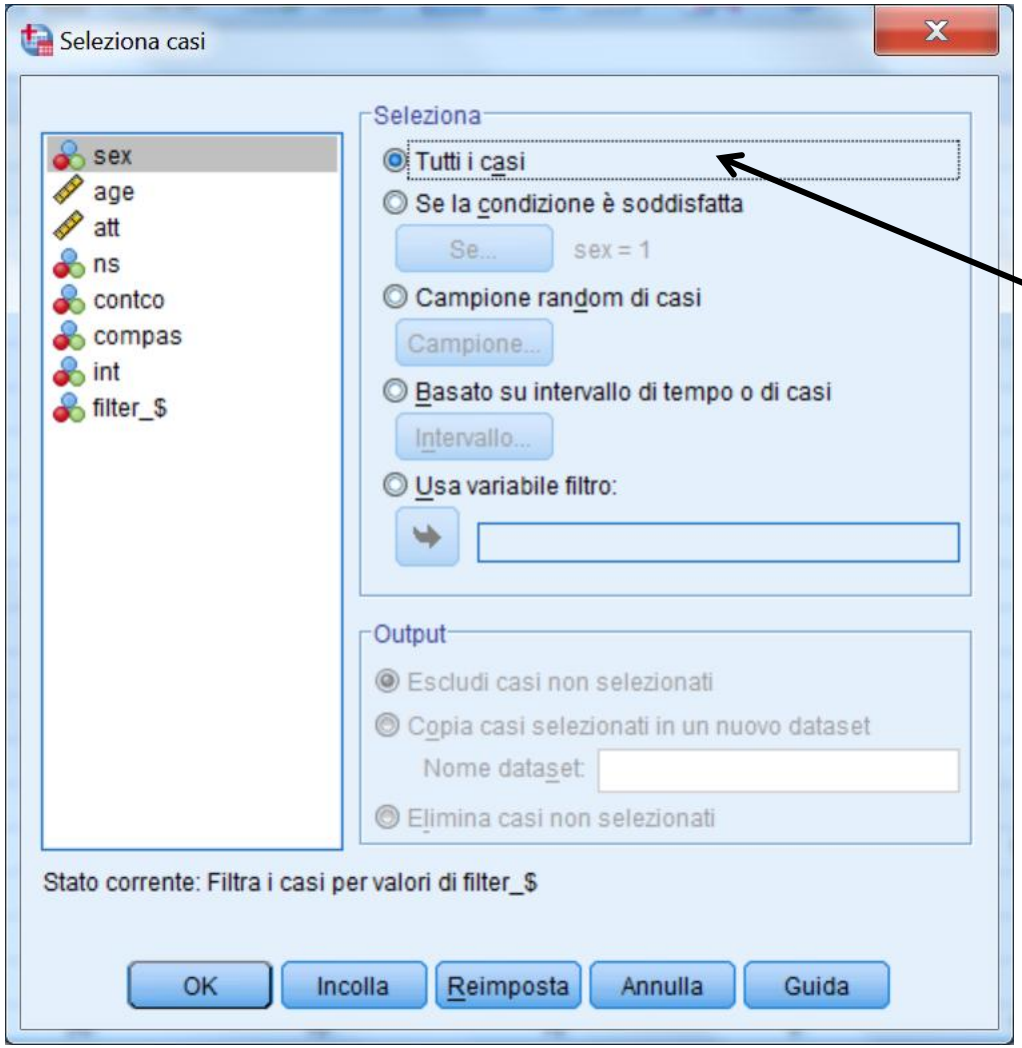
SPSS
Il menu dati





	sex	age	att	ns
1	1	43	16	
2	1	30	54	
3	1	45	29	
4	1	34	30	
5	9	99	37	
6	2	51	32	
7	9	99	31	
8	2	28	30	
9	2	26	30	
10	1	30	42	
11	1	51	43	
12	1	50	22	
13	1	29	34	
14	2	32	27	
15	1	40	50	
16	1	28	28	
17	2	26	50	
18	1	28	32	
19	2	18	42	
20	2	25	24	

SPSS
Il menu dati



Per togliere il filtro cliccare su "Tutti i casi"

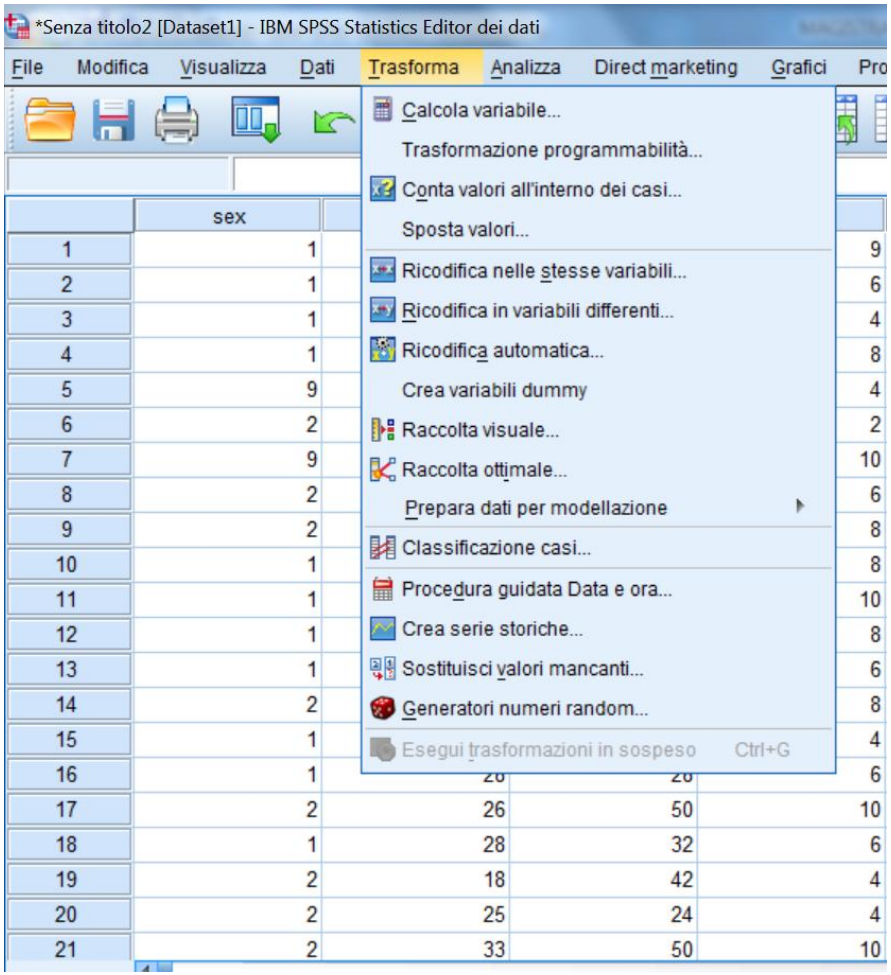
SPSS

Il menu Trasforma

Consente di modificare le variabili (o definirne delle nuove) operando trasformazioni su variabili già esistenti

Sono presenti i seguenti comandi:

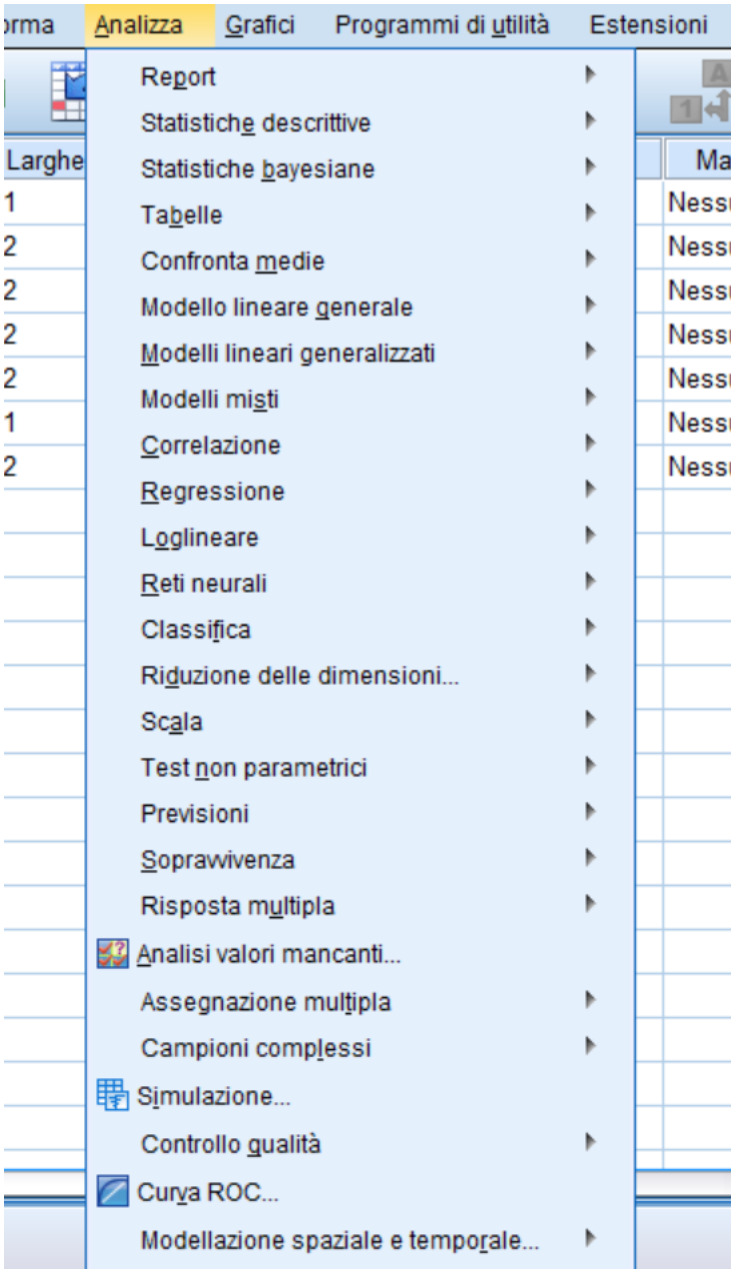
1. **Calcola variabile: consente di calcolare i valori di una variabile in base alle trasformazioni numeriche di altre variabili.**
2. **Ricodifica: è possibile scegliere tra due opzioni ricodifica nelle stesse variabili e ricodifica in variabili differenti**



SPSS

Il menu Analizza

È il menu più importante di SPSS, quello che consente di effettuare le analisi statistiche

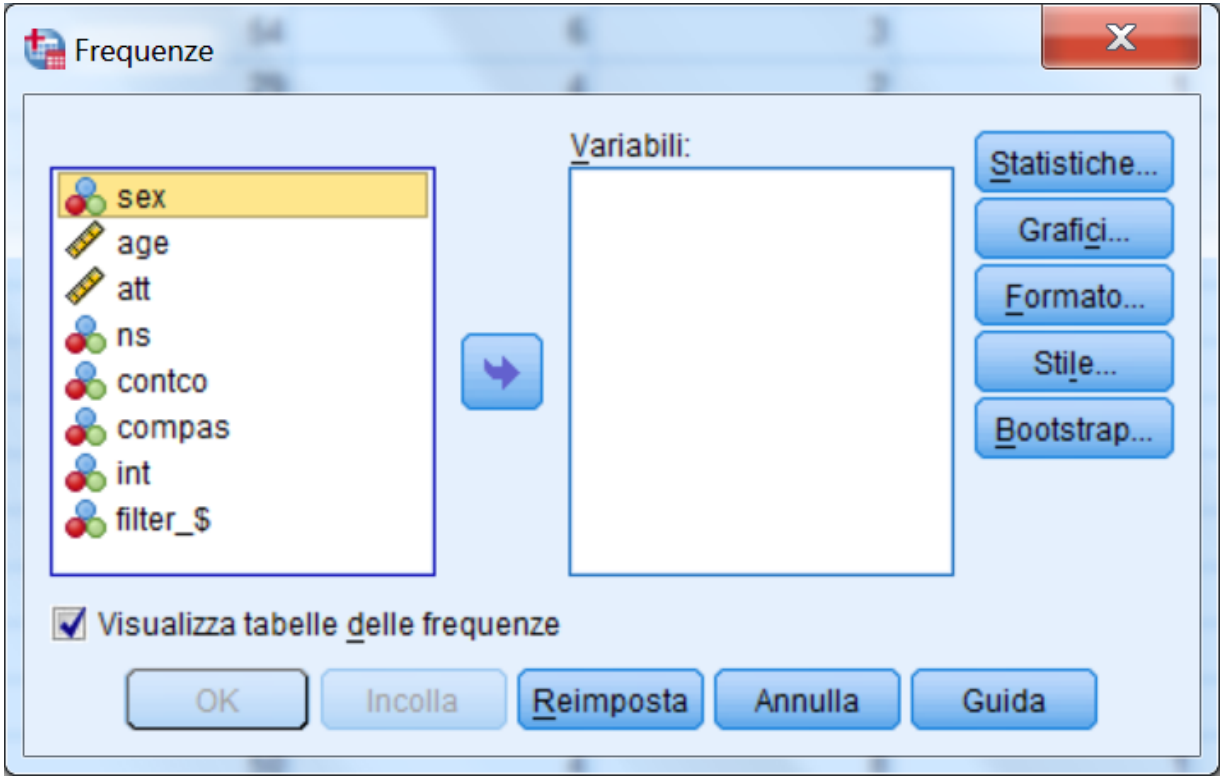


Il menù ANALIZZA si trova in tutte le finestre di SPSS

SPSS

Le finestre di Dialogo

La maggior parte delle opzioni nel menu "Analizza" consentono di aprire delle "finestre di dialogo"

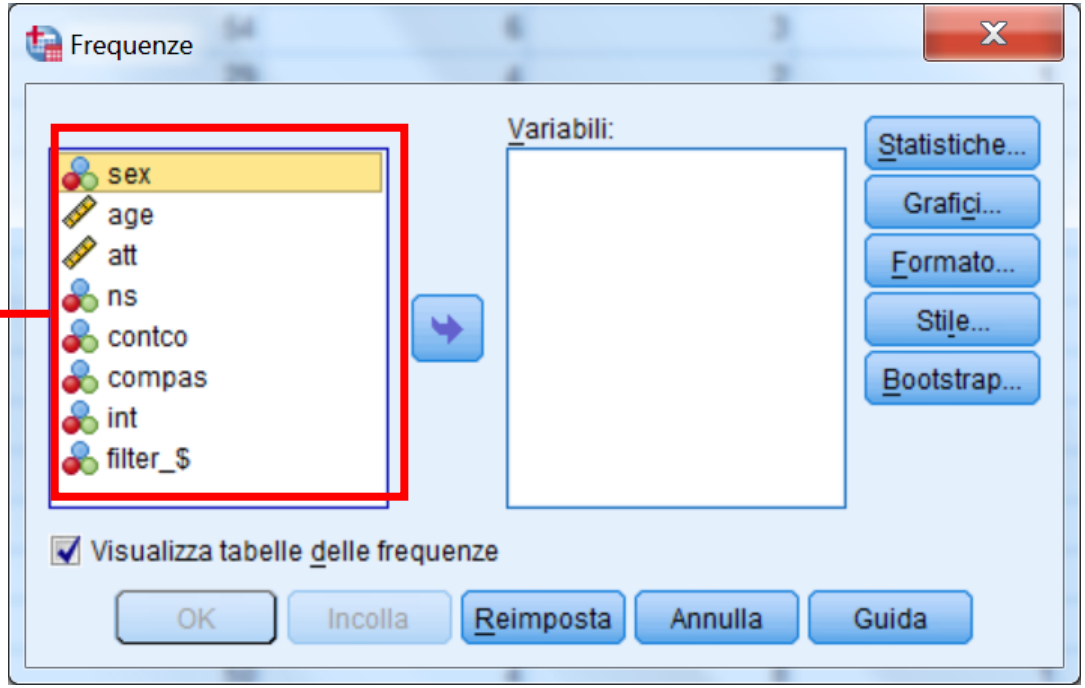


Le finestre di dialogo vengono utilizzate per selezionare le variabili da analizzare (e le diverse opzioni disponibili)

SPSS

Le finestre di dialogo sono composte da alcuni elementi fondamentali

Lista delle variabili origine

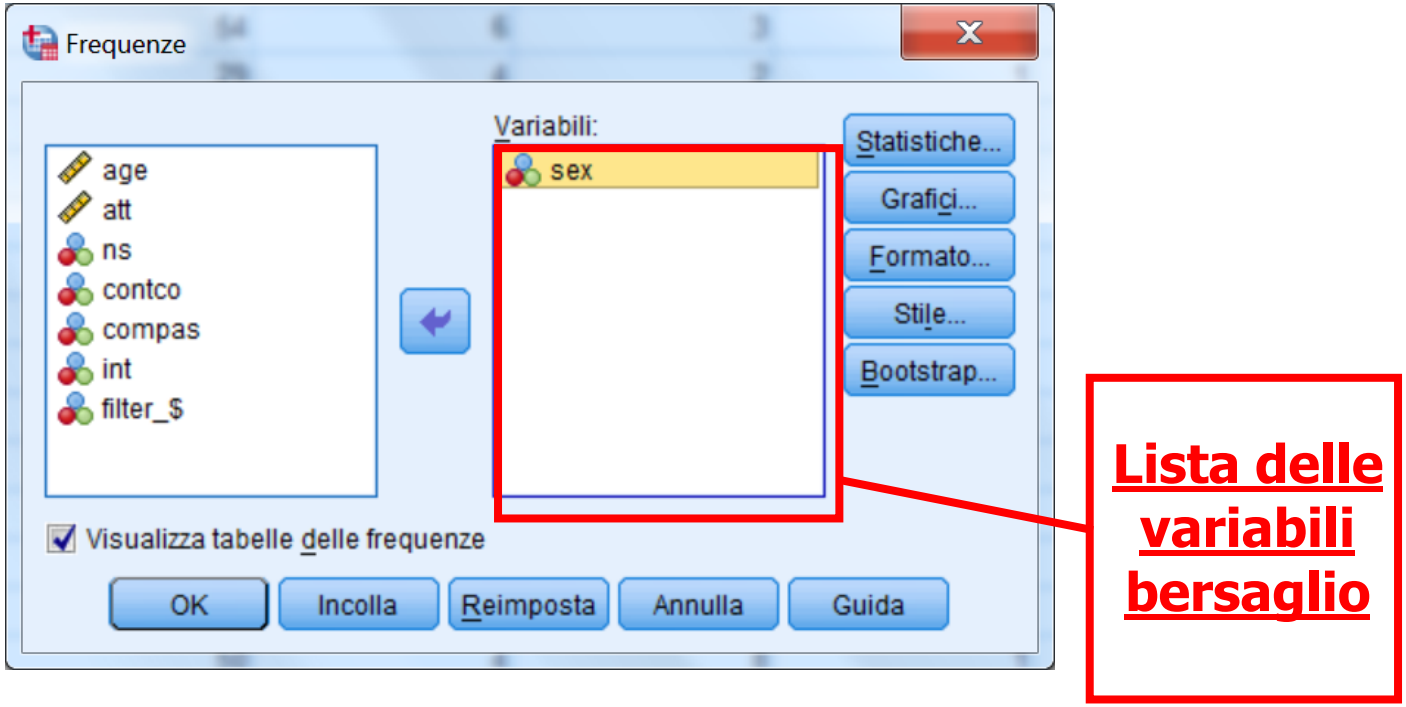


- Lista delle variabili origine: è la lista delle variabili contenute nel data file attivo al momento. Non tutte le variabili del file possono comparire in questa lista, ma solo i tipi di variabili consentite dalla procedura selezionata. Ad esempio. una variabile alfanumerica (o "stringa") può apparire soltanto in alcune procedure elementari.

SPSS

- Lista delle variabili bersaglio (o variabili attive):

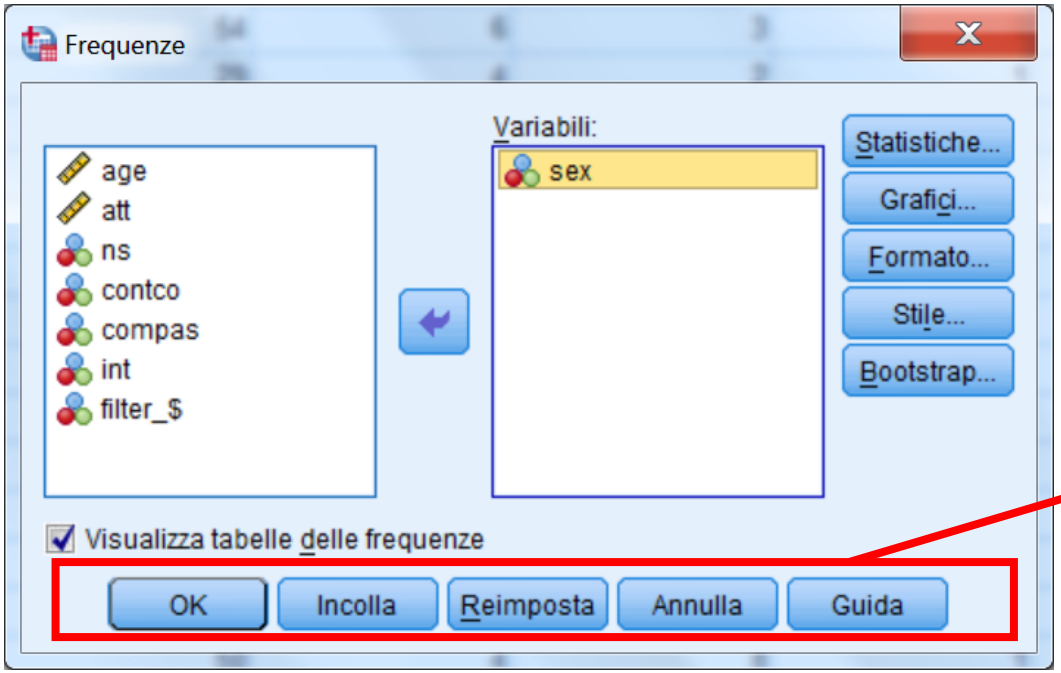
Una o più liste che indicano quali variabili sono state scelte per le analisi. Ad esempio, quali sono le variabili dipendenti e quelle indipendenti



Il comando calcola la distribuzione di frequenza delle variabili incluse nella lista delle variabili bersaglio (a destra)

SPSS

-Bottoni dei comandi: sono i pulsanti che consentono al programma di realizzare un'azione, ad esempio eseguire una procedura di analisi statistica

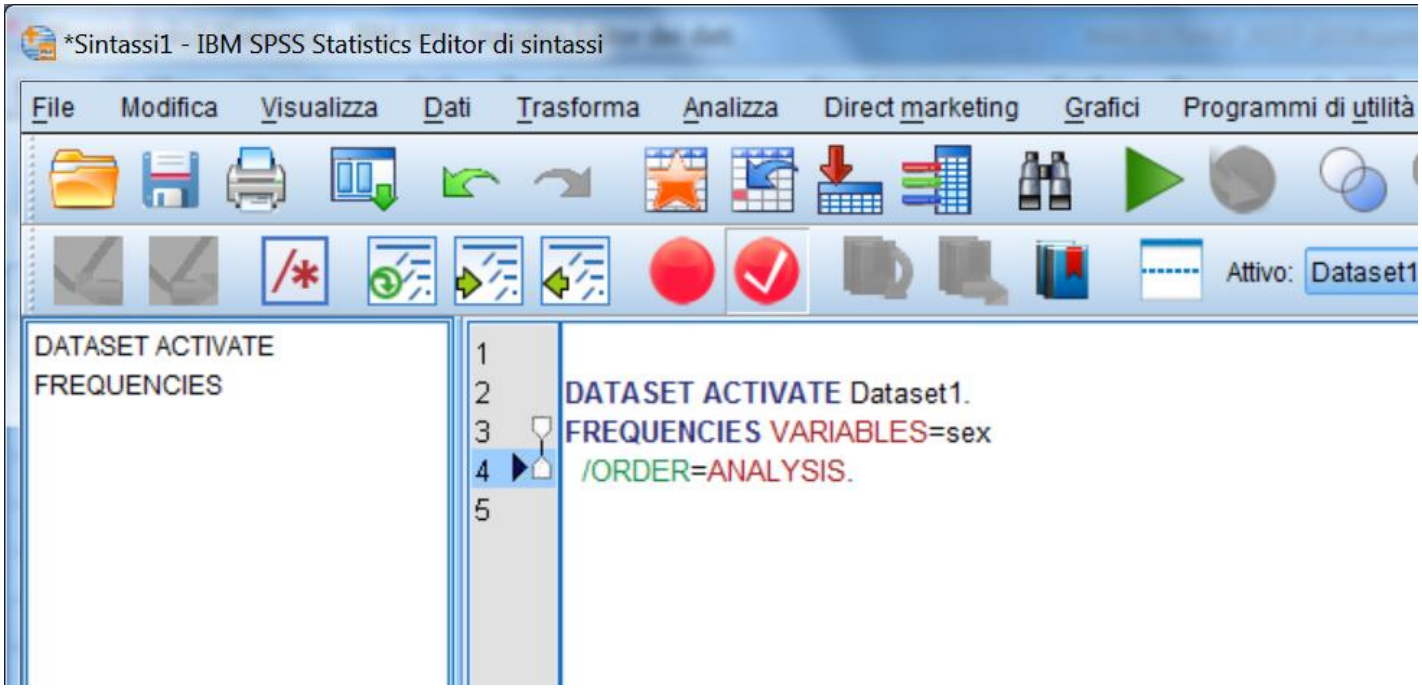


Premendo sul pulsante OK si eseguono le analisi. REIMPOSTA azzera tutte modifica apportate nella finestra di dialogo. CANCELLA chiude la finestra. AIUTO rappresenta una funzione di aiuto on-line relativa alla finestra di dialogo

SPSS

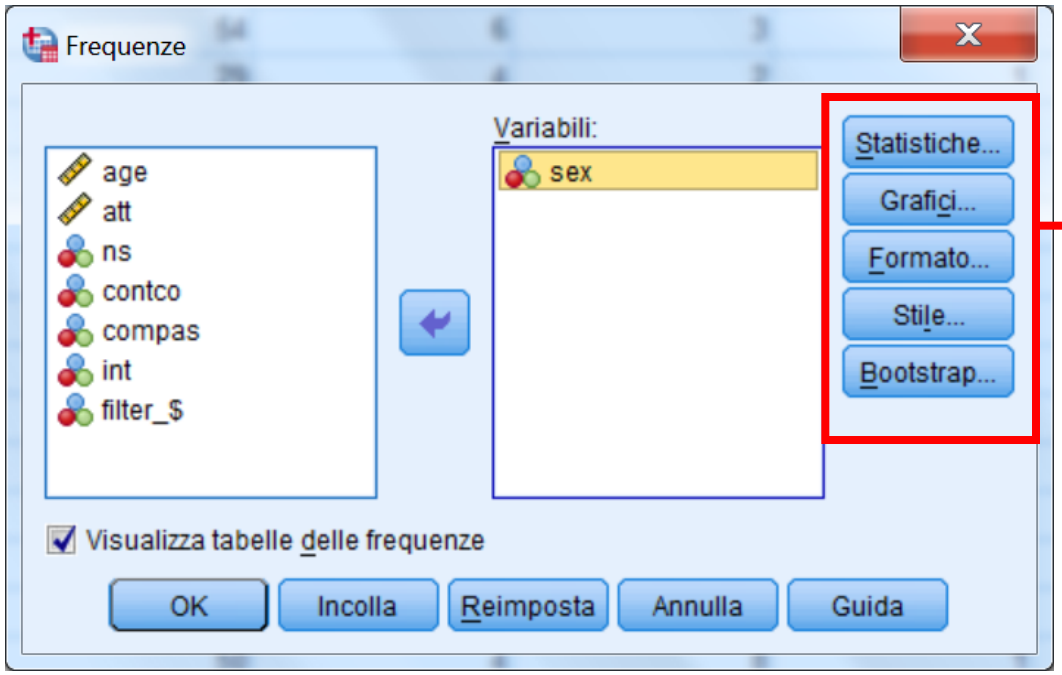
INCOLLA consente di "tradurre" i comandi e le opzioni selezionate nella finestra di dialogo nel linguaggio di programmazione di SPSS.

Le linee di sintassi vengono inserite nella finestra Sintassi attiva al momento (se non c'è nessuna finestra Sintassi aperta, ne viene creata una nuova).



SPSS

I pulsanti posti a destra nella finestra di dialogo consentono di aprire delle ulteriori finestre di dialogo in cui è possibile specificare una serie di opzioni relative alla procedura in corso



Questi pulsanti consentono di aprire ulteriori finestre di dialogo

Questi pulsanti sono diversi per ciascuna finestra di dialogo

SPSS

Le Barre degli strumenti

Ogni finestra ha la propria barra degli strumenti, che fornisce un metodo più rapido, grazie all'utilizzo di un unico pulsante, per accedere ad alcuni dei comandi utilizzati più frequentemente

Posizionandosi con il mouse sulle icone, viene fornita una breve descrizione di ciascun comando

SPSS

Analisi dei dati con SPSS

Analisi monovariate: prendono in esame una sola variabile per volta: indici di tendenza centrale, indici di dispersione (statistiche descrittive -> frequenze e/o descrittive)

Analisi bivariate: prendono in esame l'andamento congiunto di due variabili: correlazione (correlazione -> bivariata), regressione (regressione -> lineare), analisi della varianza (modello lineare generalizzato -> univariata)

Analisi multivariate: prendono in esame simultaneamente più di due variabili: analisi fattoriale (riduzione dimensione -> fattoriale)

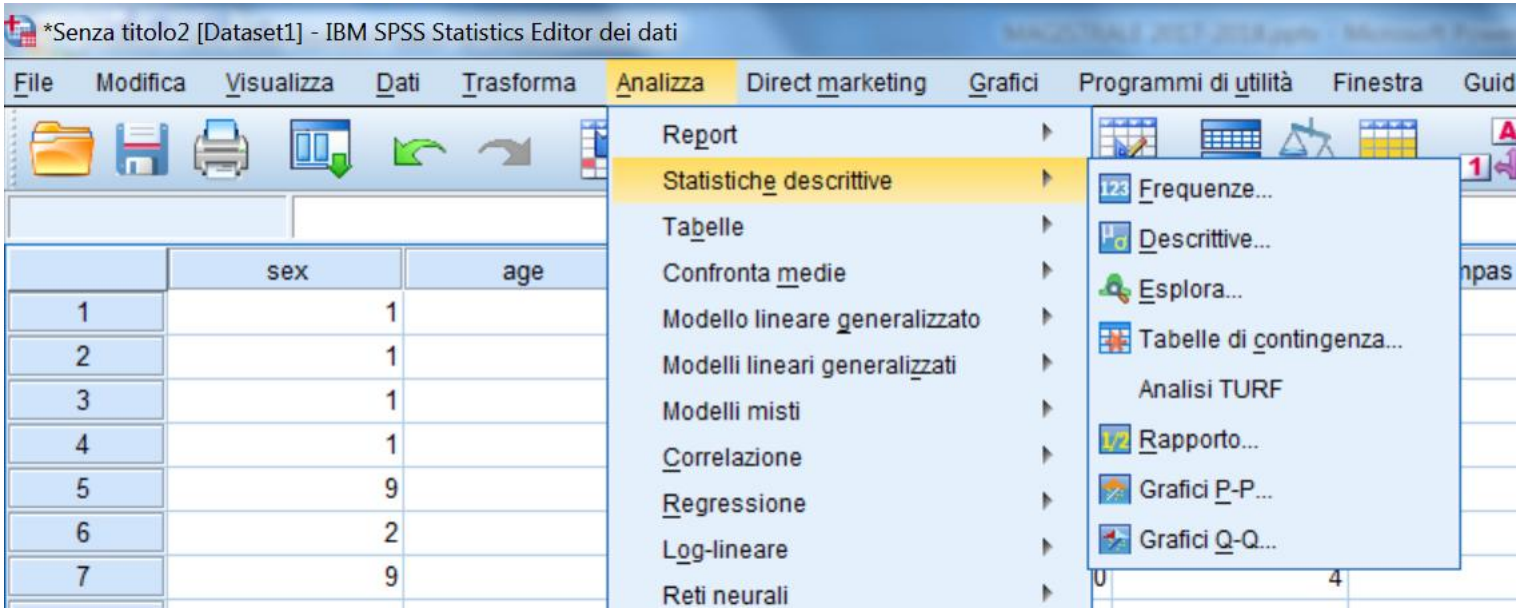
Esplorazione dei dati: data screening

SPSS consente di calcolare una serie di statistiche che riassumono l'informazione nei dati.
L'esplorazione iniziale dei dati è necessaria per esaminare se:
- ci sono errori nei dati, e quindi le variabili assumono valori fuori scala (ad esempio, un item che varia da 1 a 5 ha un punteggio di 8)
- ci sono "valori anomali" (outliers) ovvero soggetti che presentano valori estremamente elevati in una variabile
- ci sono casi con valori mancanti

L'esplorazione iniziale dei dati è necessaria anche per studiare le caratteristiche distributive delle variabili.

Esplorazione dei dati: data screening

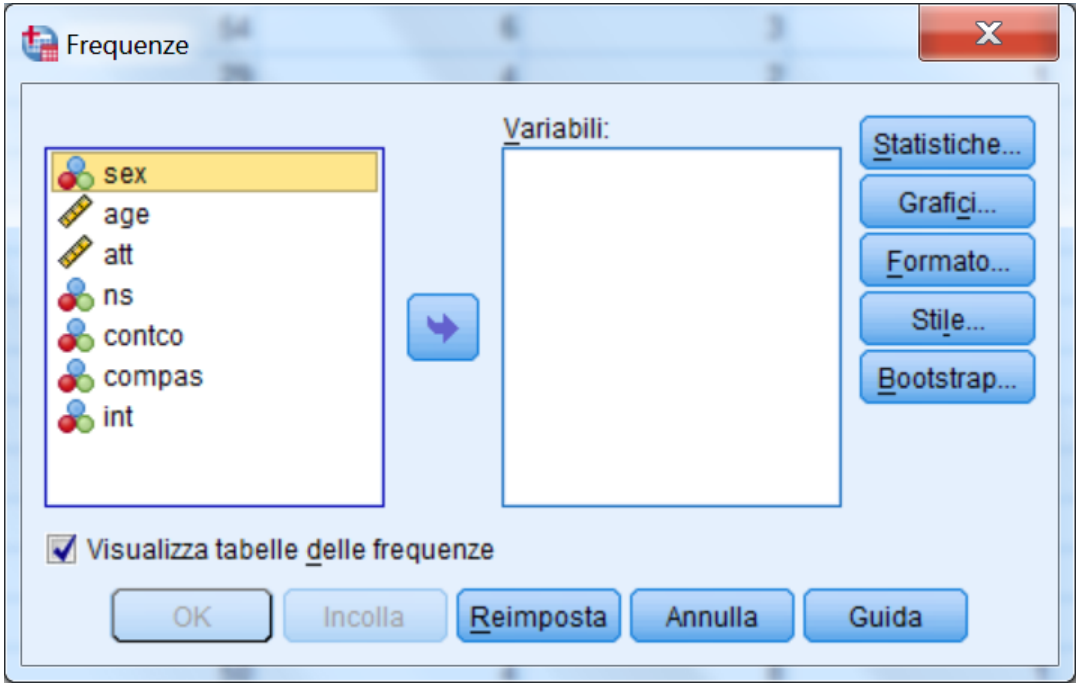
E' possibile esplorare i dati richiedendo:
- distribuzioni di frequenza (procedura Frequenze)
- statistiche descrittive come media, deviazione standard, curtosi, asimmetria (procedura Descrittive)
- tabelle di esplorazione (procedura Esplora)
- tabelle di contingenza
- rappresentazioni grafiche



SPSS

La procedura Frequenze

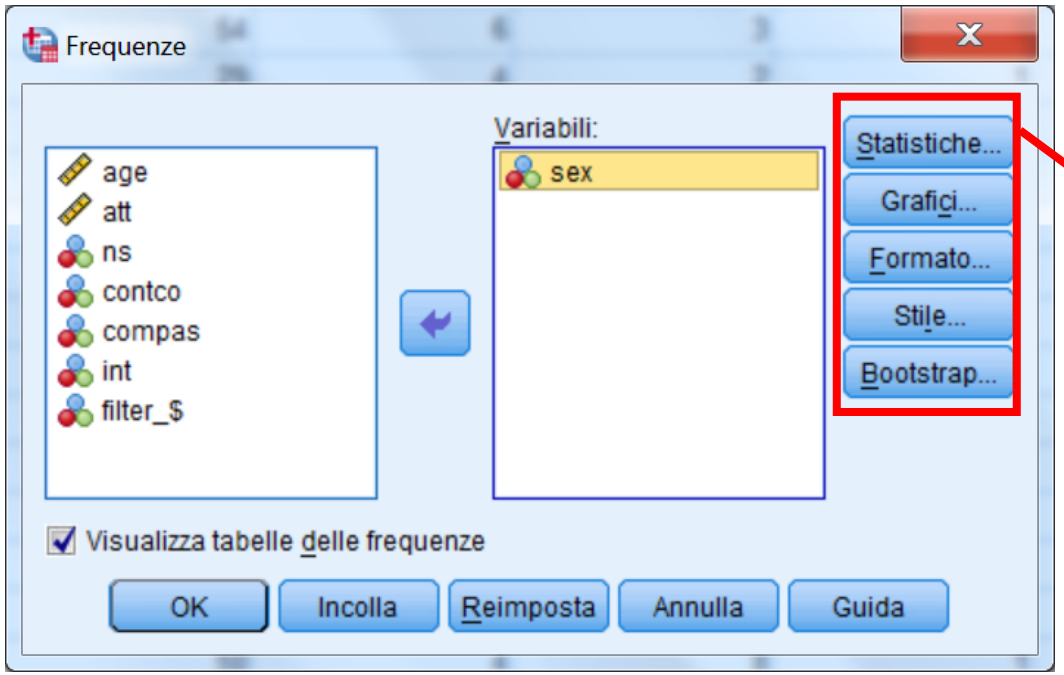
La procedura "Frequenze" consente di effettuare una serie di analisi preliminari, tramite statistiche descrittive e grafici. Selezionando la procedura frequenze si aprirà questa finestra di dialogo:



SPSS

La procedura Frequenze

Una volta selezionate la variabili di interesse (es. "Estroversione"), possiamo chiedere diversi tipi di statistiche (tramite il pulsante STATISTICHE) e di grafici (tramite il pulsante GRAFICI).

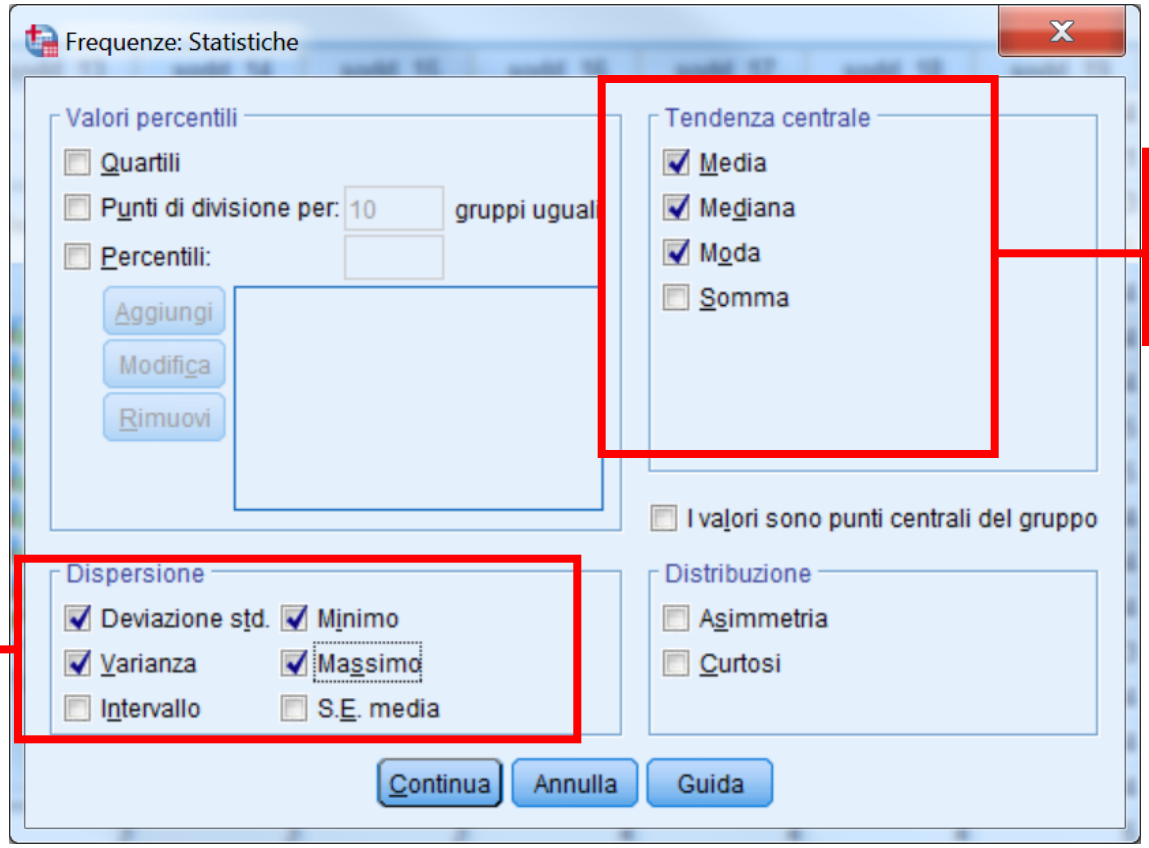


Il pulsante "Statistiche" consente di richiedere una serie di statistiche descrittive

Il pulsante FORMATO (sulla destra) consente di specificare il formato in cui i dati sono presentati nelle tabelle

SPSS

Cliccando sul pulsante "Statistiche" si aprirà questa finestra:



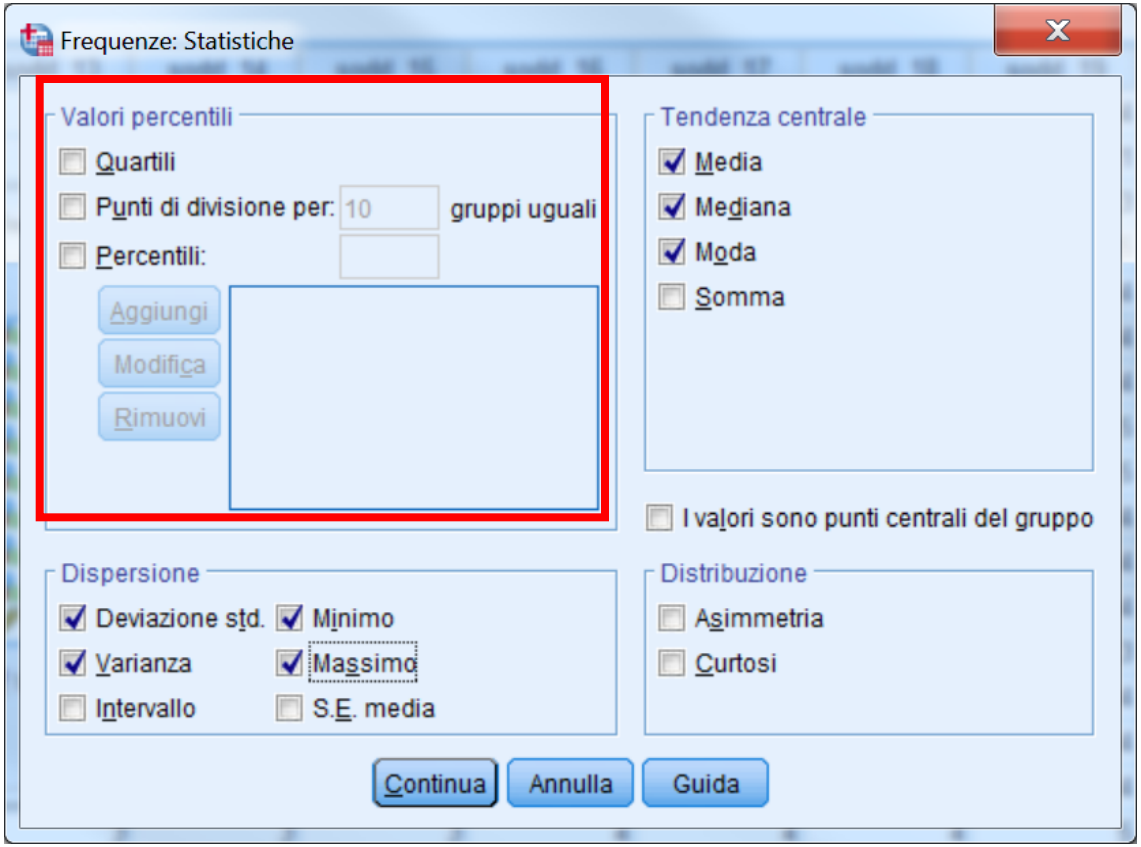
Indici di tendenza centrale

Indici di dispersione

L'opzione "Statistiche" consente calcolare una serie di statistiche, come ad gli indici di tendenza centrale e gli indici di dispersione

SPSS

È possibile inoltre calcolare i quartili e percentili.

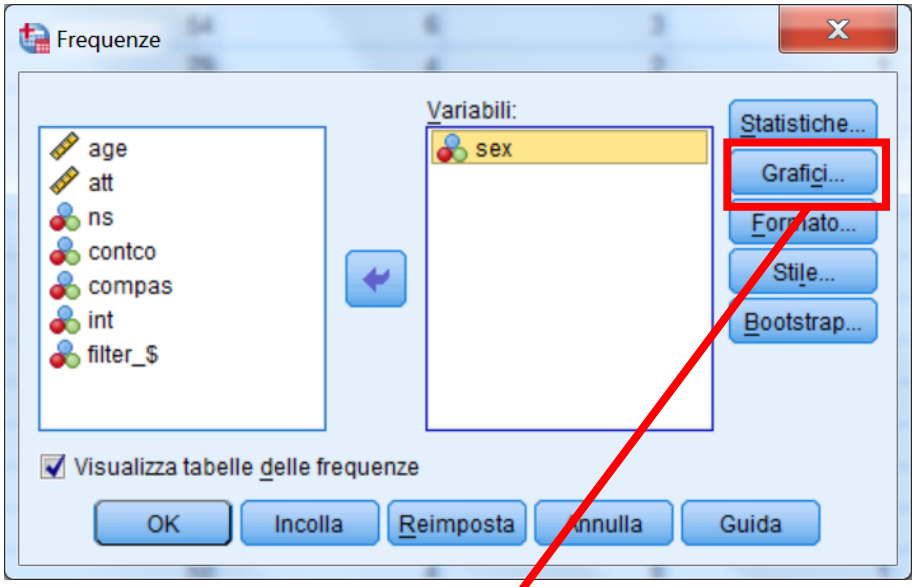


RICORDA: i Quartili indicano quei valori che dividono la distribuzione in quattro parti uguali.

I Percentili indicano quei valori che dividono la distribuzione in 100 parti uguali.

SPSS

Selezionando il pulsante "Grafici" si aprirà la finestra riportata in basso, che consente di specificare il tipo di grafico che vogliamo utilizzare per rappresentare la distribuzione di frequenze.

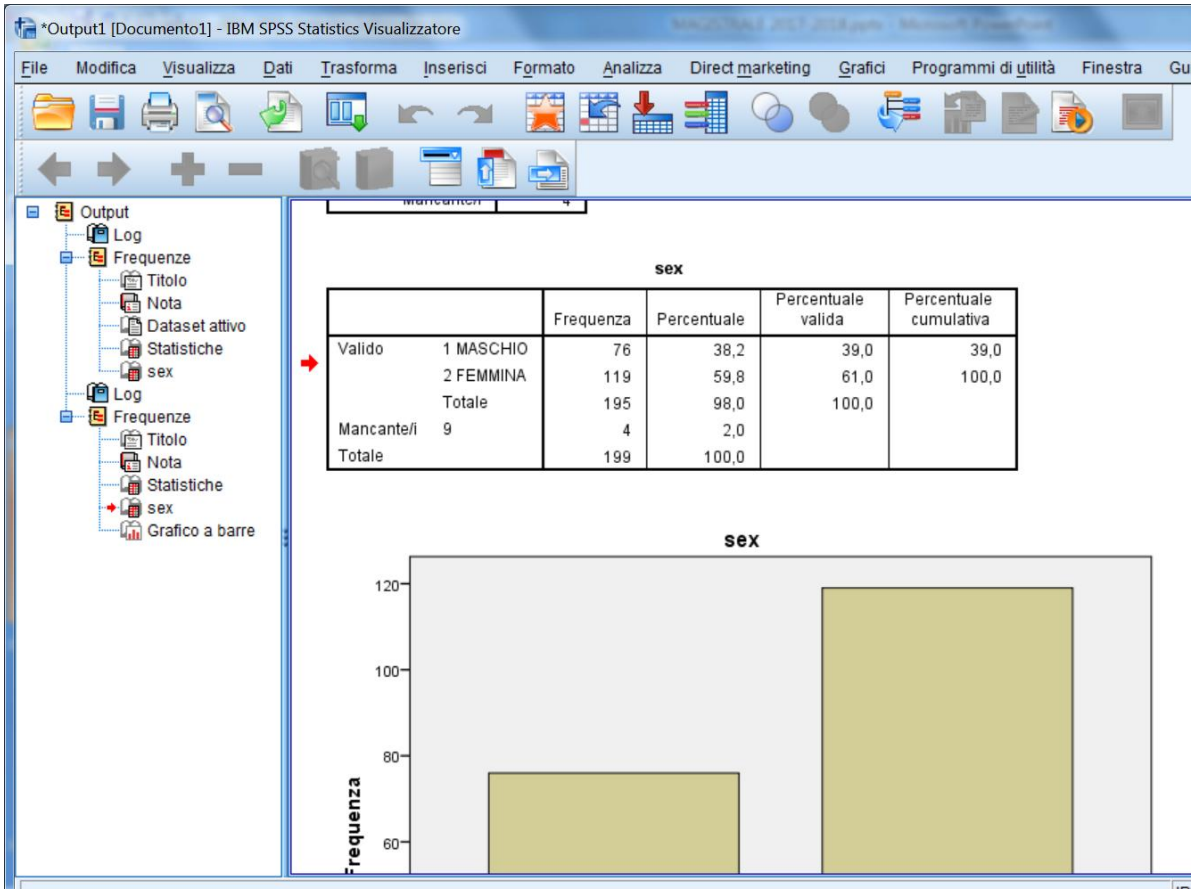


Questa finestra consente di creare diversi tipi di grafici: grafici a barre, grafici a torta e Istogrammi

SPSS
L'output della procedura Frequenze

Nell'output vengono riportate una serie di tabelle e un grafico

La tabella seguente contiene i valori delle statistiche descrittive che abbiamo richiesto nella finestra "Statistiche"



SPSS
L'output della procedura Frequenze

Nell'output vengono riportate una serie di tabelle e un grafico

La tabella seguente contiene i valori delle statistiche descrittive che abbiamo richiesto nella finestra "Statistiche"

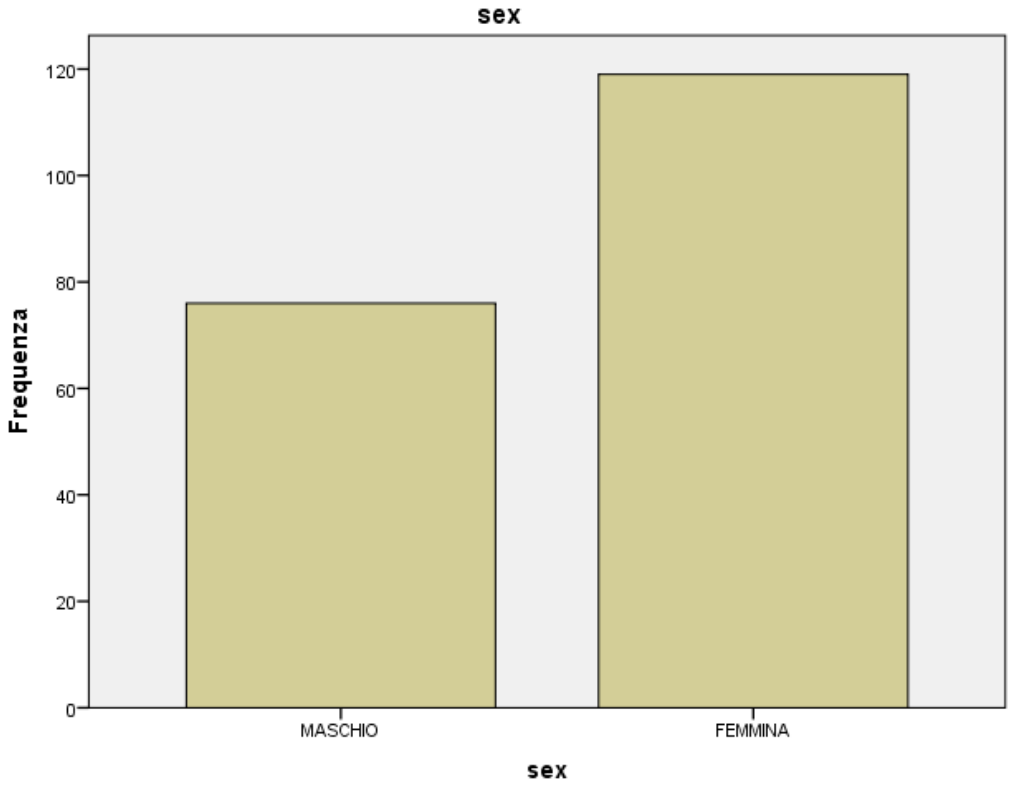
sex

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulativa
Valido	1 MASCHIO	76	38,2	39,0	39,0
	2 FEMMINA	119	59,8	61,0	100,0
	Totale	195	98,0	100,0	
Mancante/i	9	4	2,0		
Totale		199	100,0		

SPSS
L'output della procedura Frequenze

Nell'output vengono riportate una serie di tabelle e un grafico

Il grafico seguente contiene il diagramma a barre delle frequenze



SPSS
L'output della procedura Frequenze

Nell'output vengono riportate una serie di tabelle e un grafico

La tabella seguente contiene i valori delle statistiche descrittive che abbiamo richiesto nella finestra "Statistiche"

Statistiche

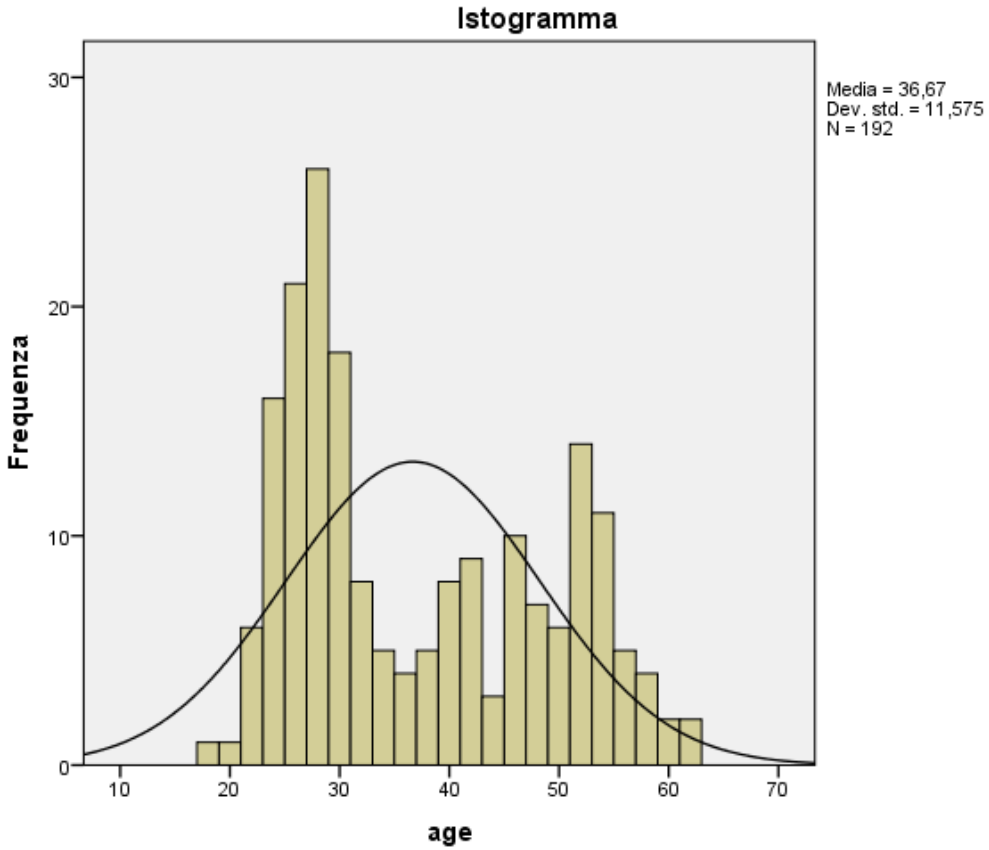
age

N	Valido	192
	Mancante/i	7
Media		36,67
Mediana		32,00
Modalità		28
Deviazione std.		11,575
Varianza		133,983
Asimmetria		,425
Errore standard dell'asimmetria		,175
Curtosi		-1,219
Errore standard della curtosi		,349
Minimo		18
Massimo		62

SPSS

L'output della procedura Frequenze

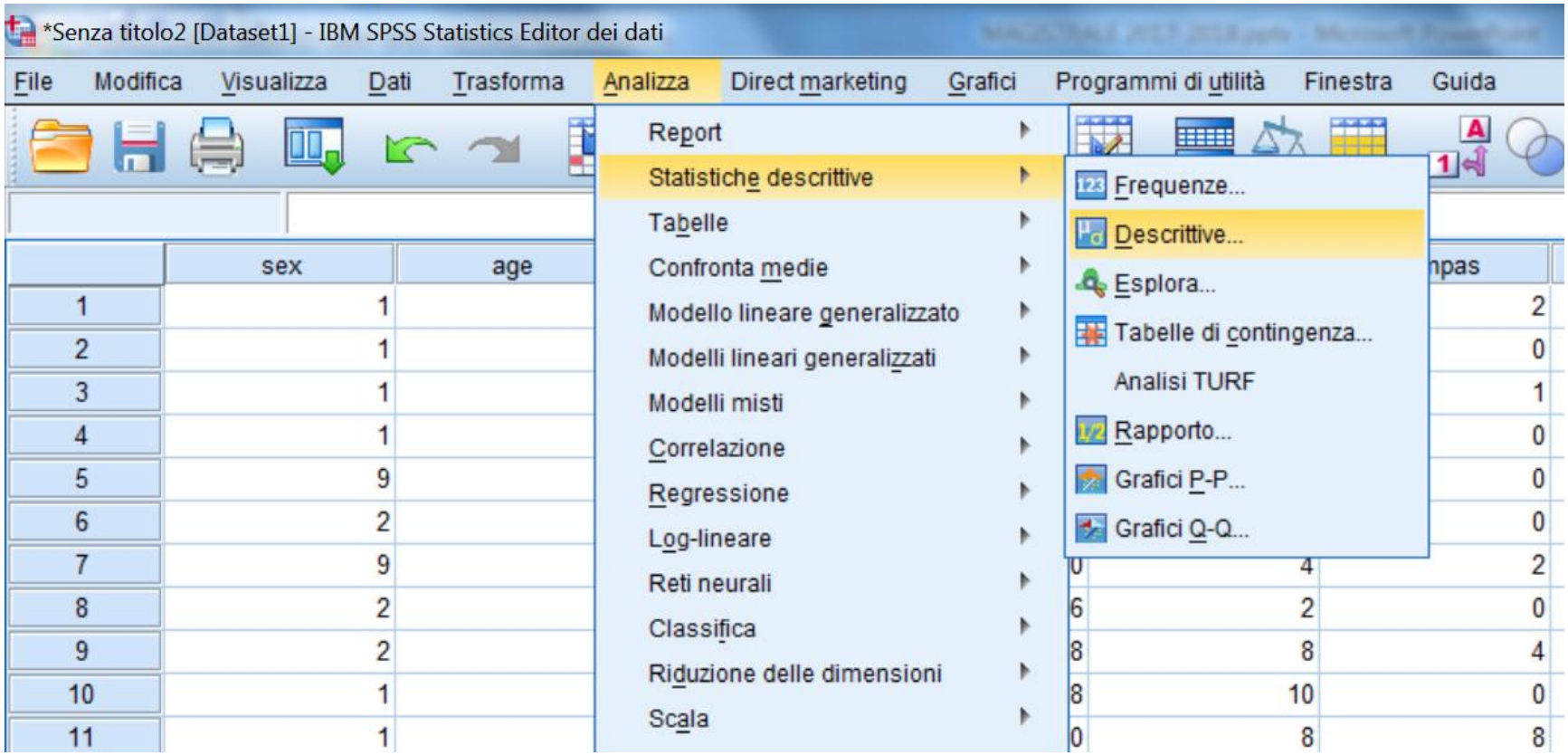
In questa figura viene riportato l'istogramma della variabile



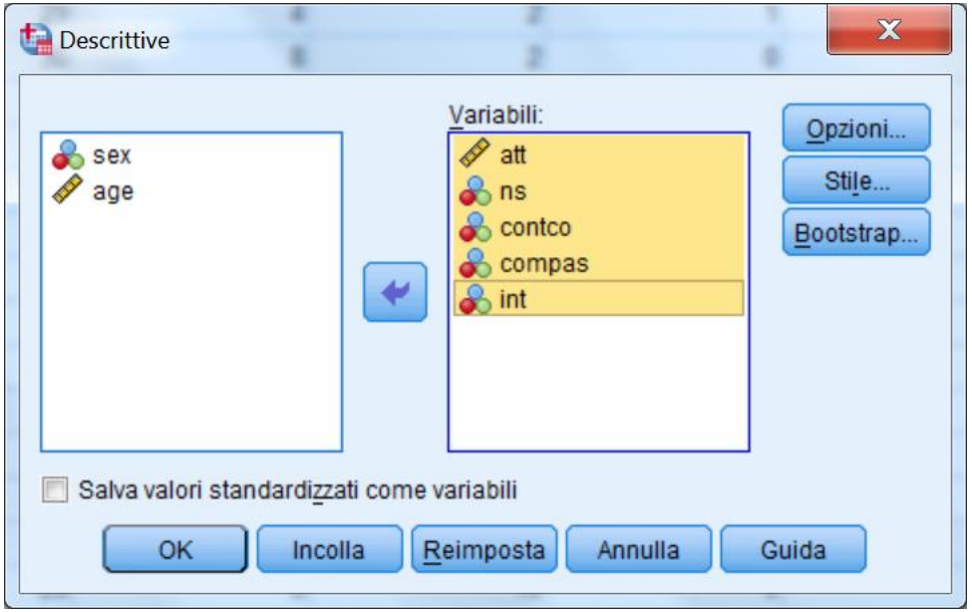
Una curva normale sovrapposta all'istogramma consente di valutare se i dati sono distribuiti normalmente

SPSS
La procedura Descrittive

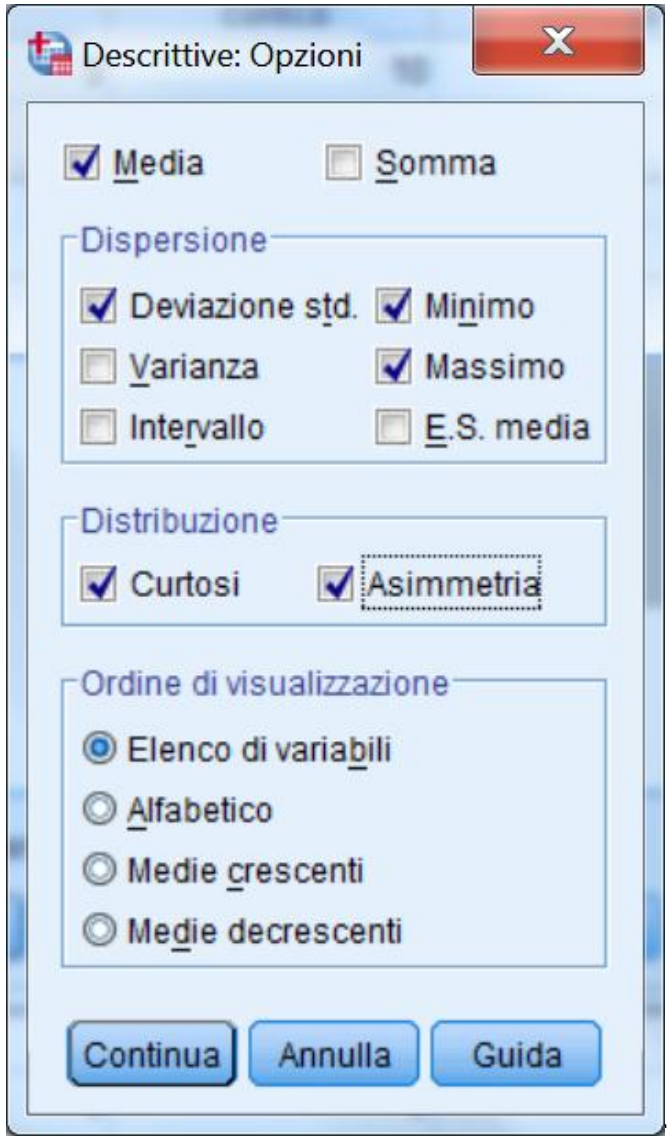
La procedura "Descrittive" consente di calcolare alcune statistiche univariate. Questa procedura è consigliabile per lo screening di file che contengono molte variabili



Chiedere le statistiche descrittive per vedere se le distribuzioni delle variabili sono normali

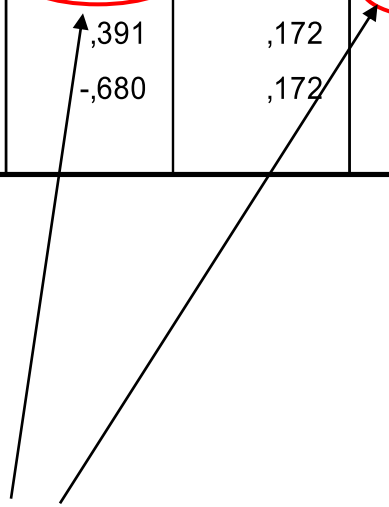


Selezionando il pulsante "Opzioni" si apre questa finestra di dialogo:



Statistiche descrittive

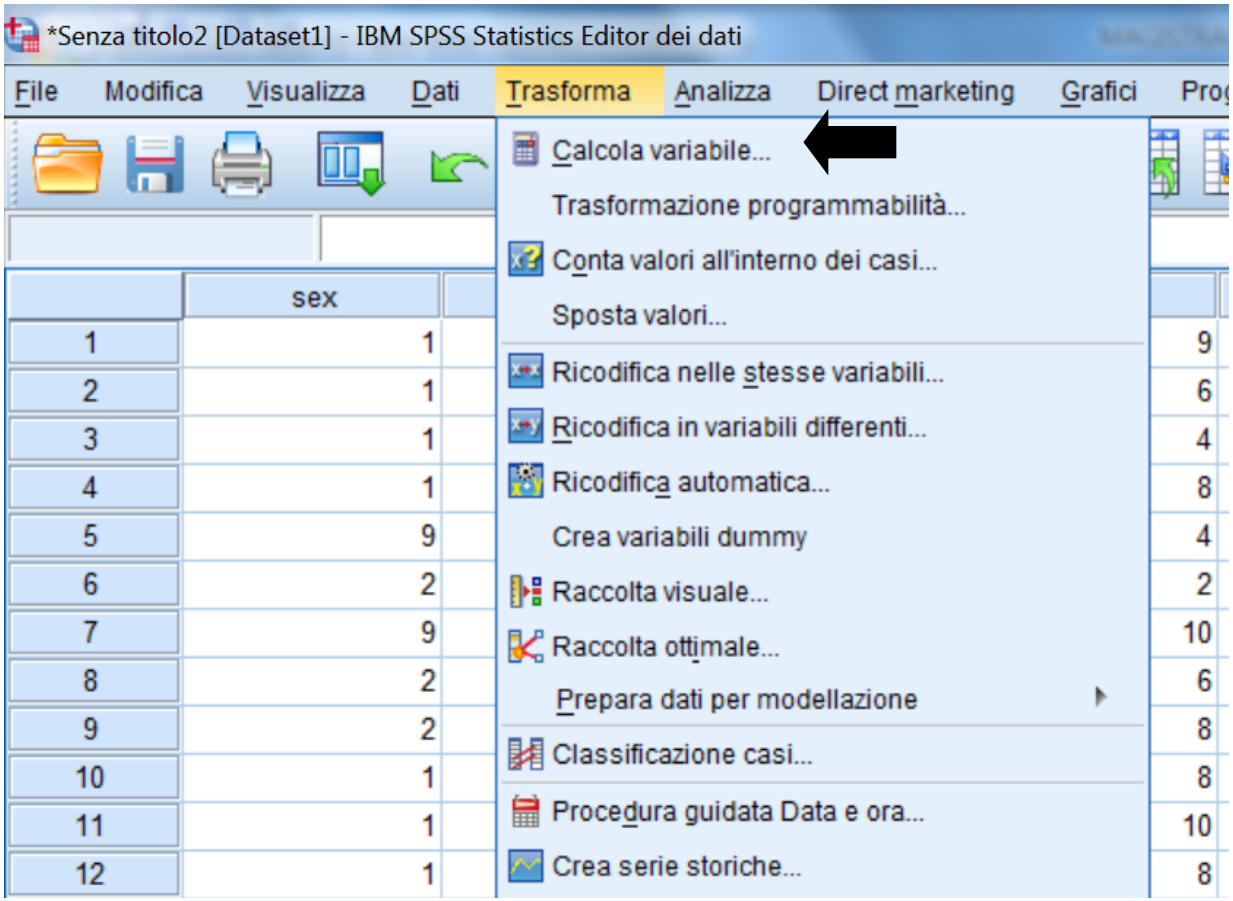
	N	Minimo	Massimo	Media	Deviazione std.	Asimmetria		Curtosi	
	Statistica	Statistica	Statistica	Statistica	Statistica	Statistica	Errore std	Statistica	Errore std
att	199	16	54	42,80	7,311	-,869	,172	,209	,343
ns	199	2	10	7,88	1,801	-,554	,172	-,339	,343
contco	199	2	10	8,68	1,863	-1,850	,172	3,444	,343
compas	199	0	8	2,64	1,969	,391	,172	-,446	,343
int	199	2	10	7,29	2,544	-,680	,172	-,630	,343
Validi (listwise)	199								



Questa variabile ha una distribuzione fortemente non normale !

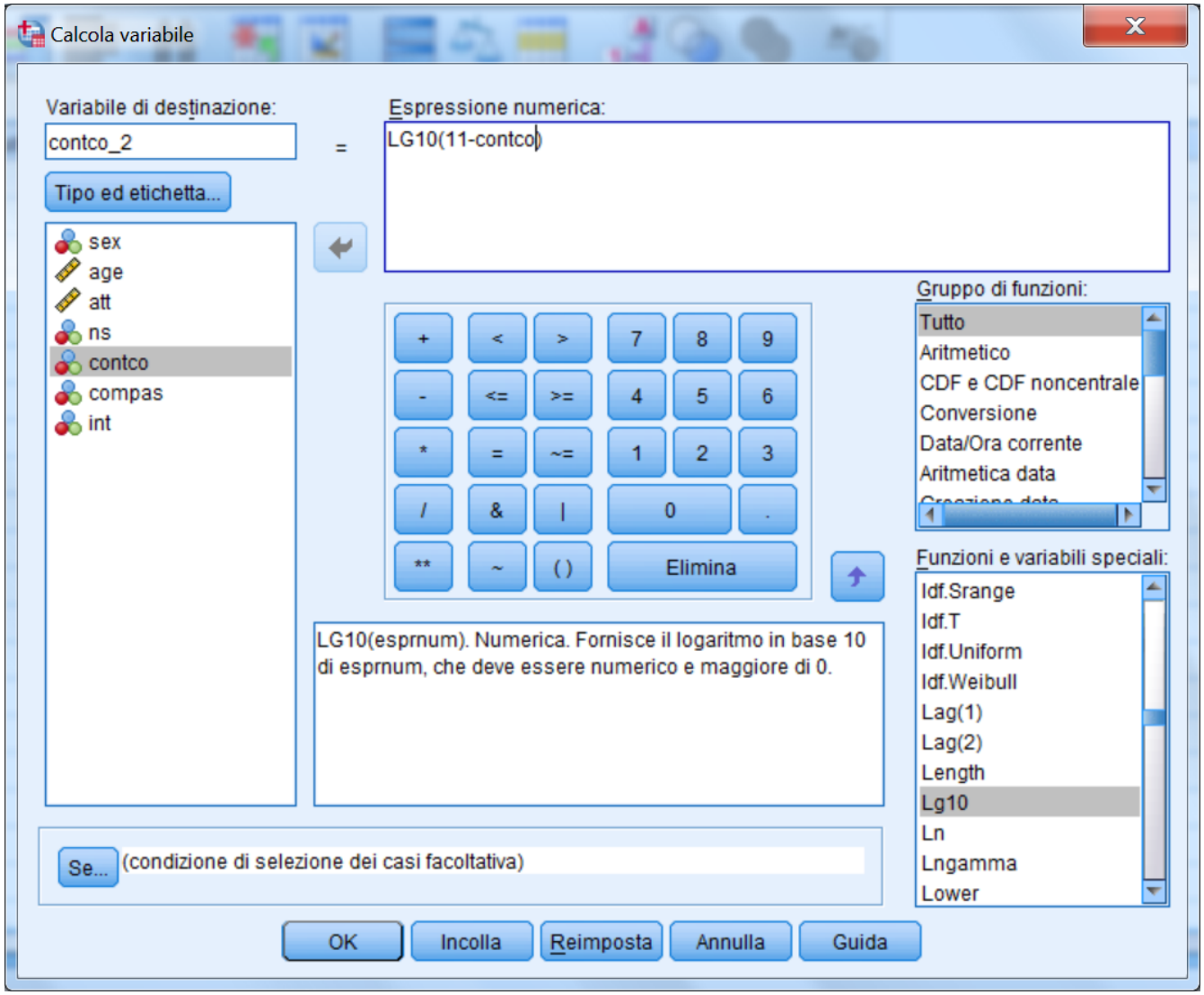
Trasformazione della variabile "contco"

Asimmetria Negativa Sostanziale (valori tra -1 e -2)	Logaritmo = X* =Log10(K-X) (qui, K = 10+1=11)



Ricodificare la variabile "contco"

Asimmetria Negativa Sostanziale (valori tra -1 e -2)	Logaritmo = X* =Log10(K-X) (K = max(X)+1=10+1=11)

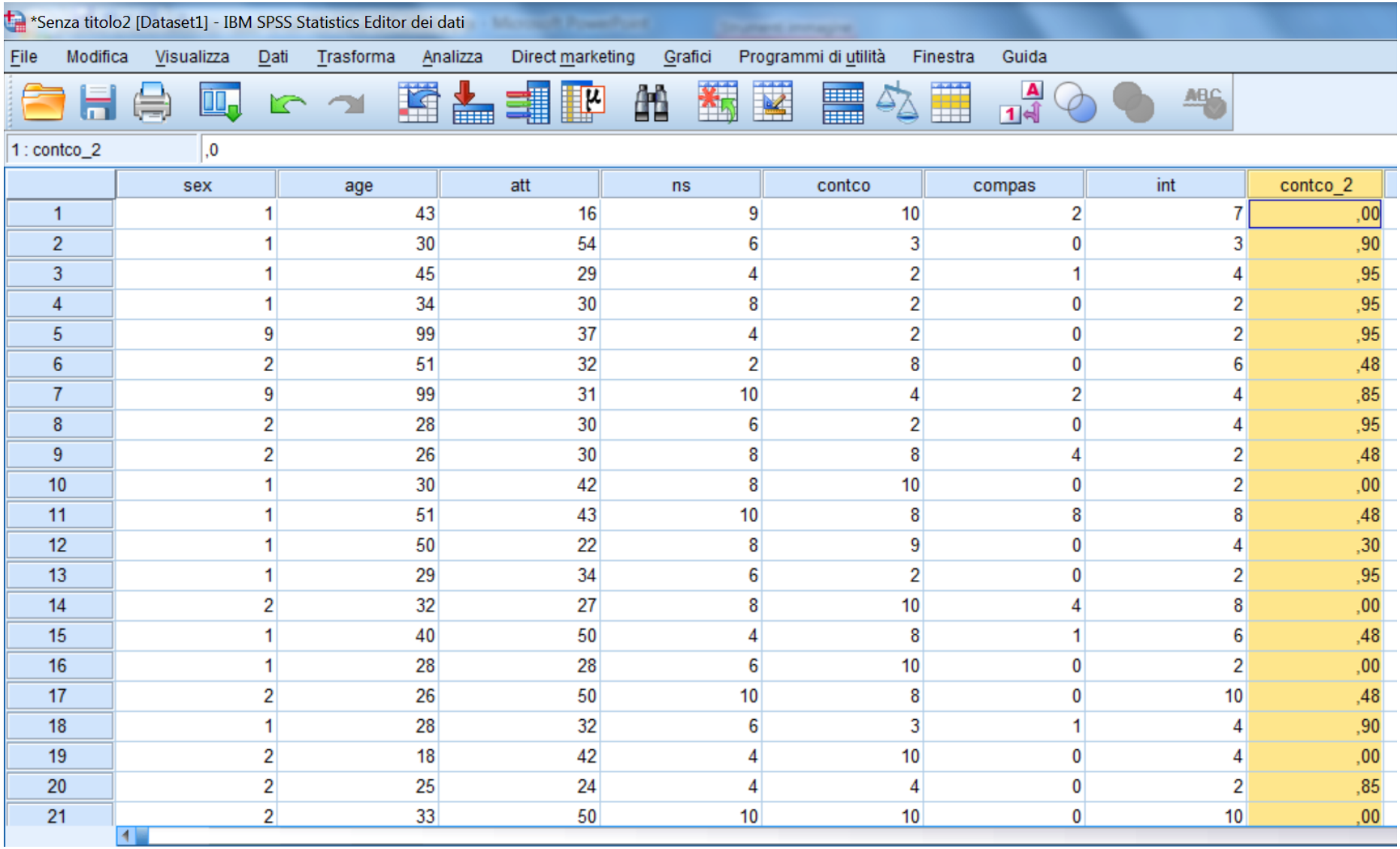


Ricodificare la variabile "contco"

Asimmetria Negativa Sostanziale (valori tra -1 e -2)	**Logaritmo = X* =Log10(K-X) (qui, K = 10+1=11)**

Comando di Sintassi:

```
COMPUTE contco_2=LG10(11-contco).
```

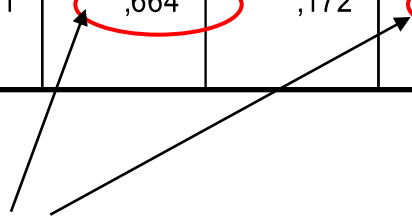


	sex	age	att	ns	contco	compas	int	contco_2
1	1	43	16	9	10	2	7	,00
2	1	30	54	6	3	0	3	,90
3	1	45	29	4	2	1	4	,95
4	1	34	30	8	2	0	2	,95
5	9	99	37	4	2	0	2	,95
6	2	51	32	2	8	0	6	,48
7	9	99	31	10	4	2	4	,85
8	2	28	30	6	2	0	4	,95
9	2	26	30	8	8	4	2	,48
10	1	30	42	8	10	0	2	,00
11	1	51	43	10	8	8	8	,48
12	1	50	22	8	9	0	4	,30
13	1	29	34	6	2	0	2	,95
14	2	32	27	8	10	4	8	,00
15	1	40	50	4	8	1	6	,48
16	1	28	28	6	10	0	2	,00
17	2	26	50	10	8	0	10	,48
18	1	28	32	6	3	1	4	,90
19	2	18	42	4	10	0	4	,00
20	2	25	24	4	4	0	2	,85
21	2	33	50	10	10	0	10	,00

Controlliamo se la normalizzazione è avvenuta chiedendo di nuovo le descrittive

Statistiche descrittive

	N	Minimo	Massimo	Media	Deviazione std.	Asimmetria		Curtosi	
	Statistica	Statistica	Statistica	Statistica	Statistica	Statistica	Errore std	Statistica	Errore std
att	199	16	54	42,80	7,311	-,869	,172	,209	,343
ns	199	2	10	7,88	1,801	-,554	,172	-,339	,343
contco	199	2	10	8,68	1,863	-1,850	,172	3,444	,343
compas	199	0	8	2,64	1,969	,391	,172	-,446	,343
int	199	2	10	7,29	2,544	-,680	,172	-,630	,343
contco_2	199	,00	,95	,2576	,29111	,664	,172	-,767	,343
Validi (listwise)	199								



Ora i valori sono accettabili !

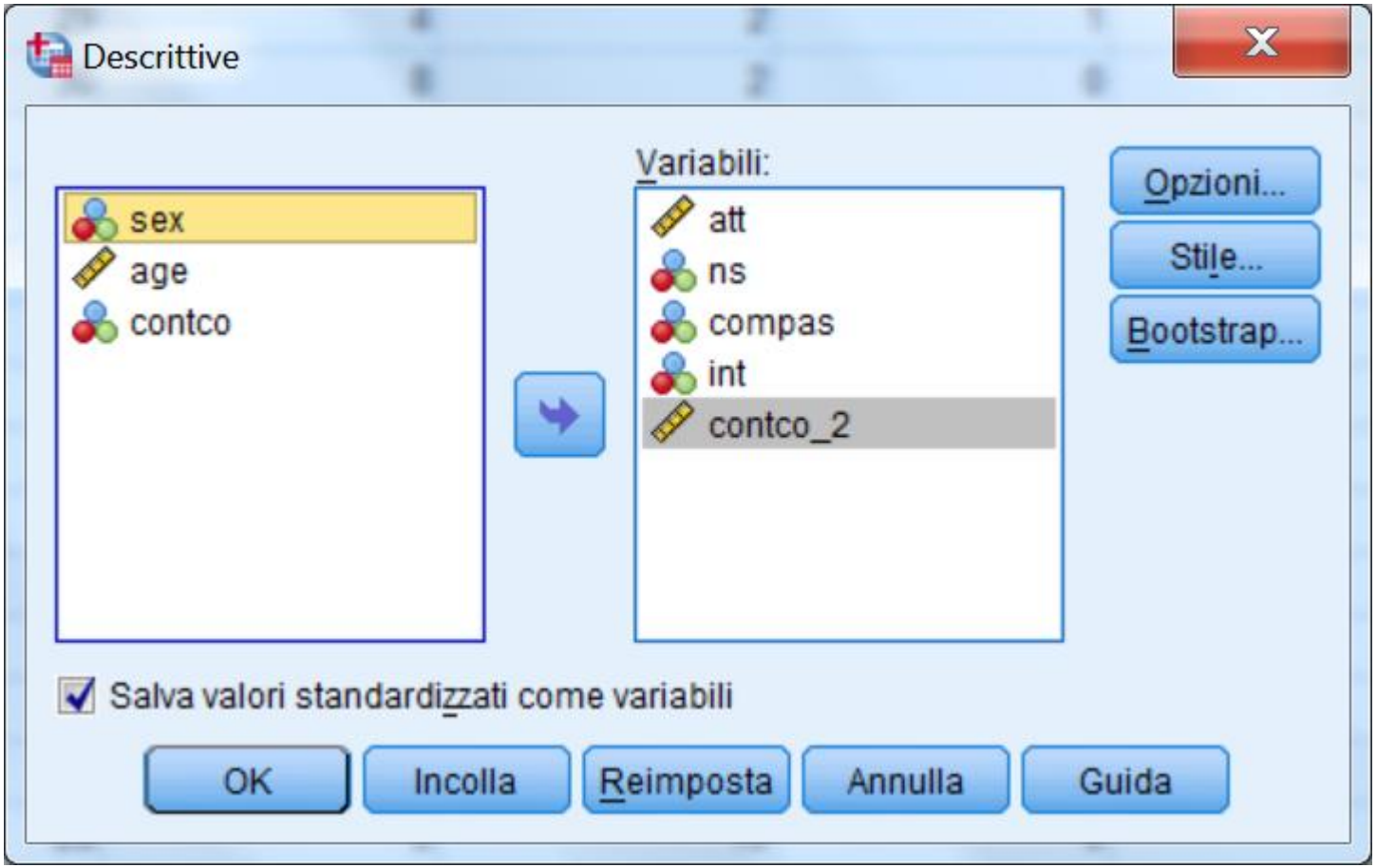
Esplorazione dei dati: gli outliers (valori anomali)

I valori anomali sono quei valori che risultano differenziarsi particolarmente nella distribuzione dei punteggi.

I valori anomali, o outliers, univariati sono quei casi che in una variabile presentano valori estremamente elevati o estremamente bassi rispetto al resto della distribuzione.
Per individuare gli outliers univariati è possibile standardizzare i punteggi relativi alla variabile in esame e chiedere una distribuzione delle frequenze.
Vengono considerati come possibili valori anomali quei punteggi che corrispondono a una z maggiore di 3 in valore assoluto.
E' necessario considerare la distribuzione nella sua interezza e vedere se i punteggi troppo alti o troppo bassi rappresentano casi isolati dal resto della distribuzione oppure no.

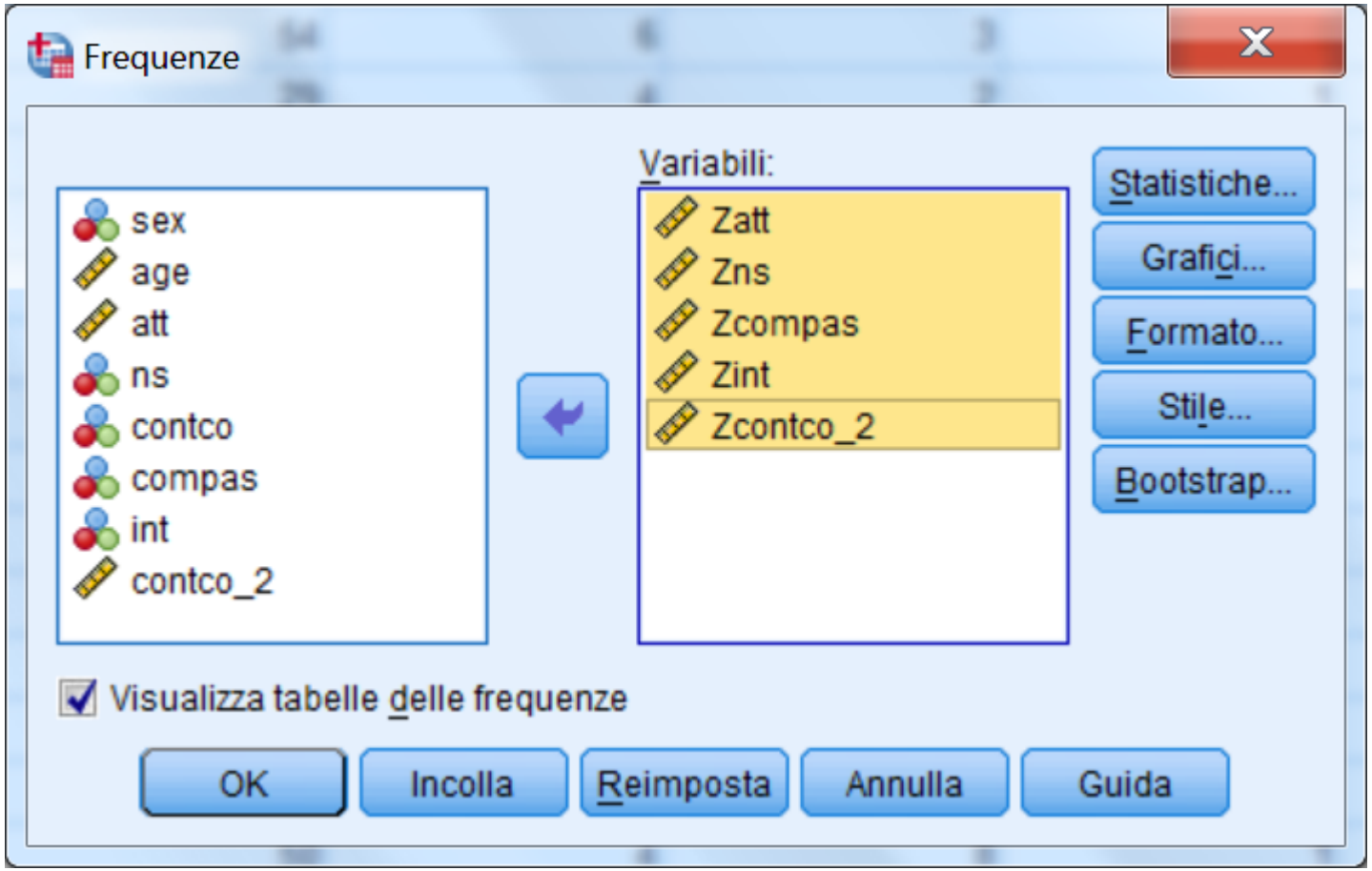
Esplorazione dei dati: gli outliers (valori anomali)

Standardizziamo (z) le variabili



Esplorazione dei dati: gli outliers (valori anomali)

Chiediamo le frequenze delle nuove z



Esplorazione dei dati: gli outliers (valori anomali)

Punteg(att)

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	-3,66545	1	,5	,5	,5
	-2,84479	1	,5	,5	1,0
	-2,57124	1	,5	,5	1,5
	-2,29769	2	1,0	1,0	2,5
	-2,16092	1	,5	,5	3,0
	-2,02414	3	1,5	1,5	4,5
	-1,88737	3	1,5	1,5	6,0
	-1,75059	5	2,5	2,5	8,5
	-1,61381	2	1,0	1,0	9,5
	-1,47704	2	1,0	1,0	10,6
	-1,34026	1	,5	,5	11,1
	-1,20349	8	4,0	4,0	15,1
	-1,06671	1	,5	,5	15,6
	-,92994	5	2,5	2,5	18,1
	-,79316	8	4,0	4,0	22,1
	-,65639	10	5,0	5,0	27,1
	-,51961	10	5,0	5,0	32,2
	-,38283	6	3,0	3,0	35,2
	-,24606	8	4,0	4,0	39,2
	-,10928	12	6,0	6,0	45,2
	,02749	7	3,5	3,5	48,7
	,16427	7	3,5	3,5	52,3
	,30104	2	1,0	1,0	53,3
	,43782	6	3,0	3,0	56,3
	,57459	12	6,0	6,0	62,3
	,71137	7	3,5	3,5	65,8
	,84815	8	4,0	4,0	69,8
	,98492	59	29,6	29,6	99,5
	1,53202	1	,5	,5	100,0
	Totale	199	100,0	100,0	

Punteg(ns)

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	-3,26703	1	,5	,5	,5
	-2,15663	9	4,5	4,5	5,0
	-1,60143	10	5,0	5,0	10,1
	-1,04623	30	15,1	15,1	25,1
	-,49103	17	8,5	8,5	33,7
	,06417	62	31,2	31,2	64,8
	,61937	14	7,0	7,0	71,9
	1,17457	56	28,1	28,1	100,0
	Totale	199	100,0	100,0	

Chi sono questi due soggetti ?

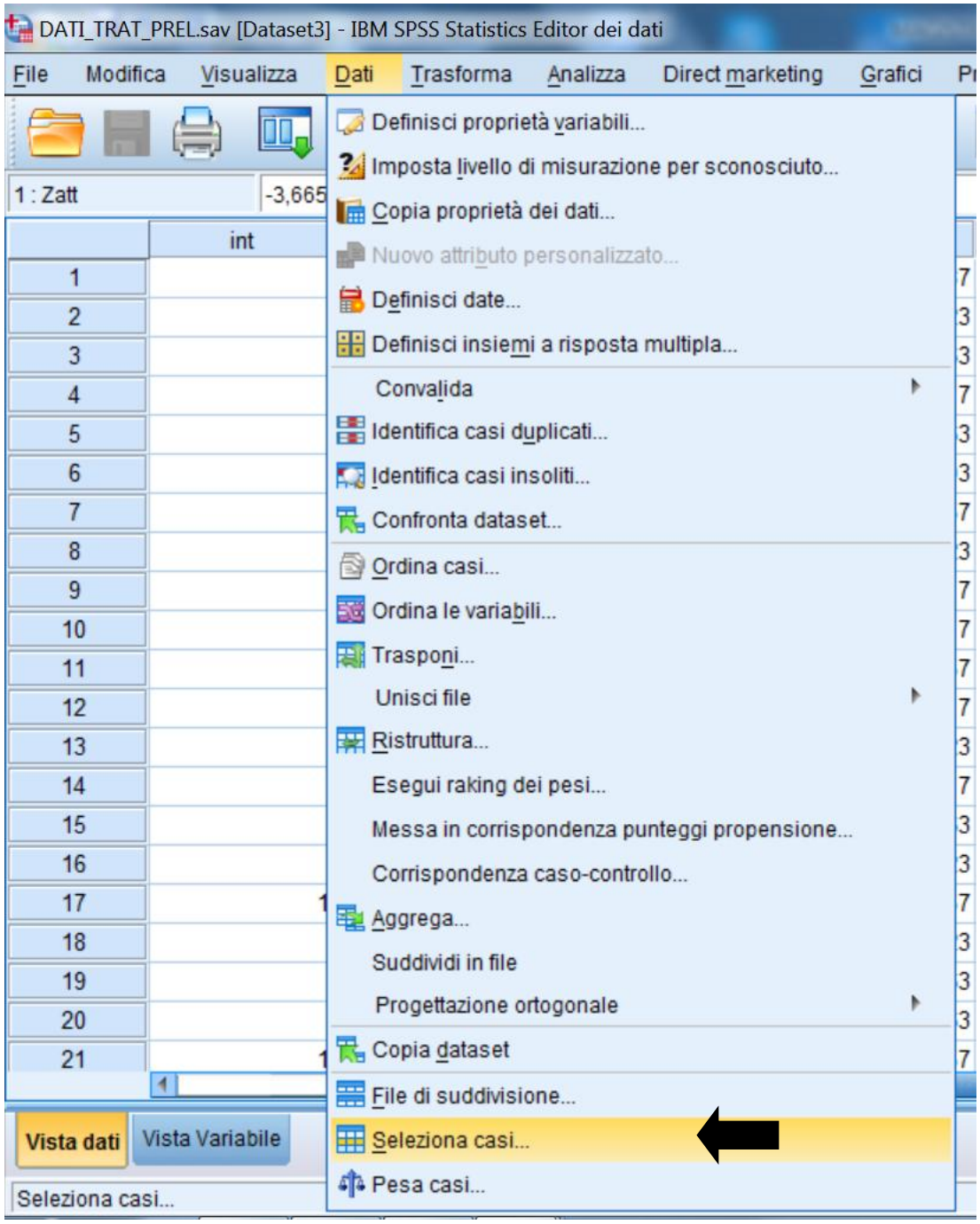
Esplorazione dei dati: gli outliers (valori anomali)

DATI_TRAT_PREL.sav [Dataset3] - IBM SPSS Statistics Editor dei dati

File Modifica Visualizza Dati Trasforma Analizza Direct marketing Grafici Programmi di utilità Finestra Guida

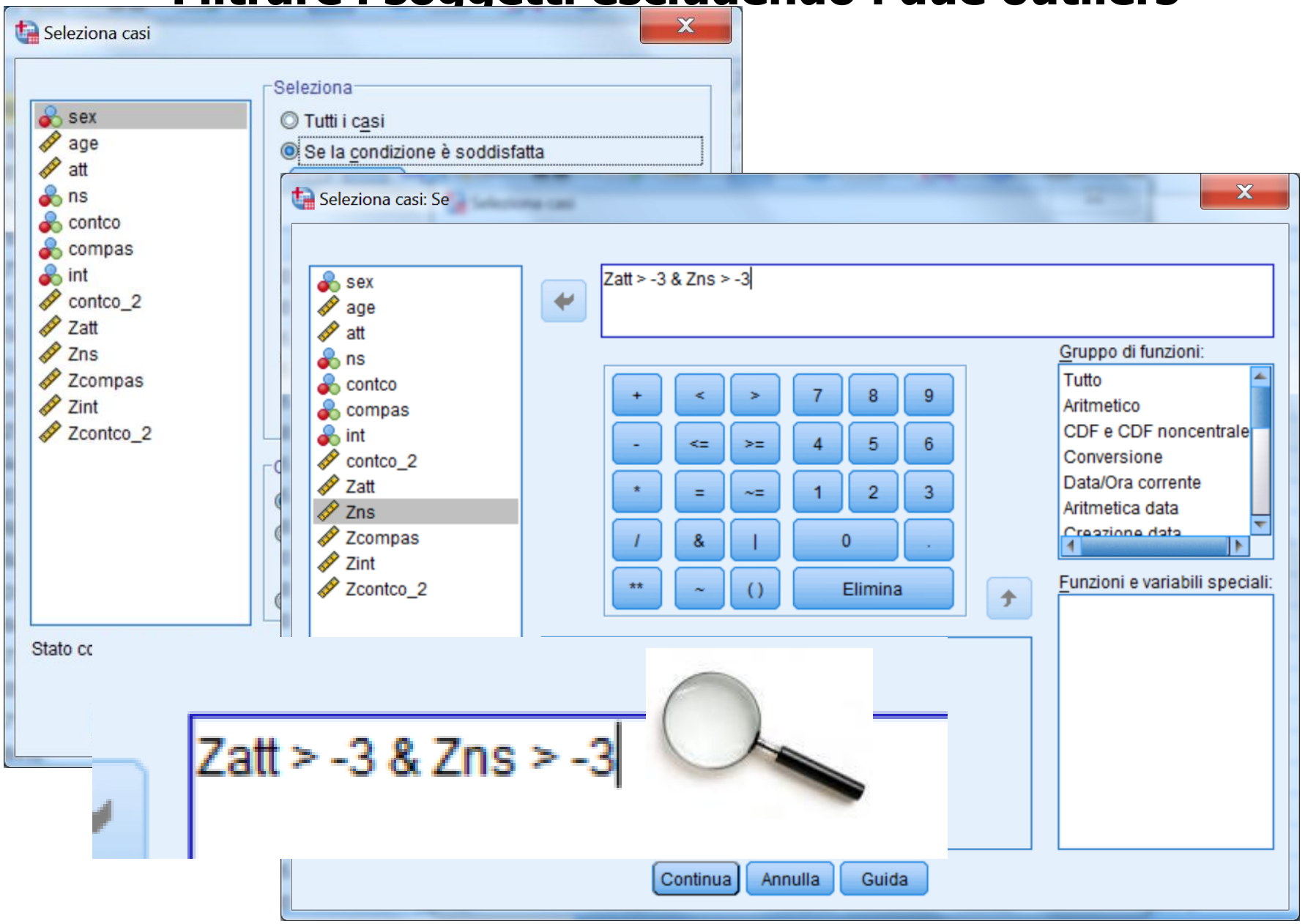
1 : Zatt -3,66544756828786

	int	contco_2	Zatt	Zns	Zcompas	Zint	Zcontco_2
1	7	,00	-3,66545	,61937	-,32667	-,11458	-,88476
2	3	,90	-1,53202	-1,04623	-1,34239	-1,68712	2,21750
3	4	,95	-1,88737	-2,15663	-,83453	-1,29399	2,39322
4	2	,95	-1,75059	,06417	-1,34239	-2,08026	2,39322
5	2	,95	-,79316	-2,15663	-1,34239	-2,08026	2,39322
6	6	,48	-1,47704	-3,26703	-1,34239	-,50772	,75423
7	4	,85	-1,61381	1,17457	-,32667	-1,29399	2,01829
8	4	,95	-1,75059	-1,04623	-1,34239	-1,29399	2,39322
9	2	,48	-1,75059	,06417	,68906	-2,08026	,75423
10	2	,00	-,10928	,06417	-1,34239	-2,08026	-,88476
11	8	,48	,02749	1,17457	2,72051	,27855	,75423
12	4	,30	-2,84479	,06417	-1,34239	-1,29399	,14933
13	2	,95	-1,20349	-1,04623	-1,34239	-2,08026	2,39322
14	8	,00	-2,16092	,06417	,68906	,27855	-,88476
15	6	,48	,98492	-2,15663	-,83453	-,50772	,75423
16	2	,00	-2,02414	-1,04623	-1,34239	-2,08026	-,88476
17	10	,48	,98492	1,17457	-1,34239	1,06482	,75423
18	4	,90	-1,47704	-1,04623	-,83453	-1,29399	2,21750

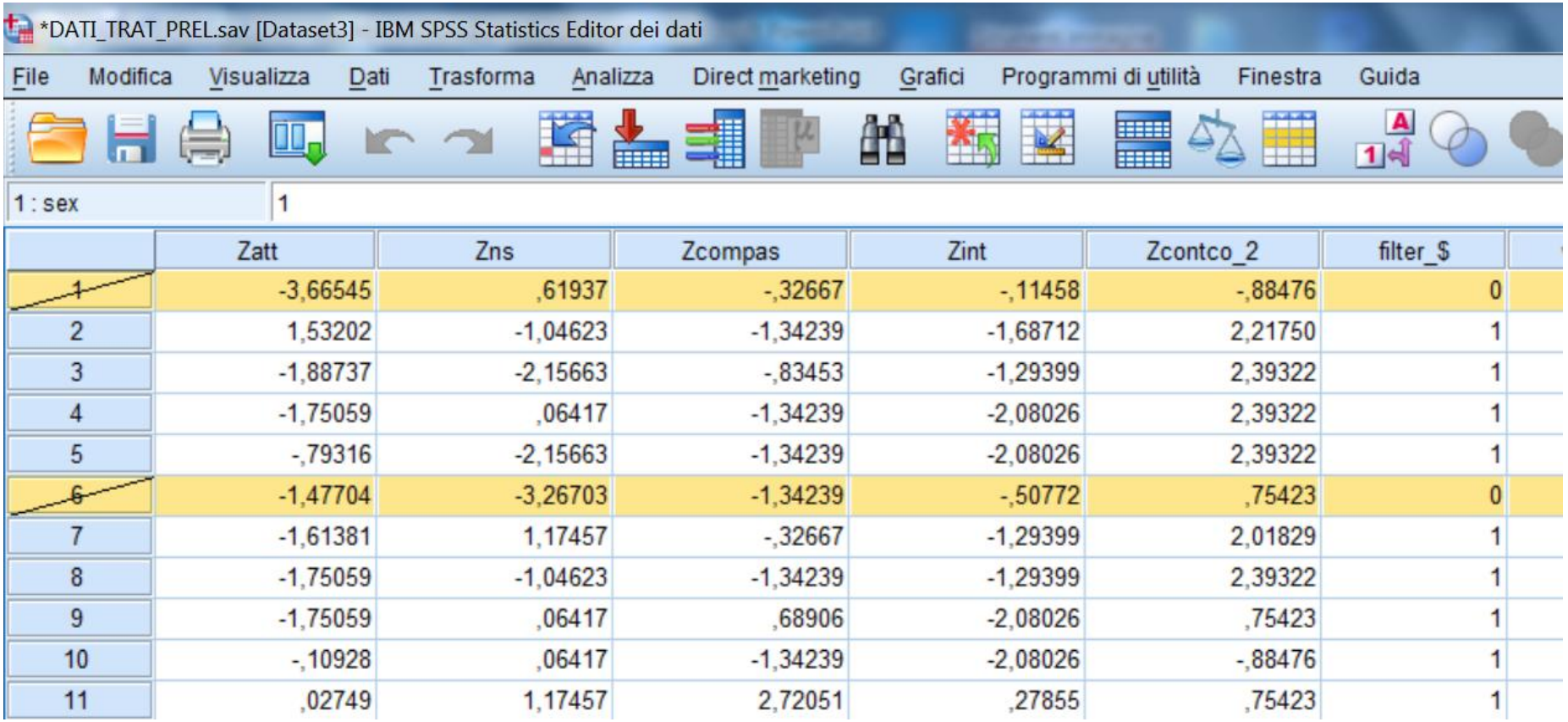


Filtrare i soggetti escludendo i due outliers

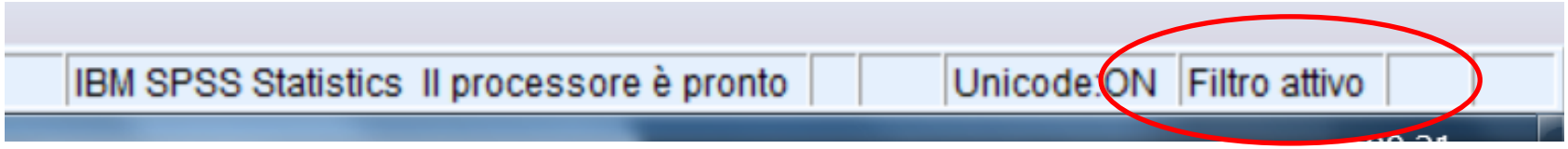
Filtrare i soggetti escludendo i due outliers



Filtrare i soggetti escludendo i due outliers



	Zatt	Zns	Zcompas	Zint	Zcontco_2	filter_$
1	-3,66545	,61937	-,32667	-,11458	-,88476	0
2	1,53202	-1,04623	-1,34239	-1,68712	2,21750	1
3	-1,88737	-2,15663	-,83453	-1,29399	2,39322	1
4	-1,75059	,06417	-1,34239	-2,08026	2,39322	1
5	-,79316	-2,15663	-1,34239	-2,08026	2,39322	1
6	-1,47704	-3,26703	-1,34239	-,50772	,75423	0
7	-1,61381	1,17457	-,32667	-1,29399	2,01829	1
8	-1,75059	-1,04623	-1,34239	-1,29399	2,39322	1
9	-1,75059	,06417	,68906	-2,08026	,75423	1
10	-,10928	,06417	-1,34239	-2,08026	-,88476	1
11	,02749	1,17457	2,72051	,27855	,75423	1



Filtrare i soggetti escludendo i due outliers

Statistiche descrittive

	N	Minimo	Massimo	Media	Deviazione std.	Asimmetria		Curtosi	
	Statistica	Statistica	Statistica	Statistica	Statistica	Statistica	Errore std	Statistica	Errore std
att	197	22	54	42,99	7,050	-,760	,173	-,290	,345
ns	197	4	10	7,91	1,759	-,451	,173	-,674	,345
compas	197	0	8	2,66	1,969	,382	,173	-,452	,345
int	197	2	10	7,30	2,555	-,687	,173	-,639	,345
contco_2	197	,00	,95	,2578	,29159	,667	,173	-,765	,345
Validi (listwise)	197								

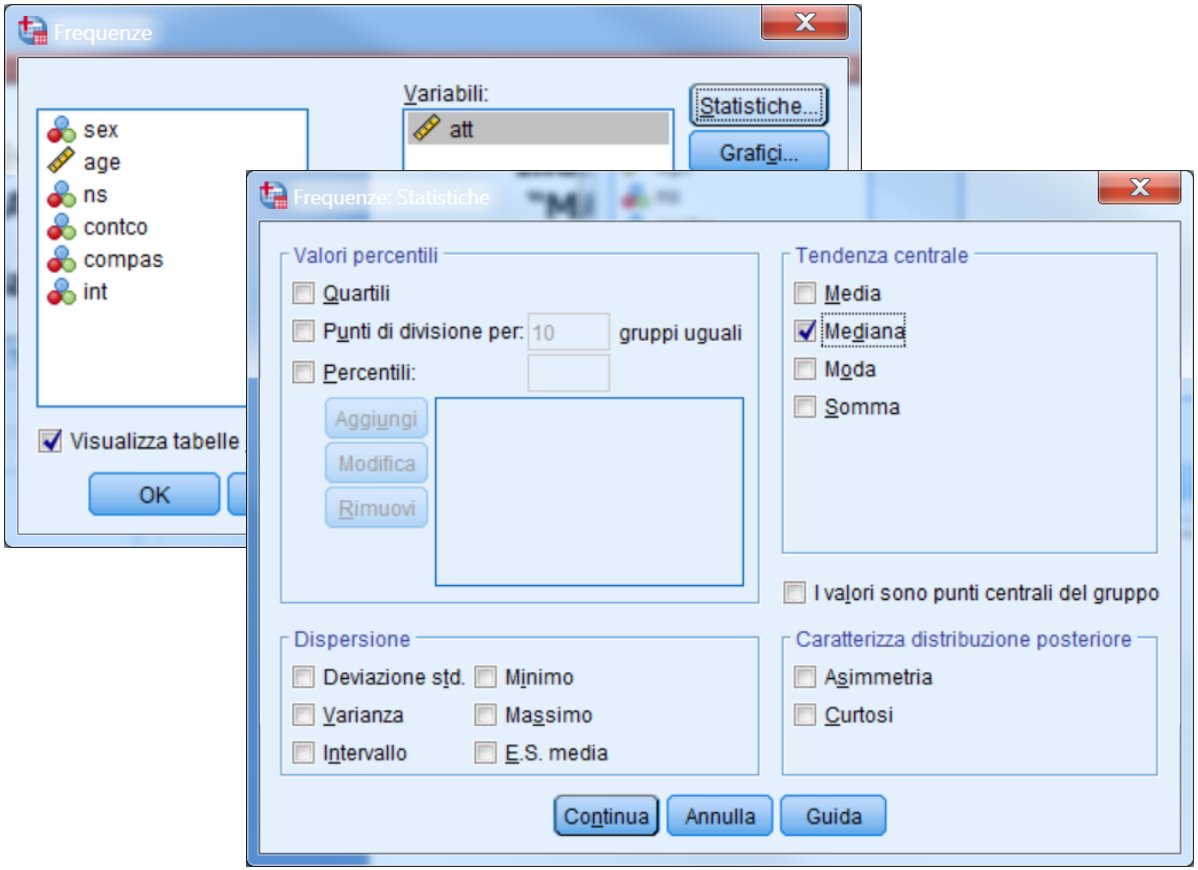
Le distribuzioni migliorano !

Individuazione degli outliers tramite l'indice "MAD" (Valore Assoluto Mediano) on SPSS

La procedura per calcolare l'indice MAD è semplice:
(a) Calcolare la mediana tramite la procedura "Frequenze";
(b) Sottrarre la mediana dal punteggio nella variabile per ogni soggetto tramite "Trasforma/Calcola variabile" in valore assoluto;
(c) Calcolare la mediana della nuova variabile ("Frequenze"): questo è il "MAD"
(d) Per ogni soggetto calcolare la formula seguente con "Calcola variabile": |X-Mdn|/(1.483*MAD)
(e) Sono da considerare outliers quei soggetti il cui valore è maggiore di 3 o di 2.5

Individuazione degli outliers tramite l'indice "MAD" (Valore Assoluto Mediano) on SPSS

Consideriamo il calcolo dell'indice MAD per ATT.

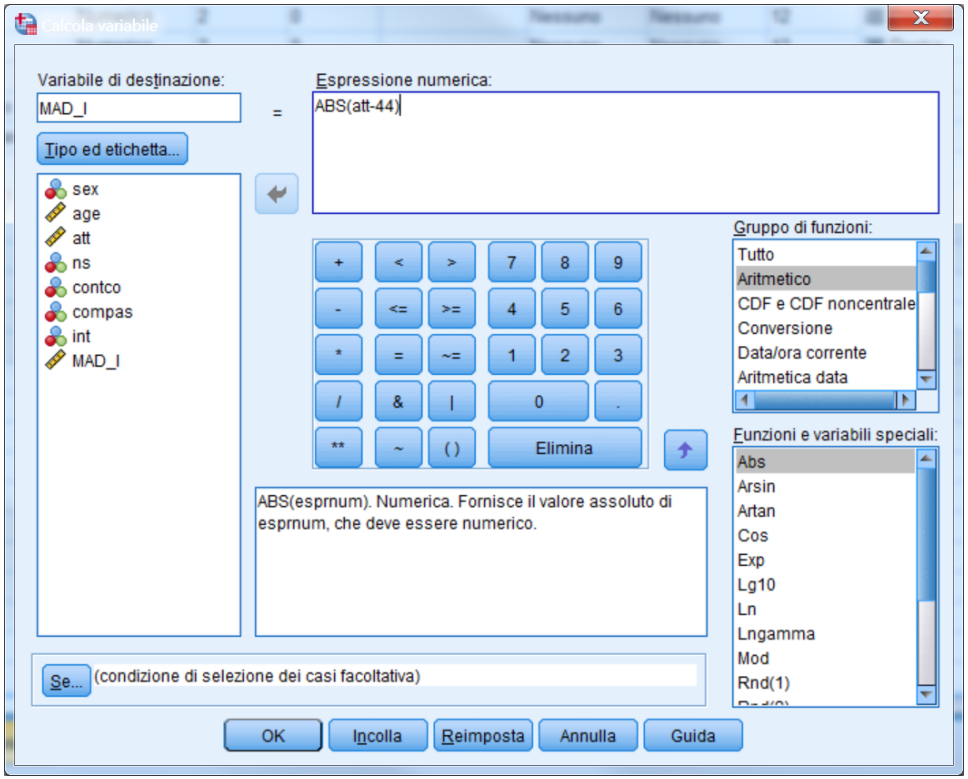


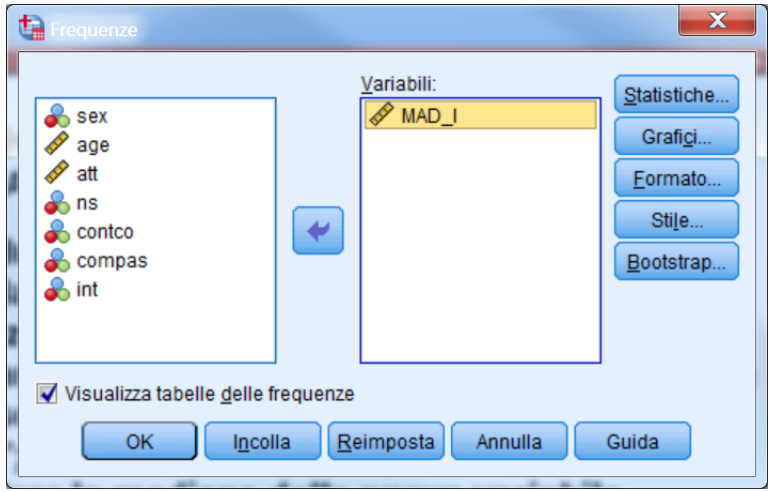
Statistiche

Mediana	44,00

Individuazione degli outliers tramite l'indice "MAD" (Valore Assoluto Mediano) on SPSS

Consideriamo il calcolo dell'indice MAD per ATT.



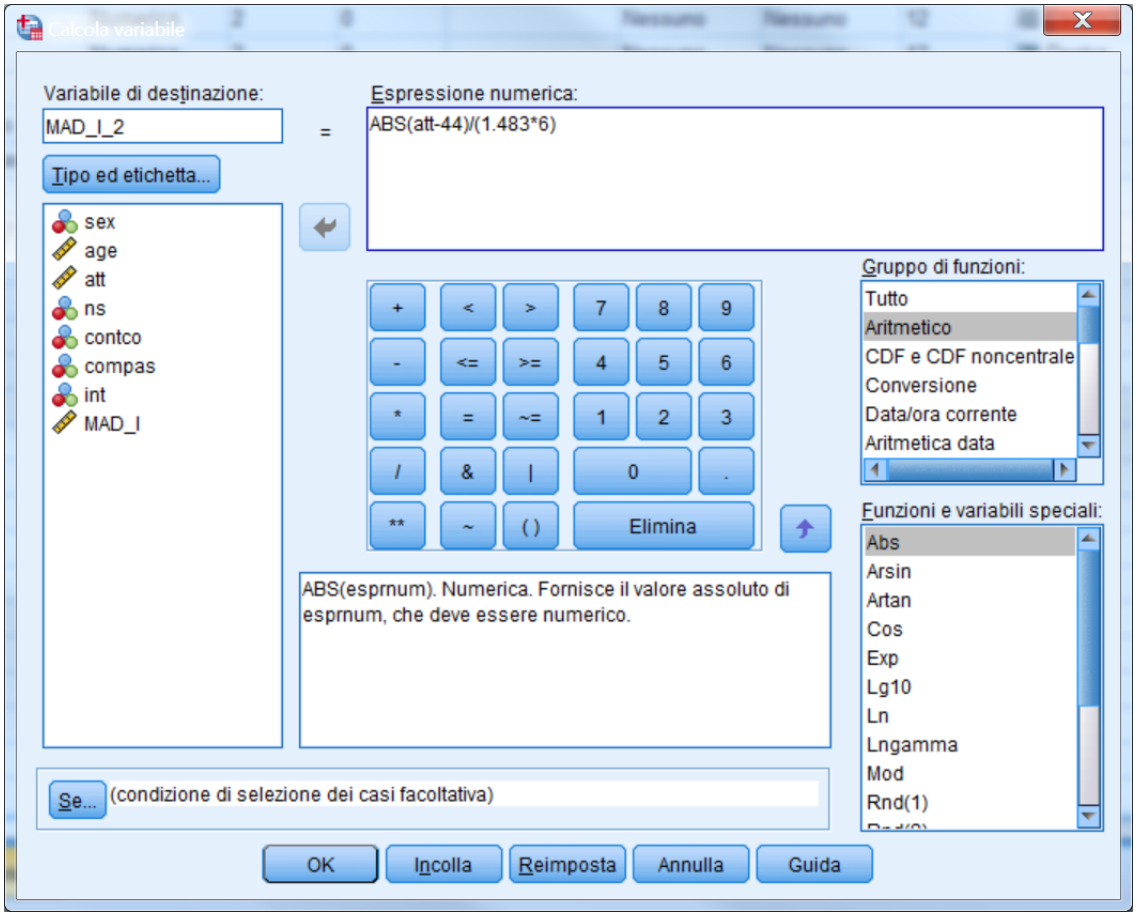


Statistiche MAD_I

Mediana	6,0000

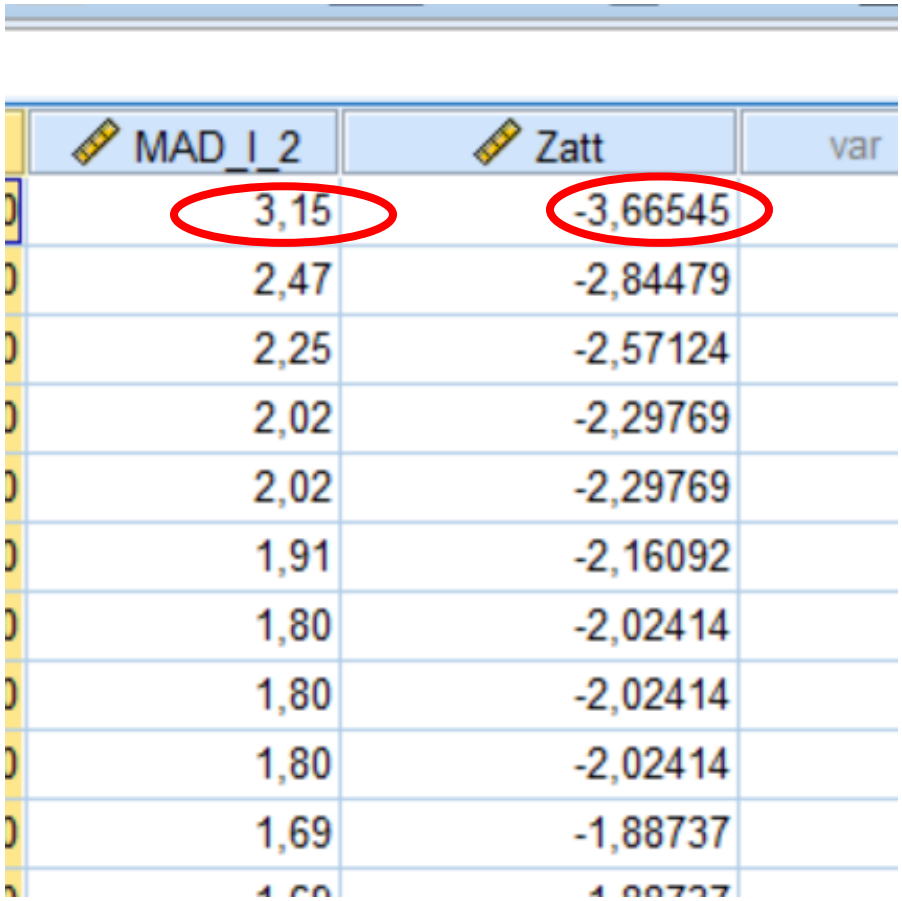
Individuazione degli outliers tramite l'indice "MAD" (Valore Assoluto Mediano) on SPSS

Consideriamo il calcolo dell'indice MAD per ATT.



Individuazione degli outliers tramite l'indice "MAD" (Valore Assoluto Mediano) on SPSS

In questo caso i due indici z e "MAD" danno risultati analoghi



MAD_I_2	Zatt	var
3,15	-3,66545	
2,47	-2,84479	
2,25	-2,57124	
2,02	-2,29769	
2,02	-2,29769	
1,91	-2,16092	
1,80	-2,02414	
1,80	-2,02414	
1,80	-2,02414	
1,69	-1,88737	

Esplorazione dei dati: la normalità multivariata

Per esaminare l'ipotesi di normalità multivariata Mardia ha sviluppato dei coefficienti di curtosi e di asimmetria multivariata. Se la distribuzione delle p variabili è normale multivariata, il coefficiente di curtosi multivariata di Mardia dovrebbe essere uguale a p(p+2) [p=numero di variabili].

Per valutare la normalità multivariata è possibile utilizzare un test grafico che si basa sull'utilizzo dei quantili della distribuzione del chi quadrato e sulla distanza generalizzata o distanza di Mahalanobis.

In SPSS la distanza di Mahalanobis è calcolabile utilizzando la procedura della regressione lineare multipla.

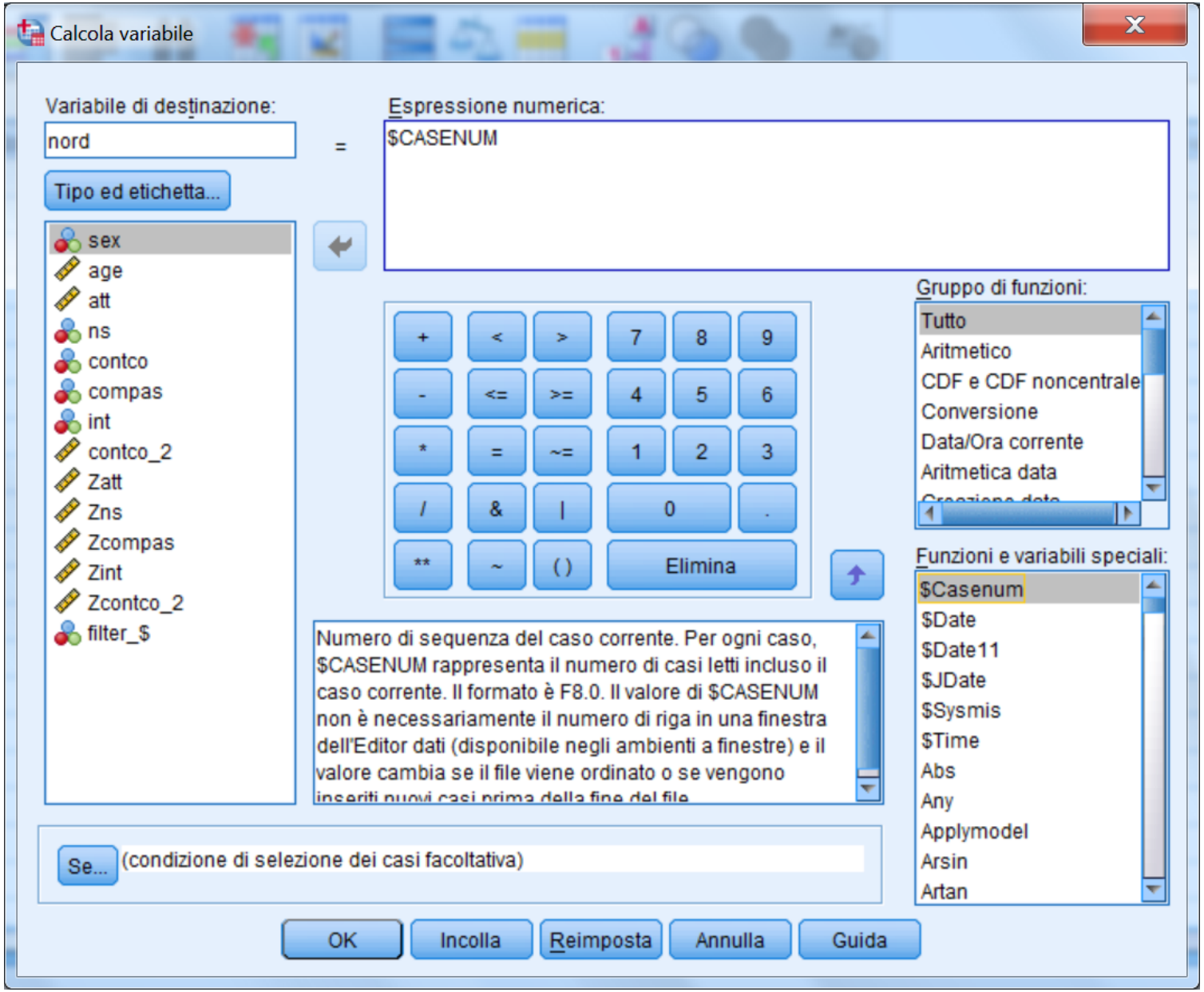
Esplorazione dei dati: la normalità multivariata e outliers multivariati

* Calcoliamo preliminarmente una nuova variabile (nord) alla quale vengono assegnati i valori della variabile di sistema $casenum: questa variabile fornisce il numero d'ordine del soggetto nel file (es., il primo soggetto nel file avrà $casenum = 1, e coì via).

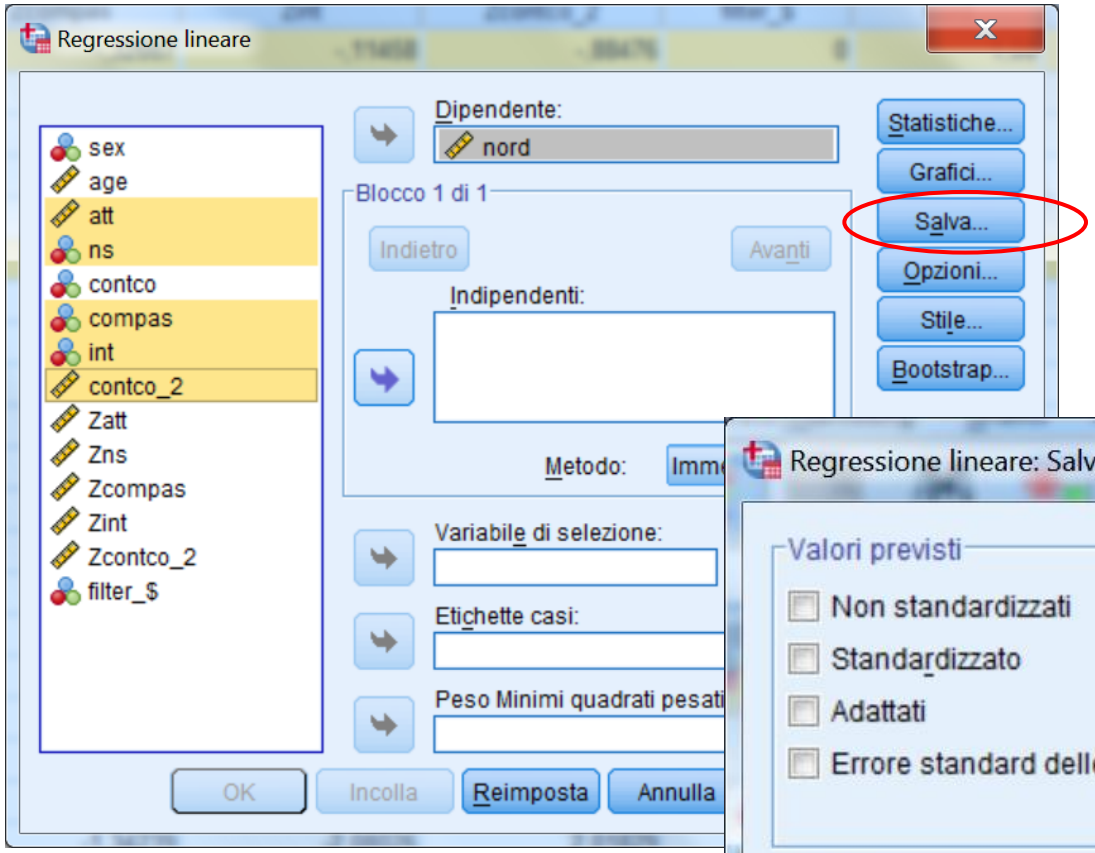
* Questa nuova variabile verrà utilizzata come variabile dipendente in una regressione multipla che ha il solo scopo di calcolare per ogni soggetto la distanza di Mahalanobis, la quale viene salvata nel file come una nuova variabile con il nome mah_1.

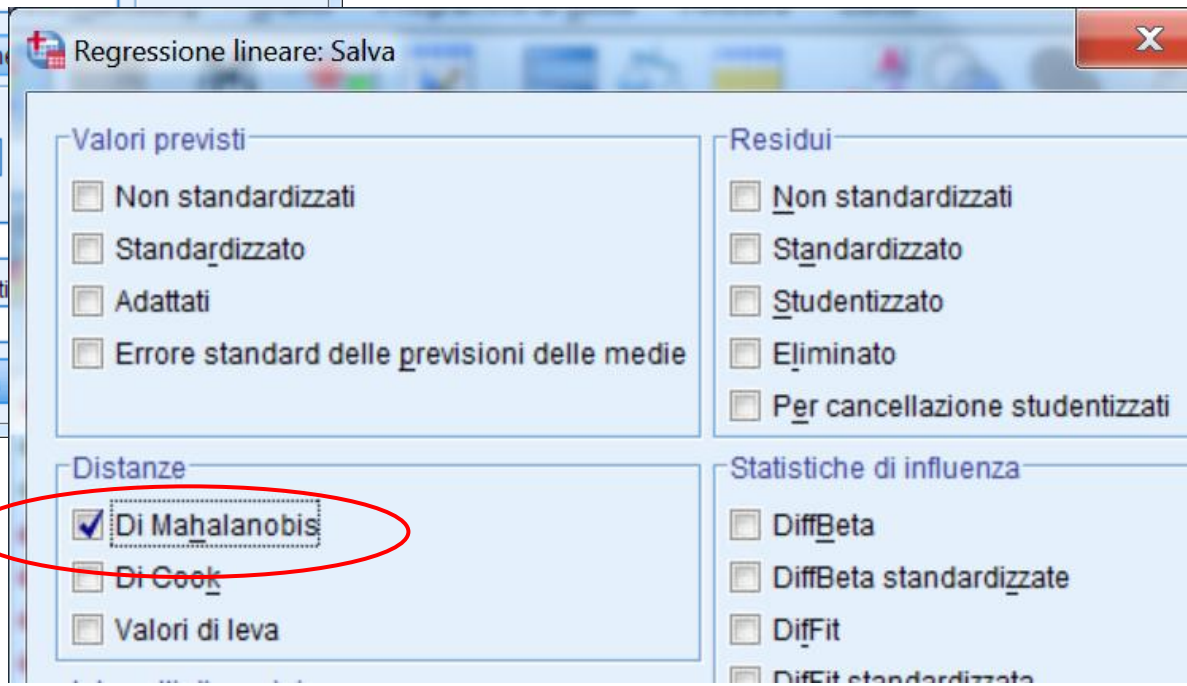
I comandi tramite le finestre di dialogo dei menù sono descritti di seguito.

Calcolo della variabile "nord"



Calcolo della distanza di Mahalanobis tramite regressione



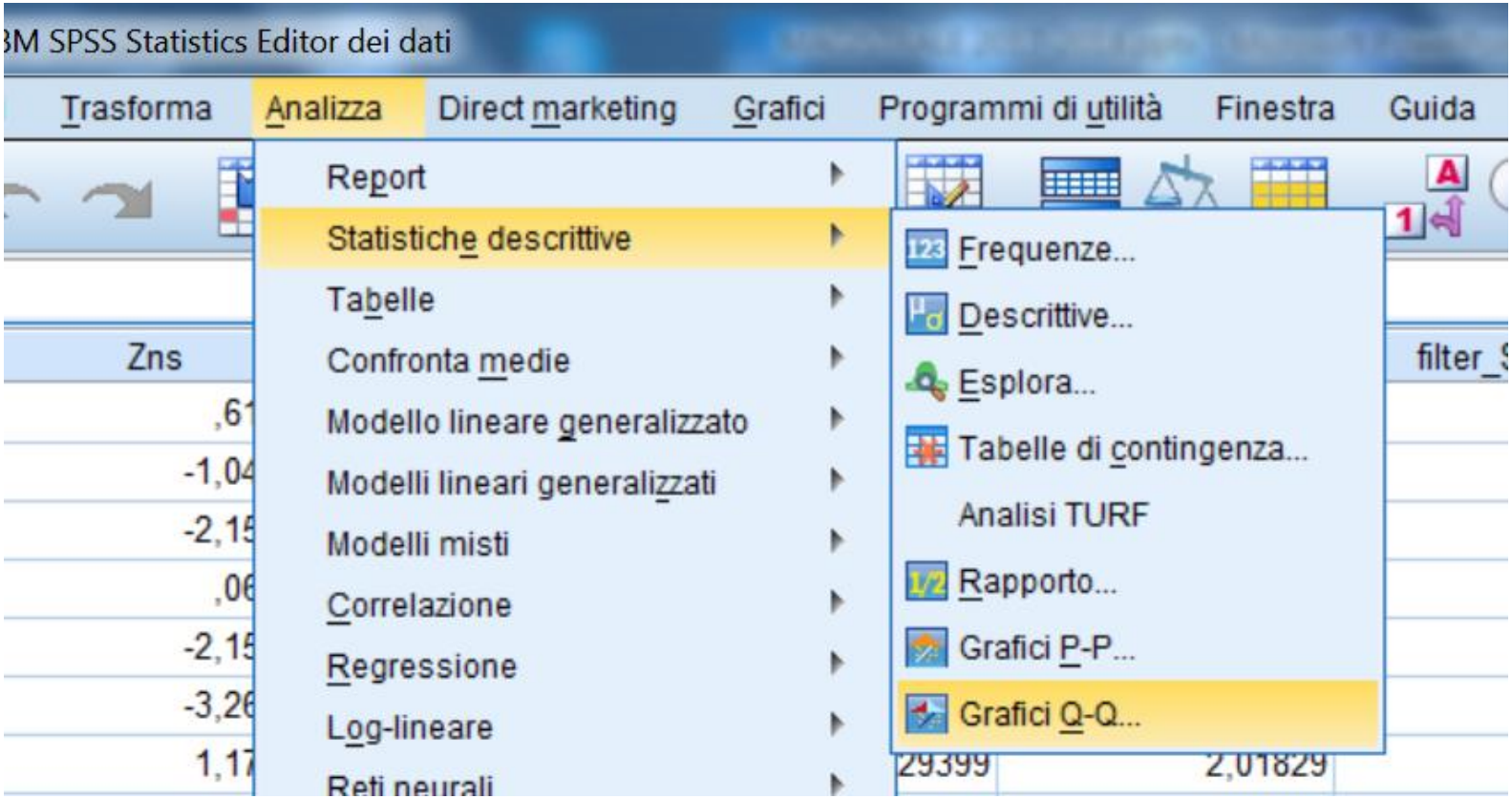


Esplorazione dei dati: la normalità multivariata e

Zcompas	Zint	Zcontco_2	filter_$	nord	MAH_1
-,32667	-,11458	-,88476	0	1,00	.
-1,34239	-1,68712	2,21750	1	2,00	21,76511
-,83453	-1,29399	2,39322	1	3,00	10,37619
-1,34239	-2,08026	2,39322	1	4,00	9,71748
-1,34239	-2,08026	2,39322	1	5,00	9,69798
-1,34239	-,50772	,75423	0	6,00	.
-,32667	-1,29399	2,01829	1	7,00	12,99378
-1,34239	-1,29399	2,39322	1	8,00	7,21699
,68906	-2,08026	,75423	1	9,00	14,27052
-1,34239	-2,08026	-,88476	1	10,00	15,05194
2,72051	,27855	,75423	1	11,00	14,89854
-1,34239	-1,29399	,14933	1	12,00	14,29484
-1,34239	-2,08026	2,39322	1	13,00	6,67944
,68906	,27855	-,88476	1	14,00	14,09949
-,83453	-,50772	,75423	1	15,00	12,84113
-1,34239	-2,08026	-,88476	1	16,00	11,75779
-1,34239	1,06482	,75423	1	17,00	10,98545
-,83453	-1,29399	2,21750	1	18,00	5,74019
-1,34239	-1,29399	-,88476	1	19,00	11,17339
-1,34239	-2,08026	2,01829	1	20,00	9,14158

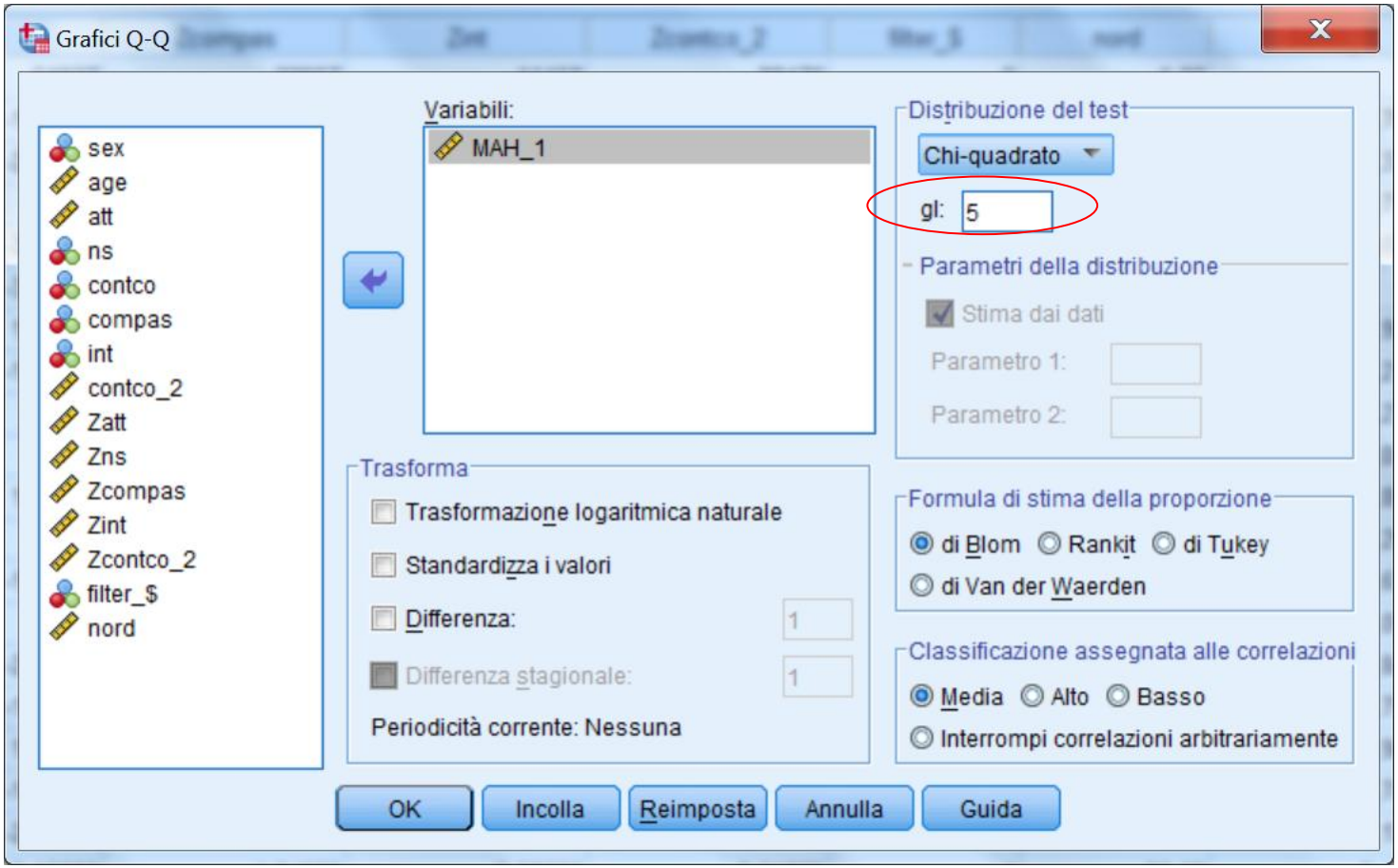
Viene creata la variabile MAH_1 nel datafile

Test grafico Q-Q Plot



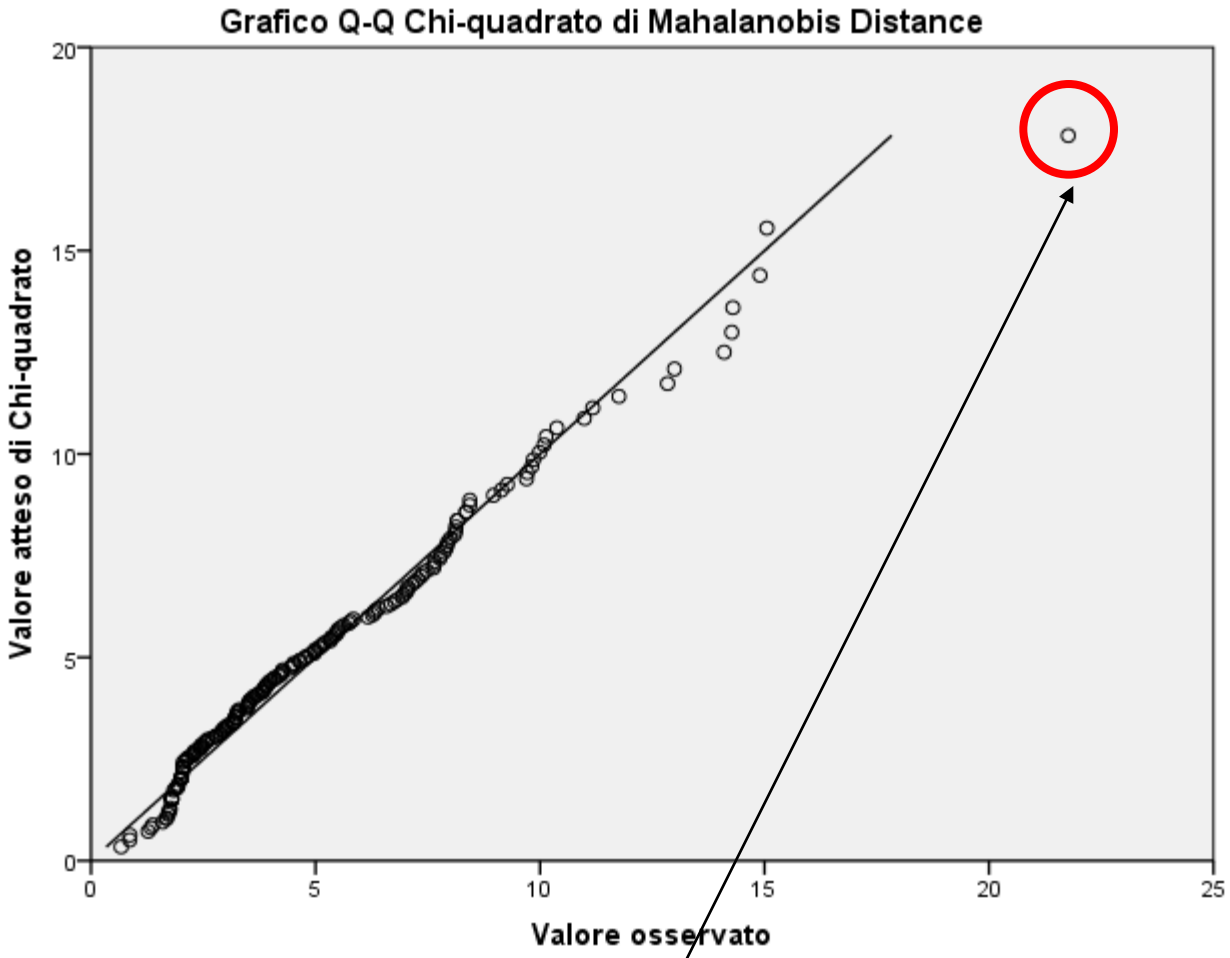
Si chiede tramite Statistiche descrittive…

Test grafico Q-Q Plot



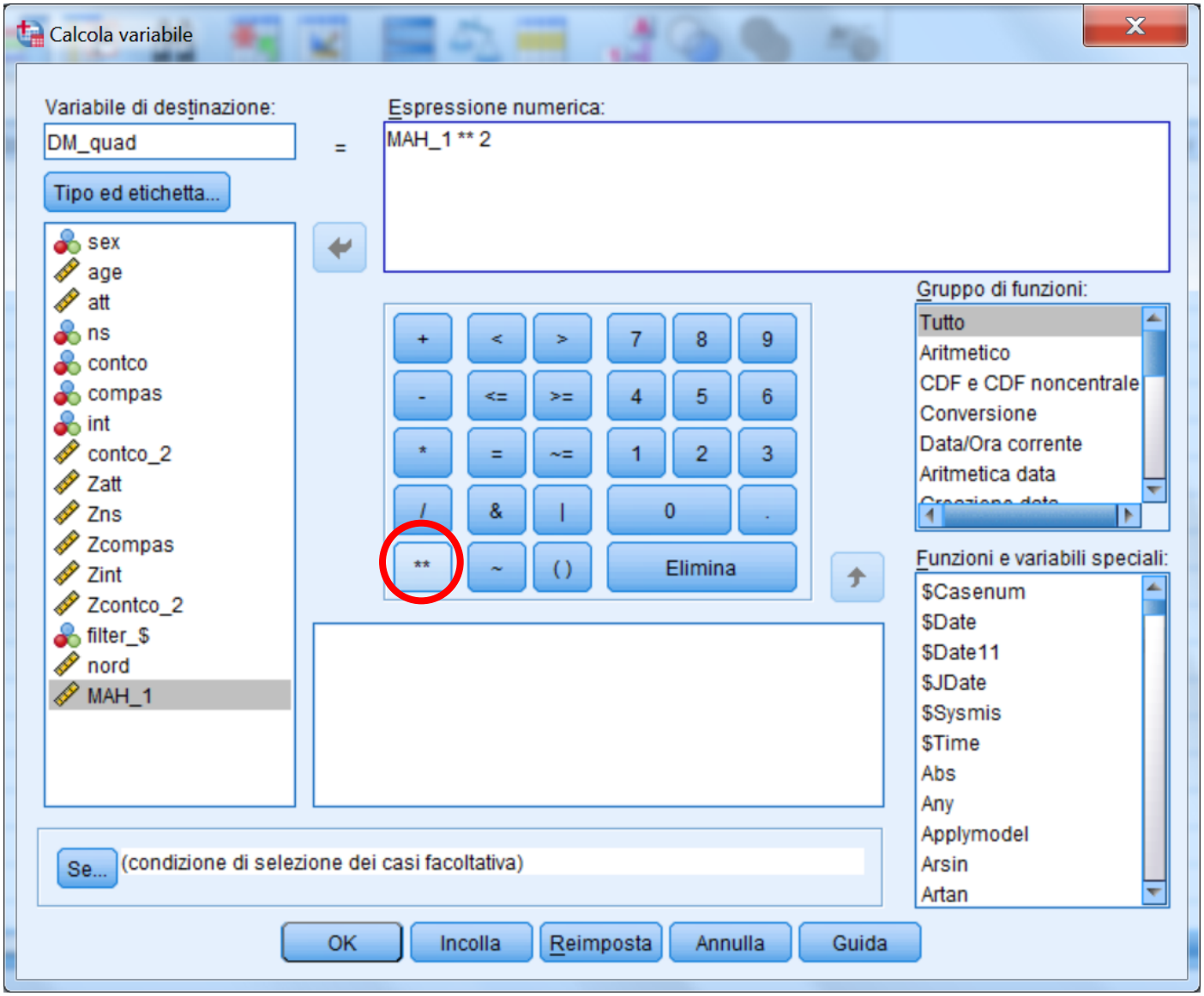
Specificare la distribuzione chi-quadrato con 5 gradi di libertà (ci sono 5 variabili)

Test grafico Q-Q Plot



E' il possibile outlier multivariato

Calcolo del coefficiente di curtosi multivariata



Formula:

$$\sum_{i=1}^{N} \left(D_i^2\right)^2 \Big/ N$$

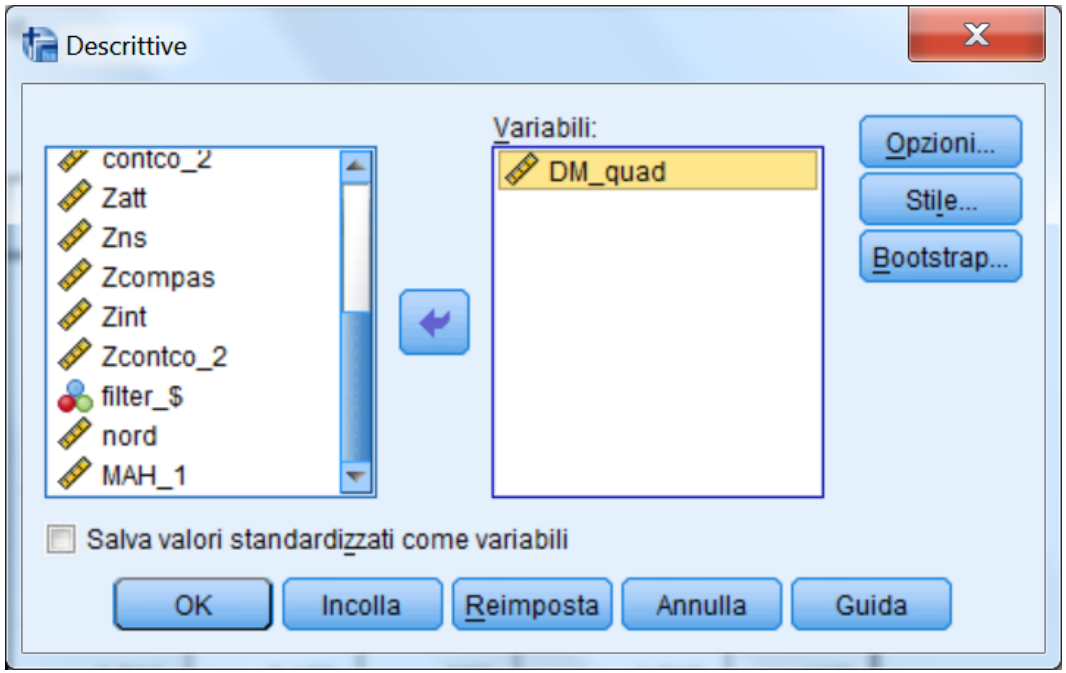
Calcoliamo il denominatore

$$D_i^2 = MAH\_1$$

Per cui:

$$\left(D_i^2\right)^2 = (MAH\_1)^2$$

Calcolo del coefficiente di curtosi multivariata



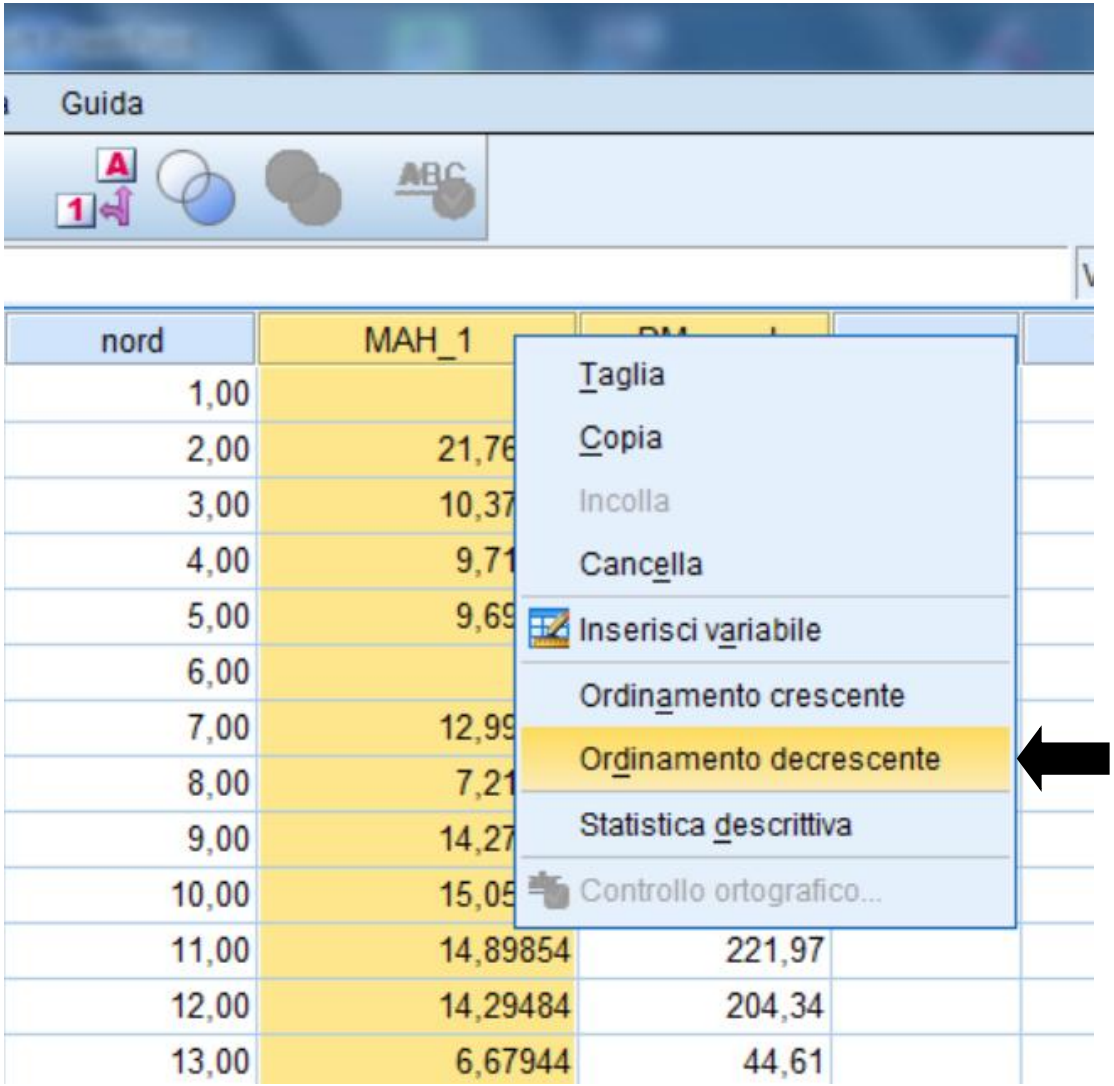
Calcolando la media della variabile DM_quad si ottiene il coefficiente di Mardia

Statistiche descrittive

	N	Minimo	Massimo	Media	Deviazione std.
DM_quad	197	,45	473,72	35,7249	53,18476
Validi (listwise)	197				

Il coefficiente è 35.72, di poco superiore a 35 (=5*7), il valore critico con 5 variabili.

Esplorazione dei dati: individuare gli outlier multivariati

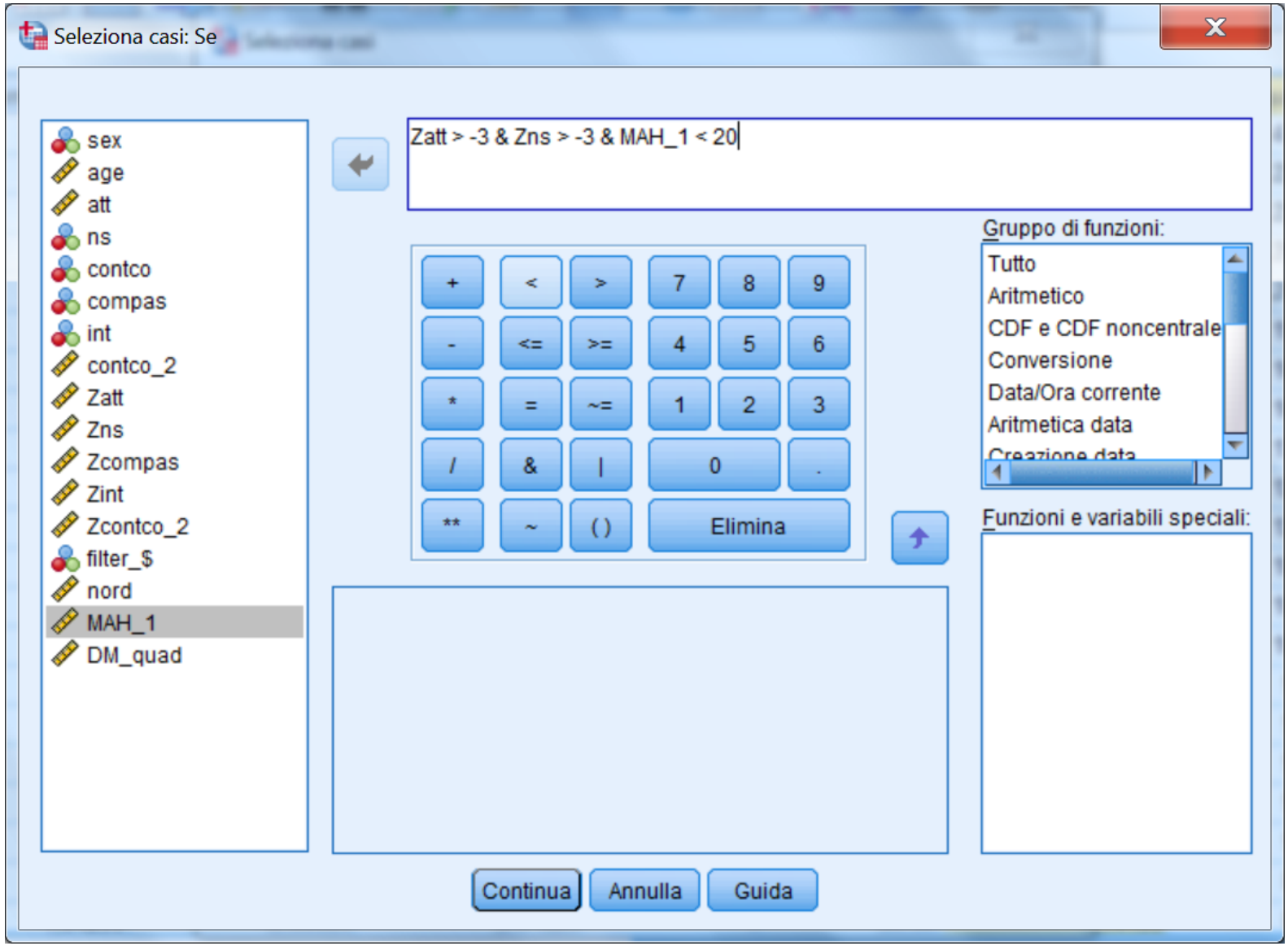


Esplorazione dei dati: la normalità multivariata e outliers multivariati

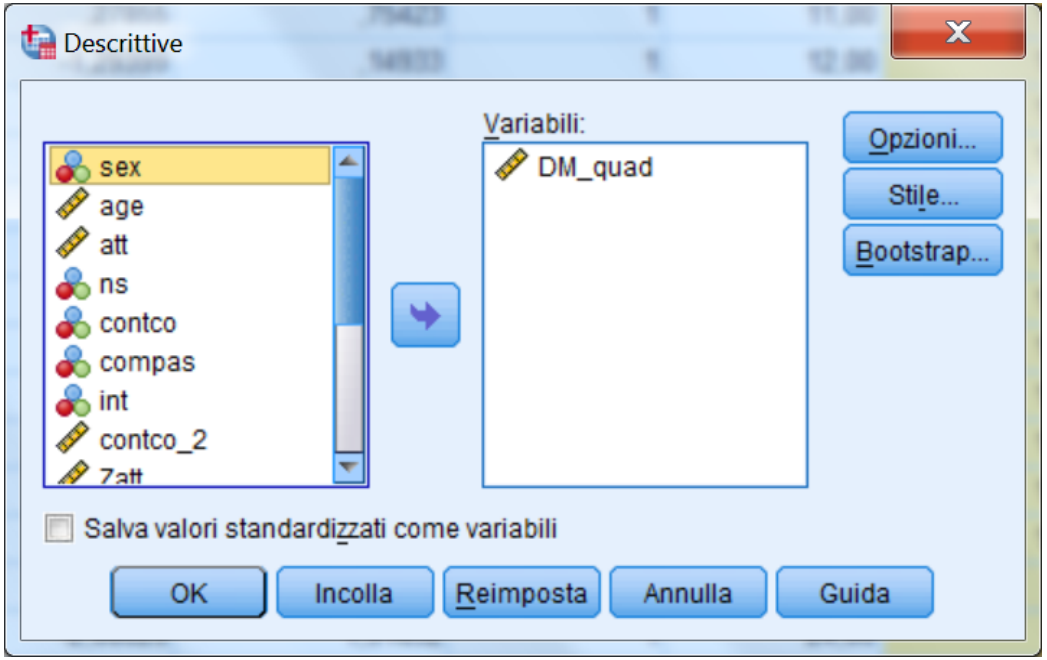
int	Zcontco_2	filter_$	nord	MAH_1	DM_quad	var
-1,68712	2,21750	1	2,00	21,76511	473,72	
-2,08026	-,88476	1	10,00	15,05194	226,56	
,27855	,75423	1	11,00	14,89854	221,97	
-1,29399	,14933	1	12,00	14,29484	204,34	
-2,08026	,75423	1	9,00	14,27052	203,65	
,27855	-,88476	1	14,00	14,09949	198,80	
-1,29399	2,01829	1	7,00	12,99378	168,84	
-,50772	,75423	1	15,00	12,84113	164,89	

Vanno considerati come outliers multivariati i casi il cui valore risulta significativo al livello $p<.001$, considerando come distribuzione di riferimento quella del chi-quadrato con p gradi di libertà (dove p = numero di variabili). Con p = 5 (abbiamo infatti 5 variabili) il livello di significatività del χ^2 è 20.51, quindi c'è un possibile outlier multivariato.

Filtrare i soggetti escludendo i due outliers uni- e l'outlier multi-variato



Calcolo del coefficiente di curtosi multivariata



Statistiche descrittive

	N	Minimo	Massimo	Media	Deviazione std.
DM_quad	196	,45	226,56	33,4903	43,06171
Validi (listwise)	196				

Il coefficiente è 33.49, ora inferiore a 35 (=5*7).
Ora i dati sono pronti per le analisi !!

ESERCIZIO 1:
TRATTAMENTI PRELIMINARI CON SPSS

Utilizzare i dati in formato testo nel file es1.xlsx

VARIABILI:
ATTEGGIAMENTO, NORME SOGGETTIVE, SENSO DI CONTROLLO, COMPORTAMENTO PASSATO, INTENZIONE.
LA VARIABILE DIPENDENTE E' "INTENZIONE"

Verificare le caratteristiche distributive delle variabili, l'eventuale presenza di outlier, ed eventualmente trasformare le variabili non normali.

Salvare il file in formato .sav

LA REGRESSIONE LINEARE

Sommario

* Scopo dell'analisi della regressione
* Regressione bivariata: Modello di base
* Regressione multipla: Modello di base
* Stima e interpretazione dei parametri
* Adeguatezza della soluzione
* Misure dell'associazione lineare tra Variabili Indipendenti (VI) e Variabile Dipendente (VD)
* Assunzioni
* Approcci analitici alla regressione
* Limiti

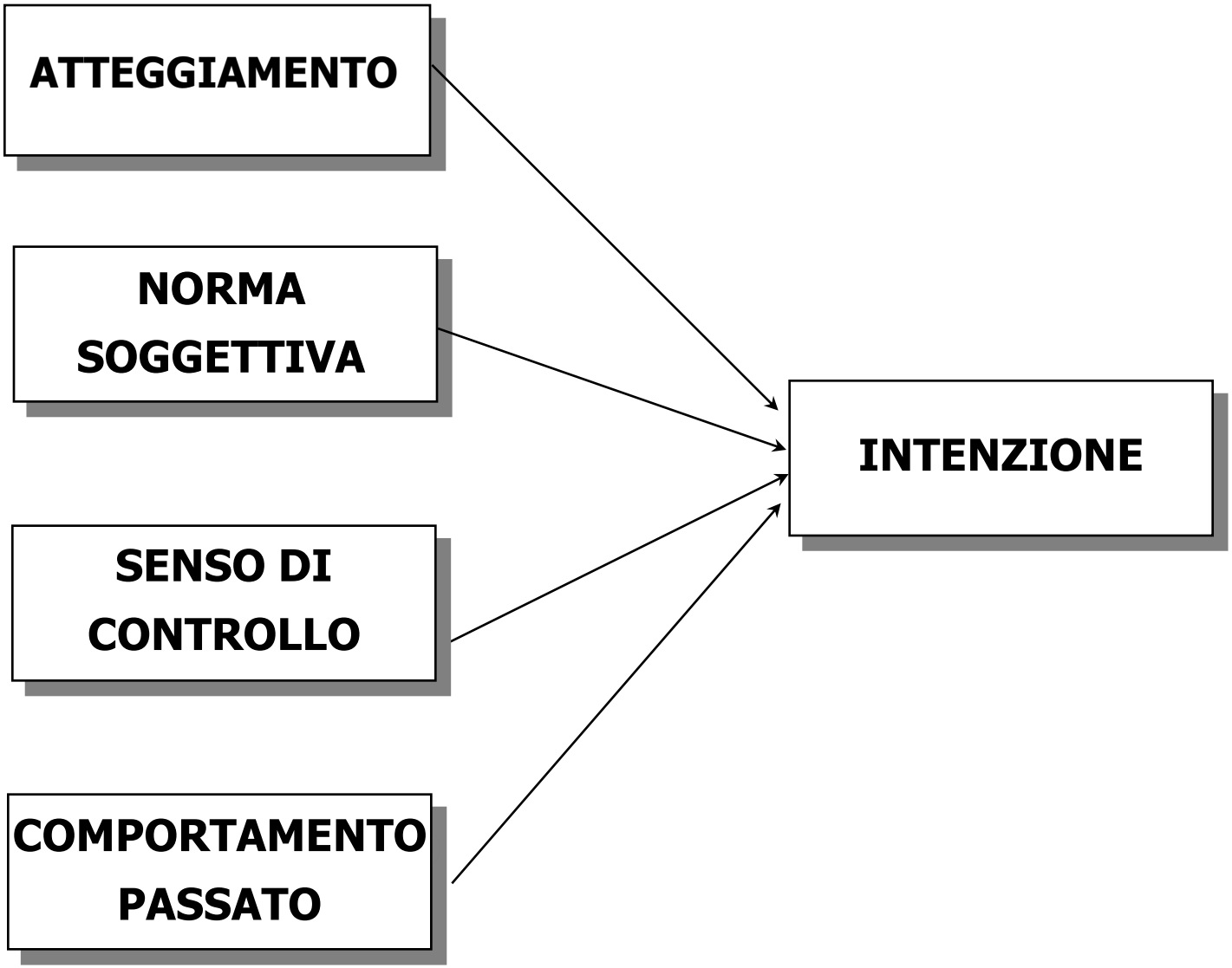
La Regressione esamina la relazione lineare tra una o più variabili esplicative (o indipendenti, VI, o "predittori") e una variabile criterio (o dipendente, VD).

Duplice scopo:

a) esplicativo: studiare e valutare gli effetti delle VI sulla VD in funzione di un determinato modello teorico

b) predittivo: individuare una combinazione lineare di VI per predire in modo ottimale il valore assunto dalla VD

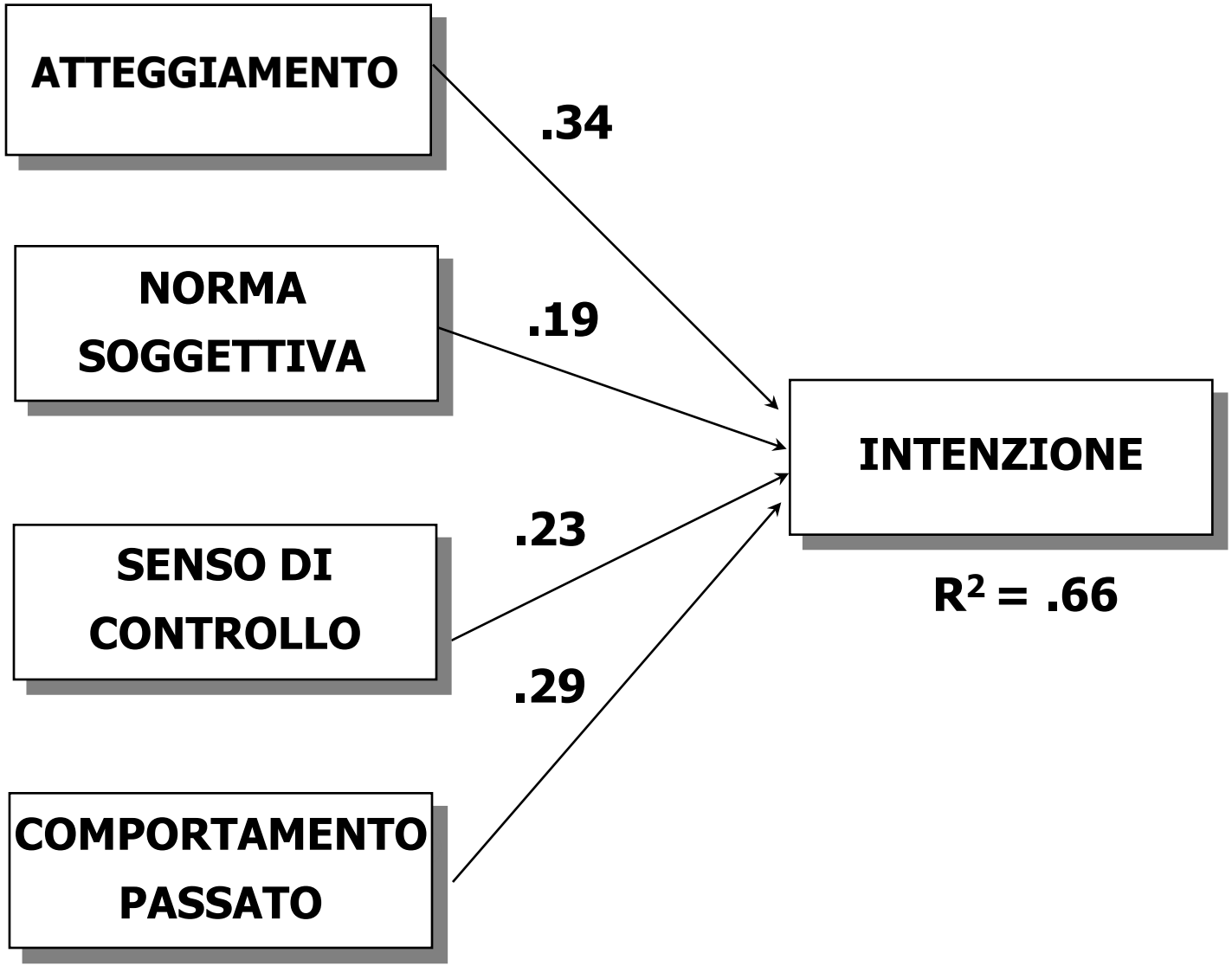
Da dove si parte: Modello concettuale



Da dove si parte: Matrice delle covarianze

	1	2	3	4	5
1.INT	6.438				
2.ATT	12.491	53.186			
3.NS	2.657	6.791	3.228		
4.CONTCO	2.650	5.534	1.149	3.453	
5.COMPAS	3.235	7.114	1.637	1.596	3.858

Dove si arriva: Modello empirico



Dove si arriva: Risultati del modello empirico

Variabile	B	Beta	T	p
Atteggiamento	.12	.34	6.38	.001
Norma Soggettiva	.28	.19	3.83	.001
Senso di Controllo	.32	.23	4.82	.001
Comport. Passato	.38	.29	5.65	.001

R^2 = .66; t= 16.74. p<.0001

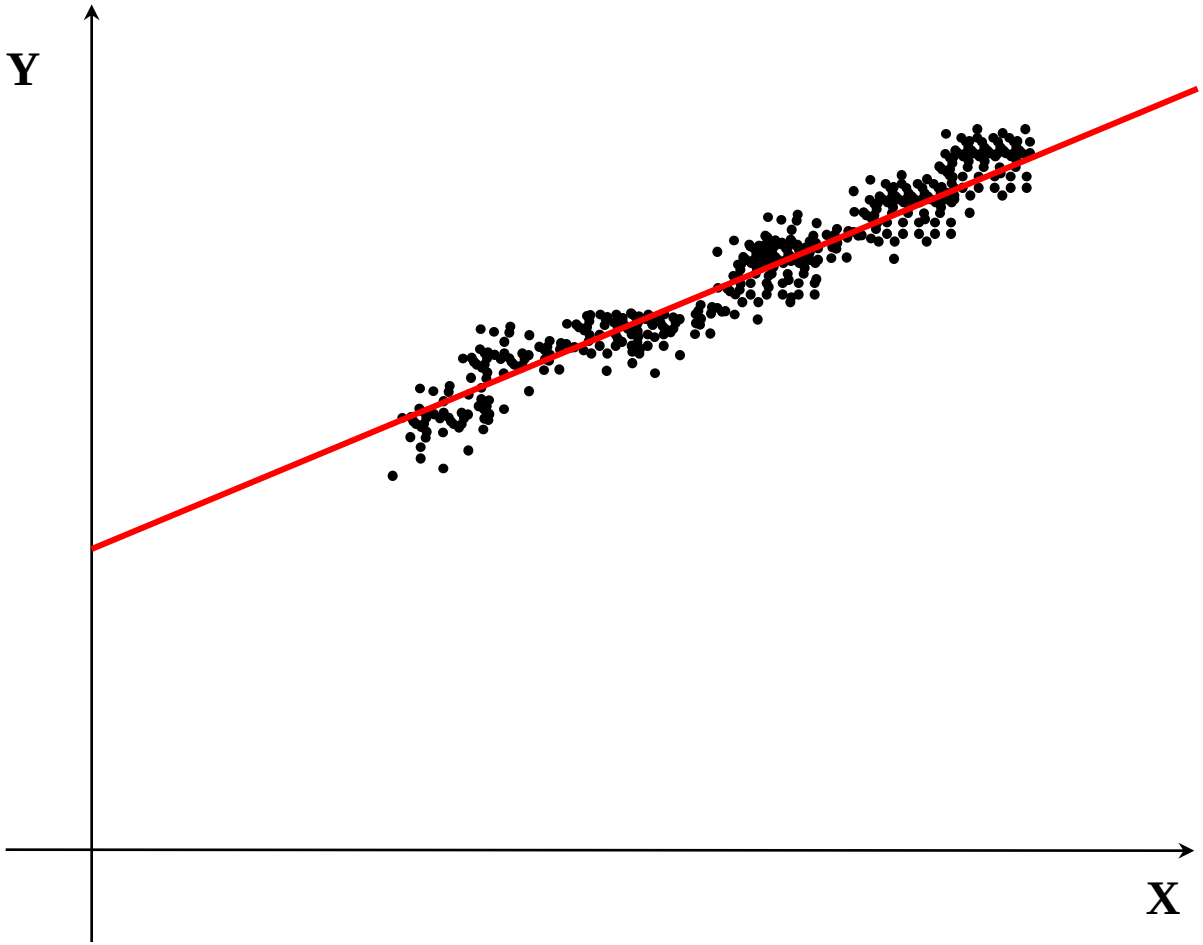
Regressione bivariata (o semplice)

Una sola variabile indipendente (VI) sulla quale "regredisce" la variabile dipendente (VD). Si ipotizza che la VI "determini" o "influenzi" o "predica" la VD.

Individuare quella retta che consente di prevedere al meglio i punteggi nella VD a partire da quelli nella VI.

Individuare la retta che "interpola" meglio la nuvola di punti (o "scatterplot") della distribuzione congiunta delle due variabili.

La retta di regressione (regressione bivariata)



Regressione bivariata (o semplice)

La relazione lineare è quella più parsimoniosa ed è quella più realistica in moltissimi casi.
L'equazione che lega Y a X è la seguente:

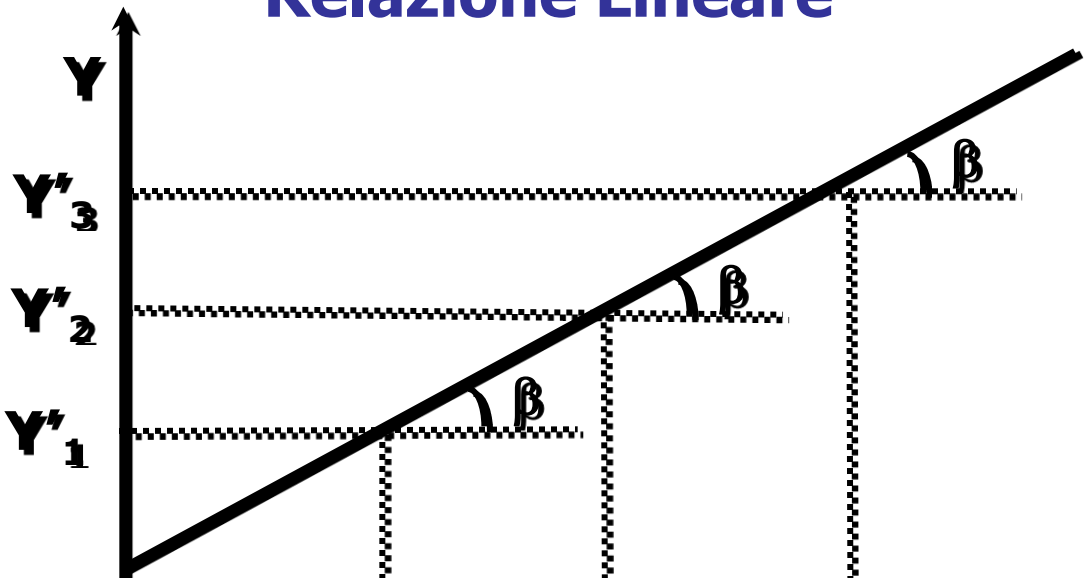
$$Y = \alpha + \beta X$$

Parametri dell'equazione:

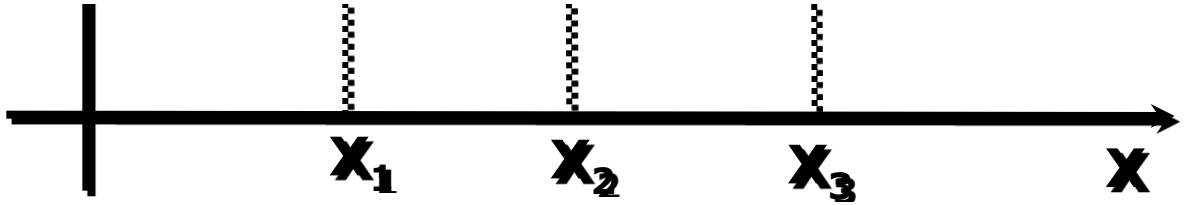
Intercetta: α, punto in cui la retta incrocia l'asse delle ordinate (altezza della linea).

Coefficiente angolare: β inclinazione della retta di regressione di Y su X; indica di quante unità cambia Y per una variazione unitaria che si verifica nella X.

Relazione Lineare

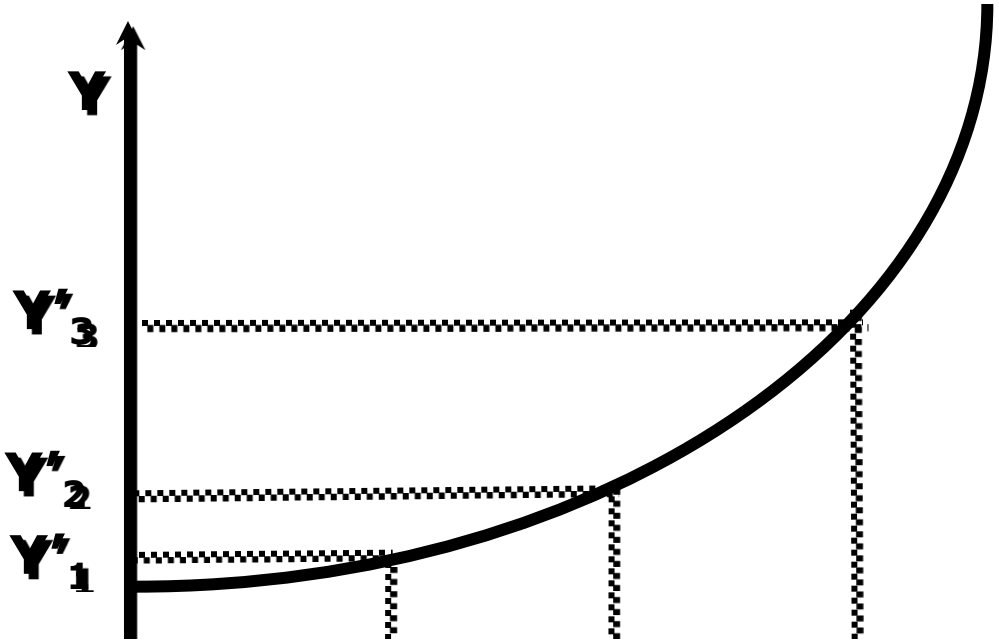


Per ogni variazione in X si determina sempre la stessa variazione in Y qualunque sia il valore di X sull'asse delle ascisse.

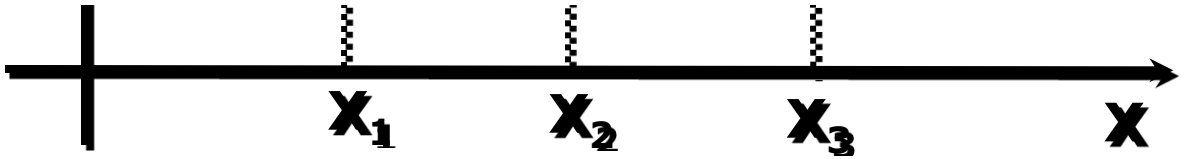


$$(X_3-X_2)=(X_2-X_1) \Rightarrow (Y'_3-Y'_2)=(Y'_2-Y'_1)$$

Relazione Non Lineare



La stessa variazione in X determina variazioni diverse in Y per diversi valori di X sull'asse delle ascisse.



$(X_3-X_2)=(X_2-X_1)$ Ma $(Y'_3-Y'_2)\neq(Y'_2-Y'_1)$

Errore o residuo

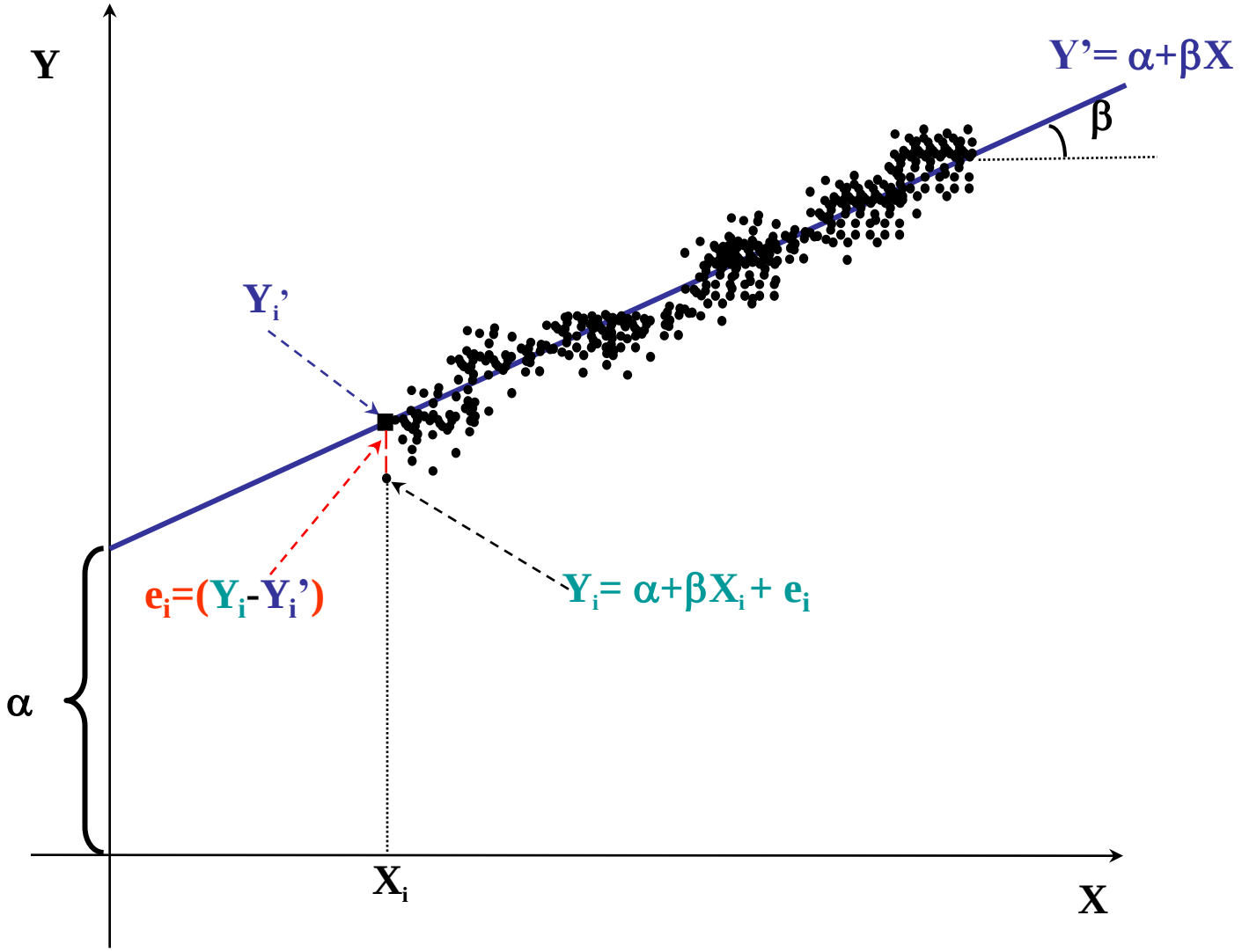
I punti sono dispersi intorno alla retta di regressione perché:
- le variabili sono misurate con errore;
- la relazione può non essere perfettamente lineare;
- predittori importanti possono essere omessi.

L'equazione quindi deve incorporare un termine di errore (o residuo) per ogni caso.

$$Y = \alpha + \beta X + e = Y' + e$$

$Y' = \alpha + \beta X$: valore "teorico" della Y, ottenuto dalla regressione.

"e": Residuo, deviazione del punteggio osservato Y dal punteggio teorico Y'.


Y
Y'= α+βX
β
Y_i'
e_i=(Y_i-Y_i')
Y_i= α+βX_i + e_i
α
X_i
X

La Stima dei parametri

Bisogna identificare la retta che meglio si adatta ai punti che descrivono la distribuzione delle Y sulle X.

La retta che interpola meglio il diagramma di dispersione, cioè quella retta che passa più vicina possibile alla nuvola dei punti, è quella che rende minima la somma delle differenze al quadrato tra le Y osservate e le Y' teoriche.

I parametri α e β vengono stimati nel campione attraverso il metodo dei minimi quadrati, ovvero il metodo che rende minimo l'errore che si commette quando Y viene "stimato" dalla equazione di regressione.

Equazione dei minimi quadrati:

$$\Sigma(Y_i - Y_i')^2 = \Sigma(Y_i-(a+bx_i))^2 = \min$$

Identifica la retta che riduce al minimo l'errore che viene commesso nello stimare Y da X.

Formule dei minimi quadrati per il calcolo di a e b:

$$b = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{cov(X,Y)}{var(X)} \qquad a = \bar{Y} - b\bar{X}$$

Il coefficiente "a" rappresenta il valore atteso di Y quando X è uguale a 0.
Il coefficiente "b" rappresenta il cambiamento atteso in Y associato a un cambio di una unità in X.

Stime standardizzate

Il coefficiente di regressione esprime la relazione tra Y e X nell'unità di misura delle 2 variabili. Per esprimere questa relazione in una scala di misura comprensibile si deve standardizzarlo.

Il coefficiente standardizzato si ottiene moltiplicando il coefficiente "grezzo" (non standardizzato) per il rapporto delle deviazioni standard della VI e della VD:

$$\beta\hat{} = b\,(s_x/s_y)$$

Nella regressione semplice è uguale al coefficiente di correlazione "semplice", ovvero: $\beta\hat{} = r_{yx}$

La regressione multipla

Una variabile dipendente che regredisce su <u>almeno</u> due variabili indipendenti. Equazione di regressione:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon_i$$

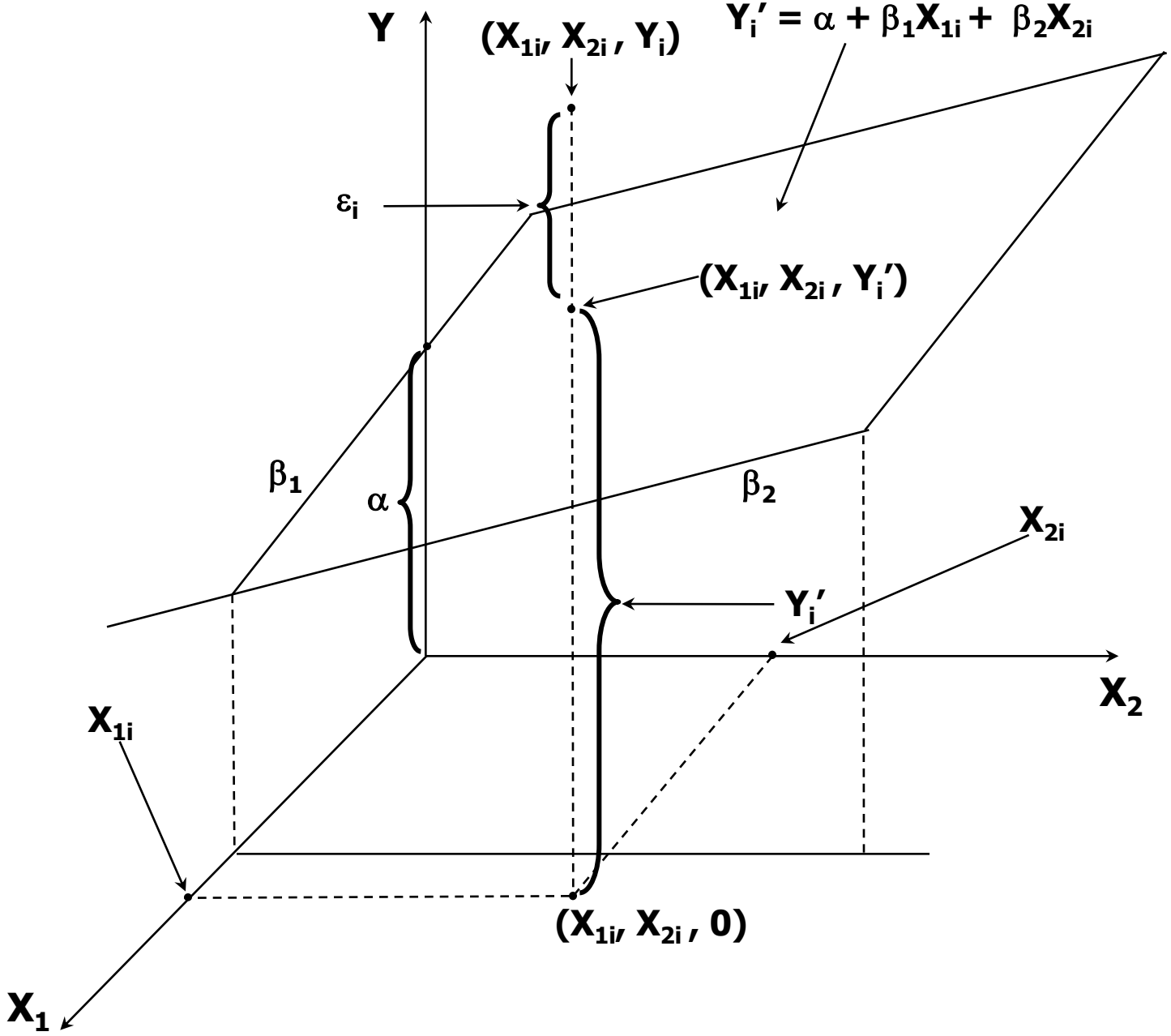
Piano di regressione (due VI);

Iperpiano (più di 2 VI).

Equazione del piano di regressione:

$$Y' = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Rappresentazione grafica: il piano di regressione



Coefficienti di regressione della regressione multipla: coefficienti "parziali" o "netti" (partial slope o partial regression coefficient).

Dipendenza della variabile Y da ciascuna delle VI X_i, al netto delle altre VI nell'equazione.
Per ogni VI rappresentano l'inclinazione della retta di regressione della variabile dipendente, ottenuta mantenendo costanti i valori delle altre VI.

Nel piano:

β_1 è l'inclinazione della retta di regressione di Y su X_1 quando si mantiene costante X_2
β_2 è l'inclinazione della retta di regressione di Y su X_2, se si mantiene costante X_1.

Stime dei coefficienti: minimi quadrati.

Individuare un iperpiano di dimensioni k che si adatti meglio ai punti nello spazio di dim. k+1 (k VI e 1 VD).

$$\Sigma [Y - (\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)]^2 = \min$$

Espressioni matriciali delle equazioni:

$y = bX + e$	equazione di regressione	(1)
$b=(X'X)^{-1} X'Y$	coefficienti di regressione	(2)
$e=Y - (Xb+a)$	residui	(3)

X'X rappresenta la codevianza tra le VI, X'Y rappresenta la codevianza tra VI e VD.

Relazioni tra una VD Y e due VI X1 e X2, espresse in termini della varianza che condividono:

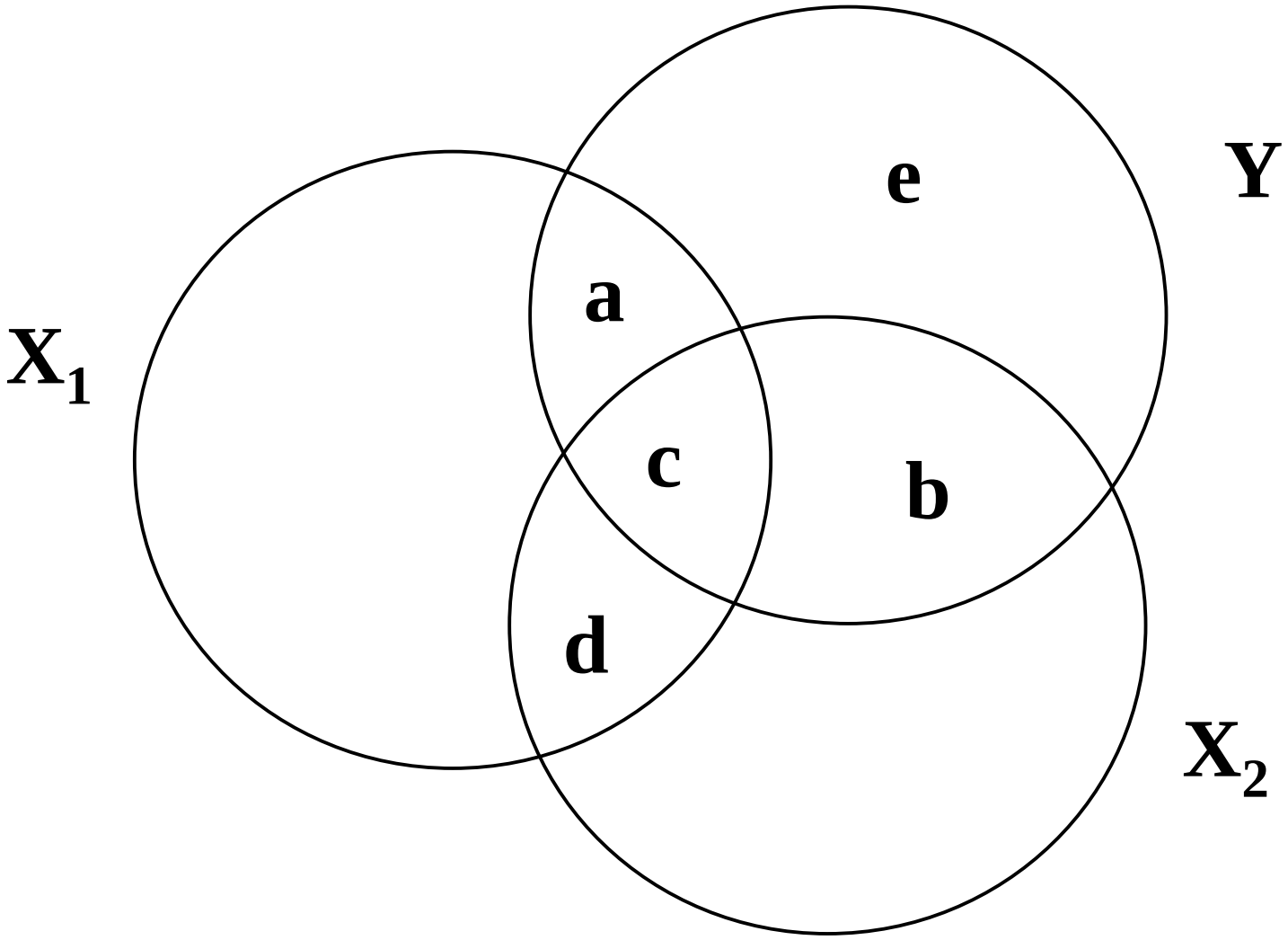
- "a+c": varianza in comune tra X1 e Y, e "a": varianza che Y condivide solo con X1;

- "c+b": varianza in comune tra X2 e Y, e "b": che Y condivide solo con X2;

- "c+d": varianza in comune tra X1 e X2;

- "e" var. che Y non condivide né con X1 né con X2;

Relazioni tra una VD Y e due VI X1 e X2



Coefficienti che misurano l'associazione tra VD e VI.

1. Coefficiente di Correlazione Semi-parziale: corr. tra X1 e Y, se X2 viene parzializzata solo da X1.

$$sr_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{1 - r_{12}^2}} \qquad sr^2_{y1.2...k} = R^2_{y.12..i...k} - R^2_{y.12..(i)...k}$$

$$sr^2_{y1.2} = a/(a+c+b+e)$$

Proporzione della varianza totale di Y spiegata unicamente da X1 al netto di X2,

$$F_i = \frac{sr^2_{y1.2...k}}{(1 - R^2)/df_{res}}, df = (1, N - k - 1)$$

Coefficienti che misurano l'associazione tra VD e VI.

2. Coefficiente di Correlazione Parziale: corr. tra X1 e Y, se X2 viene parzializzata da X1 e da Y.

$$pr_{y1.2} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}}$$

$$pr^2_{y1.2} = a/(a+e)$$

Proporzione della varianza di Y non spiegata da X2, spiegata unicamente da X1 al netto di X2.

Formula alternativa:

$$pr^2_{y1.2....k} = \frac{sr^2_{y1.2...k}}{1 - R^2_{y.12..(i)...k}}$$

Coefficienti che misurano l'associazione tra VD e VI.

3. Coefficiente di Regressione:
Inclinazione della retta di regressione di Y su X_1 per valori costanti di X_2, cambiamento atteso in Y in seguito ad un cambiamento di una unità (b) o di una deviazione standard (b^) in X_1 al netto di X_2.

$$b_{y1.2} = \frac{b_{y1} - b_{y2} b_{12}}{1 - r_{12}^2} \qquad \beta_{y1.2} \wedge = b_{y1.2} \frac{s_y}{s_1} = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2}$$

b_{y1}, b_{y2}, b_{12}: coefficienti delle regressioni bivariate rispettivamente di Y su X_1, di Y su X_2 e di X_1 su X_2.

Adeguatezza della equazione di regressione

1) $\Sigma(Y_i - \bar{Y})^2$ devianza totale delle Y_i dalla loro media.

2) $\Sigma(Y_i' - \bar{Y})^2$ devianza di Y_i spiegata dalla regressione. Scarto tra la retta dei minimi quadrati e la media: quanto migliora la previsione di Y per il fatto di conoscere X.

3) $\Sigma(Y_i - Y_i')^2$ è la devianza di Y_i non spiegata dalla regressione. Scarto di Y_i dalla retta dei minimi quadrati: quantità di errore che si commette per predire Y con Y'.

Adeguatezza della equazione di regressione

E' possibile dimostrare che:

$$r^2 = \frac{\sum (Y_i' - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\text{Devianza Spiegata}}{\text{Devianza Totale}}$$

Dividendo i due termini per n:

$$r^2 = \frac{\sum (Y_i' - \bar{Y})^2 / n}{\sum (Y_i - \bar{Y})^2 / n} = \frac{\text{Varianza Spiegata}}{\text{Varianza Totale}}$$

r^2 = <u>coefficiente di determinazione</u> = indice della proporzione della varianza totale di Y che viene spiegata dalla regressione lineare di Y su X.

Adeguatezza della equazione di regressione

$(1-r^2)$ = proporzione della varianza totale di Y che <u>non</u> è spiegata dalla regressione di Y su X.

E' possibile dimostrare infatti che:

$$(1-r^2) = \frac{\sum (Y_i - Y_i')^2}{\sum (Y_i - \bar{Y})^2} = \frac{\text{Devianza Residua}}{\text{Devianza Totale}}$$

$$\sqrt{(1-r^2)} = \underline{\text{coefficiente di alienazione}} =$$

parte di deviazione standard di Y non spiegata dalla regressione

Adeguatezza della equazione di regressione

Da $\sqrt{(1-r^2)}$ è possibile ricavare il coefficiente che rappresenta la dispersione intorno alla retta dei minimi quadrati per ogni valore di X: "errore standard della stima" ed è un indice della precisione della retta di regressione

$$S_e = \sqrt{(1-r^2)}S_y = \sqrt{\frac{\sum(Y - Y')^2}{n-2}}$$

Se r = 0, $S_e = S_y$ e la varianza d'errore coincide con la varianza totale di Y;
Se r = 1 $S_e = 0$ tutti gli Y cadono sulla retta di regressione Y', quindi l'errore è uguale a 0.

Varianza spiegata nella regressione multipla

Coefficiente di determinazione multiplo (R^2): indica la proporzione di varianza della VD spiegata dalle VI prese nel loro complesso.

$$R^2_{y.12..k} = \sum r_{yi}\hat{\beta}_{yi}$$

Nel caso di due variabili indipendenti la formula è:

$$R^2_{y.12} = r_{y1}\hat{\beta}_{y1} + r_{y2}\hat{\beta}_{y2}$$

Somma dei prodotti delle correlazioni semplici (o "di ordine zero") e dei coefficienti β^ tra VD e ogni VI.

Varianza spiegata nella regressione multipla

R^2 non diminuisce mai se si aggiungono altre VI. Correzione per il numero di VI: coefficiente corretto (Adjusted, o Shrunken).

$$AR^2 = R^2 - (1-R^2)*(k/(N-k-1))$$

Può diminuire rispetto a R^2 se le VI aggiunte forniscono un contributo mediocre alla spiegazione della varianza della VD.

Coefficiente di correlazione multiplo (R o RM): associazione tra una VD e un insieme di VI.

Coefficiente di correlazione multiplo:

$$R_{y.12\ldots k} = \sqrt{R^2_{y.12\ldots k}}$$

R è sempre maggiore/uguale a 0, ed è maggiore dei singoli coefficienti di ordine zero.

VI molto correlate: R vicino al più elevato coefficiente di correlazione semplice tra le VI e la VD.

VI poco correlate: R più elevato del più grande dei coefficienti di correlazione di ordine zero.

Verifica delle ipotesi (test di significatività)

Significatività statistica di R^2

Ipotesi statistiche: H_0: $r = 0$; H_1: $r > 0$

(equivale a H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$)

Varianza	Somme dei quadrati	Gradi di Libertà	Stime della Varianza	F
Totale	Σy^2	N-1		
Spiegata	$R^2\Sigma y^2$	k	$\frac{R^2\Sigma y^2}{K}$	$\frac{(N-k-1)R^2}{k(1-R^2)}$
Non Spiegata	$(1-R^2)\Sigma y^2$	N-k-1	$\frac{(1-R^2)\Sigma y^2}{(N-k-1)}$	

dove $y=(Y - \bar{Y})$ e k è il numero di VI.

Verifica delle ipotesi (test di significatività)

Significatività statistica dei singoli b:

$$H_0: b = 0; H_1: b \neq 0$$

$$t = (b - 0)/S_b, \text{ con } N-k-1 \text{ gradi di libertà.}$$

Stima dell'errore standard di β:

$$S_b = \frac{s_y}{s_i}\sqrt{\frac{1-R_Y^2}{N-k-1}}\sqrt{\frac{1}{1-R_i^2}} = \sqrt{\frac{S_e^2}{S_i^2(1-R_i^2)}}$$

Assunzioni alla base della regressione multipla

1. Assenza di errore di specificazione
   a. Relazione tra le X_i e Y lineare
   b. Non sono state omesse VI rilevanti
   c. Non sono state incluse VI irrilevanti
2. Assenza di errore di misurazione: variabili misurate senza errore
3. VI quantitative o dicotomiche, VD quantitativa
4. Varianza della VI è > 0
5. Campionamento casuale
6. Nessuna VI è combinazione lineare perfetta delle altre (assenza di perfetta multicollinearità)

Assunzioni alla base della regressione multipla

7. Assunzioni sui residui (o termini di errore) ε_i

a. Media uguale a zero: $E(\varepsilon_i)=0$

b. Omoschedasticità, $VAR(\varepsilon_i)=s2$

c. Assenza di autocorrelazione: $Cov(\varepsilon_i, \varepsilon_j)=0$

d. VI non correlate con gli errori: $Cov(\varepsilon_i, X_i)=0$

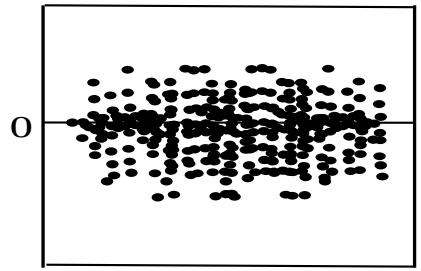
e. Normalità: Le distribuzioni dei valori di ε_i per ogni valore dato di X sono di forma normale

Violazione delle assunzioni:

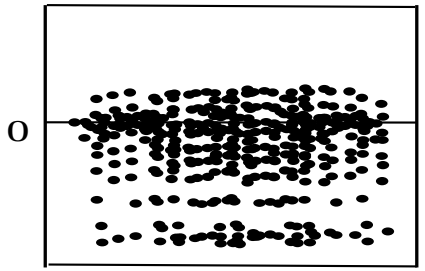
Esame della distribuzione dei residui e=(Y-Y') rispetto ai punteggi teorici Y'.

Utile per rilevare:

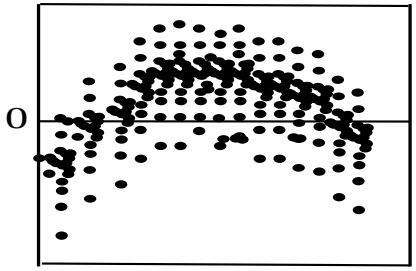
- **La non linearità della relazione tra VI e VD, e tra VI,**
- **La non omogeneità della varianza**
- **La non normalità dei residui**



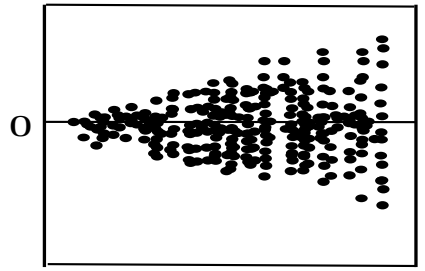
1.Assunzioni rispettate



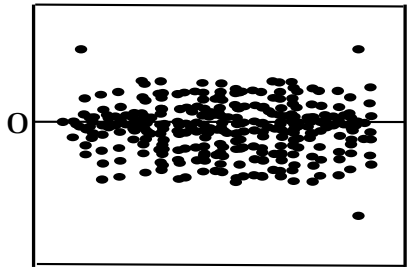
2.Non normalità



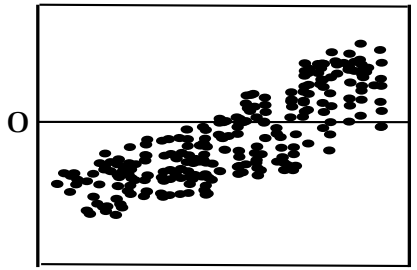
3.Non linearità



4.Eteroschedasticità



5. Casi estremi



6. Autocorrelazione

Nei riquadri 1-5: Punteggi predetti Y': in ascisse; Residui (Y-Y'): in ordinate.
Nel riquadro 6: Tempo o ordine di acquisizione: in ascisse; Residui (Y-Y'): in ordinate.

Rilevare la collinearità (correlazione elevata tra le VI):

- **Correlazioni tra le VI (se sono >.8);**
- **R^2 elevati e b bassi;**
- **Errori standard elevati;**
- **Indici di tolleranza e VIF.**

Tolleranza di una VI: quantità di varianza che *non* è spiegata dalle altre VI: $T_i = (1 - R_i^2)$
valori bassi di tolleranza indicano alta collinearità, valori alti bassa collinearità.

Variance Inflaction Factor (VIF): $VIF_i = 1/T_i = 1/(1 - R_i^2)$;

valori bassi del VIF indicano bassa collinearità, valori alti elevata collinearità.

Non indipendenza degli errori (Autocorrelazione):

Test di Durbin-Watson.

Ha un valore compreso tra 0 e 4: se i residui di osservazioni consecutive non sono correlati il test di Durbin-Watson ha un valore intorno a 2.

Se n≥ 100 e le VI almeno 2, valori compresi tra 1.5 e 2.2 possono essere considerati indicativi di assenza di autocorrelazione, quindi:
Valori inferiori a 1.5 = autocorrelazione positiva.
Valori superiori a 2.2 =autocorrelazione negativa.

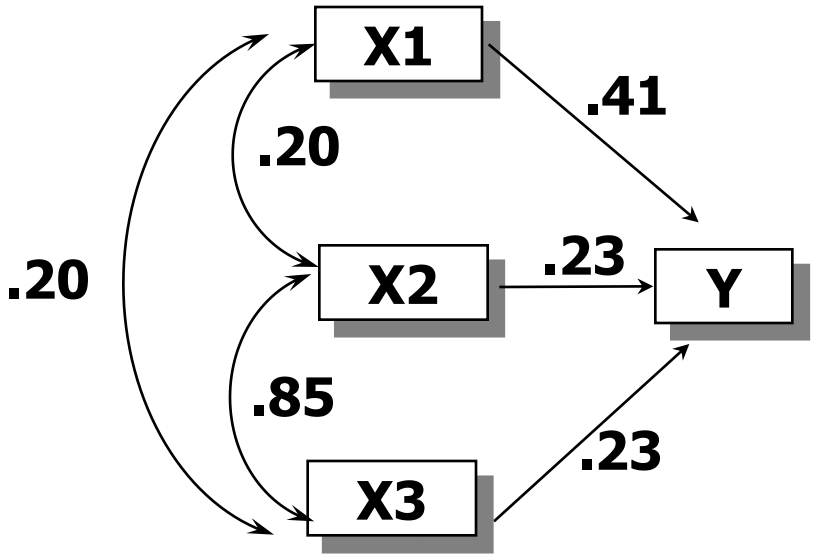
Rimedi per risolvere le violazioni: trasformazione delle variabili originali (logaritmo, reciproco, radice quadr.).

Scomposizione degli effetti

La *ridondanza* riguarda il caso in cui i coefficienti di correlazione semiparziale (sr), parziale (pr) e di regressione standardizzato (β) sono inferiori (in valore assoluto) al coefficiente di correlazione semplice r e hanno il suo stesso segno.

Allora ogni variabile indipendente porta un'informazione sulla variabile dipendente che in parte si sovrappone con quella veicolata dalle altre variabili indipendenti.

Ridondanza



r(X1,Y)=r(X2,Y)=r(X3,Y)=.50

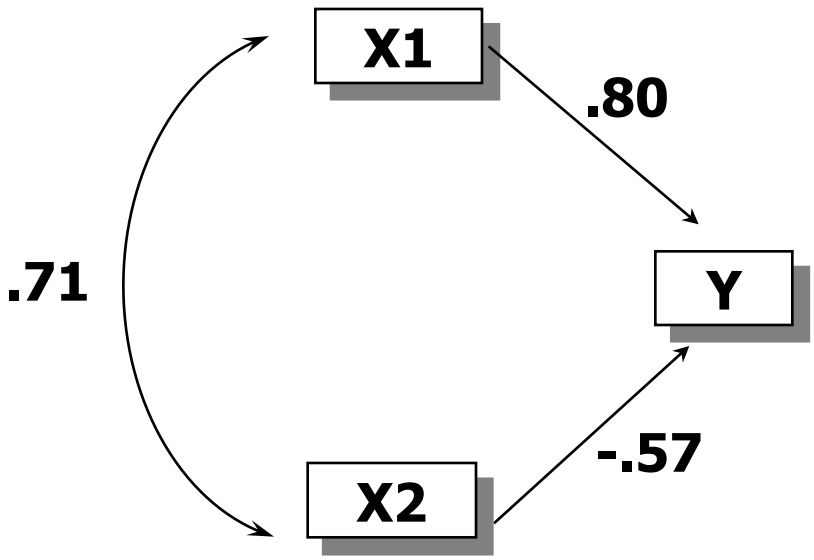
Scomposizione degli effetti

La *soppressione* riguarda il caso in cui i coefficienti sr, pr e β sono maggiori (in valore assoluto) del coefficiente di correlazione semplice r.

Il termine soppressione indica che la relazione tra le variabili indipendenti "maschera" o "sopprime" la loro reale relazione con la variabile dipendente, che potrebbe essere maggiore o addirittura di segno opposto se le variabili indipendenti *non* fossero correlate. Il soppressore è una VI la cui inclusione nella regressione aumenta l'effetto di un'altra VI sulla VD.

Un caso particolare di soppressione è il ribaltamento, dove il coefficiente parziale assume il segno opposto del coefficiente semplice.

Soppressione

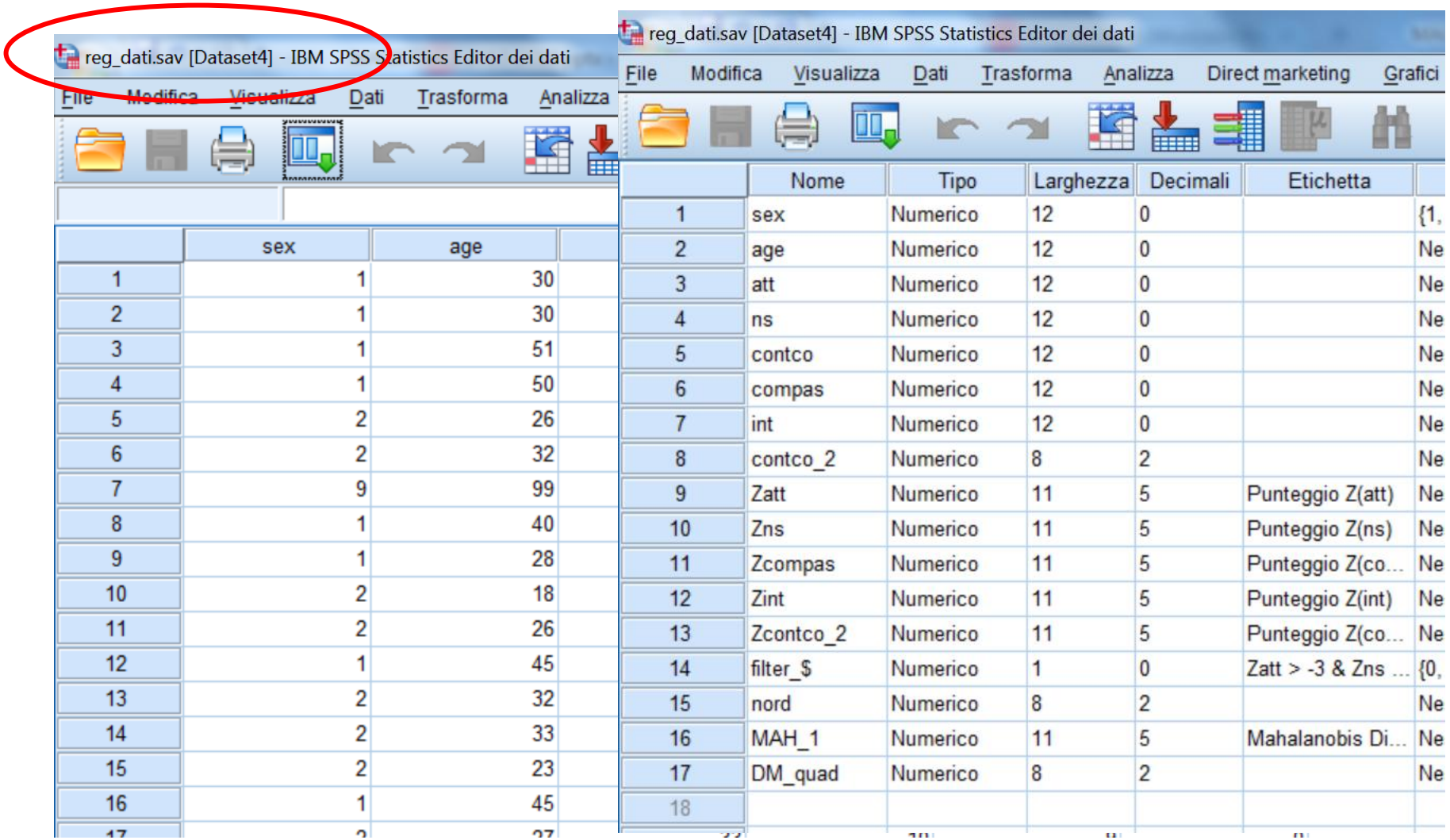


r(X1,Y)=.40; r(X2,Y)=0

REGRESSIONE

CON SPSS

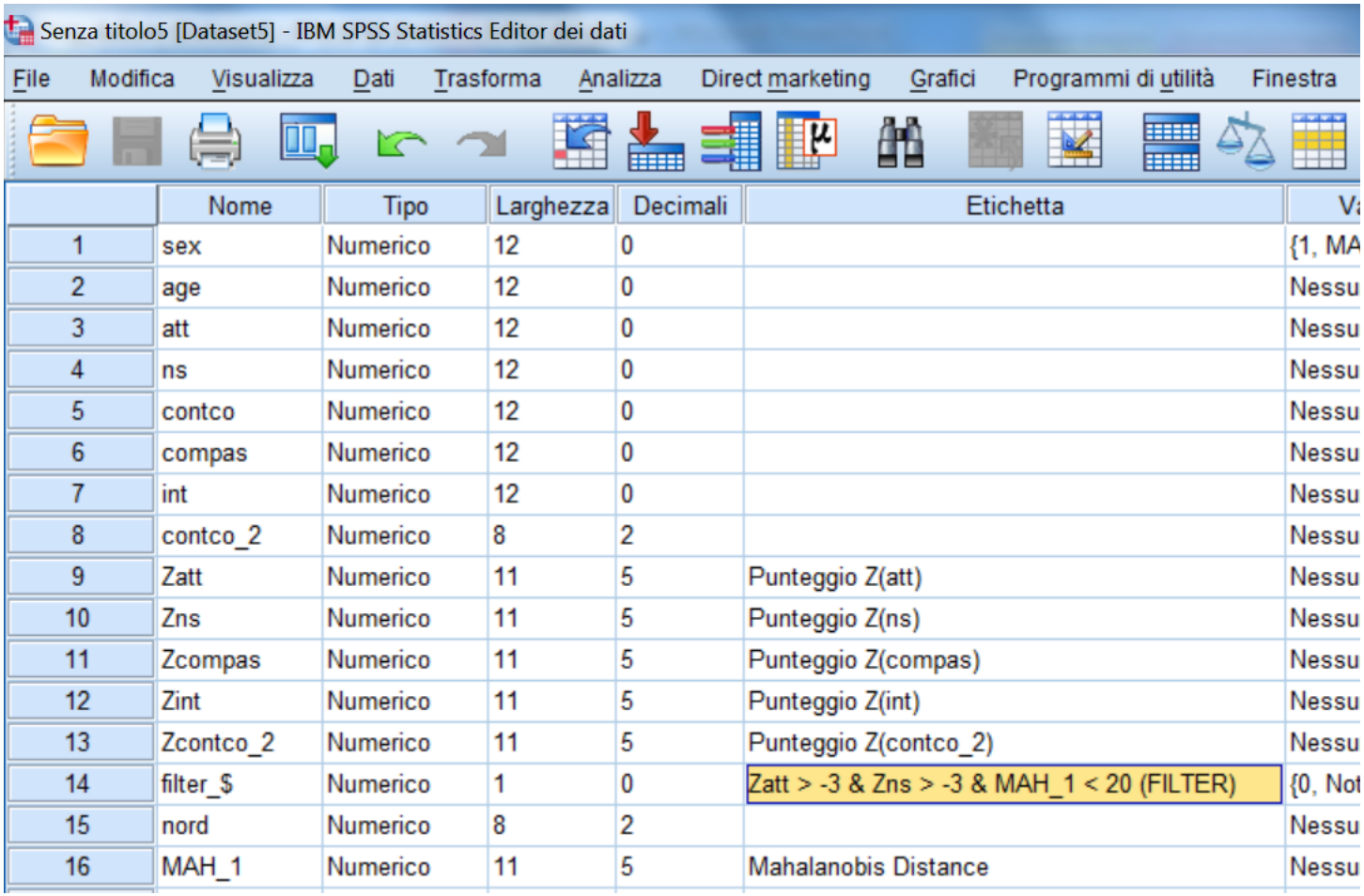
Carichiamo i dati utilizzati per l'esempio sul trattamento preliminari (rinominare il file come reg_dati.sav).



	sex	age
1	1	30
2	1	30
3	1	51
4	1	50
5	2	26
6	2	32
7	9	99
8	1	40
9	1	28
10	2	18
11	2	26
12	1	45
13	2	32
14	2	33
15	2	23
16	1	45

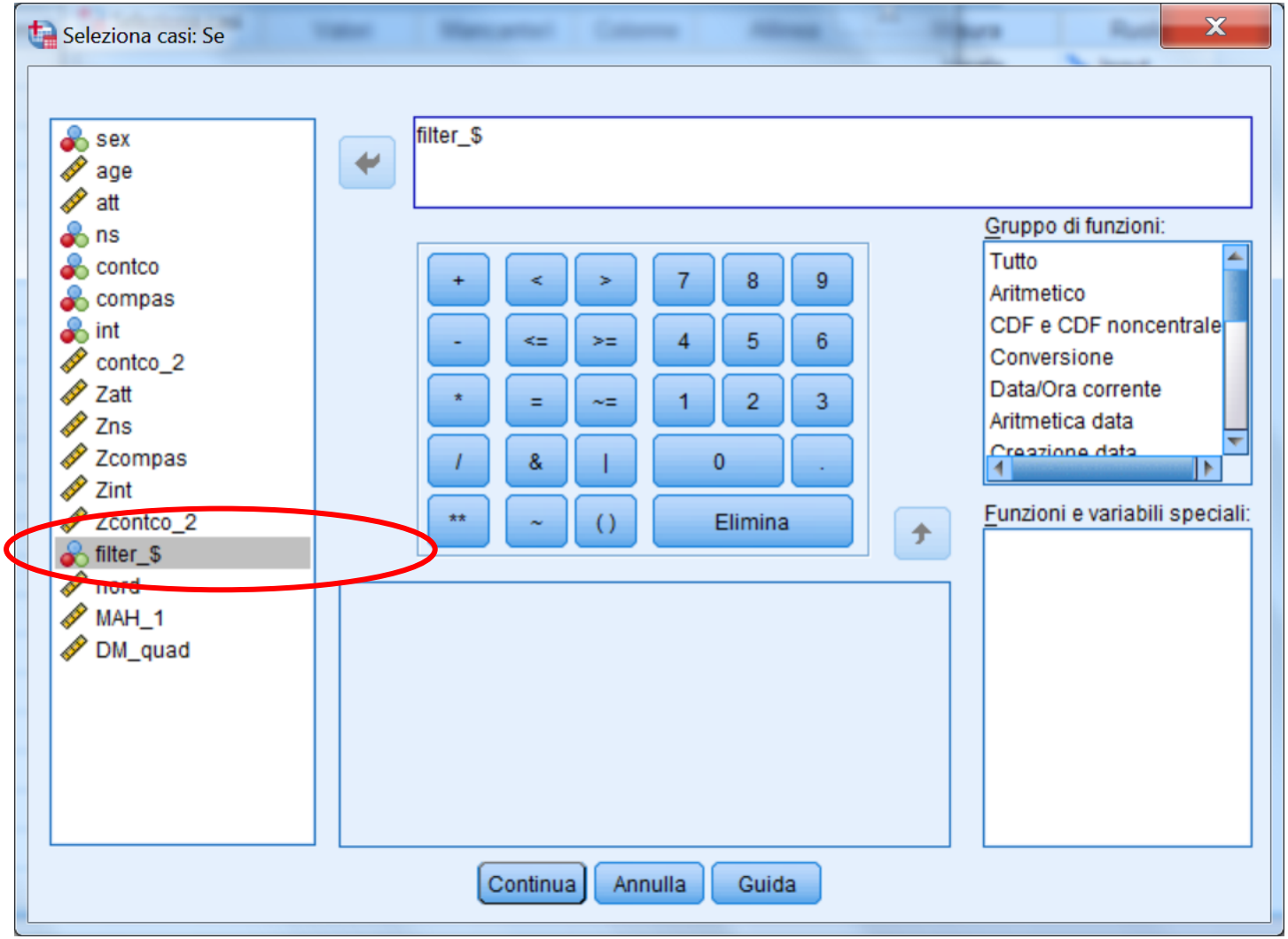
	Nome	Tipo	Larghezza	Decimali	Etichetta	
1	sex	Numerico	12	0		{1,
2	age	Numerico	12	0		Ne
3	att	Numerico	12	0		Ne
4	ns	Numerico	12	0		Ne
5	contco	Numerico	12	0		Ne
6	compas	Numerico	12	0		Ne
7	int	Numerico	12	0		Ne
8	contco_2	Numerico	8	2		Ne
9	Zatt	Numerico	11	5	Punteggio Z(att)	Ne
10	Zns	Numerico	11	5	Punteggio Z(ns)	Ne
11	Zcompas	Numerico	11	5	Punteggio Z(co...	Ne
12	Zint	Numerico	11	5	Punteggio Z(int)	Ne
13	Zcontco_2	Numerico	11	5	Punteggio Z(co...	Ne
14	filter_$	Numerico	1	0	Zatt > -3 & Zns ...	{0,
15	nord	Numerico	8	2		Ne
16	MAH_1	Numerico	11	5	Mahalanobis Di...	Ne
17	DM_quad	Numerico	8	2		Ne
18						

Riattiviamo il filtro per i 3 outliers



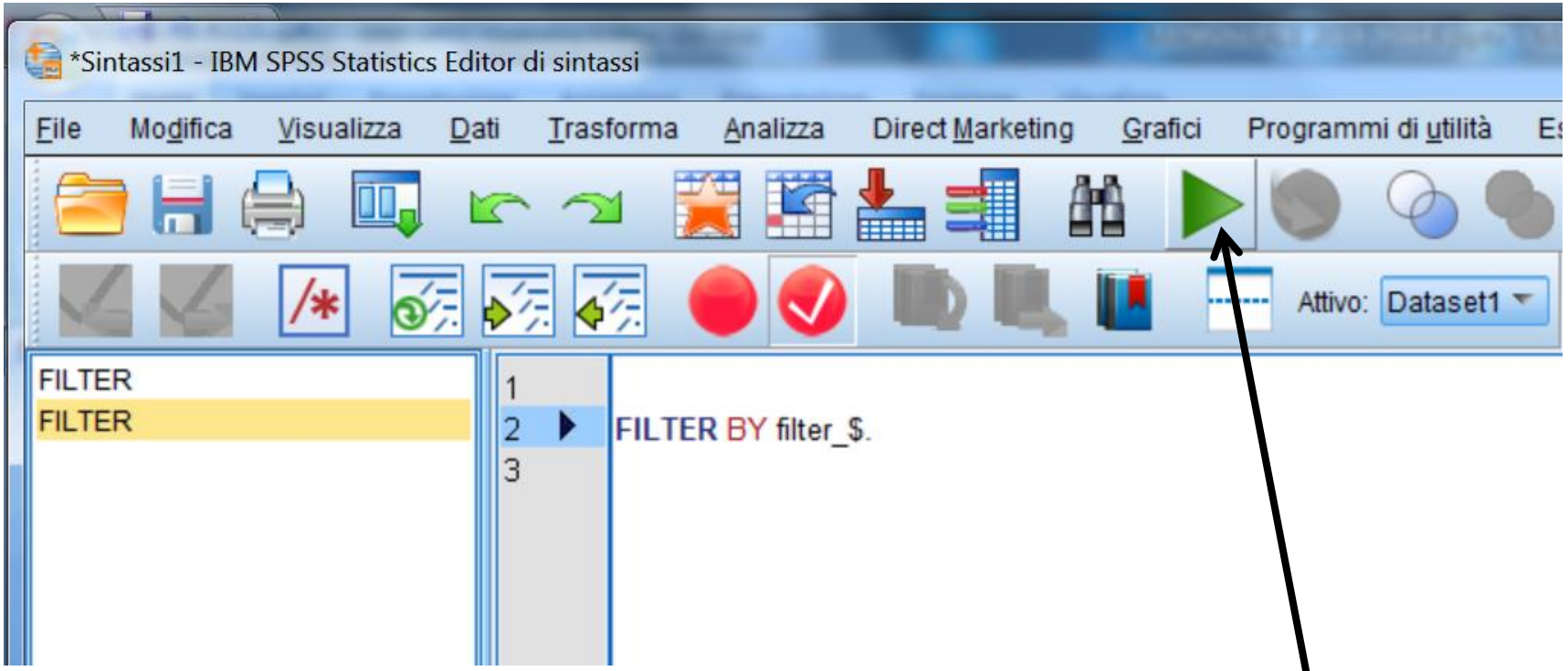
	Nome	Tipo	Larghezza	Decimali	Etichetta	Va
1	sex	Numerico	12	0		{1, MA
2	age	Numerico	12	0		Nessu
3	att	Numerico	12	0		Nessu
4	ns	Numerico	12	0		Nessu
5	contco	Numerico	12	0		Nessu
6	compas	Numerico	12	0		Nessu
7	int	Numerico	12	0		Nessu
8	contco_2	Numerico	8	2		Nessu
9	Zatt	Numerico	11	5	Punteggio Z(att)	Nessu
10	Zns	Numerico	11	5	Punteggio Z(ns)	Nessu
11	Zcompas	Numerico	11	5	Punteggio Z(compas)	Nessu
12	Zint	Numerico	11	5	Punteggio Z(int)	Nessu
13	Zcontco_2	Numerico	11	5	Punteggio Z(contco_2)	Nessu
14	filter_$	Numerico	1	0	Zatt > -3 & Zns > -3 & MAH_1 < 20 (FILTER)	{0, Not
15	nord	Numerico	8	2		Nessu
16	MAH_1	Numerico	11	5	Mahalanobis Distance	Nessu

Riattiviamo il filtro per i 3 outliers

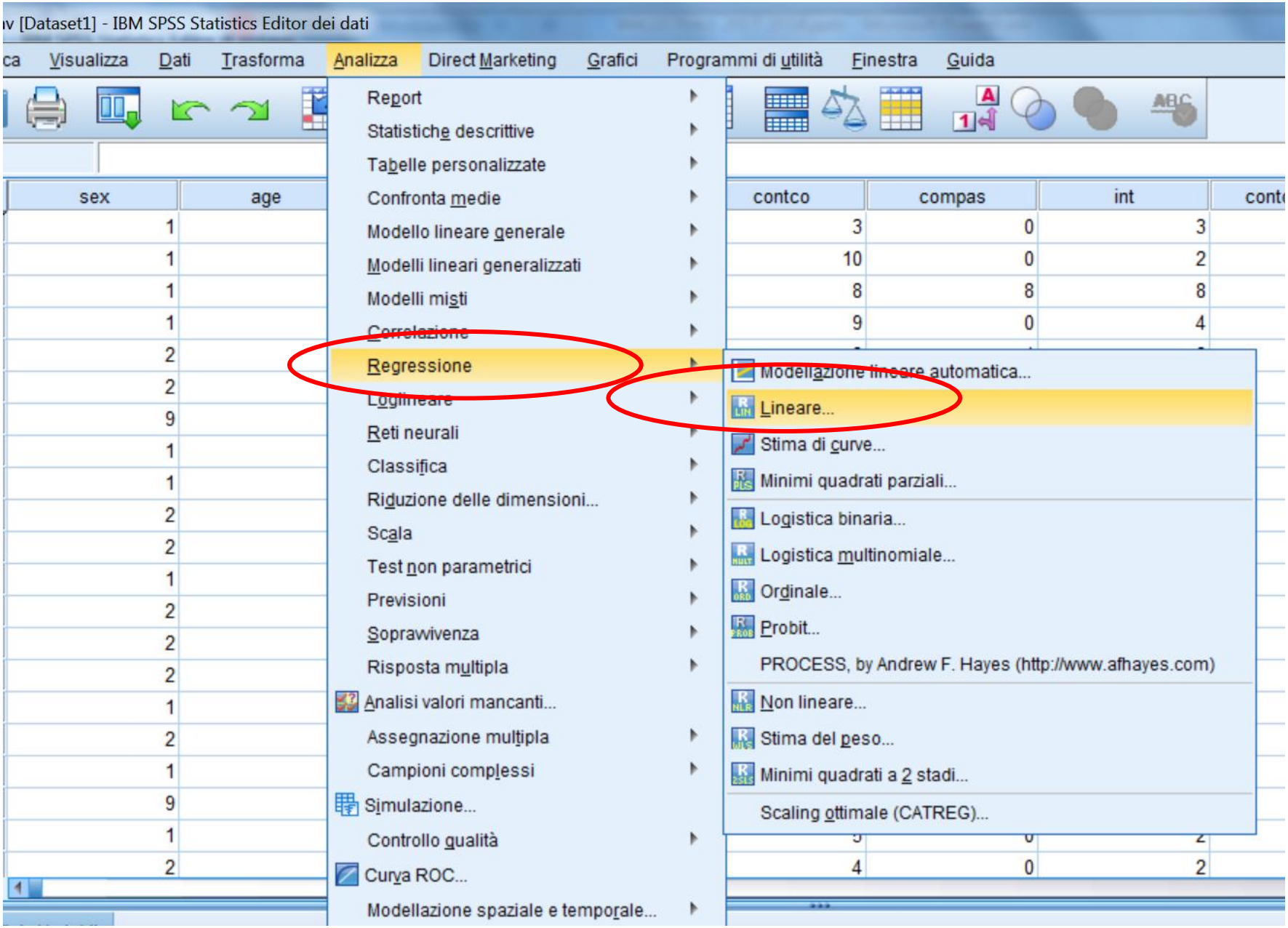


Questa procedura però cancella l'etichetta della variabile filter_$

Riattivare il filtro per i 3 outliers senza cancellare l'etichetta della variabile filter_$

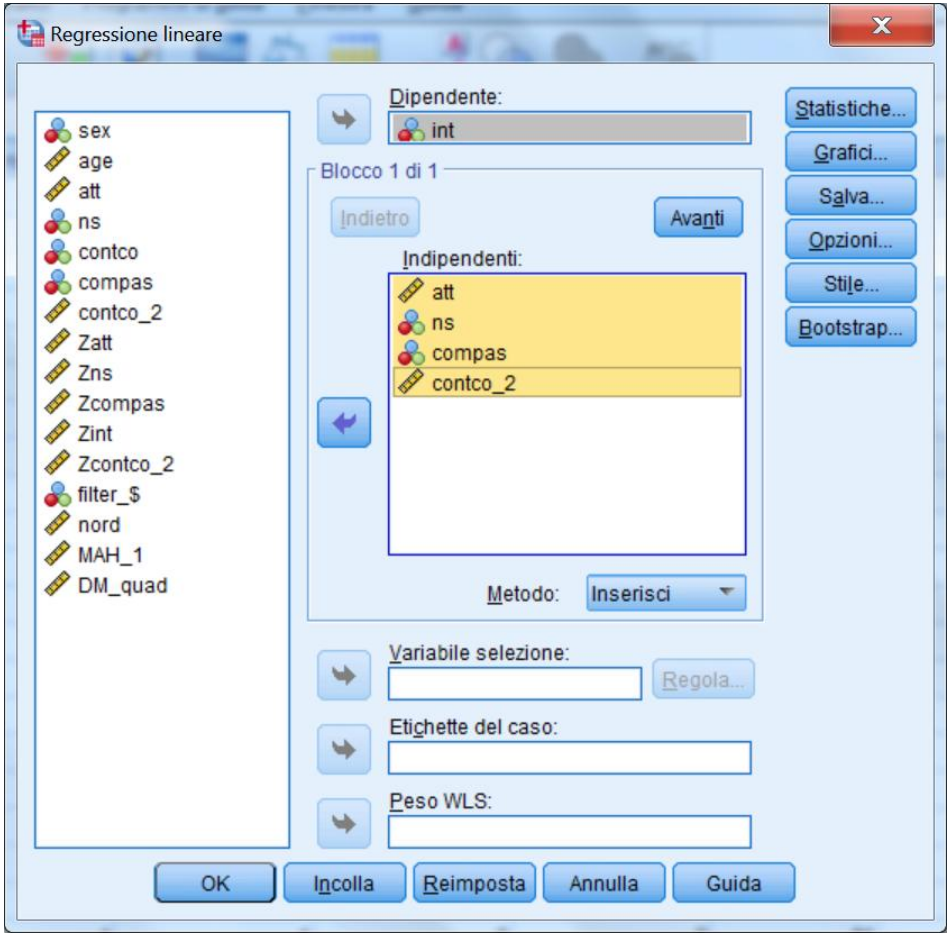


Dalla finestra Sintassi lanciare il comando FILTER BY filter_$ posizionando il cursore sulla linea del comando e cliccando sul triangolino verde.


v [Dataset1] - IBM SPSS Statistics Editor dei dati
ca Visualizza Dati Trasforma Analizza Direct Marketing Grafici Programmi di utilità Finestra Guida
Report
Statistiche descrittive
Tabelle personalizzate
Confronta medie
Modello lineare generale
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Reti neurali
Classifica
Riduzione delle dimensioni...
Scala
Test non parametrici
Previsioni
Sopravvivenza
Risposta multipla
Analisi valori mancanti...
Assegnazione multipla
Campioni complessi
Simulazione...
Controllo qualità
Curva ROC...
Modellazione spaziale e temporale...
Modellazione lineare automatica...
Lineare...
Stima di curve...
Minimi quadrati parziali...
Logistica binaria...
Logistica multinomiale...
Ordinale...
Probit...
PROCESS, by Andrew F. Hayes (http://www.afhayes.com)
Non lineare...
Stima del peso...
Minimi quadrati a 2 stadi...
Scaling ottimale (CATREG)...
sex age contco compas int
1 3 0 3
1 10 0 2
1 8 8 8
1 9 0 4
2
2
9
1
1
2
2
1
2
2
2
1
2
1
9
1
2
4 0 2

Regressione standard

Selezionare la variabile dipendente ("int") e poi tutte le variabili indipendenti ("att", "ns", "contco_2", "compas") che verranno inserite in un unico blocco. Lasciare nell'opzione "Metodo" il valore di default "Inserisci".



Strategie Analitiche per la regressione

Regressione standard:

- Quale è l'entità della relazione globale tra VD e VI?
- Quale è il contributo unico di ciascuna VI nel determinare questa relazione ?

Regressione gerarchica:

- Se la VI X1 è inserita dopo la VI X2, quale contributo aggiuntivo dà alla spiegazione della VD ?

Regressione statistica:

- Quale è la migliore combinazione lineare di VI per predire la VD in un determinato campione ?

La regressione standard

Tutte le VI vengono inserite nell'equazione simultaneamente.

Ogni VI è trattata come se fosse inserita nell'equazione dopo aver preso in considerazione tutte le altre VI.

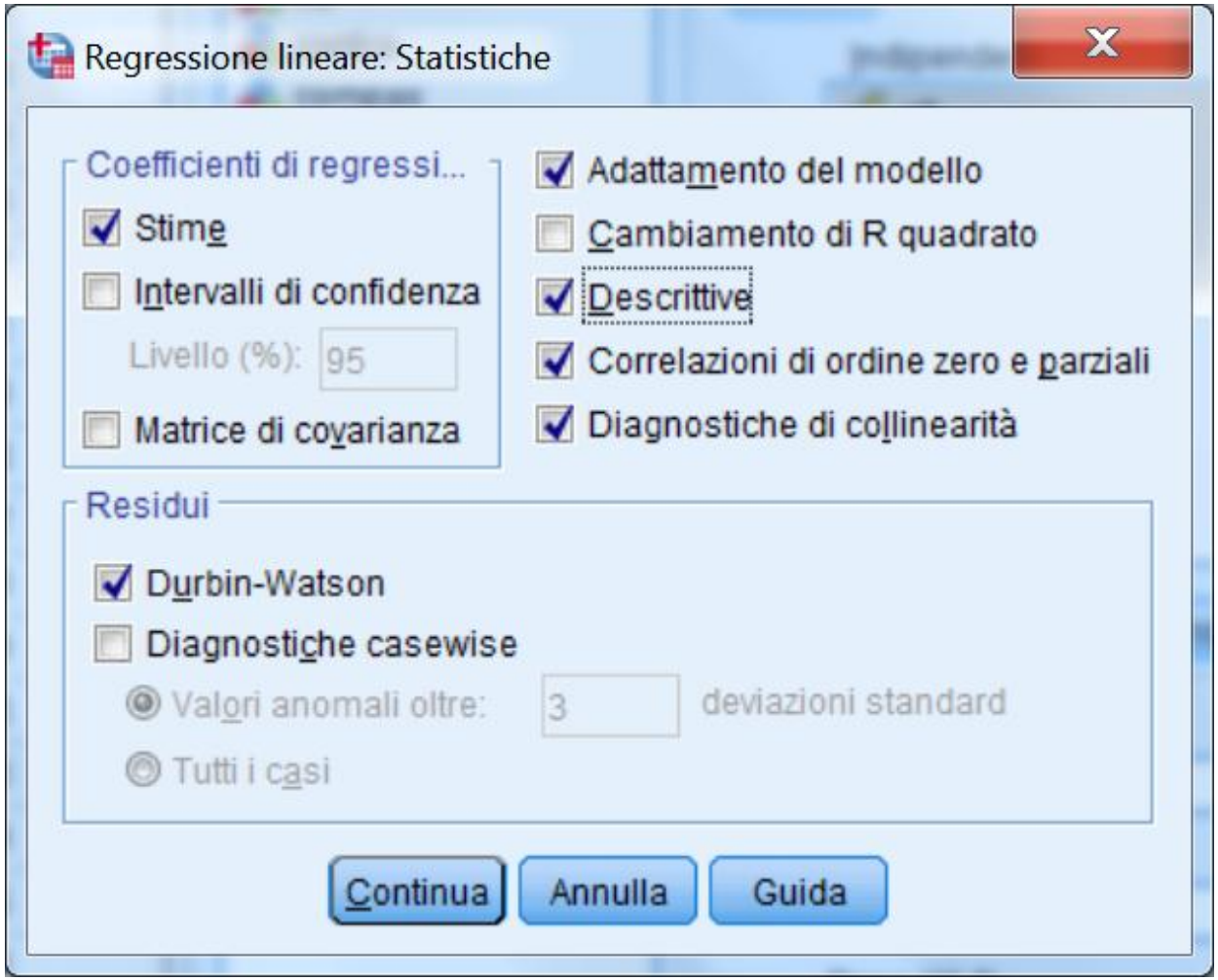
Ogni VI è valutata per quanto aggiunge, nello spiegare la VD, a quanto viene spiegato da tutte le altre VI.

Ogni VI spiega solo quella parte di varianza della VD che condivide unicamente con la VD, al netto delle VI.

La variabilità che la VD condivide simultaneamente con più VI viene ad aggiungersi all'R^2 ma non è assegnata individualmente a nessuna delle VI.

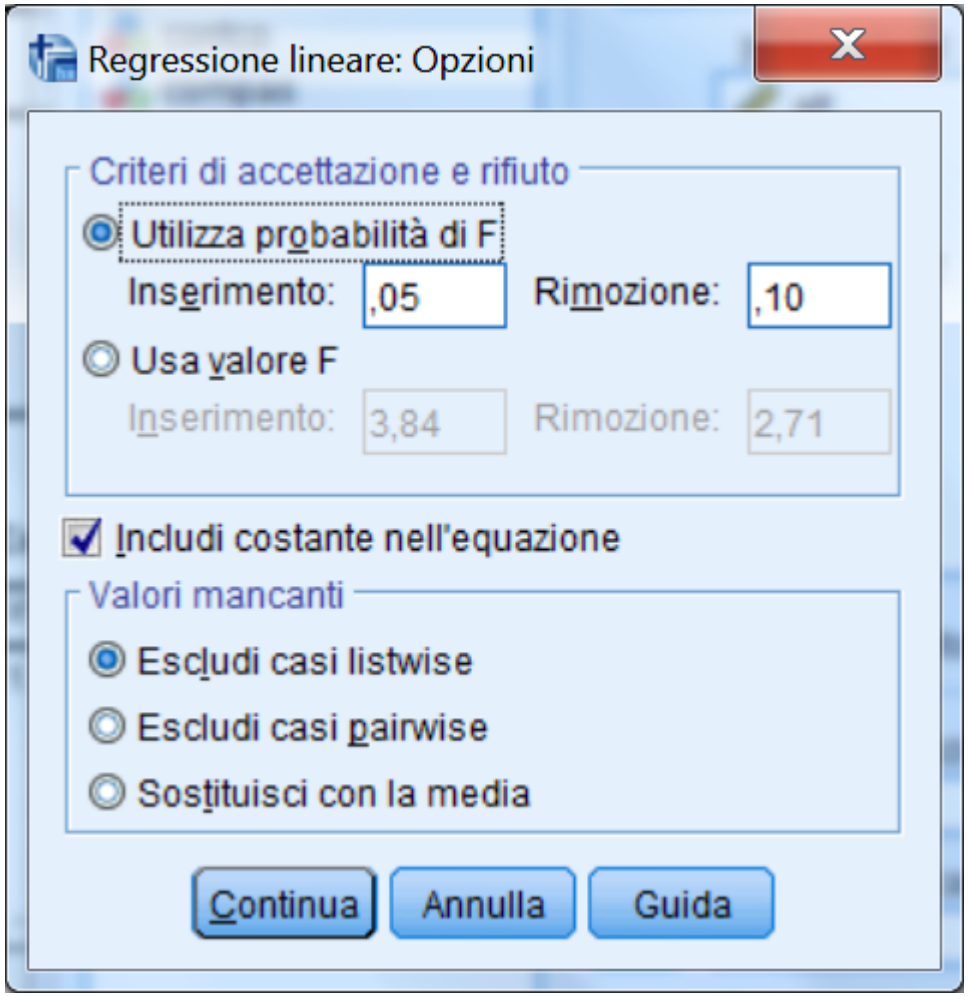
Regressione standard

Nella finestra di dialogo "Statistiche" bisogna selezionare determinati parametri per ottenere nell'output le informazioni necessarie per interpretare e valutare la soluzione.



Regressione standard

Nella finestra di dialogo "Opzioni" vengono presentate le opzioni relative al trattamento dei valori mancanti.



Pairwise

Vengono utilizzati tutti i valori disponibili
Le analisi vengono effettuate considerando tutti i soggetti che hanno valori validi sulle variabili di volta in volta considerate

Listwise

Vengono utilizzati solo quei soggetti che NON hanno alcun valore mancante. È sufficiente che un soggetto presenti un valore mancante in una sola variabile per essere escluso dalle analisi
Per molte procedure è il metodo di *default* di SPSS

Sostituzione con la media

Sostituisce i valori mancanti con la media della variabile nel campione

Statistiche descrittive

Statistica descrittiva

	Media	Deviazione std.	N
int	7,32	2,543	196
att	42,93	7,024	196
ns	7,92	1,758	196
compas	2,67	1,965	196
contco_2	,2545	,28864	196

Correlazioni

		int	att	ns	compas	contco_2
Correlazione di Pearson	int	1,000	,721	,589	,645	-,544
	att	,721	1,000	,556	,520	-,449
	ns	,589	,556	1,000	,454	-,315
	compas	,645	,520	,454	1,000	-,445
	contco_2	-,544	-,449	-,315	-,445	1,000
Sign. (a una coda)	int	.	,000	,000	,000	,000
	att	,000	.	,000	,000	,000
	ns	,000	,000	.	,000	,000
	compas	,000	,000	,000	.	,000
	contco_2	,000	,000	,000	,000	.
N	int	196	196	196	196	196
	att	196	196	196	196	196
	ns	196	196	196	196	196
	compas	196	196	196	196	196
	contco_2	196	196	196	196	196

Regressione standard

Il pannello iniziale evidenzia come che tutte le variabili siano state inserite in un unico passo

Variabili immesse/rimosse[a]

Modello	Variabili immesse	Variabili rimosse	Metodo
1	contco_2, ns, compas, att[b]	.	Inserisci

a. Variabile dipendente: int

b. Sono state immesse tutte le variabili richieste.

La varianza spiegata si trova in questa tabella

Riepilogo del modello[b]

Modello	R	R-quadrato	R-quadrato adattato	Errore std. della stima	Durbin-Watson
1	,819[a]	,671	,664	1,474	1,709

a. Predittori: (costante), contco_2, ns, compas, att

b. Variabile dipendente: int

ANOVA[a]

Modello		Somma dei quadrati	gl	Media quadratica	F	Sign.
1	Regressione	845,599	4	211,400	97,259	,000[b]
	Residuo	415,151	191	2,174		
	Totale	1260,750	195			

a. Variabile dipendente: int

b. Predittori: (costante), contco_2, ns, compas, att

Regressione standard

Per interpretare gli effetti delle VI guardare questa tabella

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sign.
		B	Errore std.	Beta		
1	(Costante)	-1,422	,816		-1,742	,083
	att	,141	,020	,390	7,045	,000
	ns	,273	,074	,189	3,676	,000
	compas	,354	,067	,273	5,274	,000
	contco_2	-1,656	,426	-,188	-3,885	,000

a. Variabile dipendente: int

Correlazioni			Statistiche di collinearità	
Ordine zero	Parziale	Parte	Tolleranza	VIF
,721	,454	,293	,563	1,775
,589	,257	,153	,653	1,530
,645	,357	,219	,642	1,557
-,544	-,271	-,161	,737	1,358

Risultati della regressione standard

sr^2 = contributo unico della VI all'R^2 nell'insieme di VI.

Somma degli sr^2: può non raggiungere il valore di R^2.

Differenza tra somma degli sr^2 e R^2: proporzione di varianza della VD spiegata simultaneamente da più VI, ma non attribuita a nessuna VI in particolare.

Dati dell'esempio:

$$\Sigma sr^2 = (.29)^2+(.15)^2+(.22)^2+(-.16)^2=.183;\ R^2 = .671;$$

$$R^2 - \Sigma sr^2 = .67 - .183 = .488$$

E' la varianza spiegata simultaneamente dalle VI

Regressione standard

Varianza unica e varianza comune spiegata dalla VI

	varianza unica	
	sr	sr^2
att	,293	0,086
ns	,153	0,023
compas	,219	0,048
contco_2	-,161	0,026
Varianza totale spiegata		0,671
Varianza unica spiegata		0,183
Varianza comune spiegata		0,488

La regressione gerarchica

Le VI vengono inserite nell'equazione secondo un ordine specificato dal ricercatore.

L'ordine di "entrata" viene assegnato dal ricercatore secondo considerazioni teoriche o logiche.

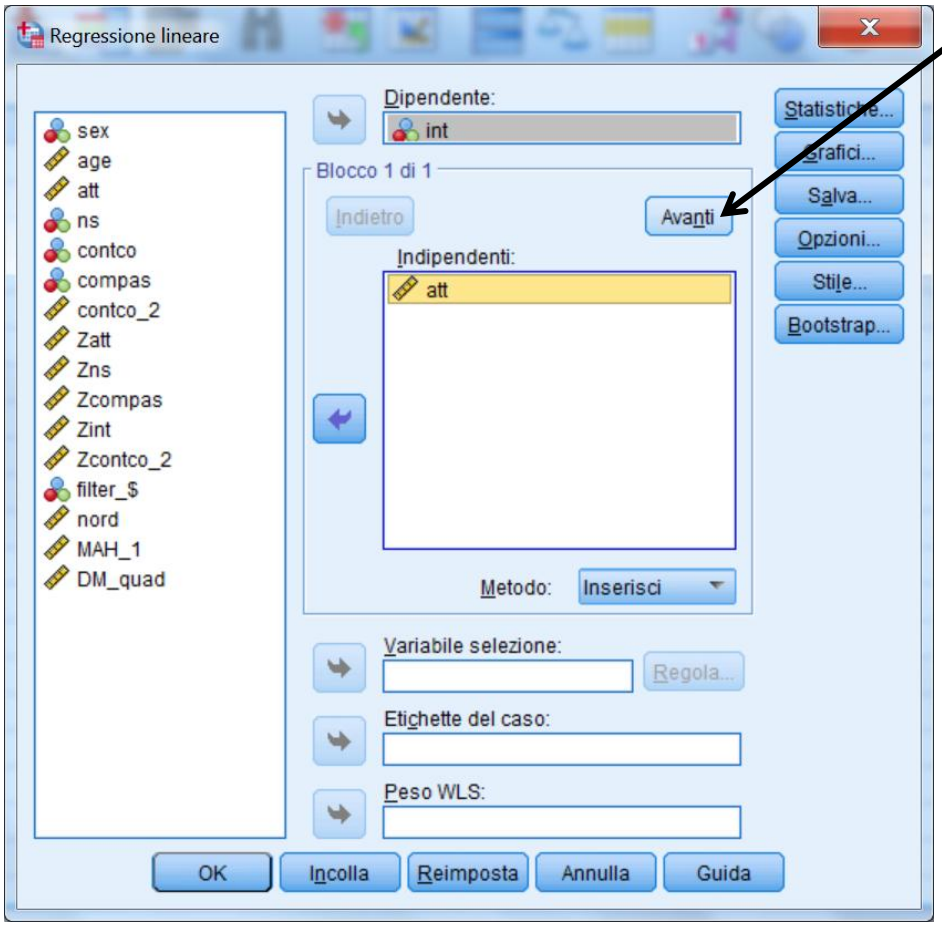
L'analisi procede attraverso "passi" sequenziali. Ogni VI è valutata per quanto aggiunge, nello spiegare la VD, rispetto a quanto è stato spiegato dalle VI inserite precedentemente. Partizione ordinata della varianza di VD spiegata dalle VI.

Contributo di una VI: può variare se la sua posizione nella gerarchia viene cambiata

Regressione gerarchica

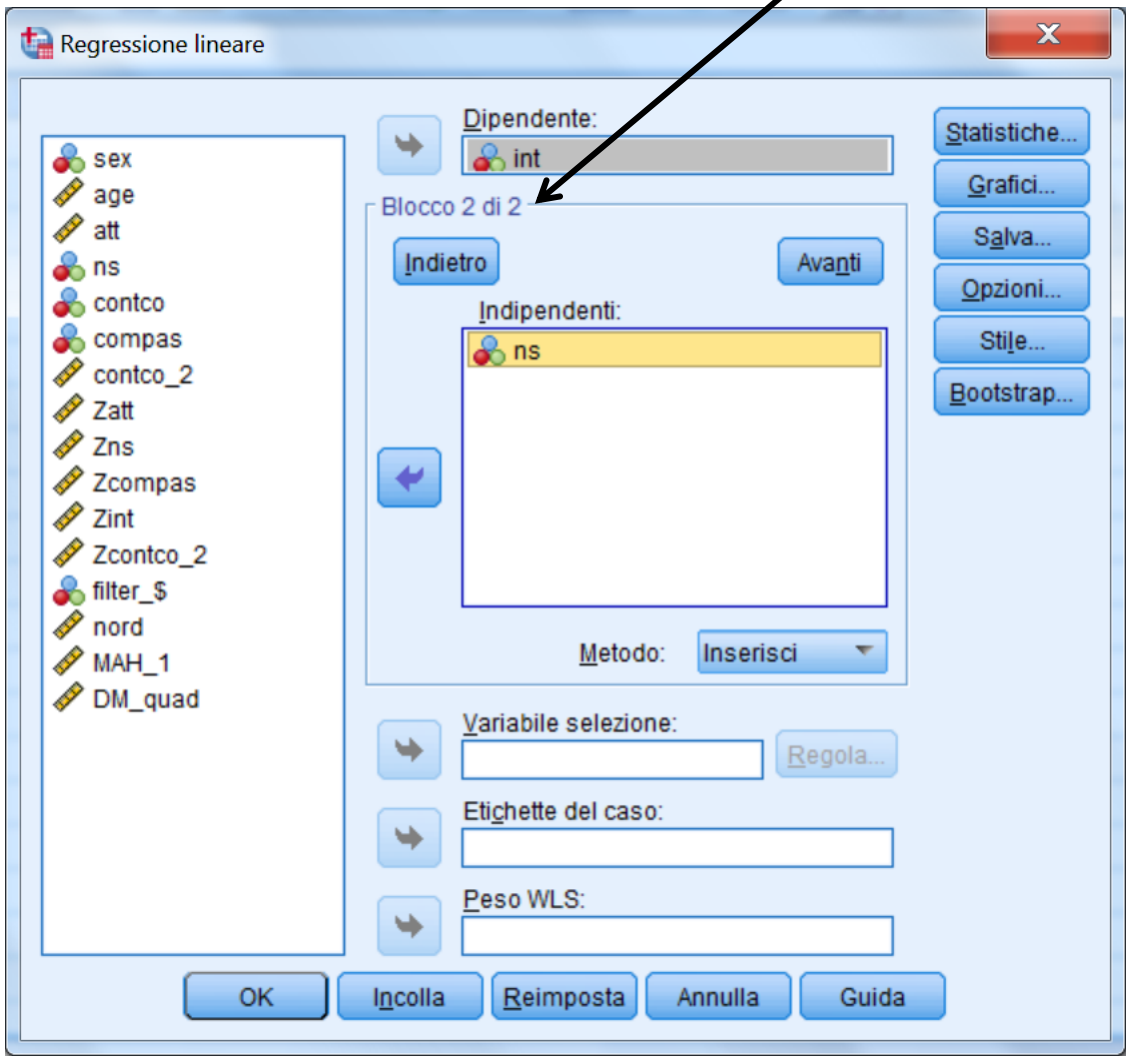
Selezionare la variabile dipendente ("int"). Quindi tutte le variabili indipendenti verranno inserite in blocchi separati, secondo un ordine consistente con il modello teorico che il ricercatore vuole esaminare.

Inserita la prima variabile ("att") cliccare sul pulsante "Avanti"



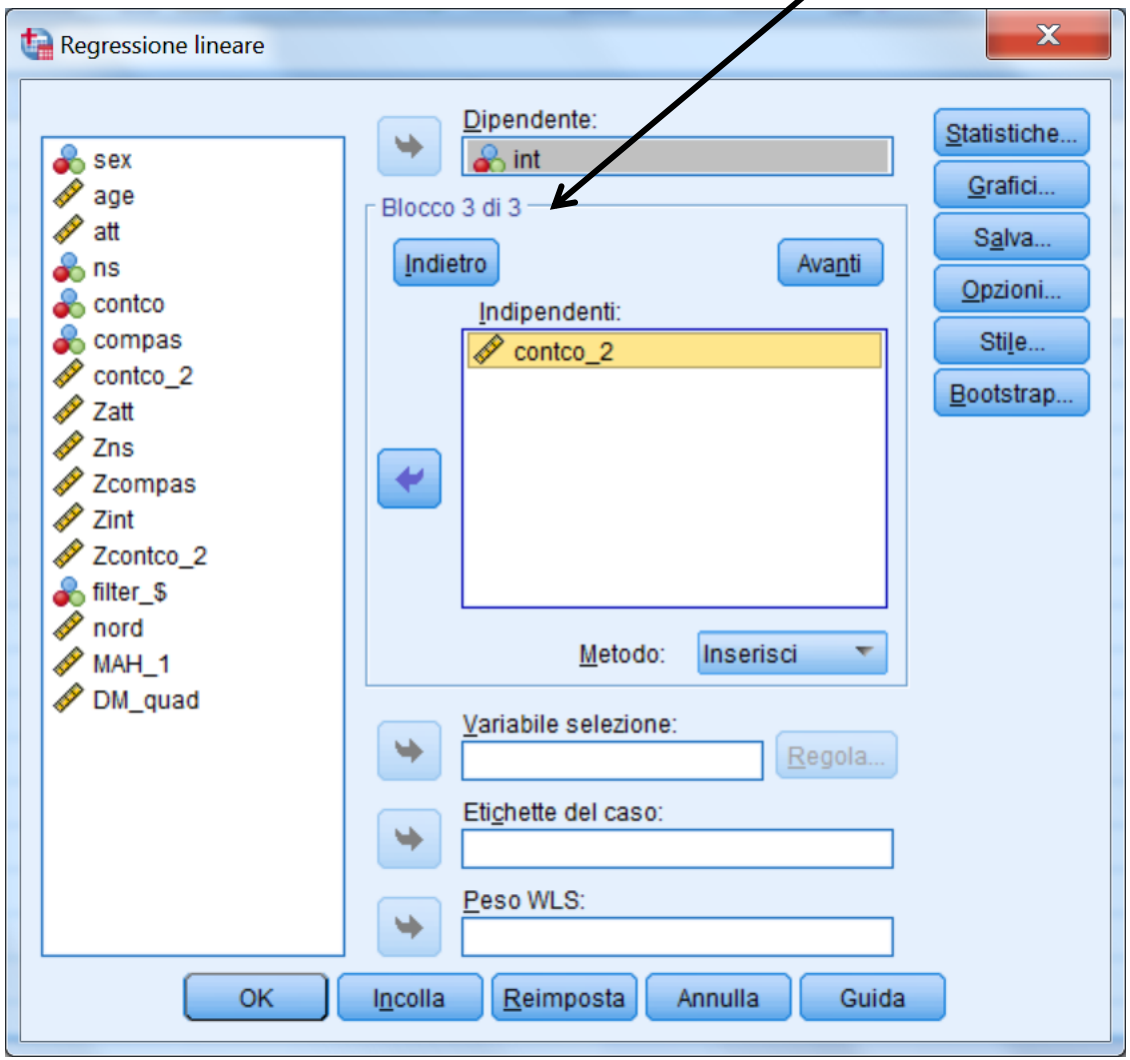
Regressione gerarchica

Inserire la seconda variabile nel "Blocco 2 di 2" ("ns") e di nuovo cliccare sul pulsante "Avanti"



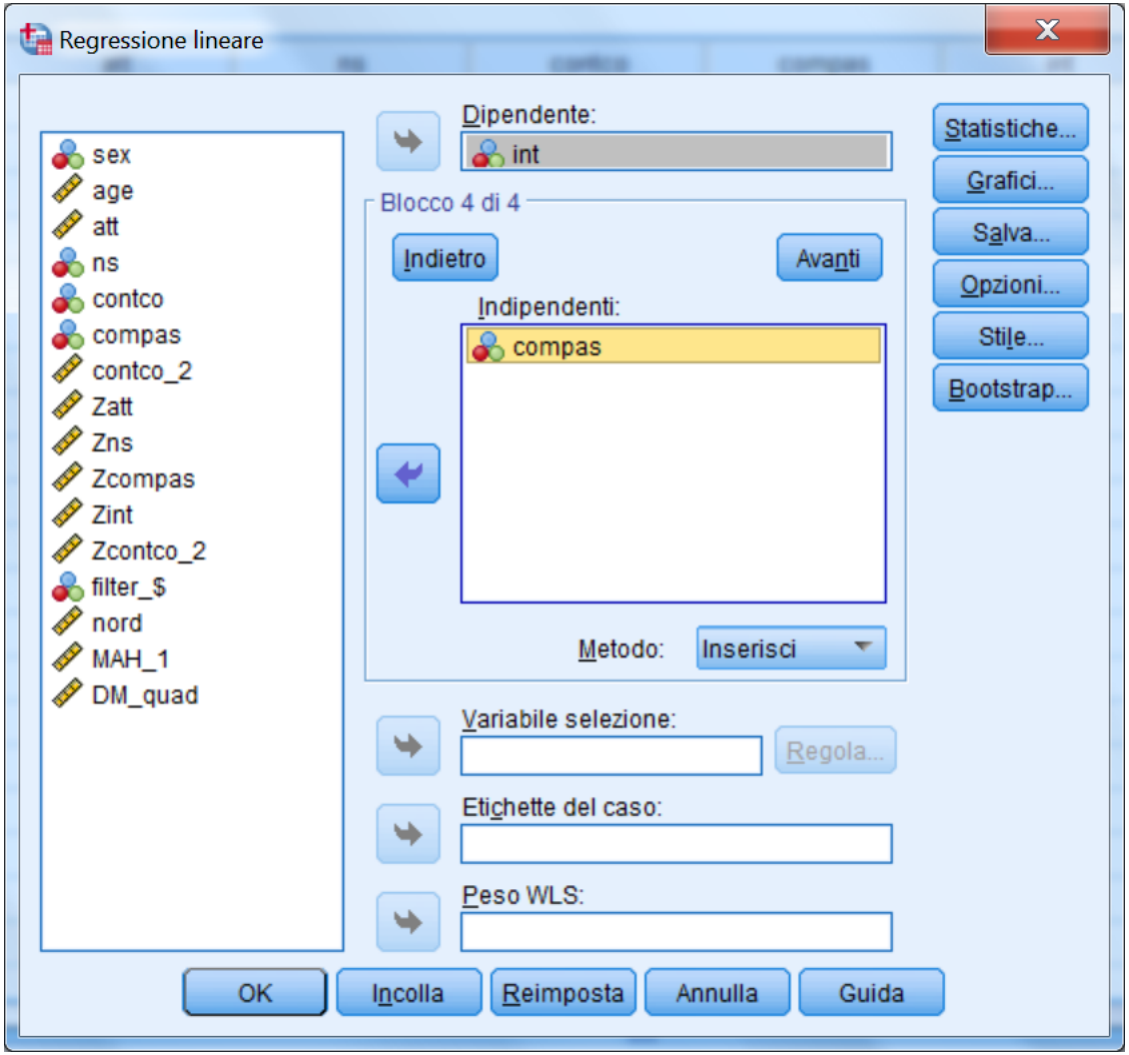
Regressione gerarchica

Inserire la terza variabile nel "Blocco 3 di 3" ("contco_2") e di nuovo cliccare sul pulsante "Avanti"



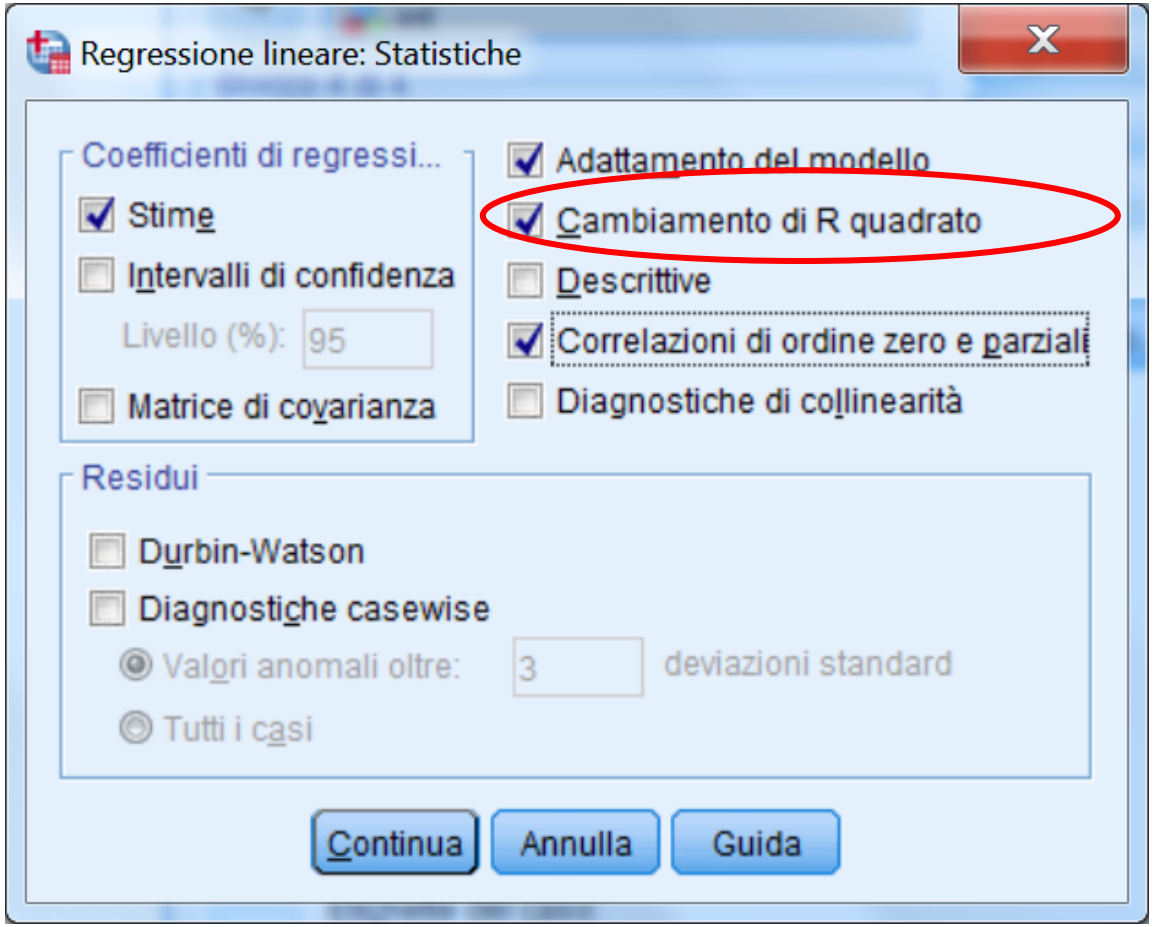
Regressione gerarchica

Inserire la quarta e ultima variabile nel "Blocco 4 di 4" ("compas").
In questi passaggi non cambiare mai il tipo di Metodo !!!



Regressione gerarchica

Nella finestra di dialogo "Statistiche" bisogna selezionare determinati parametri per ottenere nell'output le informazioni necessarie per interpretare e valutare la soluzione.



Regressione gerarchica

Il pannello iniziale riporta un riepilogo delle variabili inserite nel modello nei 4 passi della regressione: è diverso dal pannello analogo della regressione standard poiché ora non c'è più un unico blocco

Variabili immesse/rimosse[a]

Modello	Variabili immesse	Variabili rimosse	Metodo
1	att[b]	.	Inserisci
2	ns[b]	.	Inserisci
3	contco_2[b]	.	Inserisci
4	compas[b]	.	Inserisci

a. Variabile dipendente: int

b. Sono state immesse tutte le variabili richieste.

Regressione gerarchica

La varianza spiegata attraverso i diversi passi e il contributo unico delle variabili aggiunte ad ogni blocco si trova in questa tabella

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato adattato	Errore std. della stima	Statistiche delle modifiche				
					Modifica R-quadrato	Modifica F	gl1	gl2	Sign. Modifica F
1	,721[a]	,520	,517	1,766	,520	210,108	1	194	,000
2	,756[b]	,571	,567	1,674	,051	23,011	1	193	,000
3	,789[c]	,623	,617	1,574	,052	26,310	1	192	,000
4	,819[d]	,671	,664	1,474	,048	27,812	1	191	,000

a. Predittori: (costante), att

b. Predittori: (costante), att, ns

c. Predittori: (costante), att, ns, contco_2

d. Predittori: (costante), att, ns, contco_2, compas

Regressione gerarchica

La tabella dei coefficienti cambia a seconda del numero di predittori inseriti: l'ultima sezione (Modello 4) presenta risultati identici a quelli della regressione standard.

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sign.	Correlazioni		
		B	Errore std.	Beta			Ordine zero	Parziale	Parte
1	(Costante)	-3,886	,783		-4,960	,000			
	att	,261	,018	,721	14,495	,000	,721	,721	,721
2	(Costante)	-4,652	,759		-6,126	,000			
	att	,206	,021	,570	10,051	,000	,721	,586	,474
	ns	,393	,082	,272	4,797	,000	,589	,326	,226
3	(Costante)	-2,227	,856		-2,601	,010			
	att	,170	,021	,469	8,245	,000	,721	,511	,365
	ns	,358	,077	,248	4,627	,000	,589	,317	,205
	contco_2	-2,250	,439	-,255	-5,129	,000	-,544	-,347	-,227
4	(Costante)	-1,422	,816		-1,742	,083			
	att	,141	,020	,390	7,045	,000	,721	,454	,293
	ns	,273	,074	,189	3,676	,000	,589	,257	,153
	contco_2	-1,656	,426	-,188	-3,885	,000	-,544	-,271	-,161
	compas	,354	,067	,273	5,274	,000	,645	,357	,219

a. Variabile dipendente: int

Risultati della regressione gerarchica

Cambiamento di R e R^2 attraverso i riversi passi

Step	Variabile	R	R^2	R^2C	F	p
1	Atteggiamento	.72	.52	.52	210	.00
2	Norma Soggettiva	.76	.57	.05	23	.00
3	Senso di Controllo	.79	.62	.05	26	.00
4	Comport. Passato	.82	.67	.05	28	.00

sr^2: quantità di varianza aggiunta all' R^2 da ciascuna VI nel punto in cui la VI entra nell'equazione ("incremental sr^2" o cambiamento in R^2).

La somma degli sr^2 è uguale al valore di R^2.

Test statistico per valutare l'incremento nell'R^2

(Tabachnik & Fidell, 2007, p. 149)

$$F_{inc} = \frac{(R^2_{wi} - R^2_{wo})/m}{(1 - R^2)/df_{res}}$$

$R^2_{wi} = R^2$ ottenuto dall'inserimento della nuova variabile

$R^2_{wo} = R^2$ senza la nuova variabile

$m =$ numero di variabili nel nuovo blocco

$df_{res} = (N - k - 1)$

La regressione statistica

L'ordine di ingresso delle VI nell'equazione, e la decisione su quali VI vengono incluse o escluse dall'equazione di regressione sono determinati da criteri statistici

Limite: Differenze marginali rispetto a questi criteri possono influenzare in modo sostanziale l'importanza attribuita alle diverse VI

Tipi di regressione statistica

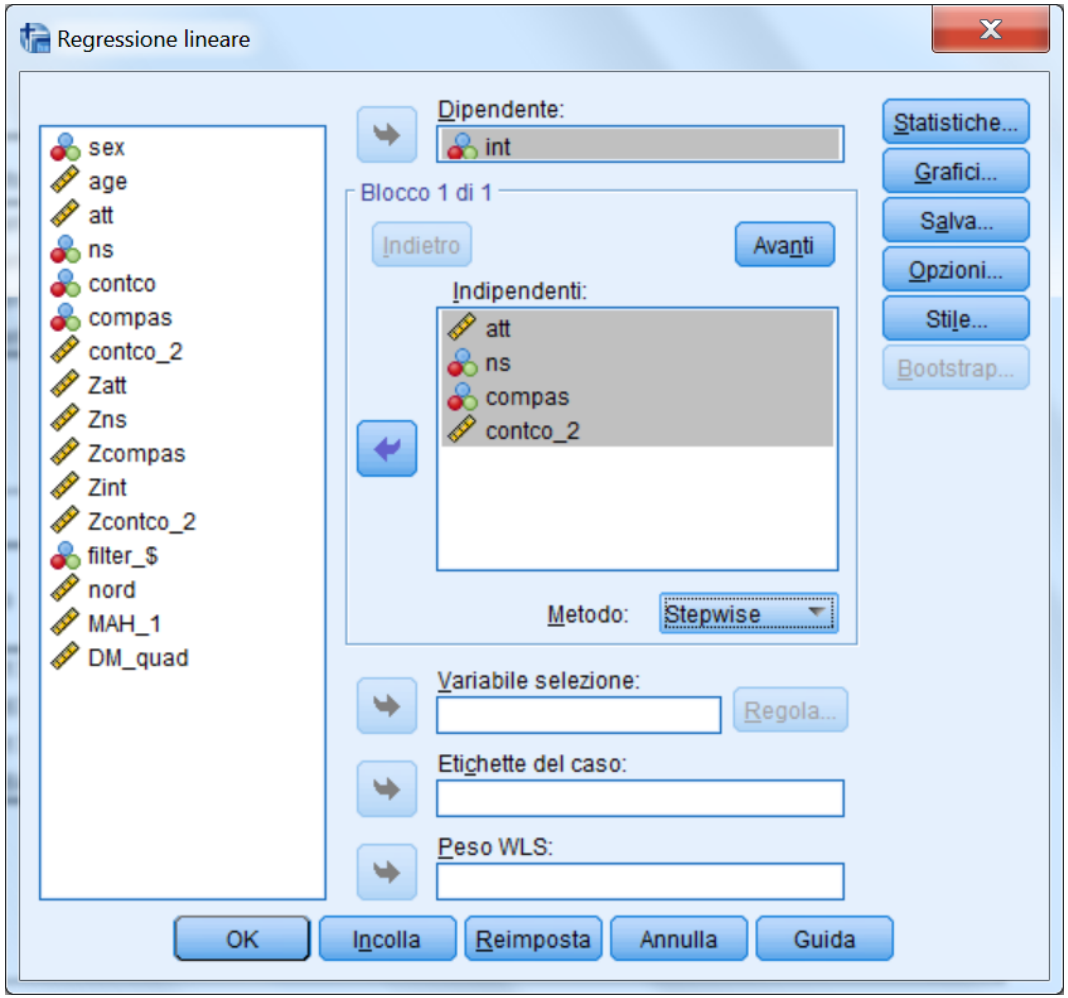
Regressione forward (in avanati): equazione inizialmente "vuota"; ad ogni step viene aggiunta la VI che presenta la correlazione più elevata con la VD. Se una VI entra in equazione, vi rimane

Regressione backward (all'indietro): l'equazione inizialmente comprende tutte le VI; ad ogni step viene eliminata la VI che non correla significativamente con la VD. Se una VI esce dall'equazione, non può più rientrarvi

Regressione stepwise: equazione inizialmente "vuota"; ad ogni step viene aggiunta la VI che correla di più con la VD. Le variabili che non forniscono più un contributo significativo vengono eliminate

Regressione Stepwise

Effettuare le stesse selezioni fatte per la regressione standard ma specificare "Stepwise" nel Metodo. Selezionare nelle Statistiche l'opzione per ottenere l'incremento dell'R^2.



Regressione Stepwise

Variabili immesse/rimosse[a]

Modello	Variabili immesse	Variabili rimosse	Metodo
1	att	.	Stepwise (criteri: Probabilità-di-F-da-inserire <= ,050, Probabilità-di-F-da-rimuovere >= ,100).
2	compas	.	Stepwise (criteri: Probabilità-di-F-da-inserire <= ,050, Probabilità-di-F-da-rimuovere >= ,100).
3	contco_2	.	Stepwise (criteri: Probabilità-di-F-da-inserire <= ,050, Probabilità-di-F-da-rimuovere >= ,100).
4	ns	.	Stepwise (criteri: Probabilità-di-F-da-inserire <= ,050, Probabilità-di-F-da-rimuovere >= ,100).

a. Variabile dipendente: int

Il pannello iniziale segnala quali variabili sono state inserite o rimosse durante la procedura Stepwise. Nella colonna metodo viene specificato quale è il metodo di inserimento/rimozione nell'equazione, e quali criteri determinano inserimento e rimozione

Regressione Stepwise

La varianza spiegata attraverso i diversi passi e il contributo unico delle variabili aggiunte ad ogni blocco si trova in questa tabella

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato adattato	Errore std. della stima	Statistiche delle modifiche				
					Modifica R-quadrato	Modifica F	gl1	gl2	Sign. Modifica F
1	,721[a]	,520	,517	1,766	,520	210,108	1	194	,000
2	,787[b]	,620	,616	1,575	,100	50,872	1	193	,000
3	,805[c]	,647	,642	1,522	,027	14,887	1	192	,000
4	,819[d]	,671	,664	1,474	,023	13,515	1	191	,000

a. Predittori: (costante), att

b. Predittori: (costante), att, compas

c. Predittori: (costante), att, compas, contco_2

d. Predittori: (costante), att, compas, contco_2, ns

La partizione della varianza è molto diversa da quella ottenibile nelle regressioni standard e gerarchica. L'ordine di importanza delle VI è quello dell'ultimo "modello"(ovvero passo): Atteggiamento, Comportamento Passato, Controllo, Norme Soggettive

Regressione Stepwise

La tabella dei coefficienti cambia a seconda dei predittori inseriti o rimossi: l'ultima sezione (Modello 4) presenta risultati identici a quelli della regressione standard e della gerarchica.

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sign.	Correlazioni		
		B	Errore std.	Beta			Ordine zero	Parziale	Parte
1	(Costante)	-3,886	,783		-4,960	,000			
	att	,261	,018	,721	14,495	,000	,721	,721	,721
2	(Costante)	-2,175	,739		-2,945	,004			
	att	,191	,019	,529	10,179	,000	,721	,591	,452
	compas	,479	,067	,370	7,132	,000	,645	,457	,316
3	(Costante)	-,657	,815		-,806	,421			
	att	,171	,019	,471	9,003	,000	,721	,545	,386
	compas	,407	,068	,315	6,027	,000	,645	,399	,258
	contco_2	-1,697	,440	-,193	-3,858	,000	-,544	-,268	-,165
4	(Costante)	-1,422	,816		-1,742	,083			
	att	,141	,020	,390	7,045	,000	,721	,454	,293
	compas	,354	,067	,273	5,274	,000	,645	,357	,219
	contco_2	-1,656	,426	-,188	-3,885	,000	-,544	-,271	-,161
	ns	,273	,074	,189	3,676	,000	,589	,257	,153

a. Variabile dipendente: int

Regressione Stepwise

Questa tabella è utile per capire quale variabile verrà inclusa nel prossimo passo. In questo caso è chiaro che tutte le variabili verranno incluse nell'analisi.

Variabili escluse[a]

Modello		Beta in	t	Sign.	Correlazione parziale	Statistiche di collinearità: Tolleranza
1	ns	,272[b]	4,797	,000	,326	,691
	compas	,370[b]	7,132	,000	,457	,730
	contco_2	-,276[b]	-5,289	,000	-,356	,798
2	ns	,194[c]	3,647	,000	,255	,654
	contco_2	-,193[c]	-3,858	,000	-,268	,737
3	ns	,189[d]	3,676	,000	,257	,653

a. Variabile dipendente: int

b. Predittori nel modello: (costante), att

c. Predittori nel modello: (costante), att, compas

d. Predittori nel modello: (costante), att, compas, contco_2

Differenti metodi ⇒ Differenti risultati

Standard ⇒ 48% di varianza non attribuibile a nessuna variabile.

Gerarchica ⇒ Norma Soggettiva spiega più varianza del comportamento passato

Stepwise ⇒ Comportamento passato variabile più importante dopo l'atteggiamento

Regressione standard: strategia analitica migliore per studi esplorativi.
Regressione gerarchica: controllo maggiore sul processo della regressione; subordinata alla formulazione di ipotesi; studi confermativo.

Conclusioni

Tecnica flessibile per studiare la relazione di dipendenza tra variabili soprattutto nelle fasi esplorative di una ricerca.

Possibilità di definire modelli a priori (nel caso della regressione *gerarchica*): estensione anche a contesti di tipo confermativo.

Lo scopo è comunque quello di spiegare al meglio una variabile dipendente (y). E' una tecnica poco adatta a rendere ragione di modelli teorici complessi, in cui ci sono diverse variabili dipendente.

Conclusioni

Limiti legati alle assunzioni statistiche:

*** Assenza di errore nelle variabili: assai irrealistica.**

*** Problema della *multicollineratià:* spesso risolvibile all'interno del modello della regressione.**

*** Impossibile considerare simultaneamente più di una variabile dipendente alla volta nello stesso modello. Modelli complessi sono esaminabili solo scindendoli in tanti pezzi separati.**

*** Risultati soggetti ad interpretazioni assai differenti a seconda del metodo di regressione scelto (standard, gerarchica, statistica).**

Accertare le condizioni di applicabilità

Scegliere l'approccio più adeguato per gli scopi del ricercatore

ESERCIZIO 2:
REALIZZAZIONE DI UN MODELLO DI REGRESSIONE CON SPSS

Utilizzare i dati in formato testo nel file ES1.SAV, risultato dell'esercizio 1.

VARIABILI:
ATTEGGIAMENTO, NORME SOGGETTIVE, SENSO DI CONTROLLO, COMPORTAMENTO PASSATO, INTENZIONE.
LA VARIABILE DIPENDENTE E' "INTENZIONE"

1) Effettuare una regressione standard, calcolando la varianza unica spiegata da ogni variabile e la varianza comune

2) Effettuare una regressione gerarchica nella quale l'ordine di entrata della VI è il seguente: comportamento passato, norme soggettive, senso di controllo, atteggiamento

L'ANALISI DELLA VARIANZA (ANOVA)

Sommario

* Il modello lineare: forma e assunzioni

* Disegni ad un fattore

* Confronti post-hoc e pianificati

* Disegni fattoriali: effetti principali ed interazione

* Potenza della verifica e ampiezza degli effetti

Scopo dell'analisi della varianza: verificare ipotesi relative a differenze tra medie di due o più popolazioni.

Variabile dipendente: su scala a intervalli o rapporti equivalenti

Variabile indipendente: categoriale.

- Una sola V.I.: Disegni a una via
- Due o più V.I.: Disegni Fattoriali
- Una sola V.D.: Analisi univariata
- Due o più V.D.: Analisi multivariata (MANOVA)

L'ANALISI DELLA VARIANZA UNIVARIATA (ANOVA): DISEGNI TRA I SOGGETTI AD UN SOLO FATTORE

Ad ogni livello della variabile indipendente corrisponde un diverso gruppo di soggetti.
In ogni condizione ci sono soggetti diversi: un soggetto esposto ad una condizione non viene esposto a nessuna altra condizione.

OBIETTIVI	
SI	NO
S1	S6
S2	S7
…	…

MODELLO LINEARE DELL'ANOVA

Il punteggio y_{ij} di un soggetto "j" nel gruppo "i" è scomponibile così:

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

- μ: **media generale ("grand mean") dei punteggi sul campione totale**
- α_i: **effetto dovuto al trattamento (livello i della variabile indipendente)**
- ε_{ij}: **è una componente "residua", di errore casuale, specifica per ogni soggetto.**

Stime campionarie dei parametri di popolazione:

$$\hat{\mu} = \bar{y}_{..}$$

media generale del campione

$$\hat{\alpha}_i = (\bar{y}_{i.} - \bar{y}_{..})$$

differenza tra la media del gruppo cui appartiene il soggetto e la media generale del campione (contributo della condizione "i" al punteggio del soggetto "j")

$$\hat{\varepsilon}_{ij} = (y_{ij} - \bar{y}_{i.})$$

differenza tra punteggio del soggetto e media del gruppo in cui è inserito (variabilità dei punteggi individuali all'interno di ogni gruppo).

Scomposizione della devianza totale

Devianza totale

$$SS_T = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$$

Somma dei quadrati degli scarti al quadrato tra i singoli punteggi e la media generale (tutti i soggetti possono essere considerati come appartenenti ad un unico campione).

Scomposizione della devianza totale

Devianza tra i gruppi (o *between)*

$$SS_B = \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})^2$$

Si calcola sostituendo ad ogni punteggio la media del gruppo cui appartiene
(come se tutti i soggetti sottoposti allo stesso trattamento avessero ottenuto esattamente lo stesso punteggio).

Scomposizione della devianza totale

Devianza entro i gruppi (o *within*)

$$SS_W = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

Somma dei quadrati degli scarti al quadrato tra i punteggi di ogni soggetto e la media del gruppo cui il soggetto appartiene.

E' possibile dimostrare che $SS_T = SS_B + SS_W$:

$$SS_T = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 \quad =$$

$$SS_B = \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})^2 \quad +$$

$$SS_W = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

GRADI DI LIBERTA' E "QUADRATI MEDI"

$$SS_T = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$$ **= n -1 (il gdl perso è quello della media totale)**

$$SS_B = \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})^2$$ **= k-1 (il gdl perso è quello della media totale)**

$$SS_W = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$ **= n-k (1 gdl perso per ogni media di gruppo)**

La scomposizione che vale per le devianze vale anche per i gradi di libertà: n-1=(k-1)+(n-k).

GRADI DI LIBERTA' E "QUADRATI MEDI"

Dividendo le devianze per i rispettivi gdl si ottengono le varianze ovvero i "quadrati medi" (mean squares).

Varianza totale (MS_T)=

Devianza totale/(n-1) = $SS_T/(n-1)$

Varianza tra i gruppi (MS_B)=

Devianza tra i gruppi/(k-1)= $SS_B/(k-1)$

Varianza entro i gruppi (MS_W)=

Devianza entro i gruppi/(n-k)=$SS_W/(n-k)$

dove: n = numero totale di soggetti

k = numero di gruppi

RAPPORTO "F"

Il rapporto tra le varianze MS_B/MS_W segue la distribuzione F (che è tabulata) quindi può essere utilizzato per esaminare ipotesi sulla significatività della differenza tra la variabilità dovuta al trattamento e quella residua.

La F testa le seguenti ipotesi statistiche:

H_0: $\mu_1=\mu_2= \ldots =\mu_k$
(Le popolazioni di provenienza dei campioni hanno medie uguali sulla V. D.)

H_1: almeno due μ diverse – $\mu_1\neq\mu_2$, o $\mu_1\neq\mu_3$, ecc.
(Almeno due campioni provengono da popolazioni con medie tra loro diverse)

RAPPORTO "F"

Varianza tra i gruppi, o *between*:
è data dalle differenze tra le medie dei gruppi sottoposti a trattamenti diversi;
riflette l'effetto della VI.

Varianza entro i gruppi, o *within*:
riflette le differenze tra i punteggi di soggetti appartenenti allo stesso gruppo,
può essere attribuita all'errore casuale.

RAPPORTO "F"

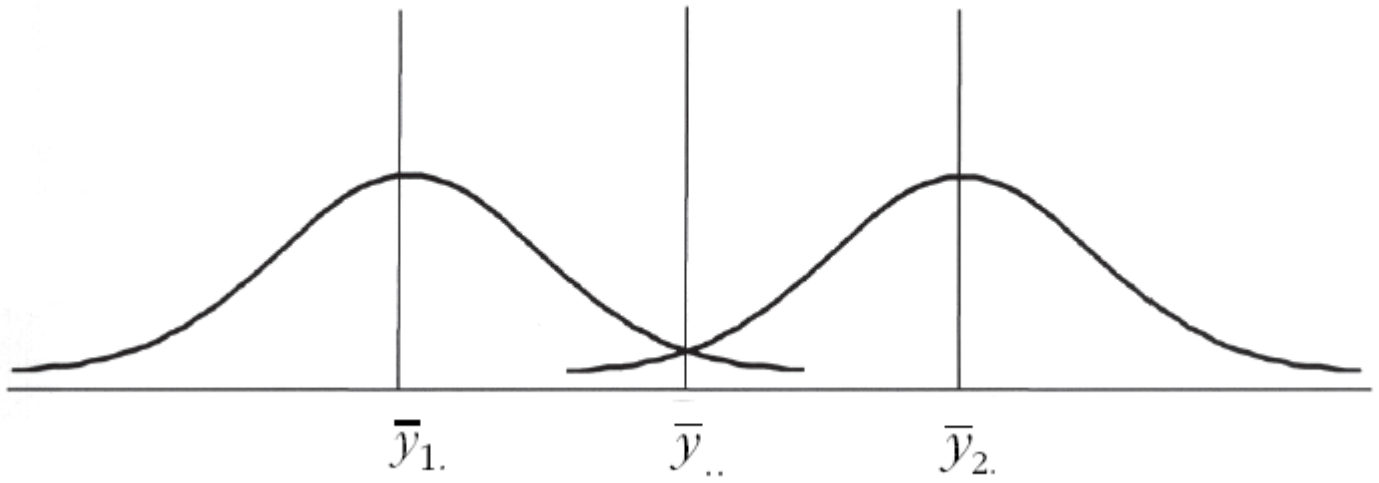
H_0 vera:
il trattamento non produce effetti,
le due varianze sono molto simili,
il rapporto F assume valori molto bassi
(vicini ad 1 o inferiori),
i punteggi dei soggetti nei diversi gruppi sono simili.

H_0 falsa:
varianza tra i gruppi (trattamento) maggiore della
varianza entro i gruppi (errore casuale),
il rapporto F assume valori elevati,
i punteggi dei soggetti nei diversi gruppi sono diversi.

RAPPORTO "F"

a) F significativa (Rifiuto H_0: $\mu_1=\mu_2= \ldots =\mu_k$)

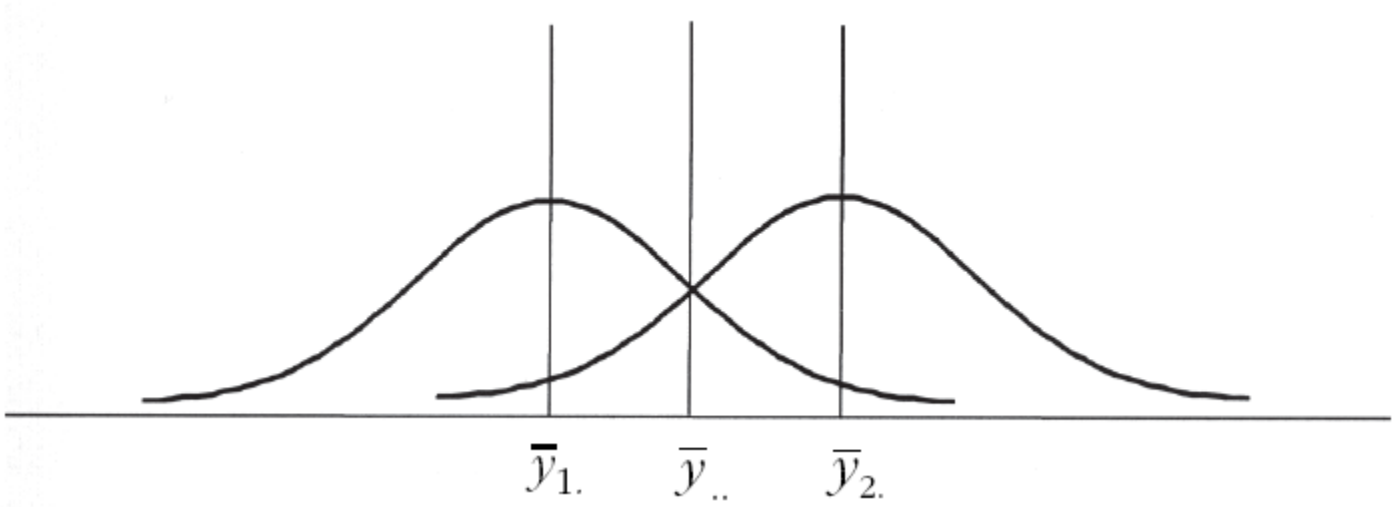
Se la varianza tra i gruppi è maggiore della varianza entro i gruppi, le medie dei gruppi saranno piuttosto <u>distanziate</u>.



RAPPORTO "F"

b) F non significativa (Non rifiuto H_0: $\mu_1=\mu_2= \ldots. =\mu_k$)

Se invece la varianza tra i gruppi non è significativamente diversa dalla varianza entro i gruppi, le medie dei gruppi saranno piuttosto ravvicinate.



ASSUNZIONI

a) gli errori (ε_{ij}) seguono la distribuzione normale ed hanno media uguale a 0.
Non normalità: ha un effetto debole sull'errore di I tipo (leggera inflazione) soprattutto nel caso in cui le celle non sono bilanciate (numero di soggetti diversi nelle differenti condizioni).

b) la varianza degli errori (σ_ε) è uguale in ogni gruppo (OMOSCHEDASTICITA').
Eteroschedasticità: La F è "robusta" anche rispetto a questa assunzione. Gli effetti più gravi si hanno nei disegni non bilanciati. L'omoschedasticità viene valutata con il test di Levene.

ASSUNZIONI

c) gli errori (ε_{ij}) sono indipendenti (il punteggio di un soggetto non è correlato con quello di altri soggetti).

Non Indipendenza delle osservazioni: può avere effetti notevoli sul livello di significatività (aumento incontrollato del livello reale di α) e sulla potenza del test. L'indipendenza viene valutata con il coefficiente di correlazione intraclasse (vedi pp. 195-197).

ASSUNZIONI

d) gli effetti hanno una natura additiva: la variabile sperimentale "aggiunge" qualcosa alla condizione-base e lo fa in maniera "identica" per tutti i soggetti.

Non additività degli effetti: aumenta debolmente l'errore sperimentale e diminuisce la potenza del test. E' un fattore di cui non ci si deve molto preoccupare.

EFFECT SIZE

La F è fortemente dipendente dalla numerosità dei gruppi considerati.

Non basta allora dimostrare che la F è statisticamente significativa per rilevare la presenza di un effetto. Bisogna dimostrare che l'effetto è importante anche da un punto di vista pratico.

Coefficienti che quantificano l'associazione tra variabile dipendente e variabile indipendente: possono essere interpretati come proporzione della varianza della variabile dipendente spiegata dalla variabile indipendente.

EFFECT SIZE

$\eta^2 = SS_B/SS_T$ **eta quadro**

$\omega^2 = [SS_B – (k\text{-}1) * MS_W]/ (SS_T + MS_W)$ **omega quadro**

Effect size nell'ANOVA:

ω^2,η^2 = .01 - .05 → **Basso**
ω^2,η^2 = .06 - .13 → **Moderato**
ω^2,η^2 = .14 → **Elevato**

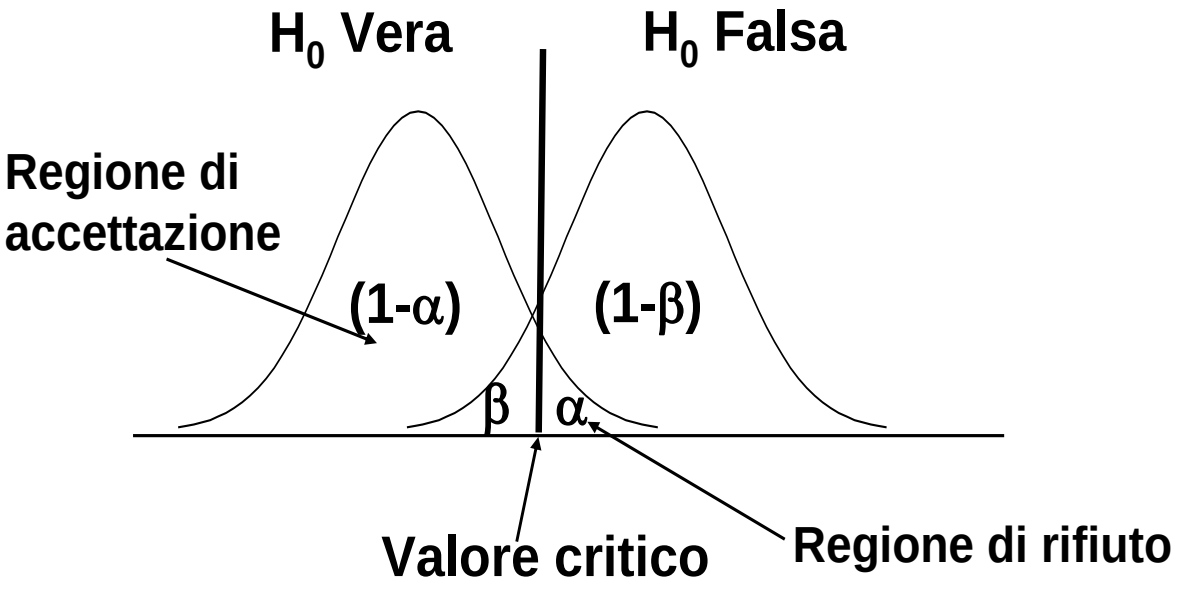
POTENZA DELLA VERIFICA

Probabilità di rifiutare l'ipotesi nulla quando essa è falsa. Probabilità di rilevare un effetto quando esso è presente.

Errore del II tipo: non rifiutare l'ipotesi nulla quando essa è falsa. La probabilità di commetterlo è indicata con il simbolo β. La potenza si indica con $1-\beta$.

Errore di I tipo (la cui probabilità è α), ed errore di II tipo sono inversamente proporzionali.

POTENZA DELLA VERIFICA



	H_0 VERA	H_0 FALSA
NON RIFIUTO H_0	Decisione Corretta ($p = 1-\alpha$)	Errore di II Tipo ($p = \beta$)
RIFIUTO H_0	Errore di I Tipo ($p = \alpha$)	Decisione Corretta ($p = 1-\beta$) Potenza della verifica

POTENZA DELLA VERIFICA

Esempio: differenze tra 2 gruppi, entrambi di 15 soggetti.

α	β	1-β
.10	.37	.63
.05	.52	.48
.01	.78	.22

Se α diminuisce da .10 a .05, β aumenta da .37 a .52, e la potenza (1-β) diminuisce da .63 a .48

POTENZA DELLA VERIFICA

La potenza della verifica dipende da tre fattori:
- livello di α
- ampiezza del campione
- effect size: quanto i gruppi differiscono effettivamente nella popolazione.

Esempio: cambiamento nella potenza in funzione di n, considerando un effect size pari a .5.

n per gruppo	1-β
10	.18
20	.33
50	.70
100	.94

COME AUMENTARE LA POTENZA DELLA VERIFICA

* **Aumentare l'effect size**
  - **Ridurre la variabilità entro i gruppi:**
    - **# gruppi più omogenei**
    - **# disegni fattoriali invece che a una via**
    - **# analisi della covarianza**
    - **# disegni within subjects**
  - **Essere sicuri che ci sia un legame forte tra variabile indipendente e variabile dipendente (validità interna dell'esperimento)**

* **Aumentare il numero di soggetti**

* **Aumentare a/Usare test a una coda**
  [soluzione poco efficiente]

STIME DELLA POTENZA POST–HOC:

Consentono di calcolare il livello (1-β) dopo aver effettuato l'analisi. Permettono di interpretare meglio i risultati (soprattutto in presenza di F non significativa, ed effect size moderato/elevato).

STIME DELLA POTENZA A PRIORI:

Consentono di stabilire (una volta identificato l'effect size che si attende nell'esperimento) quale sarà la potenza della verifica per un dato numero di gruppi (k) e di numerosità di soggetti per gruppo (nk).

STIME DELLA POTENZA A PRIORI:

Consentono anche di stabilire quanti soggetti sono necessari per ogni gruppo per ottenere un determinato livello (1-β) dato un certo valore dell'effect size.

Le stime della potenza della verifica vengono effettuate utilizzando delle apposite tabelle sviluppate da Cohen, ed opportune formule per stimare l'effect size.

Nella maggior parte delle ricerche psicologiche si considera adeguata una potenza pari a .80 (ovvero, la probabilità di commettere errore di II tipo, cioè non rifiutare l'ipotesi nulla quando è falsa, è uguale a .20). Raggiungere livelli di potenza più elevati richiede spesso troppi soggetti.

Esempio di disegno univariato ad 1 fattore

Si vuole verificare l'efficacia di programmi di formazione che prevedono:

a) l'assegnazione di obiettivi (condizione A);
b) l'assegnazione di obiettivi e un feedback sui risultati (condizione B);
c) una condizione di controllo in cui non si danno né obiettivi né feedback (condizione C).

Tre gruppi di soggetti vengono sottoposti ognuno ad una condizione diversa ottenendo i seguenti risultati (Y = numero di problemi risolti):

Tre gruppi di soggetti vengono sottoposti ognuno ad una condizione diversa ottenendo i seguenti risultati (Y = numero di problemi risolti):

Obiettivi (Y_1):	10	7	4	5	8	n = 5
Obiettivi + Feedback (Y_2):	9	10	5	4	7	n = 5
Controllo (Y_3):	3	2	2	3	1	n = 5

Disegno:

Analisi della varianza univariata (una sola V.D.) ad un fattore (una sola V.I.) tra i soggetti (un diverso gruppo di sogg. per ogni livello della V.I.)

Formulazione delle ipotesi statistiche:

H_0: $\mu_1 = \mu_2 = \mu_3$

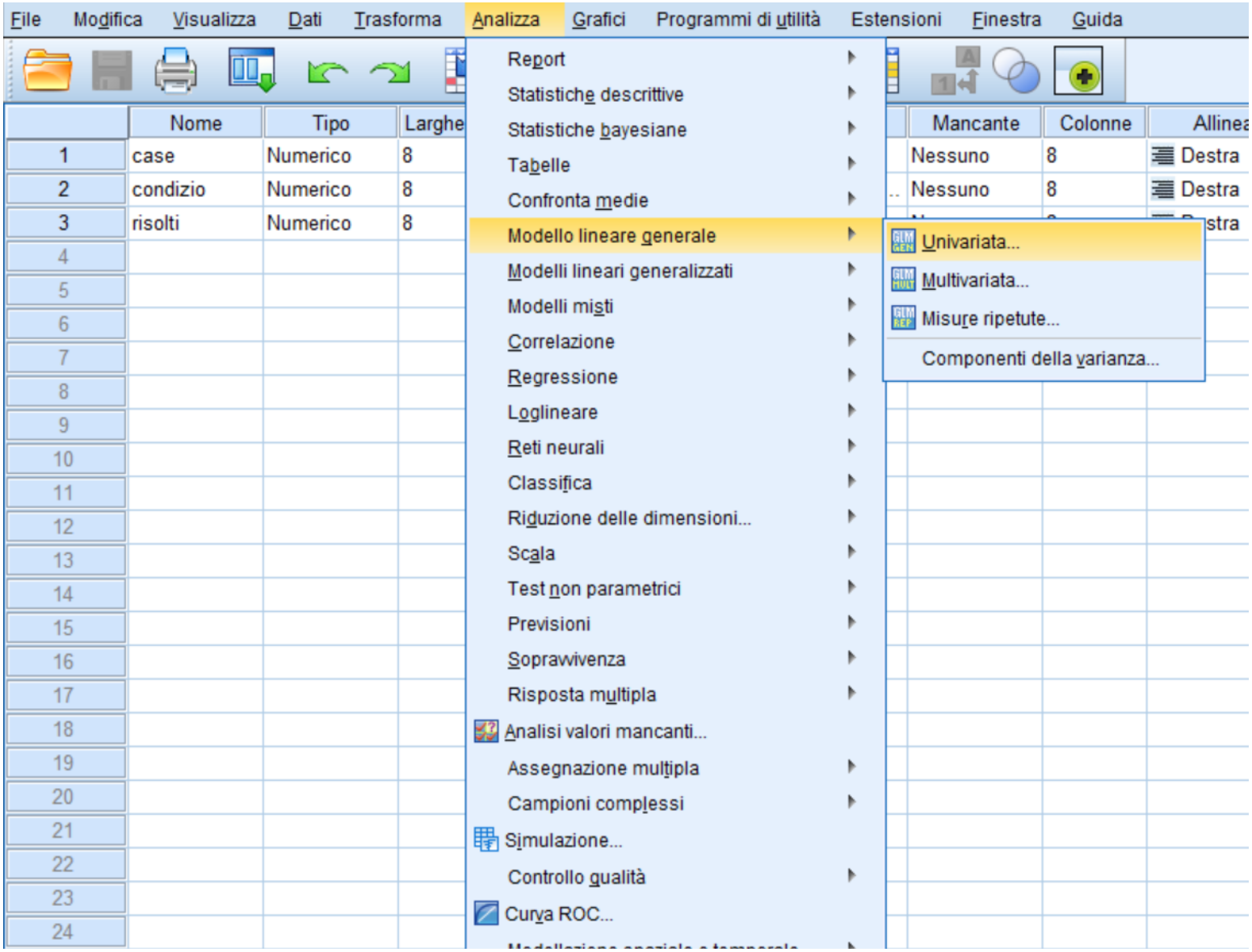
(le 3 medie sono relative a campioni che provengono dalla stessa popolazione)

H_1: $\mu_1 \neq \mu_2$, o $\mu_1 \neq \mu_3$, o $\mu_2 \neq \mu_3$

ovvero, almeno due μ diverse

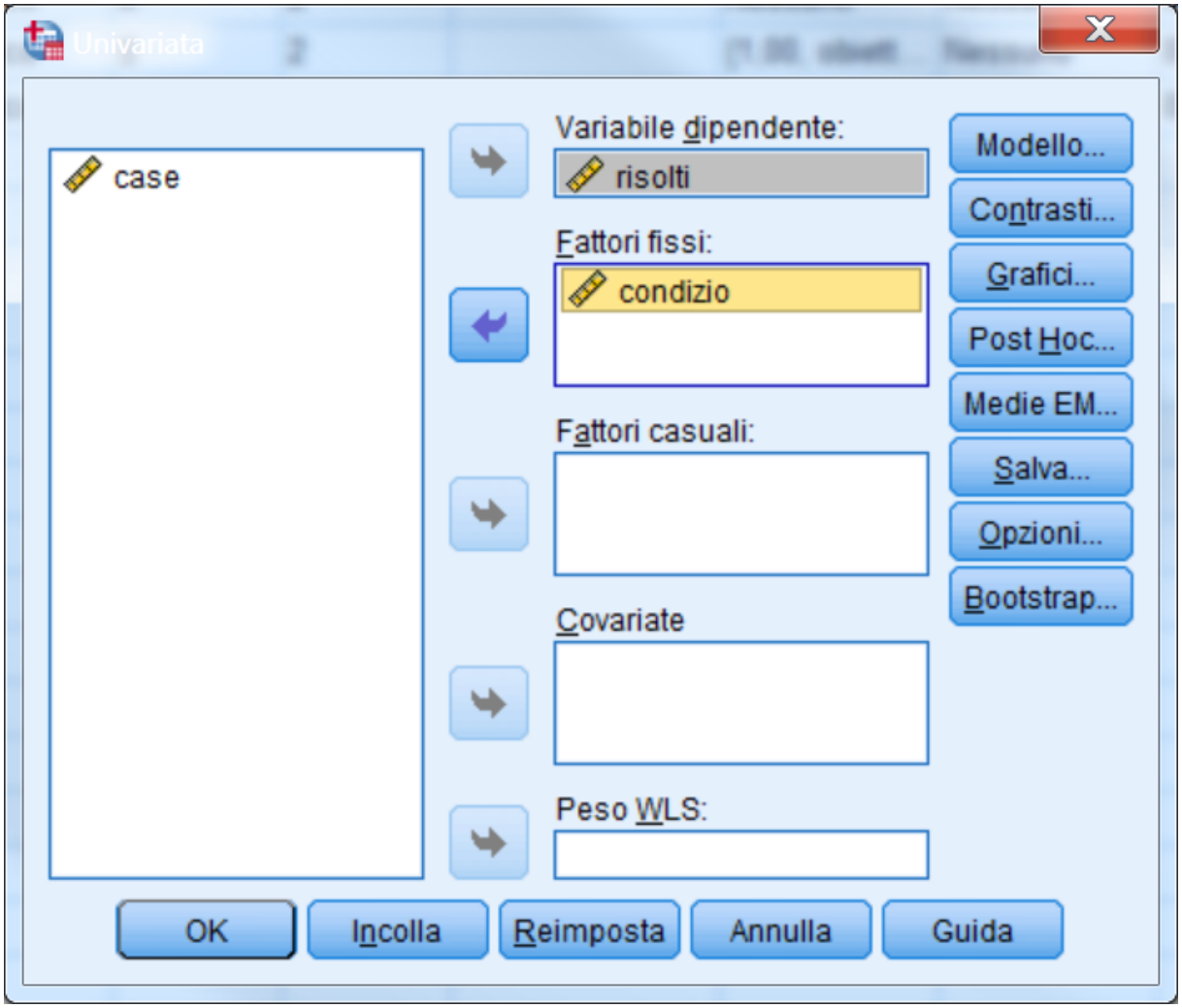
(almeno due medie sono relative a campioni che provengono da popolazioni diverse)

ANOVA IN SPSS

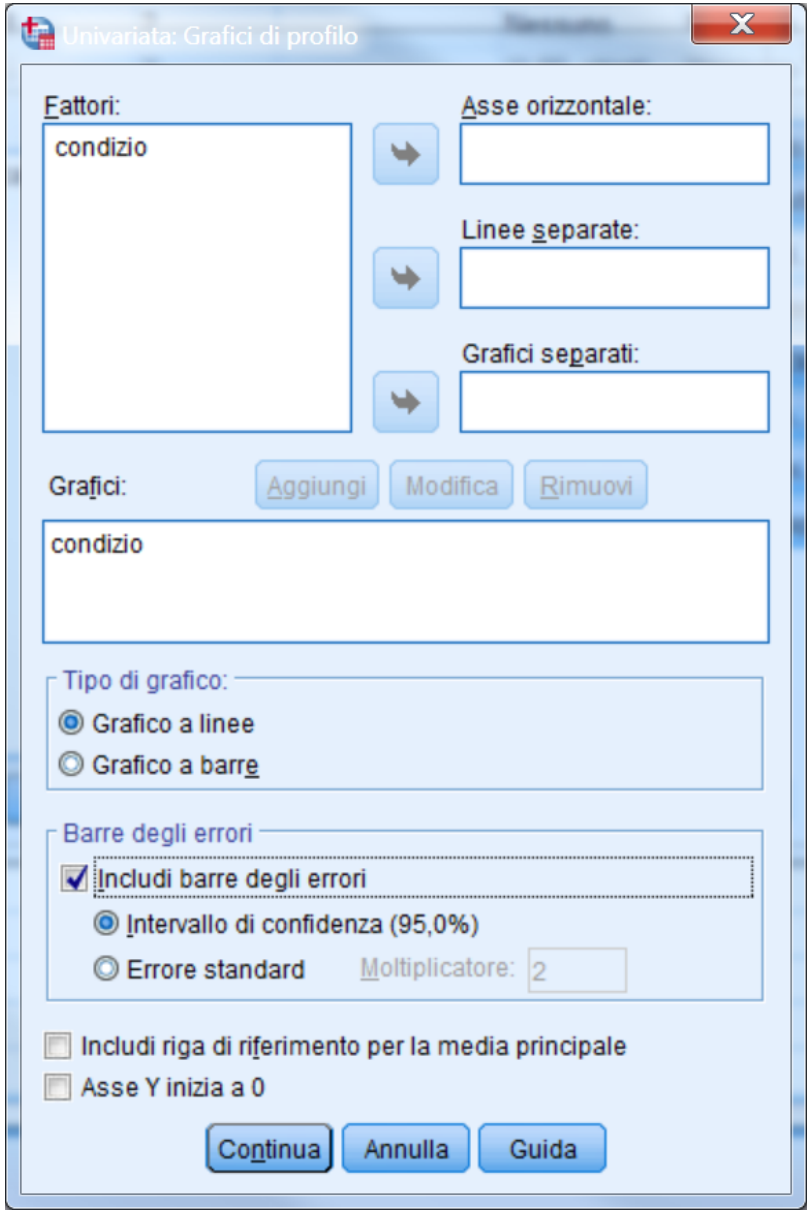


Oneway.B.sav

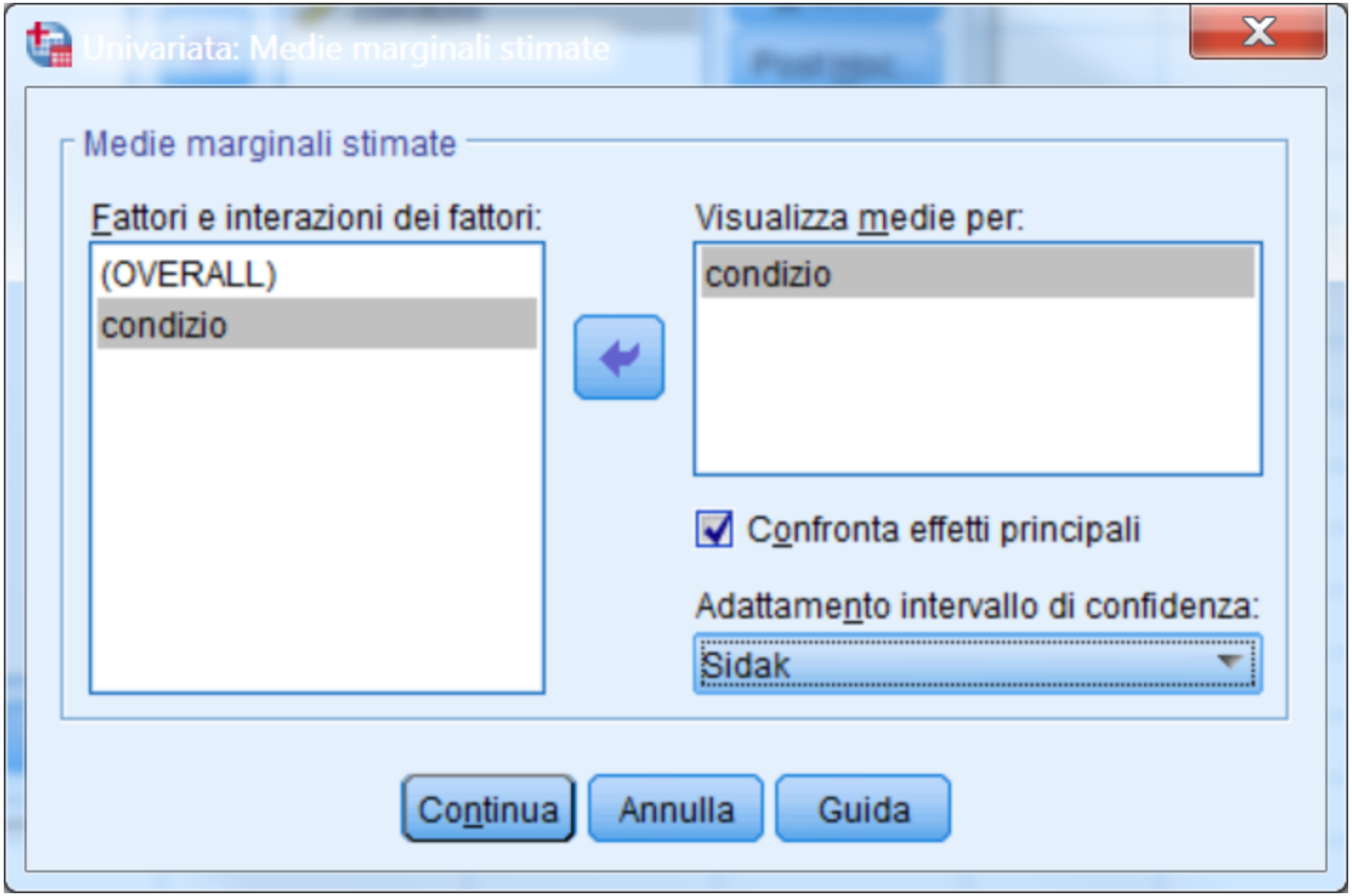
ANOVA A UNA VIA "BETWEEN" IN SPSS



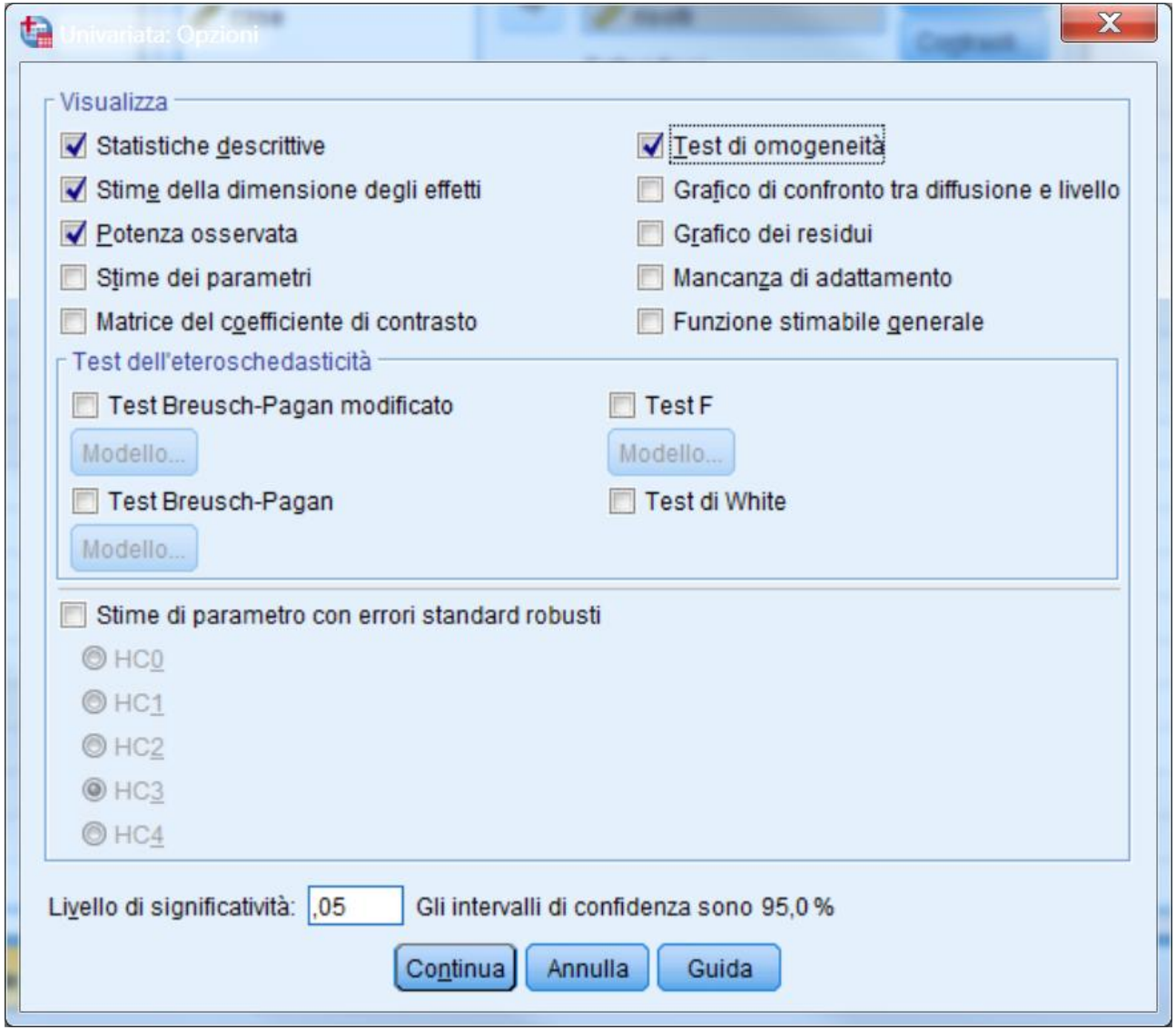
ANOVA A UNA VIA "BETWEEN" IN SPSS



ANOVA A UNA VIA "BETWEEN" IN SPSS



ANOVA A UNA VIA "BETWEEN" IN SPSS



ANOVA A UNA VIA "BETWEEN" IN SPSS

Fattori tra soggetti

		Etichetta di valore	N
condizio	1,00	obiettivi	5
	2,00	obiettivi + feedback	5
	3,00	controllo	5

Statistiche descrittive

Variabile dipendente:risolti

condizio	Media	Deviazione standard Variabile	N
1,00 obiettivi	6,8000	2,38747	5
2,00 obiettivi + feedback	7,0000	2,54951	5
3,00 controllo	2,2000	,83666	5
Totale	5,3333	2,99205	15

Test di Levene di uguaglianza delle varianze dell'errore[a]

ANOVA A UNA VIA "BETWEEN" IN SPSS

Test di Levene di eguaglianza delle varianze dell'errore[a,b]

		Statistica di Levene	gl1	gl2	Sign.
risolti	Basato sulla media	2,626	2	12	,113
	Basato sulla mediana	2,457	2	12	,128
	Basato sulla mediana e con il grado di libertà adattato	2,457	2	9,369	,139
	Basato sulla media ritagliata	2,637	2	12	,112

Verifica l'ipotesi nulla che la varianza dell'errore della variabile dipendente sia uguale tra i gruppi.

a. Variabile dipendente: risolti

b. Disegno: Intercetta + condizio

ANOVA A UNA VIA "BETWEEN" IN SPSS

Test degli effetti fra soggetti

Variabile dipendente:risolti

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.	Eta quadrato parziale
Modello corretto	73,733[a]	2	36,867	8,574	,005	,588
Intercetta	426,667	1	426,667	99,225	,000	,892
condizio	73,733	2	36,867	8,574	,005	,588
Errore	51,600	12	4,300			
Totale	552,000	15				
Totale corretto	125,333	14				

Test degli effetti fra soggetti

Variabile dipendente:risolti

Sorgente	Non centralità Parametro	Potenza osservata[b]
Modello corretto	17,147	,911
Intercetta	99,225	1,000
condizio	17,147	,911

a. R quadrato = ,588 (R quadrato corretto = ,520)
b. Calcolato usando alfa = ,05

F (8.57) significativo al 1%: bisogna rifiutare l'ipotesi nulla.

ANOVA IN SPSS

Confronti a coppie

Variabile dipendente:risolti

(I) condizio	(J) condizio	Differenza media (I-J)	Deviazione standard Errore	Sig.[a]
1,00 obiettivi	2,00 obiettivi + feedback	-,200	1,311	,998
	3,00 controllo	4,600*	1,311	,013
2,00 obiettivi + feedback	1,00 obiettivi	,200	1,311	,998
	3,00 controllo	4,800*	1,311	,010
3,00 controllo	1,00 obiettivi	-4,600*	1,311	,013
	2,00 obiettivi + feedback	-4,800*	1,311	,010

Confronti a coppie

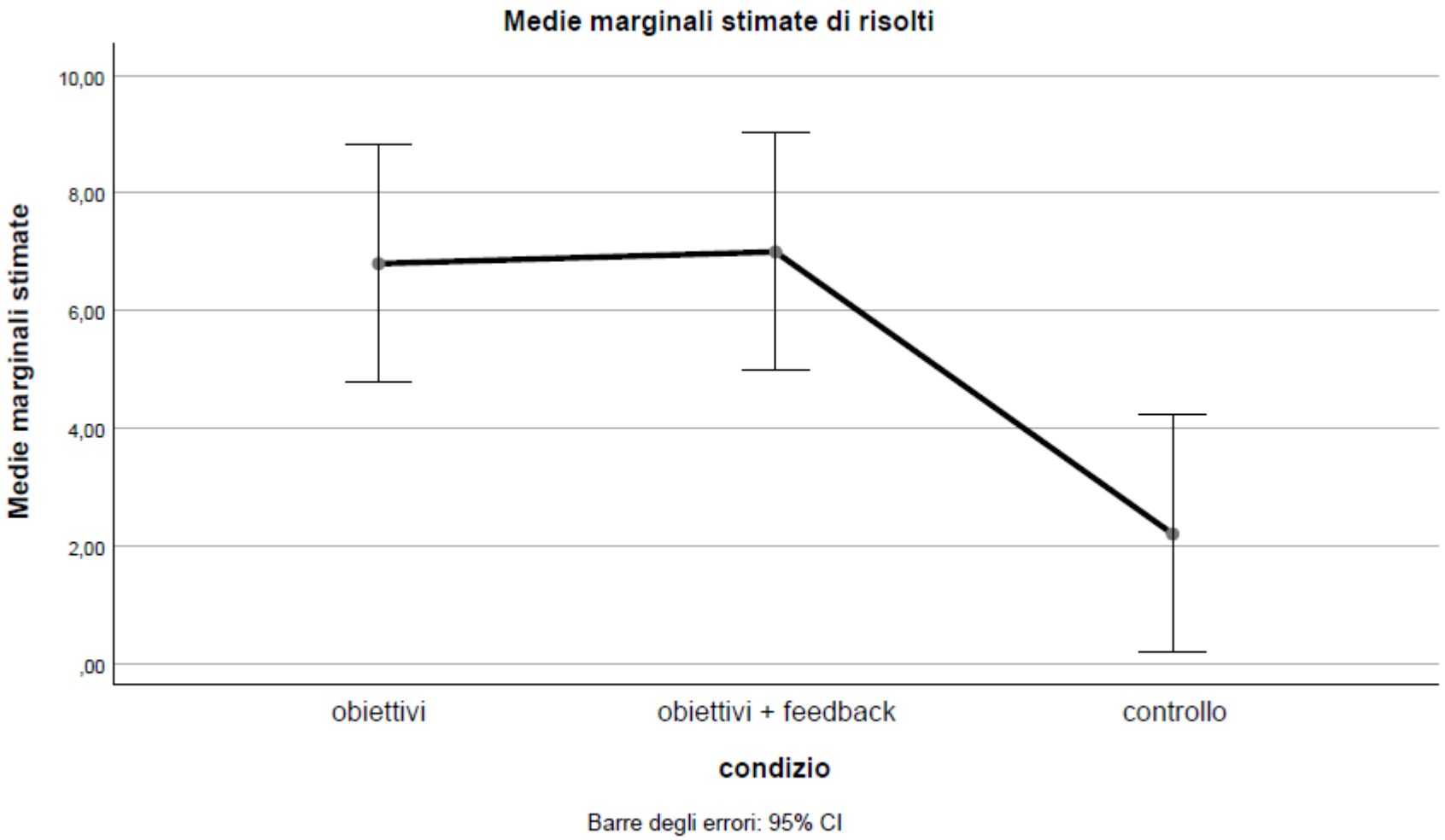
Variabile dipendente:risolti

(I) condizio	(J) condizio	Intervallo di confidenza per la differenza al 95%[a]	
		Limite inferiore	Limite superiore
1,00 obiettivi	2,00 obiettivi + feedback	-3,833	3,433
	3,00 controllo	,967	8,233
2,00 obiettivi + feedback	1,00 obiettivi	-3,433	3,833
	3,00 controllo	1,167	8,433
3,00 controllo	1,00 obiettivi	-8,233	-,967
	2,00 obiettivi + feedback	-8,433	-1,167

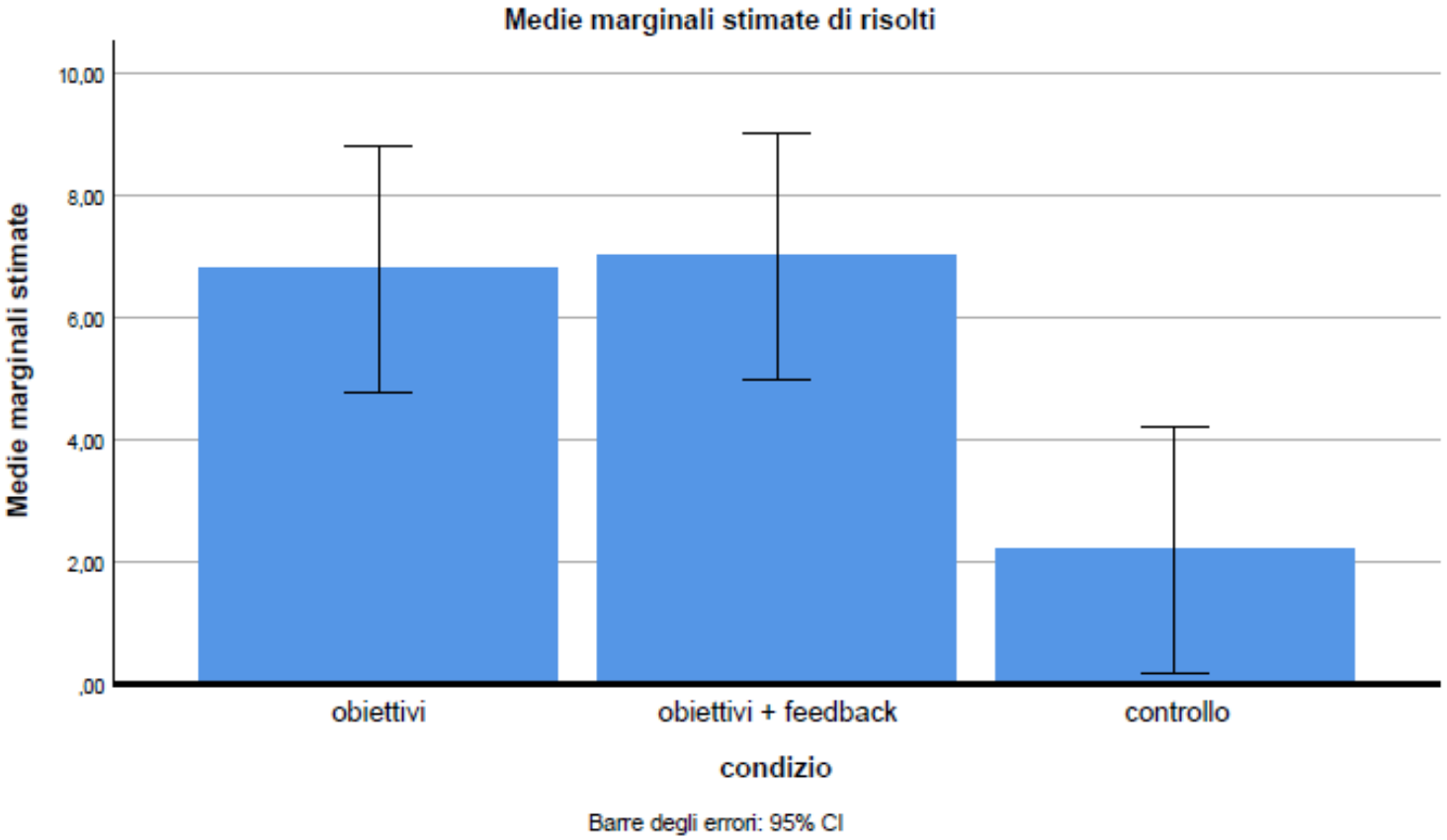
Basato sulle medie marginali stimate

a. Correzione per confronti multipli: Sidak.
*. La differenza media è significativa al livello ,05.

ANOVA A UNA VIA "BETWEEN" IN SPSS



ANOVA A UNA VIA "BETWEEN" IN SPSS



Confronti tra le medie dei gruppi

F significativo: esiste una differenza significativa tra almeno due delle medie dei gruppi messi a confronto, ma non sappiamo tra quali.

Confronto tra le medie dei gruppi con un test statistico adeguato per individuare la fonte della significatività:

a) i confronti post hoc;

b) i confronti pianificati.

a) I confronti post hoc

Ogni media in genere viene confrontata con tutte le altre. Il ricercatore non stabilisce in anticipo i confronti rilevanti ai fini della sua ipotesi.

Svantaggio: all'aumentare del numero di gruppi aumenta il numero di confronti e aumenta la probabilità di commettere l'errore di primo tipo (livello α), cioè rifiutare l'ipotesi nulla quando è vera.

Esempio. 3 possibili confronti di medie: la condizione A con la B, la A con la C e la B con la C.

a) I confronti post hoc – inflazione del livello α

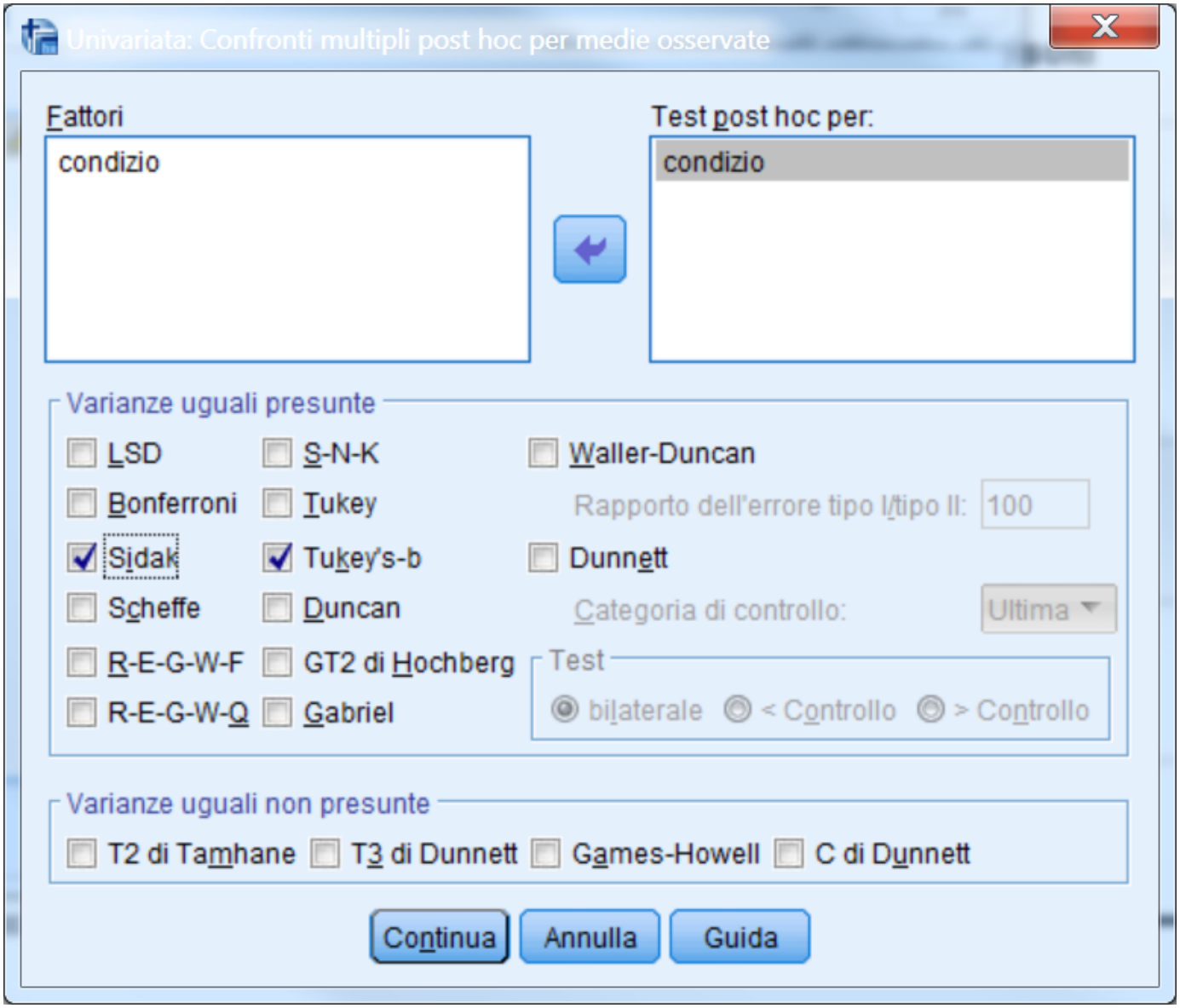
Si esamina l'ipotesi che o il primo, o il secondo, o il terzo confronto risultino significativi.

Livello α = 0.05 per ognuno dei 3 confronti:
La probabilità che almeno uno dei tre confronti risulti significativa è uguale a .05 + .05 + .05 =.15.
Livello reale di α per i 3 confronti: 3*.05 = .15.

Con k confronti il livello di probabilità che almeno uno di essi risulti significativo non è α ma kα.

Soluzione: scegliere un valore α minore di .05 (es., .05/3=.017, e in genere .05/k).

ANOVA IN SPSS



ANOVA IN SPSS

Sottoinsiemi omogenei

risolti

B di Tukey[a,b]

condizio	N	Sottoinsieme 1	Sottoinsieme 2
controllo	5	2,2000	
obiettivi	5		6,8000
obiettivi + feedback	5		7,0000

Vengono visualizzate le medie per i gruppi nei sottoinsiemi omogenei.

Si basa sulle medie osservate.

Il termine di errore è media quadratica(errore) = 4,300.

a. Utilizza dimensione del campione della media armonica = 5,000.

b. Alfa = 0,05.

b) I confronti pianificati.

Effettuare solo i confronti che appaiono più rilevanti ai fini dell'ipotesi di ricerca. Il ricercatore pianifica in anticipo quali medie (gruppi) verranno confrontate.

I confronti pianificati consentono di esaminare la differenza tra 2 medie.

Si possono confrontare 2 medie di 2 singoli gruppi, oppure "combinare" insieme le medie di più gruppi e confrontare la media "aggregata" così ottenuta con la media di un gruppo singolo, o con un'altra media "aggregata", ottenuta da più gruppi.

Il confronto comunque sarà sempre tra 2 medie.

b) I confronti pianificati sull'esempio empirico

La presenza di una consegna ben precisa (obiettivo, oppure obiettivo + feedback) rispetto all'assenza di tale consegna si accompagna a maggiore facilità nella soluzione dei problemi.

E' sufficiente un set di due confronti tra le medie (invece dei tre confronti visti per i post hoc):

nel primo si contrasta il gruppo di controllo con i gruppi "obiettivi" e "obiettivi+feedback" combinati insieme (come se fossero un unico gruppo);

nel secondo si contrastano tra loro i due gruppi "obiettivi" e "obiettivi+feedback".

b) I confronti pianificati.

Per effettuare i confronti (con il computer o manualmente) si deve attribuire ad ogni media un coefficiente, con segno positivo o negativo.

Le medie con segno diverso vengono contrastate tra loro, quelle con segno uguale vengono combinate, quelle con coefficiente 0 non vengono considerate nel confronto.

La somma dei coefficienti deve dare 0. Se anche la somma dei prodotti tra i coefficienti di un set di confronti è uguale a 0, si dice che i confronti sono tra loro ortogonali, cioè indipendenti.

Coefficienti per i dati dell'esempio:

	Obiettivi	Ob. + Feed.	Controllo	Somme
1° confronto	-1	-1	2	0
2° confronto	1	-1	0	0
Prodotti	-1	1	0	0

Sono definiti bene
(le somme sono uguali a zero per ogni riga).

Sono ortogonali
(le somme dei prodotti sono uguali a zero).

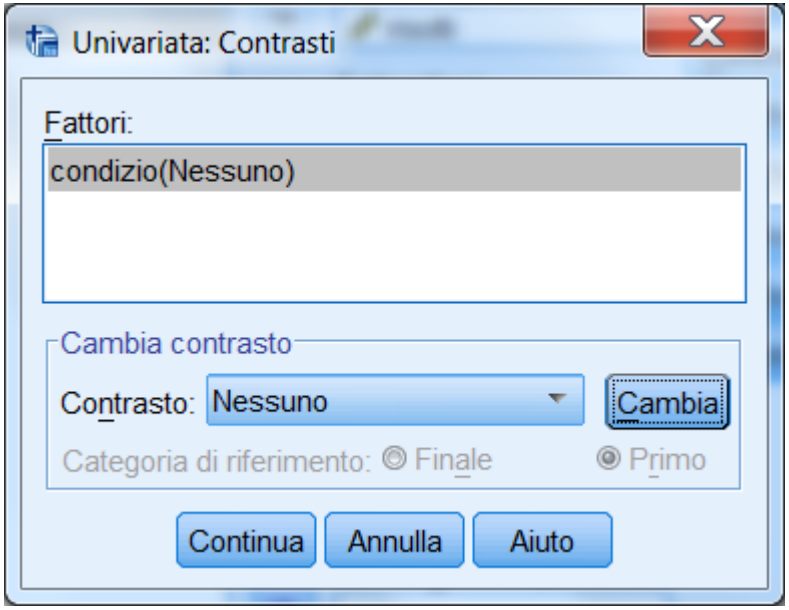
Set di confronti non ortogonali:

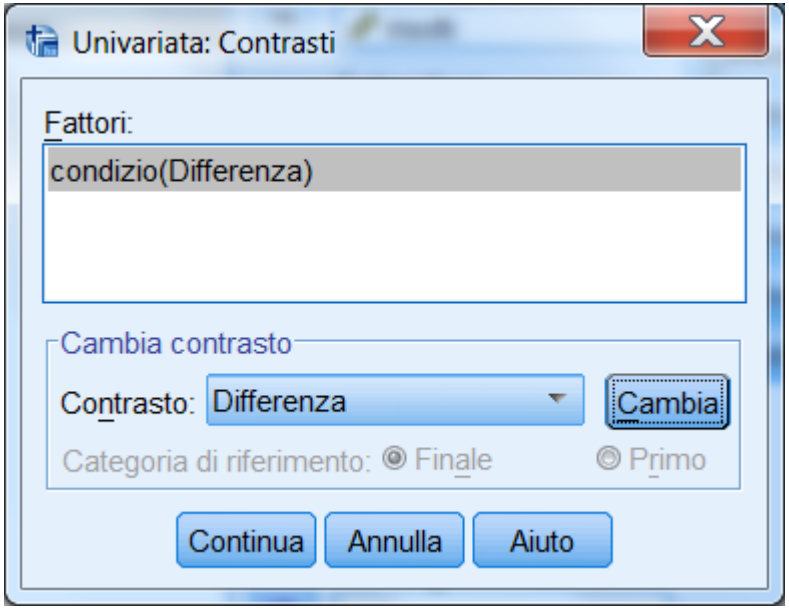
	Obiettivi	Ob. + Feed.	Controllo	Somme
1° confronto	1	-1	0	0
2° confronto	0	1	-1	0
3° confronto	1	0	-1	0

I confronti sono tutti corretti (Somme = 0), ma non sono ortogonali. Per verificare l'ortogonalità bisogna confrontare ciascuna coppia di confronti. Per ognuna di esse la somma dei prodotti dei coefficienti deve essere uguale a zero.

		Prodotti		Somme
1° vs. 2°	0 (=1*0)	-1 (=-1*1)	0 (=0*-1)	-1
1° vs. 3°	1 (=1*1)	0 (=-1*0)	0 (=0*-1)	1
2° vs. 3°	0 (=0*1)	0 (= 1*0)	1 (=-1*-1)	1

ANOVA IN SPSS





ANOVA IN SPSS

Risultati del contrasto (matrice K)

Contrasto differenza condizio			Variabile ... risolti
Livello 2 vs livello 1	Stima contrasto		,200
	Valore ipotizzato		0
	Differenza (Stima-Ipotesi)		,200
	Deviazione standard Errore		1,311
	Sig.		,881
	Intervallo di confidenza per la differenza al 95%	Limite inferiore	-2,657
		Limite superiore	3,057
Livello 3 contro precedente	Stima contrasto		-4,700
	Valore ipotizzato		0
	Differenza (Stima-Ipotesi)		-4,700
	Deviazione standard Errore		1,136
	Sig.		,001
	Intervallo di confidenza per la differenza al 95%	Limite inferiore	-7,175
		Limite superiore	-2,225

Contrasto 1: Obiettivi vs. Obiettivi&Feedback

Contrasto 2: Obiettivi+ Obiettivi&Feedback vs. Controllo

ANOVA A UNA VIA "BETWEEN" IN SPSS

Per richiedere i confronti pianificati bisogna ricorrere alla programmazione Syntax aggiungendo le seguenti linee dopo /INTERCEPT = INCLUDE:
/lmatrix condizio 1 1 –2 /lmatrix condizio 1 -1 0 oppure
/CONTRAST (condizio)=special (1 1 -2) /CONTRAST (condizio)=special (1 -1 0)

```
DATASET ACTIVATE InsiemeDati1.
UNIANOVA risolti BY condizio
  /CONTRAST(condizio)=Difference
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
/LMATRIX CONDIZIO 1 1 -2/LMATRIX CONDIZIO 1 -1 0
  /PLOT=PROFILE(condizio)
  /EMMEANS=TABLES(condizio) COMPARE ADJ(SIDAK)
  /PRINT=OPOWER ETASQ HOMOGENEITY DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /DESIGN=condizio.
```

ANOVA IN SPSS

Test di ipotesi personalizzate #2

Risultati del test

Variabile dipendente:risolti

Sorgente	Somma dei quadrati	df	Media dei quadrati	F	Sig.	Eta quadrato parziale
Contrasto	73,633	1	73,633	17,124	,001	,588
Errore	51,600	12	4,300			

Risultati del test

Variabile dipendente:risolti

Sorgente	Non centralità Parametro	Potenza osservata[a]
Contrasto	17,124	,966

a. Calcolato usando alfa = ,05

Contrasto: Obiettivi+ Obiettivi&Feedback vs. Controllo

ANOVA IN SPSS

Test di ipotesi personalizzate #3

Risultati del test

Variabile dipendente:risolti

Sorgente	Somma dei quadrati	df	Media dei quadrati	F	Sig.	Eta quadrato parziale
Contrasto	,100	1	,100	,023	,881	,002
Errore	51,600	12	4,300			

Risultati del test

Variabile dipendente:risolti

Sorgente	Non centralità Parametro	Potenza osservata[a]
Contrasto	,023	,052

a. Calcolato usando alfa = ,05

Contrasto: Obiettivi vs. Obiettivi&Feedback

Confronto 1

Fonte	SS	df	MS	F	Sig.
Contrasto	73.63	1	73.63	17.12	.001
Errore	51.60	12	4.30		

Confronto 2

Fonte	SS	df	MS	F	Sig.
Contrasto	.10	1	.10	.023	.881
Errore	51.60	12	4.30		

Il denominatore utilizzato nella F dei due confronti è sempre quello relativo alla varianza residua del test "omnibus" (Errore = 4.30).

Significato dell'ortogonalità

I confronti ortogonali forniscono informazioni indipendenti, cioè il risultato del primo non consente di ottenere indicazioni sul possibile risultato del secondo, e viceversa.
Numero massimo di confronti ortogonali = k - 1.

In un set completo di k-1 confronti ortogonali la somma delle devianze tra i gruppi dei singoli confronti è uguale alla devianza spiegata dall'effetto "omnibus" nell'ANOVA. La devianza spiegata dall'effetto viene scomposta in un certo numero di "porzioni" tra loro indipendenti (nell'esempio: 73.63 + .10 =73.73).

ESERCIZIO 3:

REALIZZAZIONE DI UN'ANOVA AD UNA VIA

Effettuare un Anova ad una via.

I dati sono nel file spss esercizio.anova.sav

VARIABILE DIPENDENTE: atte

VARIABILE INDIPENDENTE: tits

L'ANALISI DELLA VARIANZA UNIVARIATA (ANOVA): DISEGNI FATTORIALI

Vengono definiti fattoriali (o a più vie) i disegni di analisi della varianza in cui vi sono due o più variabili indipendenti.

Disegno fattoriale più semplice: Disegno "2X2", Due fattori, ciascuno con due differenti livelli ("condizioni").

Vantaggi dei disegni fattoriali

1) Aumento della potenza del test, perché viene ridotta la varianza di errore.

2) Maggiore economia nel numero dei soggetti da esaminare, mantenendo la stessa potenza del test.

3) Studio dell'interazione, cioè dell'effetto congiunto delle VI sulla VD.

EFFETTI PRINCIPALI E INTERAZIONI

Effetto principale:
effetto medio di una VI sulla VD, indipendentemente dai valori delle altre VI.

Interazione:
effetto di una VI sulla VD che si verifica solo a determinati livelli dell'altra VI;

effetto di una VI sulla VD che non è lo stesso per tutti i livelli delle altre VI.

Esempio con un disegno fattoriale 2x3

Abilità	Trattamento			
	T1	T2	T3	$m_{i.}$
Bassa	85.00	80.00	76.00	80.33
Alta	60.00	63.00	68.00	63.67
$m_{.j}$	72.50	71.50	72.00	

Medie Marginali di Colonna

Medie delle celle

Medie Marginali di Riga

Ipotesi per effetti principali e interazione

Effetti principali:

Trattamento: ipotesi sulle medie delle colonne.

$H_0: \mu_{.1} = \mu_{.2} = \mu_{.3}$

(A livello di campione: 72.5 = 71.5 = 72)

H_1: **Almeno due medie sono differenti:**

$(\mu_{.1} \neq \mu_{.2})$ o $(\mu_{.1} \neq \mu_{.3})$ o $(\mu_{.2} \neq \mu_{.3})$

Abilità: ipotesi sulle medie delle righe

$H_0: \mu_{1.} = \mu_{2.}$

(A livello di campione: 80.33 = 63.67)

$H_1: \mu_{1.} \neq \mu_{2.}$

Ipotesi per effetti principali e interazione

Interazione:

Ipotesi sulle differenze delle medie nelle diverse combinazioni delle condizioni sperimentali.

$$H_0: (\mu_A - \mu_B)_{T1} = (\mu_A - \mu_B)_{T2} = (\mu_A - \mu_B)_{T3}$$

[A livello di campione:
(85-60) = (80-63) = (76-68), cioè, 25 =17= 8 ?]

H_1: Almeno una differenza tra differenze di medie è sign.

Tutte le volte che c'è un'interazione nei dati, sarebbe fuorviante interpretare gli effetti principali senza discutere le interazioni.

Disegni fattoriali "Tra i soggetti" (Between Subjects):

I soggetti vengono assegnati casualmente ad ognuna delle singole celle (incrocio di due livelli diversi dei due fattori). Ogni soggetto è esposto solamente ad una particolare combinazione delle condizioni sperimentali. Ogni cella contiene soggetti diversi.

	FEEDBACK	
OBIETTIVI	**SI**	**NO**
SI	S_1 S_2 ...	S_6 S_7 ...
NO	S_{11} S_{12} ...	S_{16} S_{17} ...

Modello Teorico dei Disegni fattoriali "Tra i soggetti"

In un disegno fattoriale con 2 fattori "between" F1 e F2, il punteggio y_{ijk} di un soggetto "k" contenuto nella "cella" "ij" è scomponibile nel modo seguente:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \phi_{ij} + \varepsilon_{ijk}$$

$\alpha_i = \mu_{i..} - \mu$: **effetto principale di F1 (deviazione della media della i-esima riga dalla media generale)**

$\beta_j = \mu_{.j.} - \mu$: **effetto principale di F2 (deviazione della media della j-esima colonna dalla media generale)**

$\phi_{ij} = \mu_{ij.} - \mu - (\alpha_i + \beta_j)$: **effetto dell'interazione. Parte della media di una cella ij che non dipende dall'errore, e che non è spiegata né dalla media generale, né dagli effetti principali.**

ε_{ijk}: **termine residuale ("errore")**

In base al modello precedente è possibile definire le seguenti devianze:

$$SS_T = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$$ **dev. totale**

$$SS_{F1} = \sum_i \sum_j \sum_k (\bar{y}_{i..} - \bar{y}_{...})^2$$ **dev. eff. princ. di F1**

$$SS_{F2} = \sum_i \sum_j \sum_k (\bar{y}_{.j.} - \bar{y}_{...})^2$$ **dev. eff. princ. di F2**

$$SS_{F1XF2} = \sum_i \sum_j \sum_k (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$ **dev. interazione**

$$SS_W = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2$$ **devianza residua**

$$SS_T = SS_B + SS_W = SS_{F1} + SS_{F2} + SS_{F1XF2} + SS_W$$

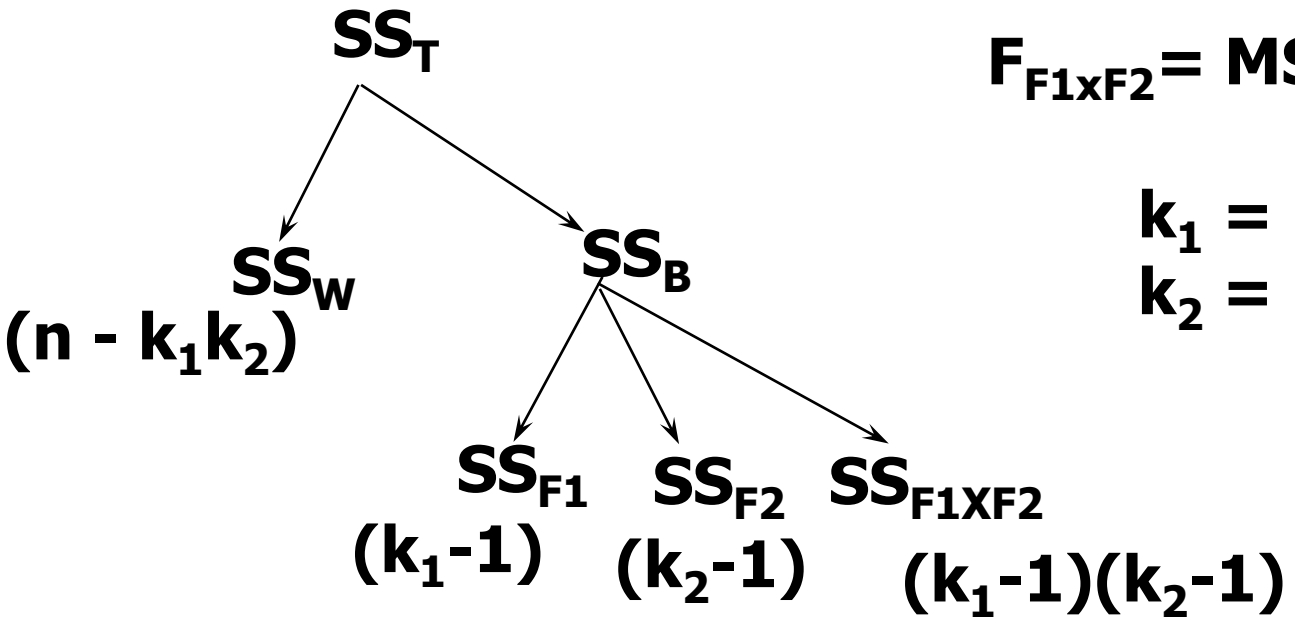
Gradi di libertà e test di significatività

$$F_{F1} = MS_{F1} / MS_W$$

$$F_{F2} = MS_{F2} / MS_W$$

$$F_{F1xF2} = MS_{F1xF2} / MS_W$$

k_1 = livelli di F1
k_2 = livelli di F2



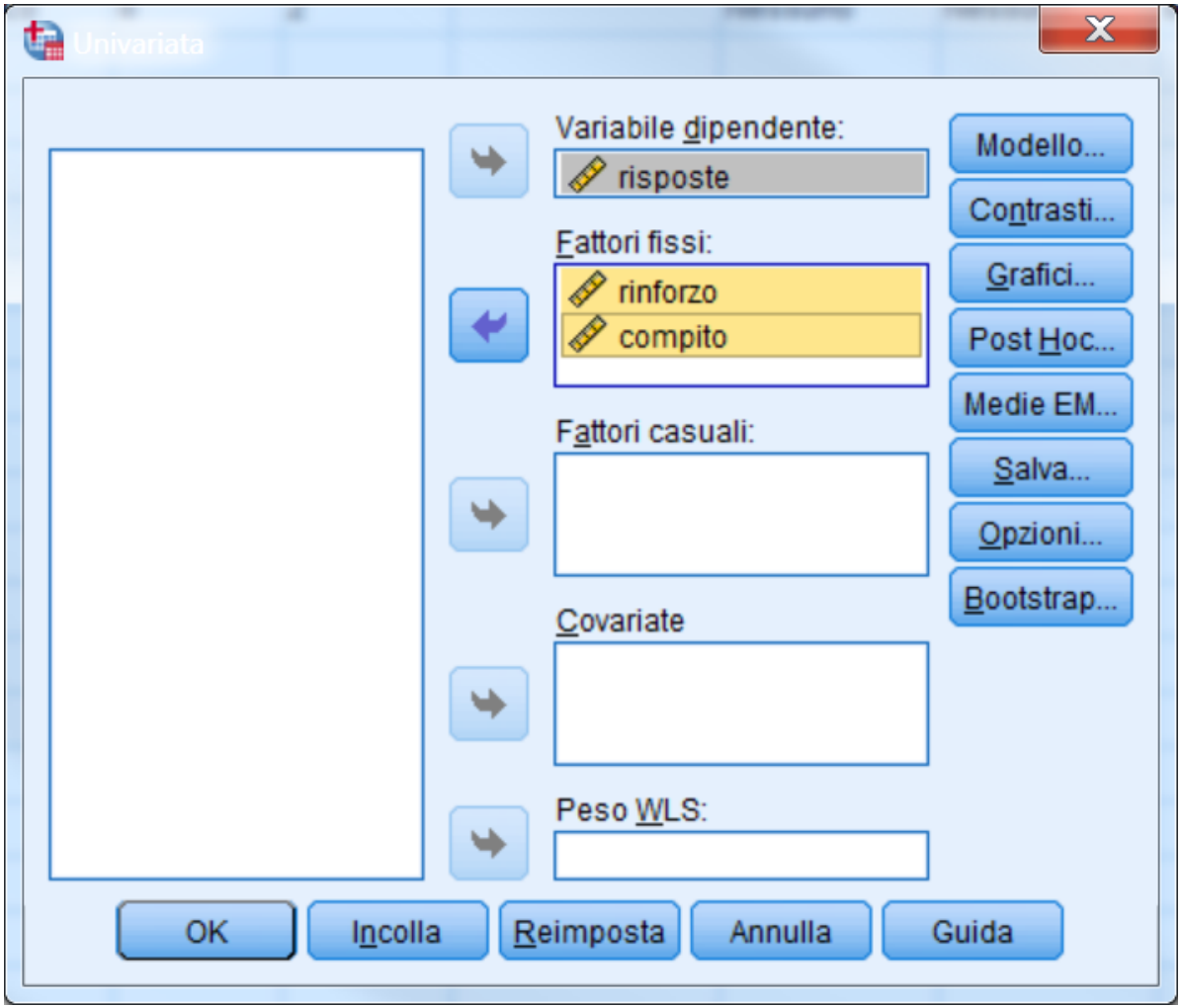
ANALISI DEI DISEGNI FATTORIALI

Esempio da Keppel et al., pp. 260 e segg.
2 fattori (o var. indipendenti): Rinforzo e Compito;
1 variabile dipendente: n. di problemi risolti.

Fattori tra soggetti

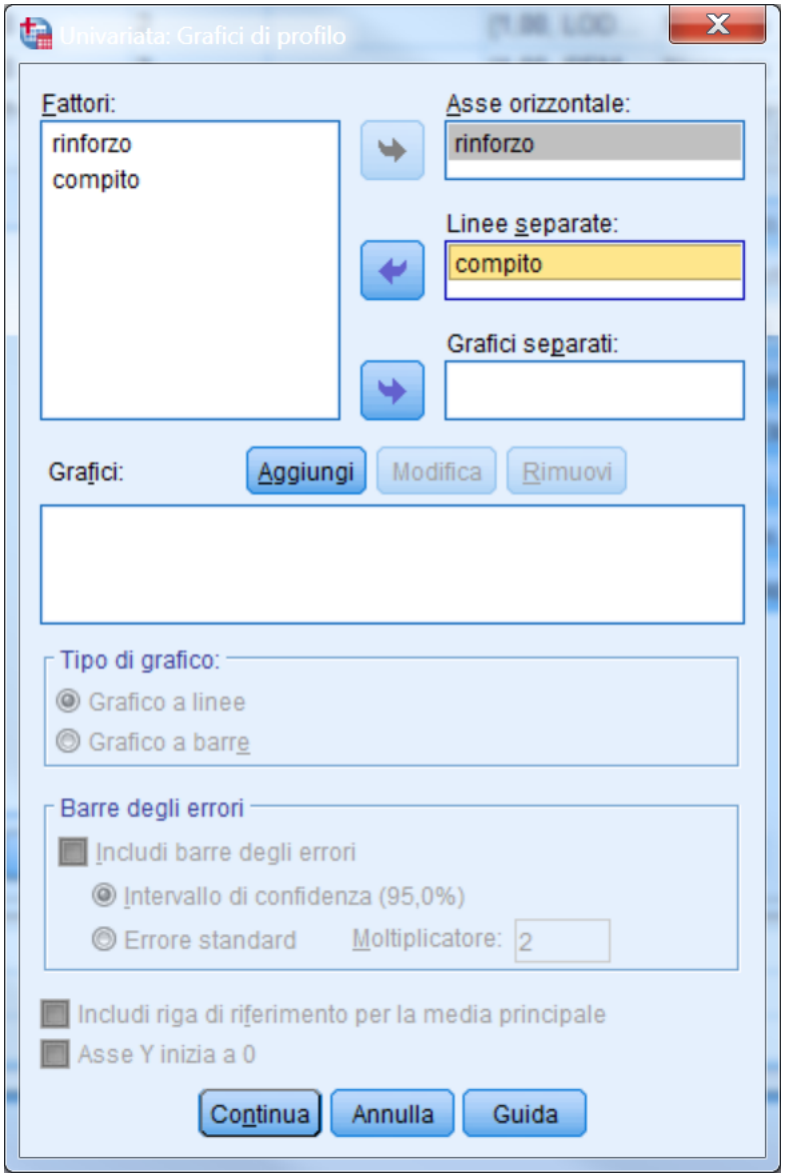
		Etichetta di valore	N
RINFORZO	1,00	LODE	10
	2,00	CRITICA	10
	3,00	SILENZIO	10
COMPITO	1,00	SEMPLICI	15
	2,00	COMPLESSI	15

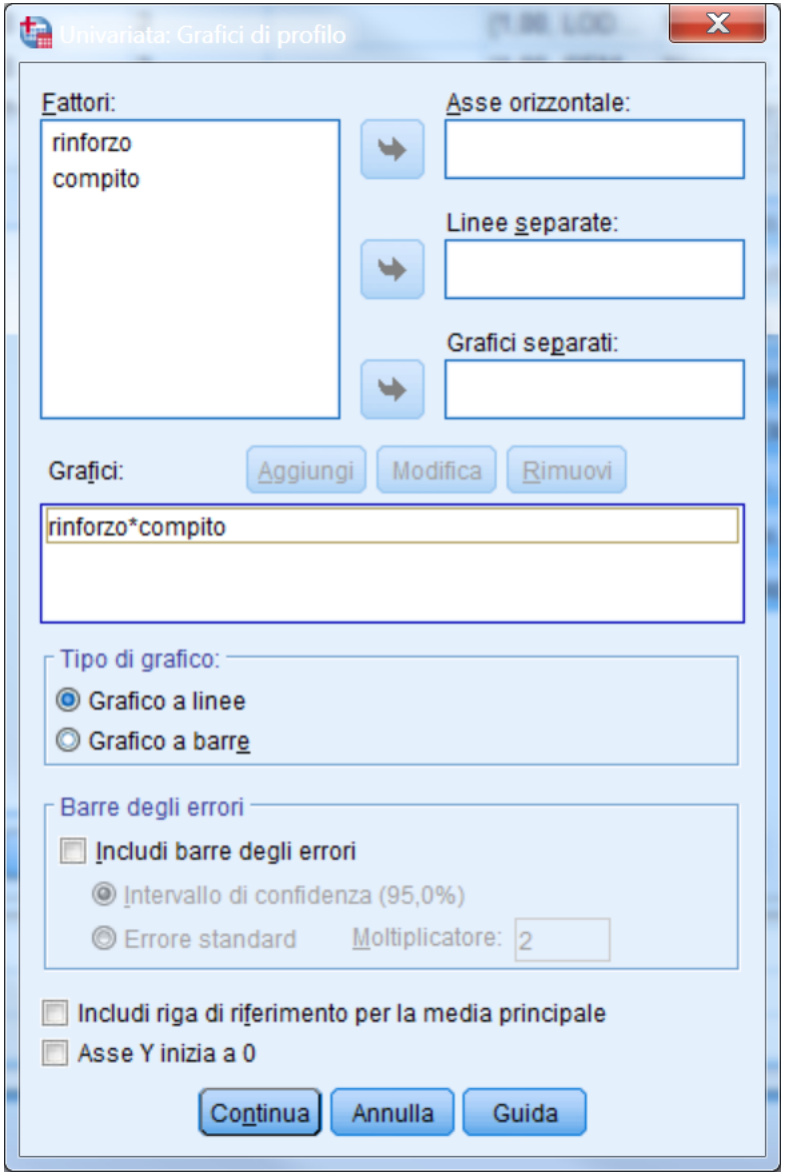
ANOVA FATTORIALE BETWEEN IN SPSS



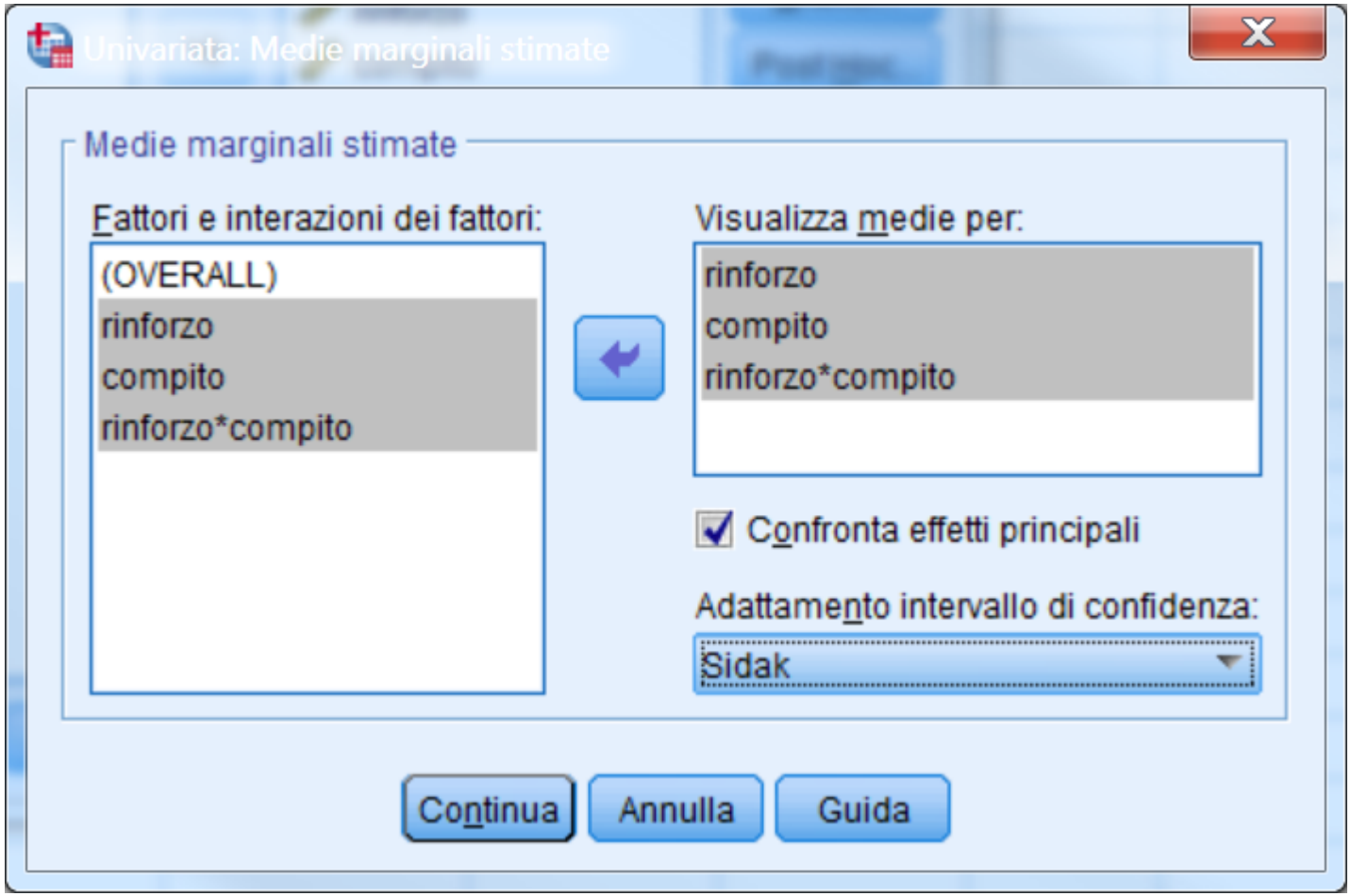
ANOVA.FAC.B.sav

ANOVA FATTORIALE BETWEEN IN SPSS

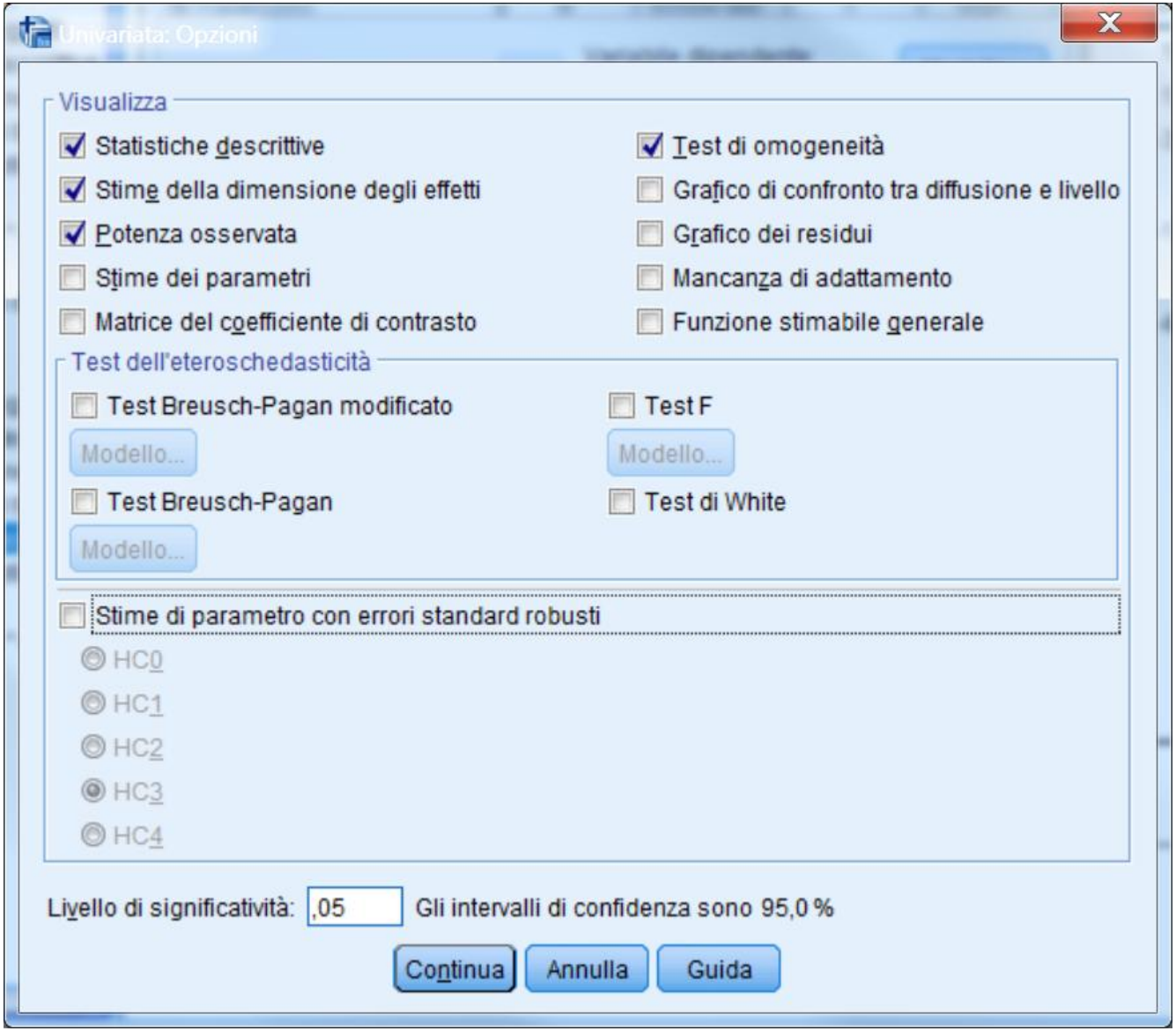




ANOVA FATTORIALE BETWEEN IN SPSS



ANOVA FATTORIALE BETWEEN IN SPSS



ANOVA FATTORIALE BETWEEN IN SPSS

Fattori tra soggetti

		Etichetta di valore	N
rinforzo	1,00	LODE	10
	2,00	CRITICA	10
	3,00	SILENZIO	10
compito	1,00	SEMPLICI	15
	2,00	COMPLESSI	15

Statistiche descrittive

Variabile dipendente:risposte

rinforzo	compito	Media	Deviazione standard Variabile	N
1,00 LODE	1,00 SEMPLICI	7,6000	1,51658	5
	2,00 COMPLESSI	7,0000	2,00000	5
	Totale	7,3000	1,70294	10
2,00 CRITICA	1,00 SEMPLICI	7,2000	2,16795	5
	2,00 COMPLESSI	2,0000	1,58114	5
	Totale	4,6000	3,27278	10
3,00 SILENZIO	1,00 SEMPLICI	4,4000	1,94936	5
	2,00 COMPLESSI	3,2000	1,92354	5
	Totale	3,8000	1,93218	10
Totale	1,00 SEMPLICI	6,4000	2,29285	15
	2,00 COMPLESSI	4,0667	2,78944	15
	Totale	5,2333	2,77530	30

ANOVA FATTORIALE BETWEEN IN SPSS

Test di Levene di eguaglianza delle varianze dell'errore[a,b]

		Statistica di Levene	gl1	gl2	Sign.
risposte	Basato sulla media	,348	5	24	,879
	Basato sulla mediana	,123	5	24	,986
	Basato sulla mediana e con il grado di libertà adattato	,123	5	20,800	,986
	Basato sulla media ritagliata	,328	5	24	,891

Verifica l'ipotesi nulla che la varianza dell'errore della variabile dipendente sia uguale tra i gruppi.

a. Variabile dipendente: risposte

b. Disegno: Intercetta + rinforzo + compito + rinforzo * compito

ANOVA FATTORIALE BETWEEN IN SPSS

Test degli effetti fra soggetti

Variabile dipendente:risposte

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.
Modello corretto	139,367[a]	5	27,873	7,964	,000
Intercetta	821,633	1	821,633	234,752	,000
rinforzo	67,267	2	33,633	9,610	,001
compito	40,833	1	40,833	11,667	,002
rinforzo * compito	31,267	2	15,633	4,467	,022
Errore	84,000	24	3,500		
Totale	1045,000	30			
Totale corretto	223,367	29			

Test degli effetti fra soggetti

Variabile dipendente:risposte

Sorgente	Eta quadrato parziale	Non centralità Parametro	Potenza osservata[b]
Modello corretto	,624	39,819	,997
Intercetta	,907	234,752	1,000
rinforzo	,445	19,219	,966
compito	,327	11,667	,906
rinforzo * compito	,271	8,933	,710

a. R quadrato = ,624 (R quadrato corretto = ,546)
b. Calcolato usando alfa = ,05

ANOVA FATTORIALE BETWEEN IN SPSS

Stime

Variabile dipendente: risposte

rinforzo	Media	Errore std.	Intervallo di confidenza 95%	
			Limite inferiore	Limite superiore
1,00 LODE	7,300	,592	6,079	8,521
2,00 CRITICA	4,600	,592	3,379	5,821
3,00 SILENZIO	3,800	,592	2,579	5,021

Confronti pairwise

Variabile dipendente: risposte

(I) rinforzo	(J) rinforzo	Differenza della media (I-J)	Errore std.	Sign.[b]	95% intervallo di confidenza	95% intervallo di confidenza per [b]
					Limite inferior	Limite superiore
1,00 LODE	2,00 CRITICA	2,700*	,837	,011	,553	4,847
	3,00 SILENZIO	3,500*	,837	,001	1,353	5,647
2,00 CRITICA	1,00 LODE	-2,700*	,837	,011	-4,847	-,553
	3,00 SILENZIO	,800	,837	,723	-1,347	2,947
3,00 SILENZIO	1,00 LODE	-3,500*	,837	,001	-5,647	-1,353
	2,00 CRITICA	-,800	,837	,723	-2,947	1,347

ANOVA FATTORIALE BETWEEN IN SPSS

Stime

Variabile dipendente: risposte

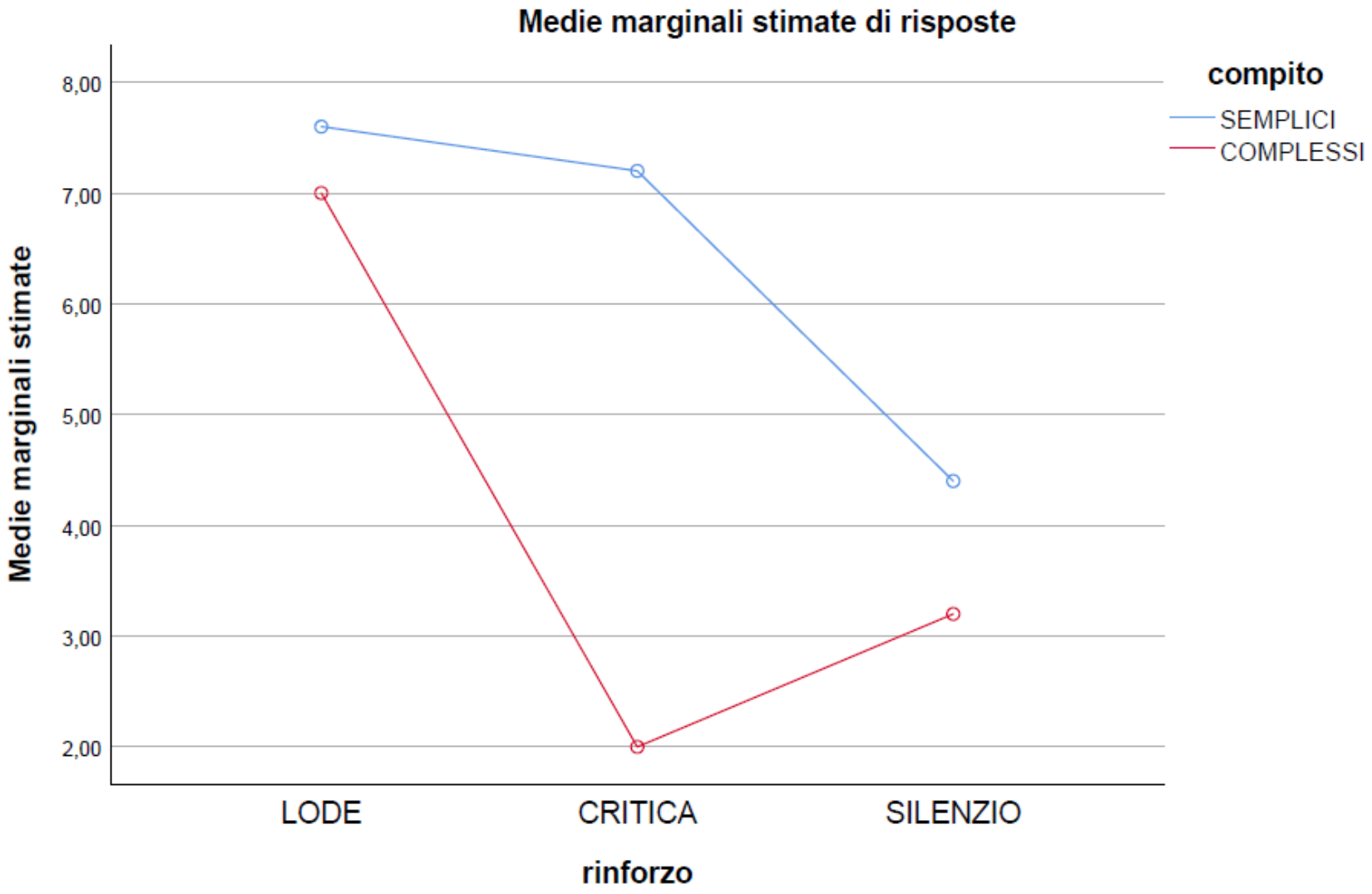
compito	Media	Errore std.	Intervallo di confidenza 95%	
			Limite inferiore	Limite superiore
1,00 SEMPLICI	6,400	,483	5,403	7,397
2,00 COMPLESSI	4,067	,483	3,070	5,064

3. rinforzo * compito

Variabile dipendente: risposte

rinforzo	compito	Media	Errore std.	Intervallo di confidenza 95%	
				Limite inferiore	Limite superiore
1,00 LODE	1,00 SEMPLICI	7,600	,837	5,873	9,327
	2,00 COMPLESSI	7,000	,837	5,273	8,727
2,00 CRITICA	1,00 SEMPLICI	7,200	,837	5,473	8,927
	2,00 COMPLESSI	2,000	,837	,273	3,727
3,00 SILENZIO	1,00 SEMPLICI	4,400	,837	2,673	6,127
	2,00 COMPLESSI	3,200	,837	1,473	4,927

ANOVA FATTORIALE BETWEEN IN SPSS



ANALISI DEI DISEGNI FATTORIALI

Risultati ottenuti da SPSS e/o tramite le formule definite per i disegni ANOVA fattoriali:

Fonte	SS	df	MS	F	Sig.
RINFORZO	67.27	2	33.63	9.61	.001
COMPITO	40.83	1	40.83	11.67	.002
RINFORZO X COMPITO	31.27	2	15.63	4.47	.022
Errore	84.00	24	3.50		
Totale	223.37	29			

1. ANALISI DEGLI EFFETTI PRINCIPALI

Effetto principale del fattore "COMPITO":

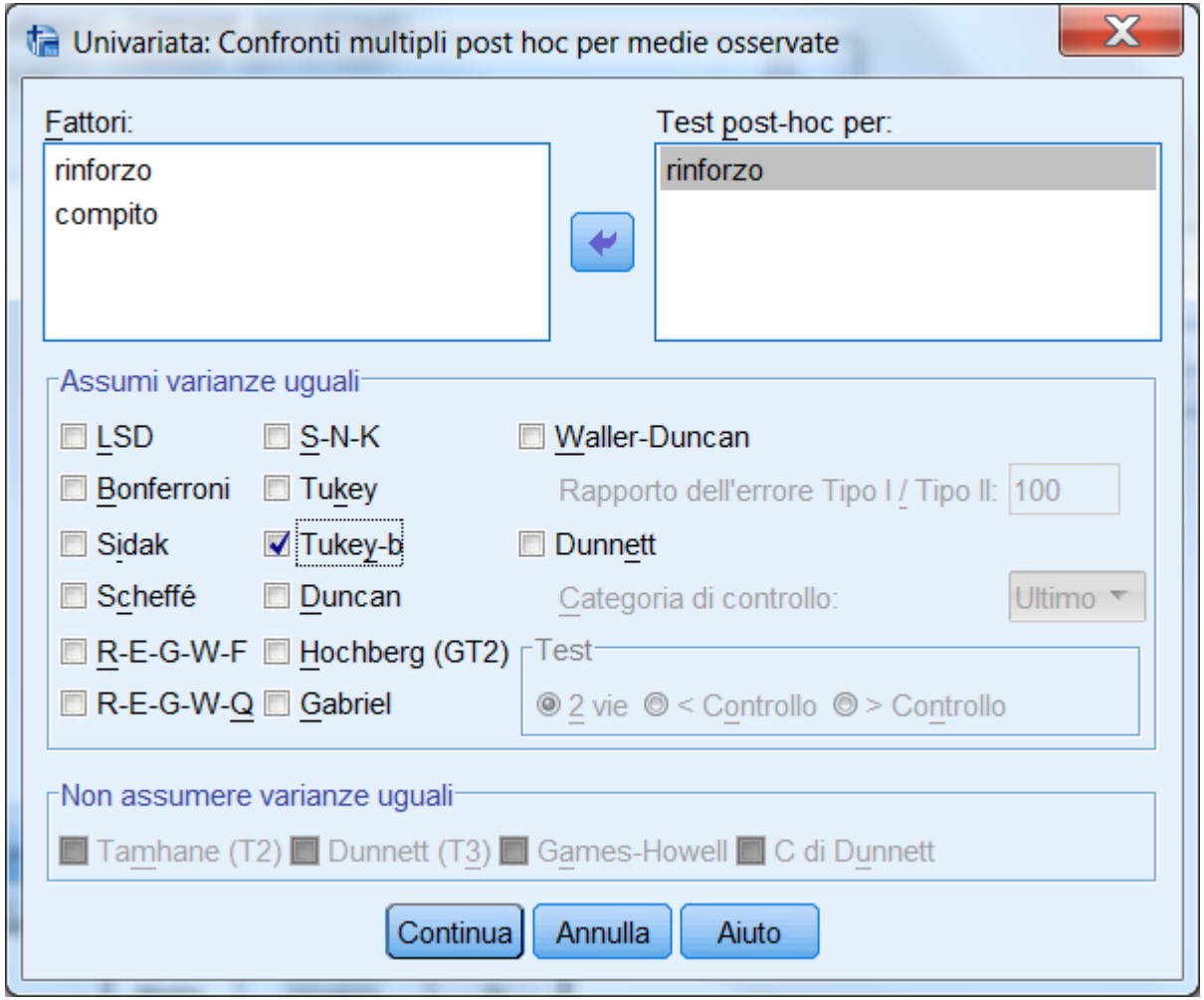
SEMPLICI	COMPLESSI
6.400	4.067

Effetto principale del fattore "RINFORZO":

LODE	CRITICA	SILENZIO
7.30	4.60	3.80

Confronti post-hoc con il metodo di Tukey-HSD: i due tipi di rinforzi Silenzio e Critica hanno medie uguali e significativamente diverse dal rinforzo Lode.

ANOVA FATTORIALE BETWEEN IN SPSS – post hoc Tukey-b



ANOVA FATTORIALE BETWEEN IN SPSS

Sottoinsiemi omogenei

risposte

B di Tukey[a,b]

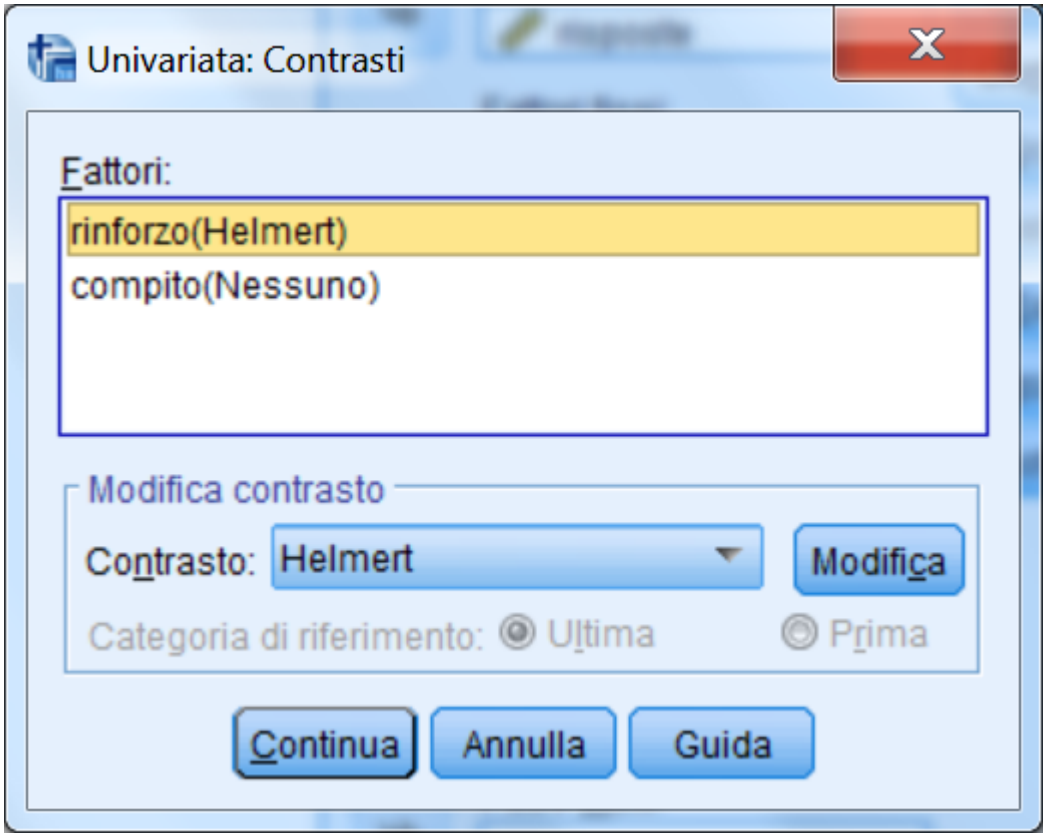
rinforzo	N	Sottoinsieme 1	Sottoinsieme 2
3,00 SILENZIO	10	3,8000	
2,00 CRITICA	10	4,6000	
1,00 LODE	10		7,3000

Sono visualizzate le medie per gruppi in sottoinsiemi omogenei.
Tali medie sono basate sulle osservazioni.
Il termine di errore è Media dei quadrati(errore) = 3,500.

a. Utilizza dimensione campionaria media armonica = 10,000
b. Alfa = ,05

ANOVA FATTORIALE BETWEEN IN SPSS – Confronti pianificati

Confronti pianificati. Possiamo confrontare le condizioni di Lode con quelle di Critica e Silenzio aggregate, e la condizione di Critica con Silenzio.



ANOVA FATTORIALE BETWEEN IN SPSS – Confronti pianificati

Risultati del contrasto (matrice K)

Contrasto di Helmert rinforzo			Variabile dipendente: risposte
Confronto tra livello 1 e successivo	Stima del contrasto		3,100
	Valore ipotizzato		0
	Differenza (stima - ipotizzato)		3,100
	Errore std.		,725
	Sign.		,000
	95% intervallo di confidenza per differenza	Limite inferiore	1,605
		Limite superiore	4,595
Confronto tra livello 2 e livello 3	Stima del contrasto		,800
	Valore ipotizzato		0
	Differenza (stima - ipotizzato)		,800
	Errore std.		,837
	Sign.		,349
	95% intervallo di confidenza per differenza	Limite inferiore	-,927
		Limite superiore	2,527

Contrasto 1: Lode vs. Critica + Silenzio

Contrasto 2: Critica vs. Silenzio

Confronti pianificati. Possiamo utilizzare anche i contrasti personalizzati tramite la Sintassi:

/CONTRAST(rinforzo)=SPECIAL (2-1-1)
/CONTRAST(rinforzo)=SPECIAL (0 1-1)

Risultati dei test

Variabile dipendente: risposte

Origine	Somma dei quadrati	gl	Media quadratica	F	Sign.	Eta quadrato parziale	Parametro di non centralità	Potenza osservata[a]
Contrasto	64,067	1	64,067	18,305	,000	,433	18,305	,984
Errore	84,000	24	3,500					

a. Calcolato utilizzando alfa = ,05

Risultati dei test

Variabile dipendente: risposte

Origine	Somma dei quadrati	gl	Media quadratica	F	Sign.	Eta quadrato parziale	Parametro di non centralità	Potenza osservata[a]
Contrasto	3,200	1	3,200	,914	,349	,037	,914	,151
Errore	84,000	24	3,500					

a. Calcolato utilizzando alfa = ,05

2. ANALISI DELL'INTERAZIONE

Nel nostro esempio l'interpretazione degli effetti principali può condurre a conclusioni errate.

RINFORZO	COMPITO	Media
LODE	SEMPLICI	7.6
	COMPLESSI	7.0
CRITICA	SEMPLICI	7.2
	COMPLESSI	2.0
SILENZIO	SEMPLICI	4.4
	COMPLESSI	3.2

La variabile Rinforzo produce un effetto sulla Variabile Risposte che è differente a seconda dei livelli della variabile Compito.

ANOVA FATTORIALE BETWEEN IN SPSS

3. rinforzo * compito

Variabile dipendente:risposte

rinforzo	compito	Media	Deviazione standard Errore	Intervallo di confidenza 95%	
				Limite inferiore	Limite superiore
1,00 LODE	1,00 SEMPLICI	7,600	,837	5,873	9,327
	2,00 COMPLESSI	7,000	,837	5,273	8,727
2,00 CRITICA	1,00 SEMPLICI	7,200	,837	5,473	8,927
	2,00 COMPLESSI	2,000	,837	,273	3,727
3,00 SILENZIO	1,00 SEMPLICI	4,400	,837	2,673	6,127
	2,00 COMPLESSI	3,200	,837	1,473	4,927

2. ANALISI DELL'INTERAZIONE

Analisi degli EFFETTI SEMPLICI:

Serve per identificare le combinazioni dei fattori che danno un'interazione significativa.

Effetti Semplici ("Simple Effects"):
esame dei valori della variabile dipendente associati ai valori di una VI, quando i valori dell'altra VI sono mantenuti costanti.

Analisi degli EFFETTI SEMPLICI:

- Disegno fattoriale semplificato effettuando tanti disegni "monofattoriali" quanti sono i livelli della VI che viene mantenuta costante.
- Se c'è un'interazione significativa, gli effetti semplici relativi ad una VI sono diversi nei livelli della VI che viene controllata.
- Gli Effetti Semplici consentono di evidenziare l'effetto di modulazione che una VI ha sulla relazione tra un'altra VI e la VD.
- L'analisi degli effetti principali annulla tale effetto, poiché confronta le medie marginali, nelle quali i livelli dell'altra variabile indipendente vengono sommati tra di loro.

ANOVA FATTORIALE BETWEEN IN SPSS

Poiché l'interazione è risultata significativa interpretare gli effetti principali isolatamente sarebbe inappropriato. Attraverso l'analisi degli effetti semplici possiamo vedere come l'effetto di un fattore sulla VD non è lo stesso per i diversi livelli dell'altro fattore.
Non è possibile ottenere gli effetti semplici dal menu di Spss. Per ottenerli è necessario ricorrere al linguaggio di programmazione Syntax.

Gli effetti semplici relativi al fattore Compito nei diversi livelli del fattore Rinforzo possono essere richiesti tramite la seguente sintassi:

```
UNIANOVA   risposte  BY rinforzo compito   /METHOD = SSTYPE(3)   /INTERCEPT = INCLUDE
 /EMMEANS = TABLES(rinforzo*compito) COMPARE (COMPITO) ADJ(SIDAK)
  /CRITERIA = ALPHA(.05)   /DESIGN = rinforzo compito rinforzo*compito .
```

Per ottenere gli effetti semplici relativi al fattore Rinforzo nei diversi livelli del fattore Compito dobbiamo utilizzare la seguente sintassi:

```
UNIANOVA
  risposte  BY rinforzo compito   /METHOD = SSTYPE(3)   /INTERCEPT = INCLUDE
  /EMMEANS = TABLES(rinforzo*compito) COMPARE (rinforzo) ADJ(SIDAK)
   /CRITERIA = ALPHA(.05)   /DESIGN = rinforzo compito rinforzo*compito .
```

ANOVA FATTORIALE BETWEEN IN SPSS

Test univariati

Variabile dipendente: risposte

rinforzo		Somma dei quadrati	gl	Media quadratica	F	Sign.
1,00 LODE	Contrasto	,900	1	,900	,257	,617
	Errore	84,000	24	3,500		
2,00 CRITICA	Contrasto	67,600	1	67,600	19,314	,000
	Errore	84,000	24	3,500		
3,00 SILENZIO	Contrasto	3,600	1	3,600	1,029	,321
	Errore	84,000	24	3,500		

Variabile dipendente: risposte

rinforzo		Eta quadrato parziale	Parametro di non centralità	Potenza osservata[a]
1,00 LODE	Contrasto	,011	,257	,078
	Errore			
2,00 CRITICA	Contrasto	,446	19,314	,988
	Errore			
3,00 SILENZIO	Contrasto	,041	1,029	,164
	Errore			

Ogni F verifica gli effetti semplici di compito all'interno di ciascuna combinazione di livello degli altri effetti visualizzati. Questi test si basano sui confronti pairwise linearmente indipendenti tra le medie marginali stimate.

a. Calcolato utilizzando alfa = ,05

Analisi degli EFFETTI SEMPLICI nell'esempio empirico:

Analisi degli effetti semplici per il fattore "Compito" mantenendo costante il fattore "Rinforzo" (l'analisi del fattore "Rinforzo" mantenendo costante il fattore "Compito" dà risultati analoghi).

RINFORZO		SS	df	MS	F	Sig.
LODE	Contrasto	.90	1	.90	.26	.62
	Errore	84.00	24	3.50		
CRITICA	Contrasto	67.60	1	67.60	19.31	.000
	Errore	84.00	24	3.50		
SILENZIO	Contrasto	3.60	1	3.60	1.03	.32
	Errore	84.00	24	3.50		

La devianza Between che viene scomposta è data dalla somma della devianza del fattore "COMPITO" (40.83) più la devianza dell'interazione (31.27), ovvero: .90+67.60+3.60=72.1 = 40.83+31.27. La devianza Within è quella del disegno fattoriale completo (84.00).

ESERCIZIO 4:

REALIZZAZIONE DI UN'ANOVA FATTORIALE

Effettuare una Anova fattoriale .

I dati sono nel file spss esercizio.anova.sav

VARIABILE DIPENDENTE: inte

VARIABILI INDIPENDENTI: tits marcpast

L'ANALISI FATTORIALE ESPLORATIVA (EFA)

Sommario

* Concetti fondamentali

* Equazioni fondamentali

* Metodi di estrazione dei fattori

* Metodi per stabilire il numero di fattori

* Metodi di rotazione dei fattori

* Assunzioni statistiche e prerequisiti

Analisi Fattoriale Esplorativa

Scopo dell'Analisi Fattoriale è quello di studiare le relazioni in un insieme di variabili per:

a) individuare "*dimensioni latenti*" che spieghino le relazioni tra le variabili

questo solitamente porta a…

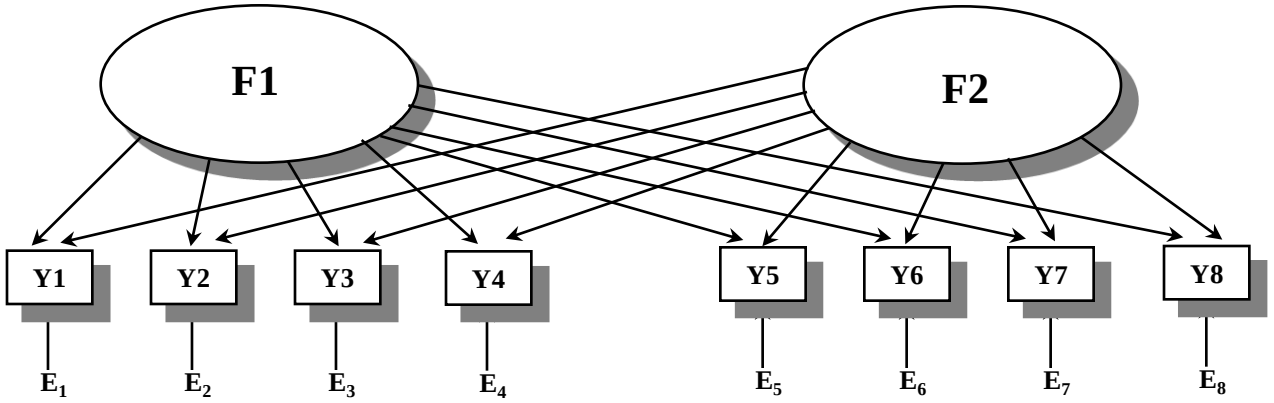
b) ridurre l'informazione in un insieme di dati

Da dove si parte…. Dati non strutturati

```
1.00
 .36 1.00
 .25  .37 1.00
 .33  .43  .41 1.00
 .05  .16  .12  .12 1.00
 .04  .05  .16  .06  .31 1.00
 .08  .06  .12  .14  .31  .24 1.00
 .02  .10  .17  .04  .29  .34  .29 1.00
```

Non viene formulata nessuna ipotesi su cosa genera le correlazioni tra le variabili. Si osserva semplicemente che alcune variabili sono più correlate tra loro di altre.

Dove si arriva…. Dati strutturati



Le relazioni tra le variabili osservate sono ricondotte alla presenza di fattori latenti.

E' un'ipotesi teorica sottoponibile a verifica empirica

Dove si arriva…. Dati strutturati

	F1	F2
Y1	0.516	-0.061
Y2	0.659	-0.010
Y3	0.539	0.119
Y4	0.685	-0.027
Y5	0.047	0.531
Y6	-0.039	0.570
Y7	0.033	0.481
Y8	-0.034	0.594

Le relazioni tra variabili osservate e fattori sono le saturazioni fattoriali.

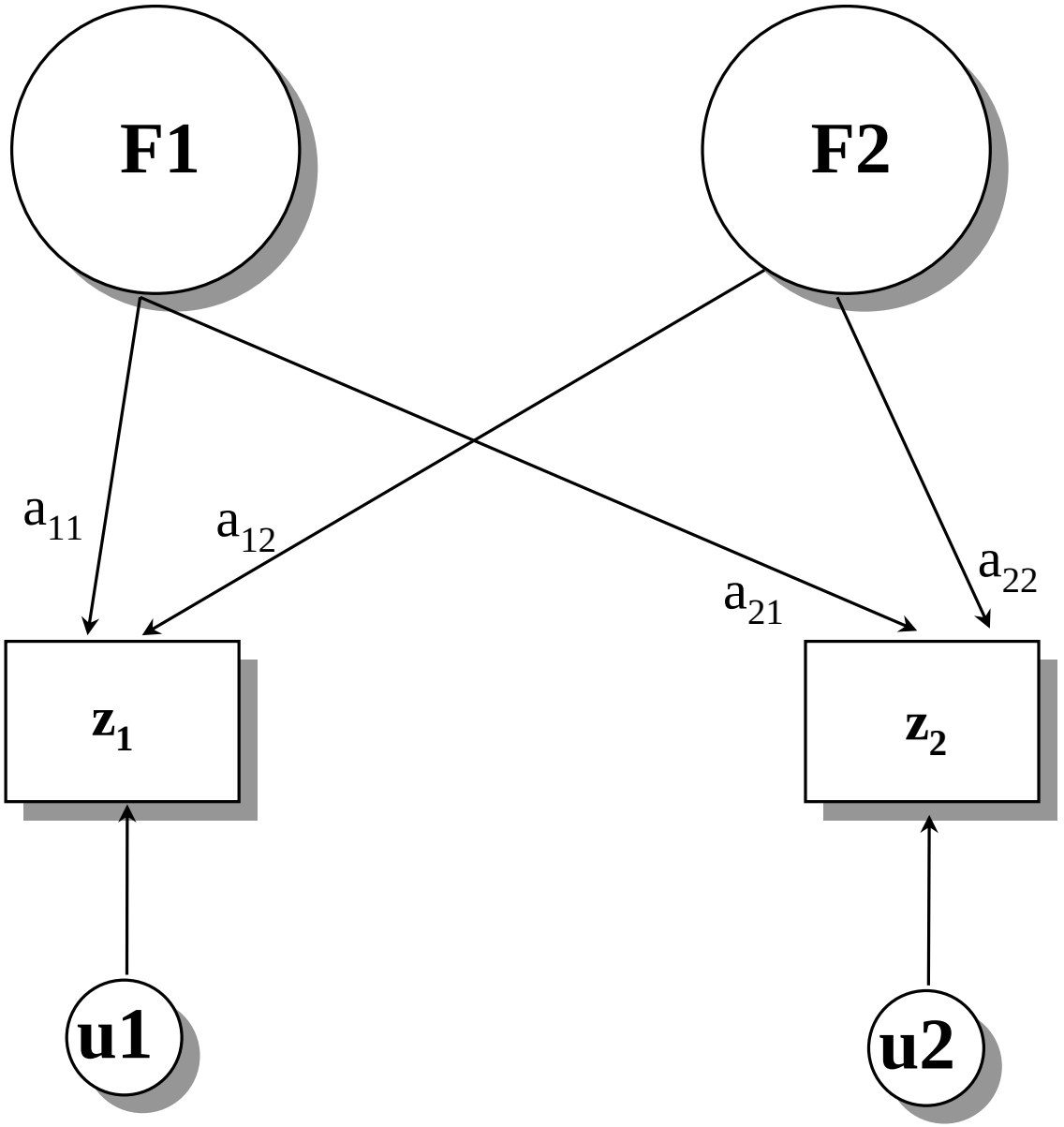
Modello teorico dell'analisi fattoriale

Esame della varianza che le variabili hanno in comune, ovvero della varianza comune.

Ipotesi di base:

La correlazione tra le variabili è determinata da dimensioni non osservabili (i fattori) che "causano" le variabili osservate.

Modello dell'analisi Fattoriale – Rappresentazione Grafica



Modello dell'analisi Fattoriale

- F = fattori comuni; rappresentano la variabilità condivisa tra le variabili in analisi. Possono influenzare più di una variabile osservata.

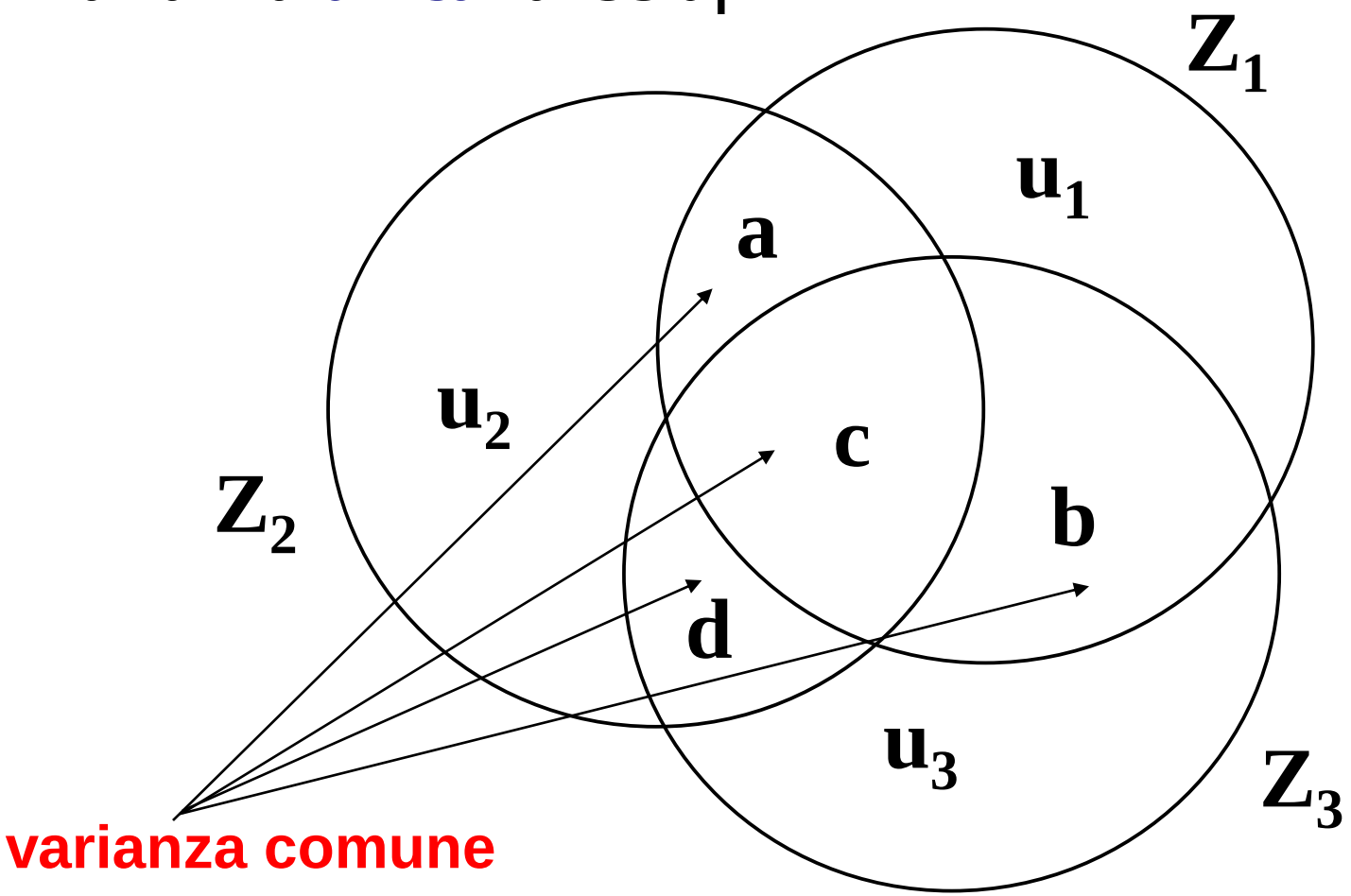
- a = saturazioni; relazioni tra variabili e fattori.

- u = termine unico o "unicità della variabile". Parte di varianza non condivisa. Dovuta a cause sistematiche specifiche, o all'errore casuale di misurazione.

Rappresentazione grafica della varianza comune

Parte di varianza comune delle 3 var: area (a+b+c+d).

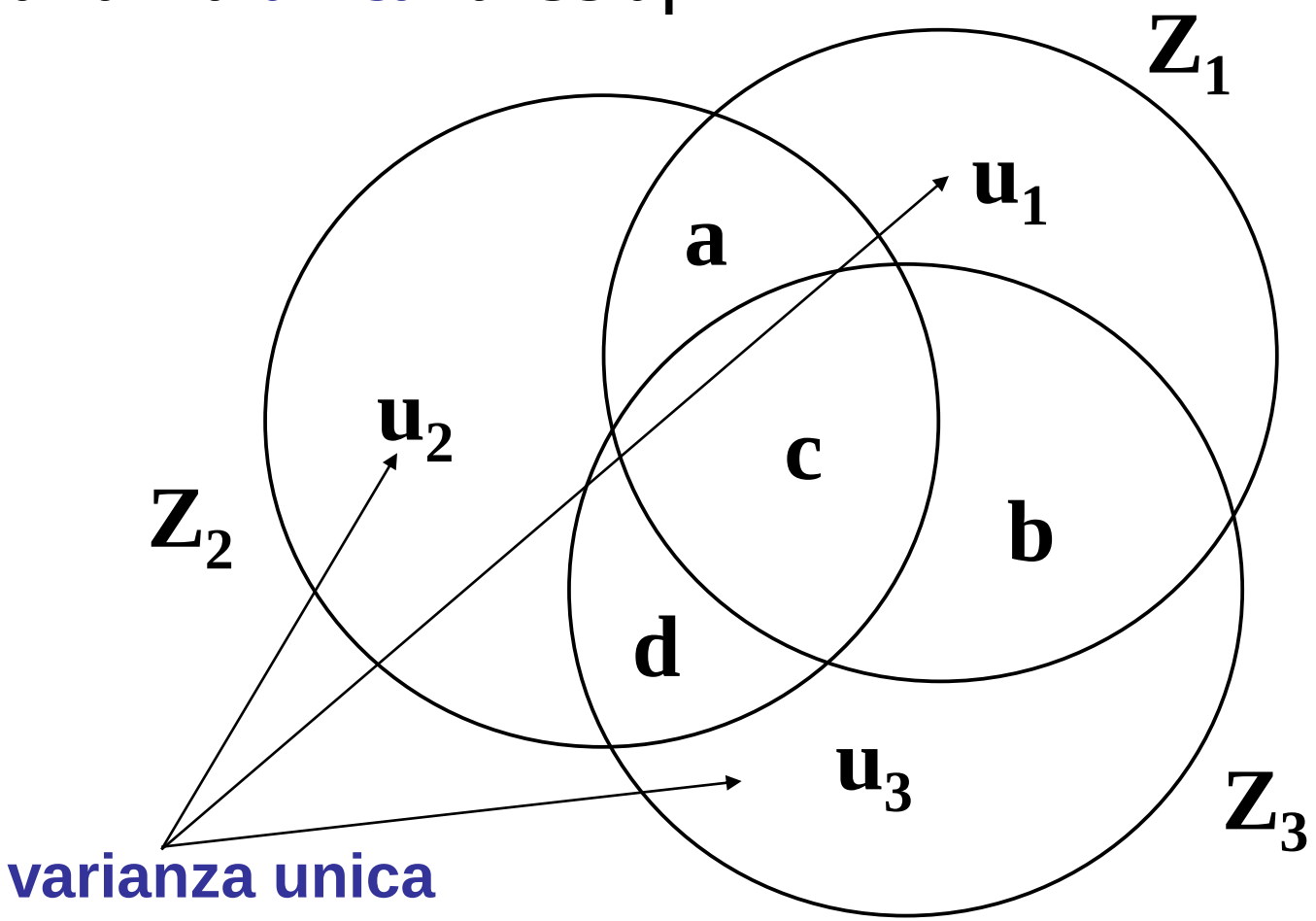
Varianza unica: aree u_i.



Rappresentazione grafica della varianza comune

Parte di varianza comune delle 3 var: area (a+b+c+d).

Varianza unica: aree u_i.



Modello dell'analisi fattoriale

Punteggio (standardizzato) di un soggetto in una variabile = somma "ponderata" del punteggio ottenuto dallo stesso soggetto:

a) nei fattori comuni;

b) in una componente unica.

a) e b) sono individuati tramite l'analisi fattoriale.

Equazione del modello teorico dell'analisi fattoriale

$$z_{ik} = F_{i1}\, a_{k1} + F_{i2}\, a_{k2} + \ldots + F_{im} a_{km} + u_{ik} \qquad (1)$$

z_{ik} = punteggio standardizzato per la persona i nella variabile k

F_{ij} = punteggi standardizzati per la persona i nei fattori comuni j

a_{kj} = saturazioni fattoriali della variabile k nei fattori comuni j

u_{ik} = punteggio standard. per la persona i nella componente unica associata alla variabile k

Espressione matriciale dell'equazione (1):

$$Z = FA' + U$$

Z: matrice dei punteggi standardizzati nelle variabili,

F: matrice dei punteggi nei fattori comuni,

A: matrice delle saturazioni delle variabili nei fattori,

U: matrice delle componenti uniche delle variabili.

Scomposizione della varianza di ogni variabile:

Varianza totale= 1 = $h^2 + u^2$

Comunalità = h^2.
Parte di varianza totale spiegata dai fattori comuni

Unicità o varianza unica = $u^2 = 1 - h^2$.
Parte di varianza totale non spiegata dai fattore comuni

Assunzioni:

$Cov(u_i,F_j)=0$, per ogni i e per ogni j
$Cov(u_i,u_j)=0$ per ogni i diversa da j
$Cov(F_i,F_j)$ diversa da 0 solo nelle soluzioni "oblique"

Espressioni matriciali

In base alle assunzioni e considerando che:

$$R = Z'Zn^{-1} \text{ e che } Z = FA' + U$$

si ha che:

$$R = AA' + U^2$$

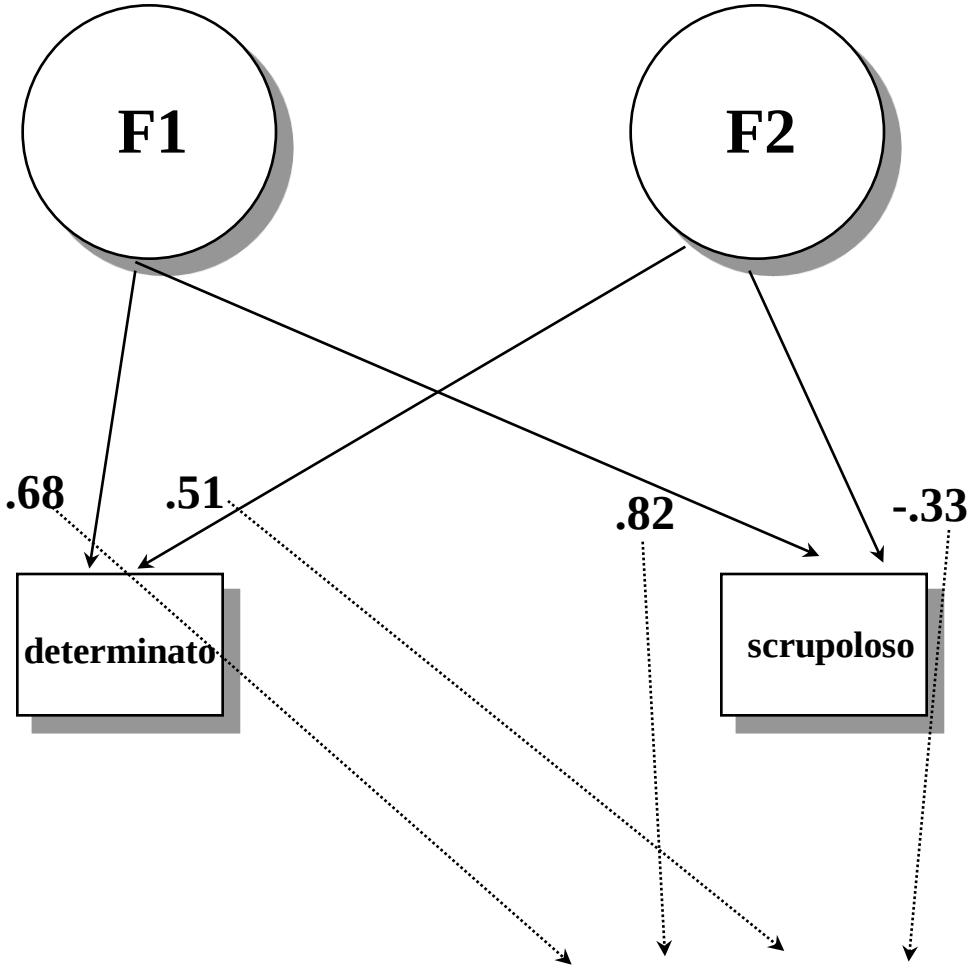
A = matrice delle saturazioni nei fattori comuni
U^2 = matrice diagonale delle varianze uniche.
AA': rende conto degli elementi fuori della diagonale principale, e della comunalità di ogni variabile.
U^2: contribuisce a rendere conto degli elementi sulla diagonale principale di R.

La correlazione tra due variabili i e j può essere riprodotta dalla somma dei prodotti delle loro saturazioni in ciascuno dei fattori comuni:

$$r_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2} + .. + a_{im}a_{jm} = \sum a_{ir}a_{jr}, \text{ se } i \neq j$$

$$r_{ii} = a_{i1}a_{i1} + a_{i2}a_{i2} + .. + a_{im}a_{im} + u_i^2 = \sum a_{ir}a_{ir} + u_i^2, \text{ se } i = j$$

Rappresentiamolo con un diagramma….



r^ (determinato, scrupoloso) = (.68*.82)+(.51*-.33)= .56 - .17 = .39

Correlazione residua

Differenza tra la correlazione osservata e la correlazione riprodotta tramite le saturazioni.

r(determinato, scrupoloso) = .40

r^(determinato, scrupoloso) = .39

r residua (e) = .40-.39 = .01

$$R=AA'+U^2$$

"Equazione fondamentale dell'analisi fattoriale"
(Thurstone, 1947).

Mette in relazione il punto di partenza dell'Analisi fattoriale con il suo punto di arrivo.

- Per riprodurre le correlazioni tra le variabili che stanno fuori la diagonale principale sono necessari solo i fattori comuni.

- Per "riprodurre" anche gli elementi sulla diagonale principale (varianza totale delle variabili) sono necessarie anche le unicità.

"Equazione fondamentale dell'analisi fattoriale"
Equazione che definisce la Struttura di R (Thurstone, 1947).

$$R^* = AA' \qquad (1)$$

$$R = AA' + U^2 \qquad (2)$$

R*: matrice delle correlazioni che contiene le comunalità sulla diagonale principale.

Come ricavare A, la matrice delle saturazioni nei fattori comuni, in maniera tale che il numero di fattori comuni sia strettamente minore del numero di variabili osservate.

Una soluzione di questo problema è rappresentata dal calcolo delle componenti principali (vedi oltre).

Calcolo di alcuni elementi che caratterizzano la matrice di correlazione

- radici caratteristiche di R (autovalori, L)
- vettori ad essi associati (autovettori, V)

Autovalori e autovettori di una matrice

Gli autovalori sono scalari di enorme importanza nell'analisi multivariata (es., nell'analisi fattoriale).

Per identificare gli autovalori di A è necessario effettuare alcuni calcoli sulla matrice, per i quali si rimanda al libro di testo.

In una matrice quadrata ci sono tanti autovalori quante sono le righe (ovvero le colonne) della matrice.

Ogni autovettore relativo ad un autovalore è un vettore che ha una colonna e tante righe quante quelle della matrice

Autovalori e autovettori di una matrice

Esempio: data la matrice seguente:

$$\mathbf{A} = \begin{bmatrix} 1 & .50 \\ .50 & 1 \end{bmatrix}$$

Gli autovalori sono: $\lambda_1 = 1.5$, e $\lambda_2 = .5$.

Gli autovettori x_1 relativo a λ_1 e x_2 relativo a λ_2 sono:

$$\mathbf{x}_1 = \begin{bmatrix} .707 \\ .707 \end{bmatrix}; \mathbf{x}_2 = \begin{bmatrix} .707 \\ -.707 \end{bmatrix}$$

Autovalori e autovettori di R

Elementi che sintetizzano l'informazione relativa alla varianza delle variabili, e alla correlazione tra le variabili.

Il calcolo di questi elementi è un passo preliminare per il calcolo delle soluzioni di analisi fattoriale.

Ogni autovalore è associato ad un autovettore.

Scomposizione della matrice di correlazione R
Se si considerano V la matrice degli autovettori e L la matrice degli autovalori, allora è possibile dimostrare che R= VLV'.

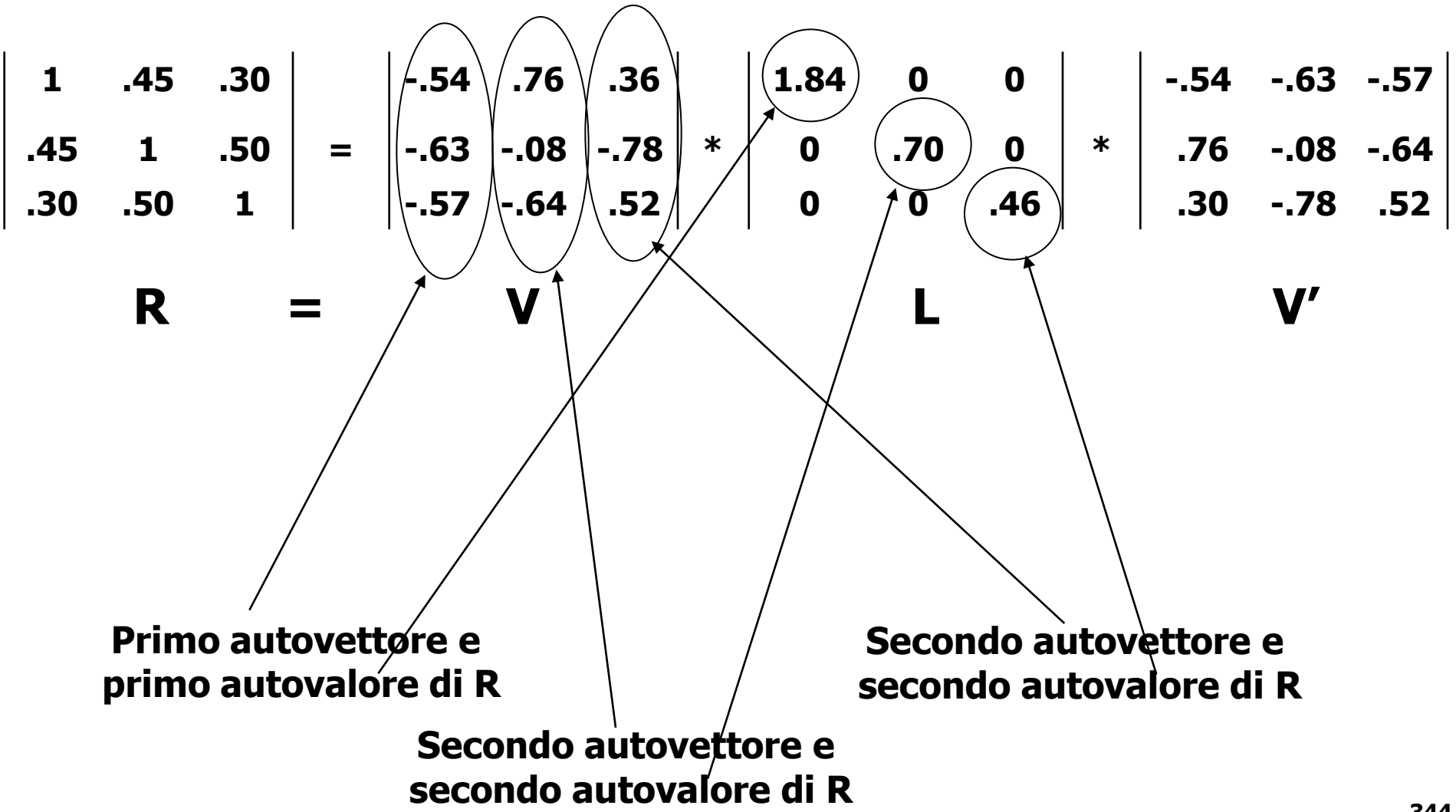
Una volta calcolate le matrici V e L, è possibile ricavare da V e da L la matrice A. In particolare:

$$A = V\sqrt{L}$$

E' possibile dimostrare che: $R = V\sqrt{L}(V\sqrt{L})' = AA'$.

Scomposizione della matrice di correlazione R

R= VLV'



Una volta calcolate le matrici V e L, è possibile ricavare da V e da L la matrice A delle saturazioni fattoriali.

In particolare, $A = V\sqrt{L}$

Nell'esempio precedente la matrice A è data da:

$$\begin{bmatrix} -.73 & .64 & .24 \\ -.85 & -.07 & -.53 \\ -.77 & -.54 & .35 \end{bmatrix} = \begin{bmatrix} -.54 & .76 & .36 \\ -.63 & -.08 & -.78 \\ -.57 & -.64 & .52 \end{bmatrix} * \begin{bmatrix} 1.36 & 0 & 0 \\ 0 & .84 & 0 \\ 0 & 0 & .68 \end{bmatrix}$$

A **V** **$\sqrt{L}$**

Primo autovalore di R: quello più elevato di tutti, associato al primo fattore che spiega una proporzione di varianza maggiore degli altri.

Secondo autovalore: quello più elevato dopo il primo, associato al secondo fattore.

La grandezza degli autovalori rappresenta una progressione decrescente che corrisponde alla progressione della varianza spiegata dai fattori associati ad essi.

Nella matrice precedente:
1.84, .70, .46

Autovalori e varianza spiegata

- Somma delle saturazioni elevate al quadrato per ogni fattore (colonna) = autovalore associato al fattore;

- Autovalore diviso per il numero di variabili osservate in analisi = proporzione di varianza spiegata dal fattore;

- Somma delle saturazioni al quadrato per ogni variabile (riga) = comunalità delle variabili.

Autovalori e varianza spiegata

	F1	F2	h^2
Determinato	.68	.51	.72
Dinamico	.74	.48	.78
Energico	.78	.33	.72
Affidabile	.80	-.41	.81
Responsabile	.84	-.43	.89
Scrupoloso	.82	-.33	.78
Autovalori	3.66	1.08	
Proporzione di Varianza Spiegata	.61	.18	

Metodi di Estrazione dei Fattori

Metodi che cercano di rendere conto di R tramite fattori che ne spiegano il massimo di varianza.

Metodi che cercano invece di rendere conto di R massimizzandone la "riproduzione".

Metodi che richiedono una stima iniziale delle comunalità.

Metodi che utilizzano solo gli elementi al di fuori della diagonale principale e richiedono una stima del numero di fattori da estrarre.

Metodi di Estrazione dei Fattori

Nelle analisi che utilizzano stime delle comunalità viene analizzata la matrice di correlazione R*, con la stima delle comunalità ($\hat{h}_j^2$) sulla diagonale principale.

Nelle analisi che non utilizzano stime delle comunalità viene analizzata la matrice R_1 nella quale non si considerano gli elementi fuori della diagonale principale.

Analisi delle Componenti Principali (ACP)

- Identifica una serie di combinazioni lineari ortogonali delle variabili originali X_i ($c_i=X_iV$, con V = autovettori di R) tali che spieghino più varianza possibile delle variabili originali X_i, e che riducano la complessità dei dati iniziali.

- L'ACP analizza la varianza totale delle variabili (analizza R con valori 1 sulla diagonale principale). La varianza unica è assorbita dai fattori comuni. Nella soluzione ci sono solo "fattori comuni" (le componenti principali).

- Le saturazioni si basano sul calcolo diretto degli autovalori e degli autovettori di R: $A= V\sqrt{L}$

Analisi delle Componenti Principali (ACP)

- Le saturazioni fattoriali risultano gonfiate dalla presenza di varianza comune e varianza unica.
- L'ACP estrae il massimo della varianza per ogni componente, cioè massimizza la varianza spiegata ad ogni estrazione.
- La prima componente è la combinazione lineare dei dati originali che spiega più varianza, la seconda è quella che spiega più varianza dopo la prima, ecc.
- Le componenti principali sono semplici trasformazioni lineari delle variabili originali che forniscono un sommario empirico dei dati. La matrice R è perfettamente replicata se vengono estratte tante componenti quante sono le variabili.

Analisi dei Fattori Principali (AFP o PAF)

- Massimizza lo stesso criterio della ACP, ma con stime della comunalità inserite nella diagonale principale.

- Analizza solo la varianza attribuibile ai fattori "comuni" (ovvero la comunalità) per ottenere una soluzione non contaminata dalla varianza unica.

- Estrae il massimo di varianza per ogni fattore, ma considera solo la varianza dovuta ai fattori comuni, quindi spiega meno varianza della ACP.

Analisi dei Fattori Principali (AFP o PAF)

- Primo passo: rimuovere dalla diagonale principale di R la varianza unica (cioè, $u^2 = 1-h^2$).

- Stima iniziale delle comunalità delle variabili:
* coefficiente di correlazione multipla al quadrato (SMC)
* correlazione più elevata
* media delle correlazioni

- Le saturazioni si basano sul calcolo diretto degli autovalori e degli autovettori di R_1: $A = V\sqrt{L}$, dove R_1 è la matrice delle correlazioni con le stime delle comunalità sulla diagonale principale

Analisi dei Fattori Principali (AFP o PAF)

	fp10	fp15	fp26	fp30
fp10	1,000	,368	,256	,344
fp15	,368	1,000	,390	,444
fp26	,256	,390	1,000	,418
fp30	,344	,444	,418	1,000

Matrice di correlazione originale

fp10	,180
fp15	,288
fp26	,231
fp30	,294

Stima delle Comunalità

	fp10	fp15	fp26	fp30
fp10	**,180**	,368	,256	,344
fp15	,368	**,288**	,390	,444
fp26	,256	,390	**,231**	,418
fp30	,344	,444	,418	**,294**

Matrice di correlazione analizzata da PAF

Analisi dei Fattori Principali (AFP o PAF)

Calcolo iterativo delle comunalità

Nell'AFP le stime delle comunalità rappresentano soltanto un *valore iniziale* che viene cambiato e ricalcolato nel corso dell'estrazione dei fattori.

Le stime iniziali servono per estrarre gli autovalori e gli autovettori di R_1 e quindi per individuare la matrice A delle saturazioni.

Questa soluzione iniziale consente di calcolare empiricamente le comunalità: le comunalità empiriche vengono quindi sostituite alle stime iniziali e il processo di estrazione dei fattori viene ripetuto, dando origine a nuove saturazioni e quindi a nuove comunalità. Il processo si interrompe quando i valori empirici delle comunalità diventano stabili.

Analisi dei Fattori Principali (AFP o PAF)

Calcolo iterativo delle comunalità

Durante il processo di iterazione delle comunalità, il numero di fattori deve rimanere costante.

Una volta che la soluzione fattoriale si è stabilizzata, i valori finali delle comunalità possono essere ricavati dalla soluzione stessa, elevando al quadrato le saturazioni di ogni variabile in ogni fattore comune e sommando tali quadrati.

Ci sono dei problemi che possono presentarsi nel processo iterativo di calcolo delle comunalità. In alcuni casi si può assistere a valori di comunalità che eccedono 1 (tale problema viene definita spesso "*Heywood case*").

Minimi Quadrati (ULS e GLS)
(Minimum residuals /Minres)

Minimizza le differenze al quadrato tra gli elementi della matrice di correlazione osservata (R), e quella riprodotta (R^) utilizzando i fattori estratti.

Minimizza le correlazioni residue (R- R^) cioè la parte di correlazione tra le variabili non spiegata dai fattori.

Funzione dei minimi quadrati ordinari (ULS) minimizzata nel processo di estrazione dei fattori:

$$\Sigma_j \Sigma_k (r_{jk} - r_{jk}\hat{})^2$$

Minimi Quadrati (ULS e GLS)
(Minimum residuals /Minres)

Massimizza la riproduzione dei coefficienti fuori della diagonale principale di R.

Si inizia il processo stabilendo il numero di fattori.

Si stimano le saturazioni iniziali con l'ACP.

Le saturazioni vengono modificate iterativamente finché lo scarto tra R e R^ non è molto piccolo.

Minimi quadrati generalizzati (GLS): si introduce un fattore di ponderazione, per cui le variabili con fattore unico più elevato hanno peso minore.

Maximum Likelihood (Massima verosimiglianza)

Calcola le saturazioni che rendono massima la probabilità di riprodurre la matrice R, ovvero identifica la soluzione che meglio riproduce R.

Stima le saturazioni della popolazione che hanno la massima verosimiglianza (ovvero la massima probabilità) nel riprodurre R, quindi che rendono minima la differenza tra matrice osservata e riprodotta.

Si considerano gli elementi fuori della diagonale principale, si fornisce il numero di fattori da estrarre.

Maximum Likelihood (Massima verosimiglianza)

La stima delle saturazioni avviene attraverso la minimizzazione di una funzione (F_{ML}) delle matrici delle correlazioni osservate R e riprodotte R^.

$$F_{ML} = tr(RC^{-1}) + \ln|C| - \ln|R| - n$$

dove C =AA'+U^2, è la matrice riprodotta dalla soluzione (A matrice delle saturazioni, U^2 stima della varianza unica), *n* è il numero di variabili, | | indica il determinante e tr() la traccia della matrice.

Per calcolare la funzione è necessario che:

$$(n-k)^2 > (n + k)$$

dove n = numero di variabili, k = numero di fattori.

NUMERO MASSIMO DI FATTORI

Il numero massimo di fattori che è possibile estrarre dipende dai gradi di libertà che sono determinati dal numero di parametri da stimare e dal numero di correlazioni non ridondanti.

$$\text{Gradi di libertà} = [(n-k)^2 - (n+k)]/2,$$

Deve valere sempre: $(n-k)^2 > (n+k)$

n = numero di variabili; k = il numero di fattori.

Esempio: n = 8

Se k = 1, $(8-1)^2-(8+1)= 49-9=40$, gdl=20

Se k = 2, $(8-2)^2-(8+2)= 36-10=26$, gdl=13

Se k = 3, $(8-3)^2-(8+3)= 25-11=14$, gdl=7

Se k = 4, $(8-4)^2-(8+4)= 16-12=4$, gdl=2

Se k = 5, $(8-5)^2-(8+5)= 9-13=-4$, gdl=-2

Con 8 variabili osservate si possono estrarre al massimo 4 fattori.

Test di bontà dell'adattamento (goodness of fit)

Si ottiene dalle funzioni ML e GLS che vengono minimizzate se le variabili seguono la distribuzione normale multivariata.

Ipotesi nulla: R = R^

Il test segue la distribuzione del χ^2

Gradi di libertà del test: $df=[(n-k)^2 - (n+k)]/2$

χ^2 non significativo: il modello che ipotizza k fattori è consistente con i dati empirici, non si può rifiutare l'ipotesi nulla H_0:R=R^, quindi non vi sono più fattori da estrarre

χ^2 significativo: il modello che ipotizza k fattori è consistente con i dati empirici, quindi è necessario procedere all'estrazione di fattori ulteriori.

Test fortemente dipendente dal numero di casi.

Stabilire il numero dei fattori da estrarre

Decisione che ha conseguenze cruciali sulla soluzione fattoriale.

Salvaguardare la parsimonia della soluzione, e la sua adeguatezza (capacità di riprodurre R).

Metodi per stabilire il numero di fattori

- **Mineigen (Kaiser-Guttman rule)**
- **Scree test degli autovalori (Cattell e Vogelman)**
- **Test statistico, indici di bontà dell'adattamento**
- **Percentuale di varianza spiegata**
- **Massima correlazione residua**

Mineigen (Kaiser-Guttman rule)

Estrae tutti quei fattori che hanno un autovalore maggiore di 1 quando viene analizzata la matrice R completa (con 1 sulla diagonale principale).

I fattori devono spiegare almeno la stessa varianza spiegata dalle variabili osservate.

Il numero di autovalori maggiori di 1 è uguale approssimativamente ad un numero compreso tra 1/3 e 1/5 del numero delle variabili.

Criterio inappropriato per soluzioni diverse dall'ACP

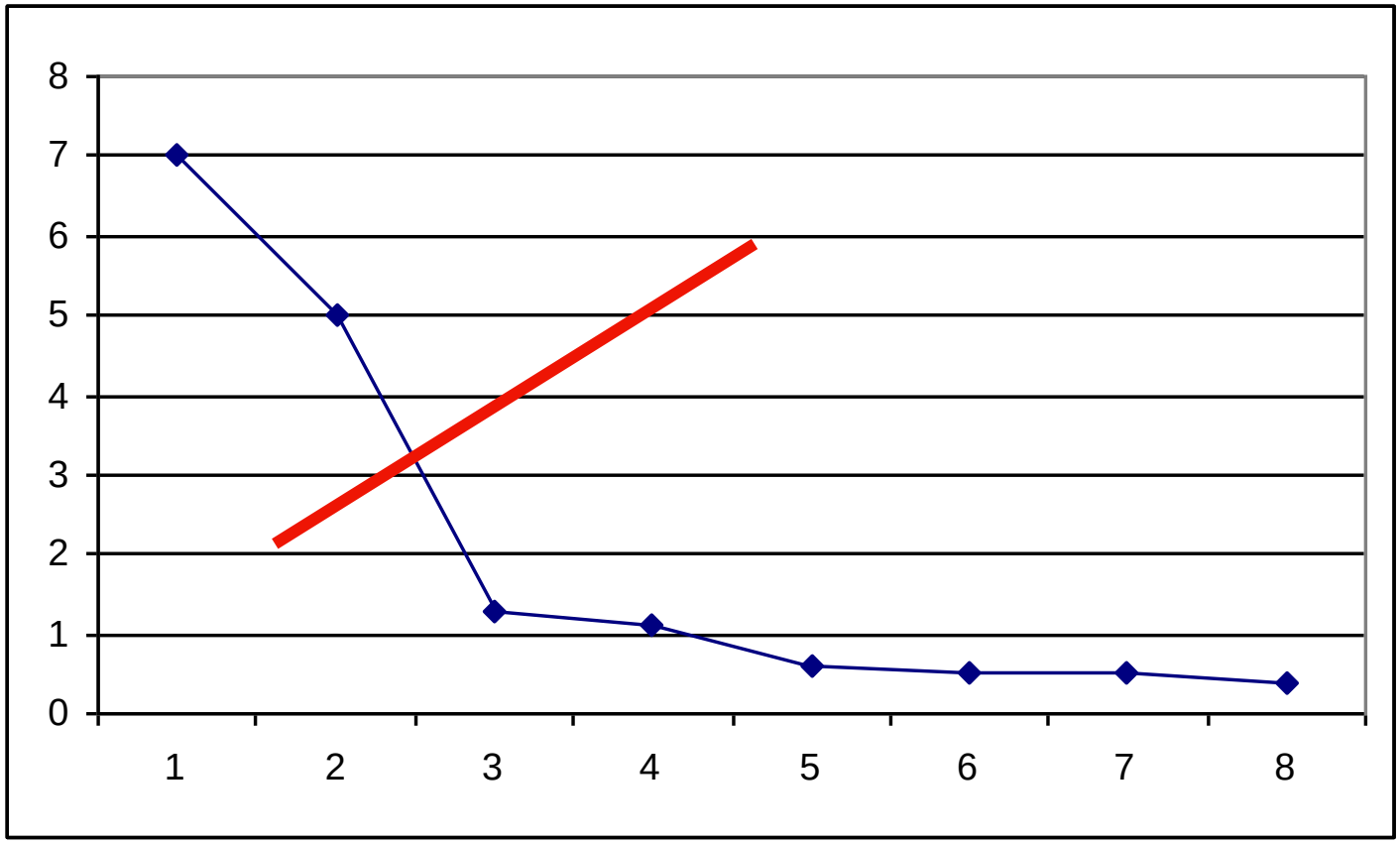
Scree test degli autovalori (Cattell e Vogelman)

I primi fattori sono i più attendibili e i più validi, poiché spiegano una percentuale di varianza maggiore rispetto ai rimanenti, e avranno autovalori più grandi degli altri.

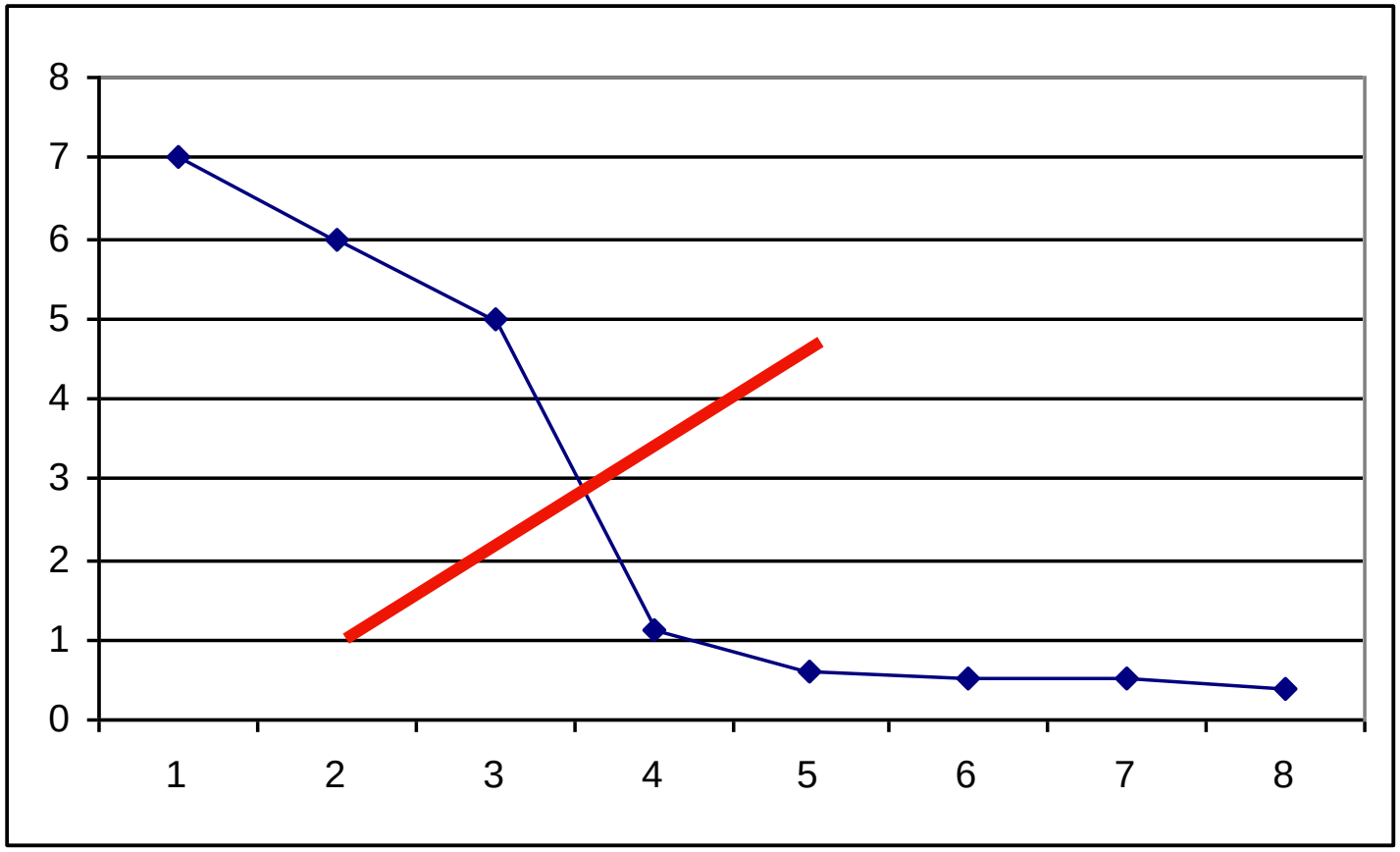
<u>Progressione decrescente degli autovalori</u>: grafico in cui ogni autovalore è in ordinata, e il numero del fattore ad esso relativo in ascissa.

Il processo di estrazione si interrompe nel punto in cui la curva degli autovalori decrescenti cambia pendenza e diventa sostanzialmente piatta. Vanno presi quei fattori i cui autovalori sono al di sopra della linea piatta formata dagli autovalori dei fattori più piccoli.

Applicazione più attendibile quando il campione è grande, le comunalità elevate, e ogni fattore satura diverse variabili.



2 fattori



3 fattori

Analisi parallela

E' un procedimento che si basa sul confronto tra gli autovalori della matrice di correlazione campionaria e gli autovalori ottenuti da una matrice calcolata su un set di dati casuali generati artificialmente.

I risultati di questi studi suggeriscono di mantenere nella soluzione quei fattori associati ad un autovalore superiore a quello associato ad un fattore omologo estratto nei dati artificiali.

Analisi parallela

Supponiamo che i primi 5 autovalori dei dati reali siano: 5.72 1.51 1.03 0.50 0.40, e quelli ricavati dai dati "artificiali" siano: 1.64 1.45 1.31 1.19 1.09.

Verranno mantenuti quei fattori reali che presentano un autovalore maggiore di quello del corrispondente fattore dei dati artificiali, quindi nel nostro caso verranno mantenuti 2 fattori perché solo per i primi due fattori gli autovalori nei dati reali sono maggiori degli autovalori associati ai corrispondenti fattori nei dati artificiali [5.72>1.64;1.51>1.45;1.03<1.31;0.50< 1.19; 0.40< 1.09]

Sintassi SPSS per l'analisi parallela:
https://people.ok.ubc.ca/brioconn/nfactors/nfactors.html

Test statistico e indici di bontà dell'adattamento

Il test statistico associato ai metodi di estrazione ML e GLS (chi-quadrato) da un punto di vista puramente statistico, è il migliore. Da un punto di vista pratico, però, questo test tende ad essere fortemente dipendente dall'ampiezza del campione.

Gli indici alternativi di bontà dell'adattamento (che introdurremo quando affronteremo i modelli confermativi) possono spesso dare risultati più verosimili: tra questi indici l'SRMR e l'RMSEA sembrano i più affidabili.
Come regola pratica il ricercatore dovrebbe considerare più indici alternativi per ciascuna soluzione, e privilegiare le soluzioni nelle quali i diversi indici mostrano maggiore convergenza.

Percentuale di varianza spiegata

Contributo minimo di un fattore alla spiegazione della varianza, oppure proporzione di varianza spiegata dall'ultimo fattore. Metodo troppo soggettivo.

Replicabilità della soluzione

I fattori "validi" sono quelli che risultano più facilmente replicabili su campioni diversi da quelli nei quali sono stati individuati.

I fattori "spuri" risultano poco generalizzabili e sono determinati sostanzialmente dall'errore campionario.

Massima correlazione residua

Per ogni elemento di R fuori della diagonale principale si può definire un residuo che è uguale a (r- r^), ovvero correlazione osservata meno correlazione riprodotta. La matrice dei residui quindi si ottiene nel modo seguente: E = (R-R^).

Se dopo aver effettuato l'estrazione di un certo numero di fattori tutti i residui sono minori di |.10|, non è necessario continuare il processo di estrazione: il nuovo fattore estratto avrebbe saturazioni molto basse.

Rotazione dei fattori

E' un'operazione che rende la soluzione fattoriale più interpretabile senza cambiarne le fondamentali proprietà matematiche (capacità di riprodurre R, % var. spiegata).

Esistono infinite matrici T che trasformano una matrice di saturazioni non ruotata A in modo che:

$$AT = B, \text{ e } R = BB'$$

T è la matrice di trasformazione (gli elementi sono seni e coseni di un generico angolo di rotazione "ϕ"), B è la matrice ruotata

Rotazioni ortogonali: i fattori ruotati non sono correlati. Rotazioni oblique: i fattori ruotati possono essere correlati tra loro.

Rotazione dei fattori

T è la matrice di <u>trasformazione</u> (gli elementi sono seni e coseni di un generico angolo di rotazione "ϕ"). Nel caso di due fattori T è come la matrice seguente:

$$\mathbf{T} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

Nella rotazione dei fattori, AT=B.
R=BB', ma B=AT, quindi R=(AT)(AT)′=ATT′A′
E′ possibile dimostrare che TT′=T′T=I, quindi:

R=BB′=AA′

Questo fenomeno viene definito indeterminatezza della soluzione fattoriale: esistono infinite matrici A tali che R=AA', cioè che riproducono una data matrice R altrettanto bene.

Soluzioni fattoriali

	Soluzione iniziale		Soluzione ortogonale		Soluzione obliqua		
	F1	F2	F1	F2	F1	F2	h^2
Determinato	.68	.51	.17	.83	-.08	.88	.72
Dinamico	.74	.48	.24	.85	-.01	.89	.78
Energico	.78	.33	.36	.77	.16	.76	.72
Affidabile	.80	-.41	.87	.23	.91	-.02	.81
Responsabile	.84	-.43	.91	.24	.95	-.02	.89
Scrupoloso	.82	-.33	.83	.30	.85	.07	.78
Varianza Spiegata	61%	18%	42%	37%	42%	37%	

La rotazione dei fattori – La struttura semplice (Thurstone, 1947)

Guida il processo di rotazione dei fattori. Si pone l'obiettivo di rendere più interpretabile la soluzione massimizzando il numero di zeri nelle righe e nelle colonne della matrice delle saturazioni.

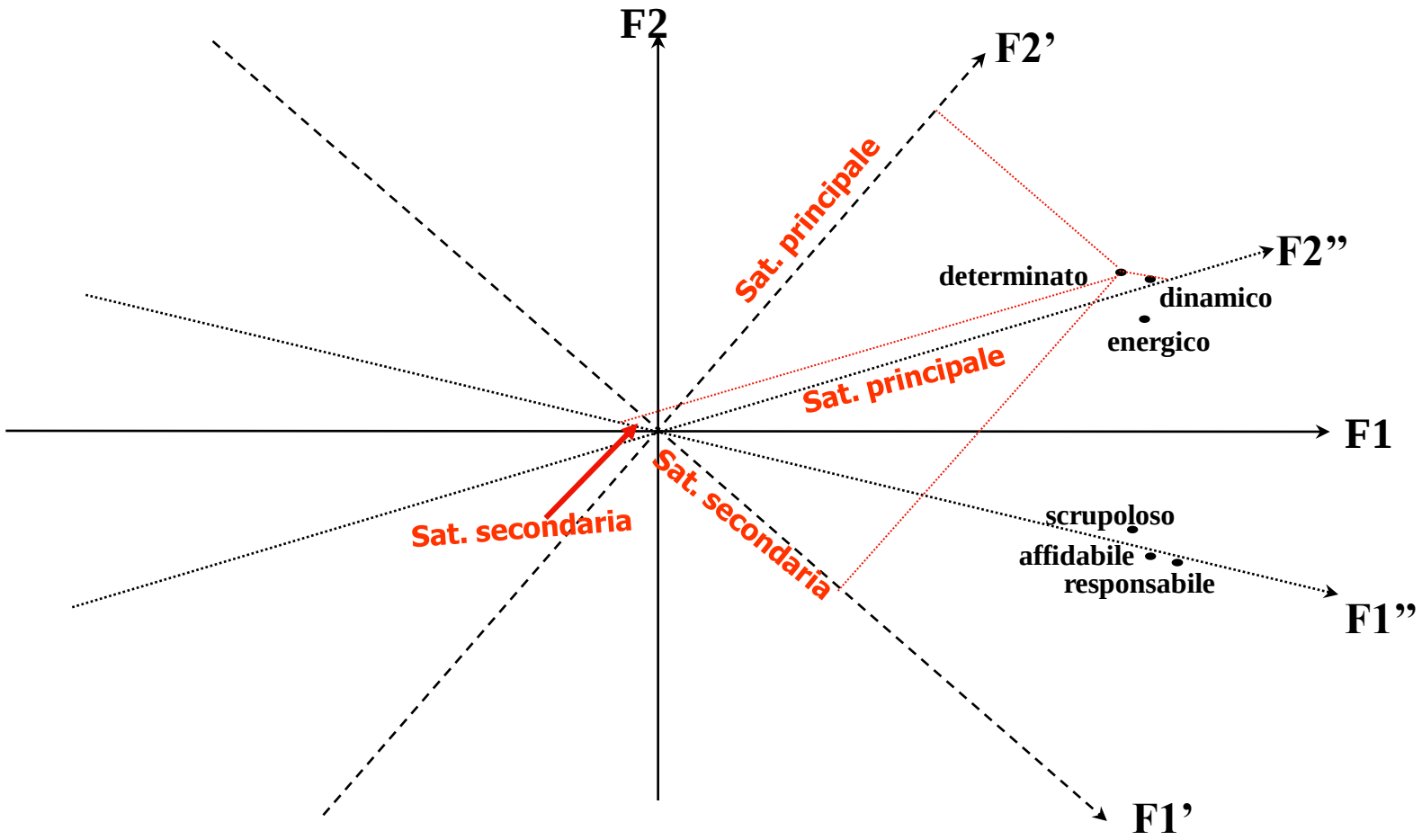
Ogni fattore deve saturare una minoranza di variabili; ogni variabile deve essere spiegata da pochi fattori (possibilmente uno solo).

La rotazione dei fattori – La struttura semplice (Thurstone, 1947)

Fa in modo che le variabili cadano il più vicino possibile agli assi fattoriali. Gli spazi "interstiziali" tendono ad essere più vuoti degli spazi vicini agli assi.

Numero di saturazioni prossime a 0 in un fattore: indice della semplicità del fattore.

La struttura fattoriale più semplice possibile è quella in cui le variabili hanno saturazioni uguali a 0 in tutti i fattori tranne che in un unico fattore comune.



Assi fattoriali prima della rotazione

Assi fattoriali dopo la rotazione ortogonale

Assi fattoriali dopo la rotazione obliqua

Rotazioni ortogonali - Varimax

Aumenta la semplicità dei fattori.

Massimizza la varianza delle saturazioni delle variabili all'interno di ogni fattore (nelle colonne di A).

Per ogni fattore, tende a far diventare le saturazioni elevate più elevate e quelle più basse ancora più basse.

La variabilità delle saturazioni è massimizzata, e la varianza redistribuita.

Varimax tende a produrre fattori che presentano alcune saturazioni elevate, poche intermedie e molte basse. Risultati più chiari e più generalizzabili, e fattori diversi separati meglio.

Rotazioni ortogonali - Quartimax

Massimizza la semplicità delle variabili a scapito dei fattori.

Massimizza la varianza delle saturazioni di ogni variabile per riga.

Concentra più varianza possibile per ogni variabile su un solo fattore, creando fattori generale.

Rotazioni oblique

Oblimin: Fa in modo che le variabili abbiano saturazioni il più possibile vicine a 0 in tutti i fattori tranne uno. Massimizza una funzione che comprende anche le covarianze tra i fattori.

Promax: Parte da una rotazione ortogonale, e la modifica per renderla più semplice, consentendo che i fattori siano correlati.

Rotazioni di Procuste: La matrice originale viene ruotata verso una matrice "bersaglio" che ha certe caratteristiche ipotizzate dal ricercatore. La soluzione iniziale viene ruotata in modo da renderla più simile possibile alla matrice bersaglio.

Nelle soluzioni ortogonali:
l'impatto del fattore sulla variabile
è uguale alla correlazione tra variabile e fattore
(saturazione fattoriale).

Nelle soluzioni oblique
è possibile distinguere tra:

- correlazione tra variabile e fattore

- impatto del fattore sulla variabile (contributo unico del fattore al netto degli altri fattori)

Nelle soluzioni oblique

La variabile osservata può condividere una parte di varianza simultaneamente con più fattori.

La correlazione tra variabile e fattore comprende sia il contributo unico del fattore sia il contributo condiviso con gli altri fattori.

Per questo ci sono due diverse matrici che riassumono le relazioni tra variabili e fattori

Matrice Pattern (P)

Impatto diretto di ciascun fattore sulle variabili, al netto dell'impatto degli altri fattori.

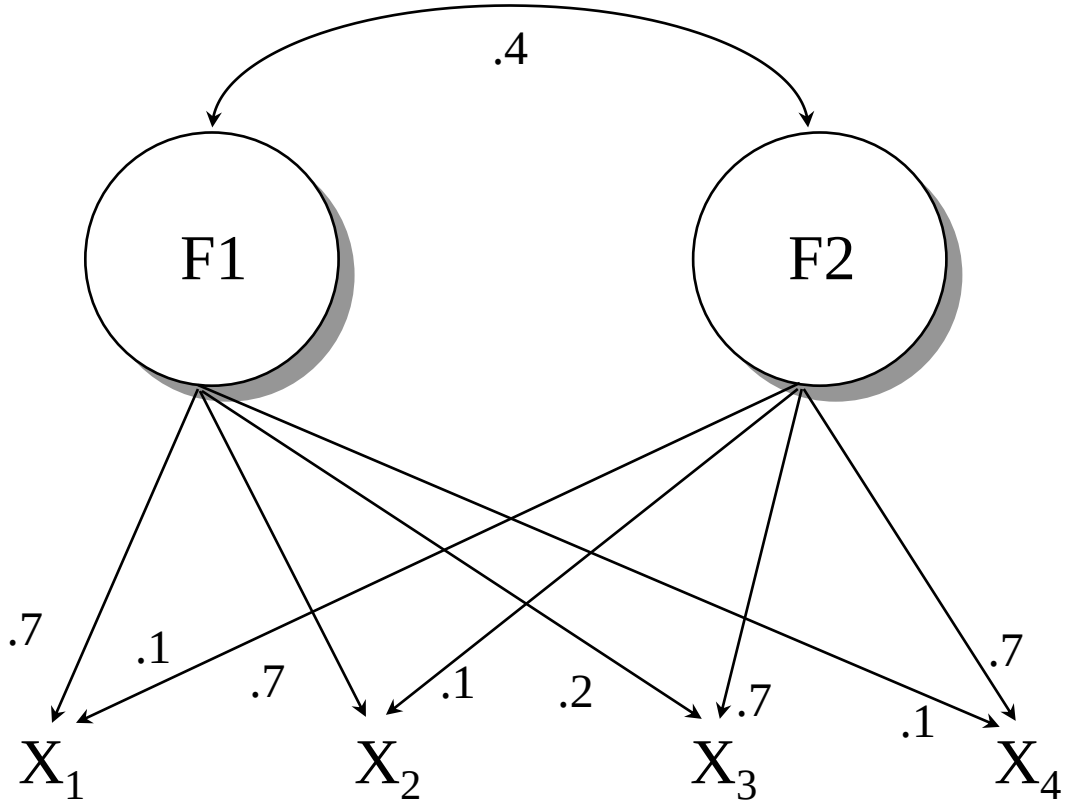
Influenza *unica* di ciascun fattore sulle variabili (pesi beta, β).

Matrice Struttura (S)

Correlazioni tra le variabili e i fattori.

Risultano tanto più "gonfiate" quanto più è elevata la correlazione tra i fattori.

Per interpretare i fattori nelle rotazioni oblique si esamina la matrice pattern.



Effetto diretto di F1 su X1 = .7 (P)
Effetto di F1 su X1 dovuto alla correlazione tra F1 e F2= .4*.1=.04
Effetto totale di F1 su X1 = .7 + .04 = .74 (S)

Riprodurre R dopo una rotazione obliqua:

Matrici Struttura (S) e Pattern (P)

$$S = P\Phi.$$

Φ= Matrice delle correlazioni tra i fattori

$$R^*=SP' \quad e \quad R=SP' + U^2$$

$$R^*=P\Phi P' \quad e \quad R=P\Phi P' + U^2$$

Nelle rotazioni ortogonali invece:

$$S = P = A, \Phi=I, \text{ quindi}$$

$$R^* = P\Phi P' = AA'$$

Varianza spiegata dopo la rotazione obliqua

- Moltiplicare P e S elemento per elemento
- Sommare i prodotti per colonna
- Dividere i totali di colonna per il numero di variabili e moltiplicare per 100.

Variabili	PATTERN (P)		STRUCTURE (S)		PRODOTTO (P * S)		
	F1	F2	F1	F2	F1	F2	h^2
X1	0,70	0,10	0,74	0,38	0,52	0,04	0,56
X2	0,70	0,10	0,74	0,38	0,52	0,04	0,56
X3	0,20	0,70	0,48	0,78	0,10	0,55	0,65
X4	0,10	0,70	0,38	0,74	0,04	0,52	0,56
% Var.					29	29	

Somma per riga dei prodotti: comunalità.

Interpretazione dei fattori e grandezza delle saturazioni

I fattori si interpretano in base alle variabili con le quali presentano correlazioni (saturazioni) più elevate.
Regola pratica: livello soglia di circa |.30| (circa 9% di varianza in comune tra fattore e variabile).

a) |.71| (50% varianza comune): eccellente;
b) |.63| (40% varianza comune): molto buona;
c) |.55| (30% varianza comune): buona;
d) |.45| (20% varianza comune): sufficiente;
e) |.32| (10% varianza comune): scarsa.

Saturazioni sotto |.30| inadeguate.

Assunzioni e prerequisiti - Fattorializzabilità di R

- Test di sfericità di Bartlett:
  H_0: R = I (I = matrice identità).
Se significativo, e il campione è sufficientemente ampio, la matrice è fattorializzabile.

- Indice di adeguatezza campionaria KMO:

$$KMO = \Sigma\Sigma r^2 / (\Sigma\Sigma r^2 + \Sigma\Sigma p^2)$$

r = correlazioni tra ogni coppia di variabili
p = correlazioni tra ogni coppia di variabili, parzializzate rispetto a tutte le altre variabili

Assunzioni e prerequisiti - Fattorializzabilità di R

- Test di adeguatezza campionaria di Kaiser (KMO):

Interpretazione dei valori del KMO:

>0.90: eccellenti;
0.80-0.90: buoni;
0.70-0.80: accettabili;
0.60-0.70: mediocri;
<0.60: scarsi/non accettabili (l'analisi è sconsigliata)

Assunzioni e prerequisiti

Livelli di misura e distribuzione delle variabili: Almeno intervalli equivalenti. Anche ordinali se il numero di categorie ordinabili di una variabile è sufficiente (es., da 5 in su), e se la distribuzione delle variabili è normale.

Coefficienti di correlazione: Coefficiente di correlazione di Pearson (dà stime più stabili). Variabili dicotomiche o ordinali: coefficienti di correlazione "speciali" (tetracorici e policorici: sono ottenibili in Prelis, non in SPSS).

Assunzioni e prerequisiti

Normalità multivariata: Se le distribuzioni sono normali la soluzione è migliore. Richiesta Con il metodo di estrazione della Maximum Likelihood.

Linearità: Necessaria perché l'analisi si basa sui coefficienti di Pearson. Metodi di AF non lineari: basati su coefficienti speciali (utili per dati non normali).

Outliers tra i casi e tra le variabili:
Casi estremi univariati e multivariati possono distorcere i risultati. Variabili "outlier": non correlano con le altre variabili in analisi, e vanno a definire fattori "residuali" e poco attendibili (saturati soltanto da quella variabile).

Assunzioni e prerequisiti

Numero di variabili:

- Numero di variabili 3 o 4 volte superiore al numero dei fattori
- Non meno di 3 variabili "marker" per ogni fattore che si vuole identificare (fattori "sovradeterminati").

Ampiezza e qualità del campione:

- Campioni piccoli producono stime poco stabili di r
- Consigliabile non scendere mai sotto i 100 soggetti e non avere mai meno di cinque casi per ogni variabile.
- Variabilità delle variabili e/o dei fattori: sufficientemente ampia. Campione molto selezionato ed omogeneo: riduzione della variabilità e quindi delle correlazioni; mancata individuazione di fattori, minore percentuale di var. spiegata, saturazioni più basse.

Ambiti di applicazione dell'Analisi Fattoriale

- Costruzione di test psicologici
- Costruzione di scale e questionari

Esame della qualità di strumenti di misura per:

* identificare indicatori adeguati e non adeguati
* identificare fattori misurati in maniera non adeguata (es., da una sola variabile)

Esplorazione di dati ☹

Problemi

Applicazione poco attenta delle opzioni di *default*

Fiducia in una *erronea* tradizione consolidata

Scarsa conoscenza del modello statistico di base

Decisioni da prendere

Adeguatezza delle variabili

Fattorializzabilità di R

Tecnica per l'estrazione

Numero di fattori

Tecnica per la rotazione

Interpretazione dei fattori

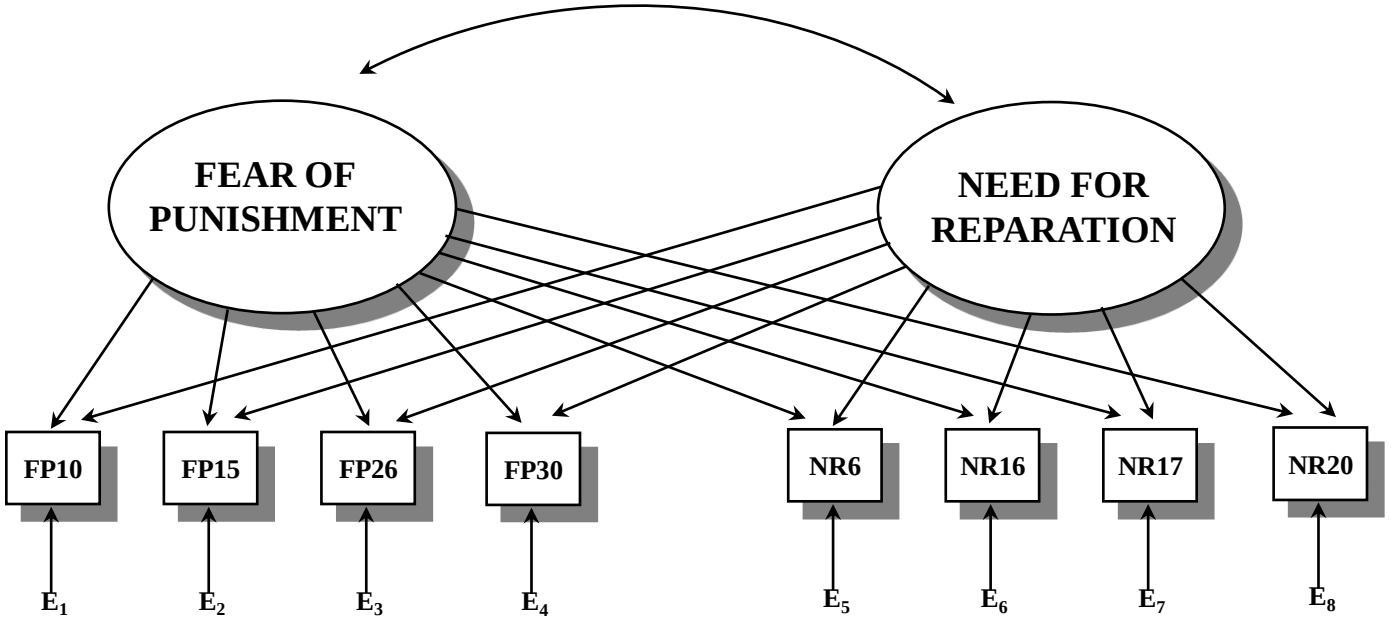
Adeguatezza della soluzione

Come ottenere buone soluzioni

- Numero di indicatori per ogni fattore (> 3)
- Almeno 100 soggetti
- Campione *non* selezionato
- Non utilizzare l'analisi delle componenti principali ma un metodo fattoriale vero
- Rotazione obliqua (ortogonale solo se i fattori non correlano)
- Più metodi per scegliere il numero di fattori: non utilizzare il criterio dell'autovalore >1

ANALISI FATTORIALE ESPLORATIVA (EFA) CON SPSS

MODELLO CONCETTUALE



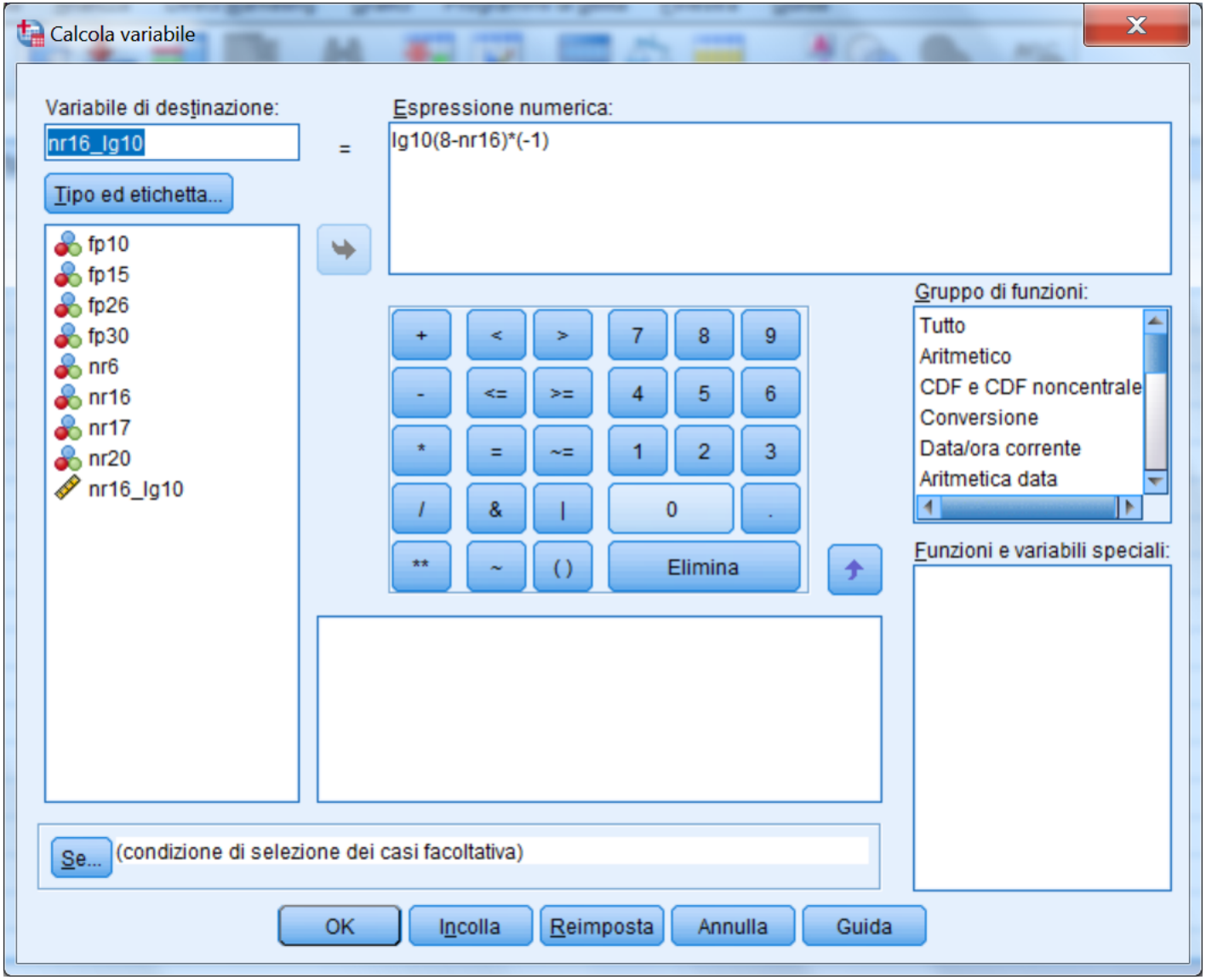
efa_dati.sav

EFA IN SPSS – analisi preliminari

Statistiche descrittive

	N	Minimo	Massimo	Media	Deviazione std.	Asimmetria		Curtosi	
	Statistica	Statistica	Statistica	Statistica	Statistica	Statistica	Errore std.	Statistica	Errore std.
fp10	819	1	6	2,66	1,582	,481	,085	-,942	,171
fp15	819	1	6	3,17	1,492	,066	,085	-,945	,171
fp26	819	1	6	3,68	1,393	-,310	,085	-,599	,171
fp30	819	1	6	3,15	1,553	,059	,085	-1,103	,171
nr6	819	1	6	4,37	1,379	-,852	,085	,186	,171
nr16	819	1	6	4,67	1,261	-1,097	,085	1,061	,171
nr17	819	1	6	4,70	1,272	-1,100	,085	,951	,171
nr20	819	1	6	5,04	1,163	-1,496	,085	2,321	,171
Numero di casi validi (listwise)	819								

EFA IN SPSS – analisi preliminari

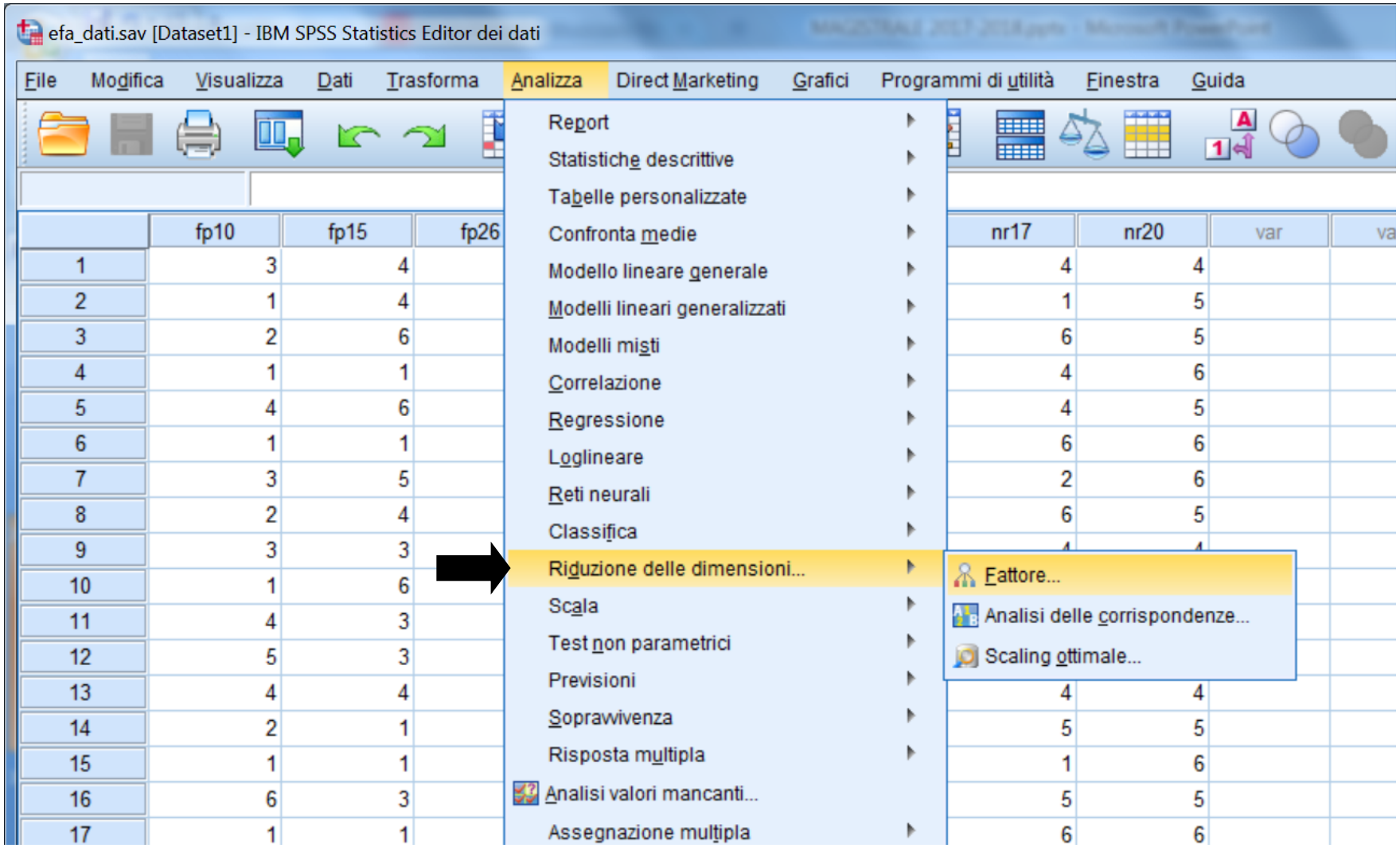


EFA IN SPSS – analisi preliminari

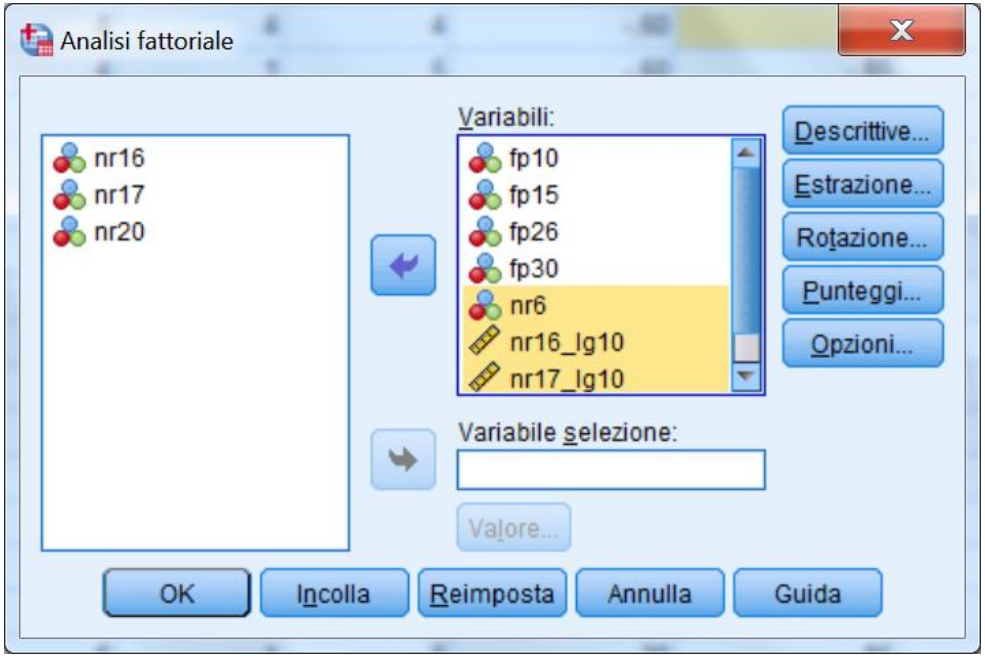
Statistiche descrittive

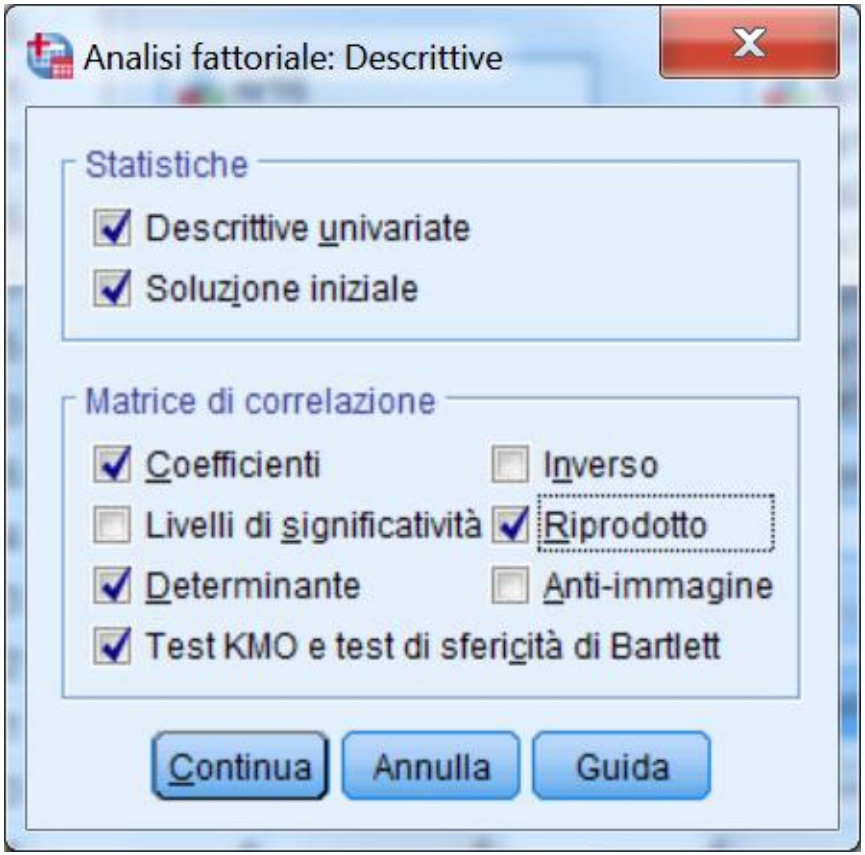
	N	Minimo	Massimo	Media	Deviazione std.	Asimmetria		Curtosi	
	Statistica	Statistica	Statistica	Statistica	Statistica	Statistica	Errore std.	Statistica	Errore std.
nr16	819	1	6	4,67	1,261	-1,097	,085	1,061	,171
nr17	819	1	6	4,70	1,272	-1,100	,085	,951	,171
nr20	819	1	6	5,04	1,163	-1,496	,085	2,321	,171
nr16_lg10	819	-,85	-,30	-,4936	,15479	-,306	,085	-,641	,171
nr17_lg10	819	-,85	-,30	-,4891	,15665	-,350	,085	-,686	,171
nr20_lg10	819	-,85	-,30	-,4435	,14996	-,721	,085	-,294	,171
Numero di casi validi (listwise)	819								

EFA IN SPSS

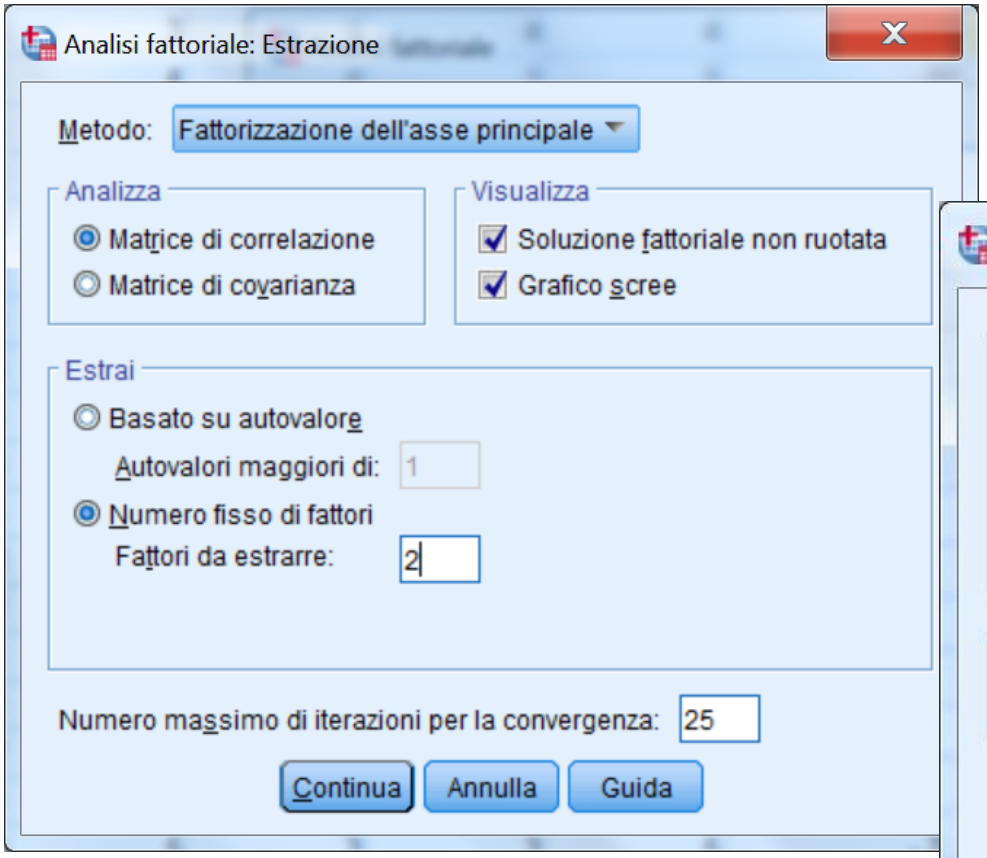

efa_dati.sav [Dataset1] - IBM SPSS Statistics Editor dei dati
File Modifica Visualizza Dati Trasforma Analizza Direct Marketing Grafici Programmi di utilità Finestra Guida
Report
Statistiche descrittive
Tabelle personalizzate
Confronta medie
Modello lineare generale
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Reti neurali
Classifica
Riduzione delle dimensioni...
Scala
Test non parametrici
Previsioni
Sopravvivenza
Risposta multipla
Analisi valori mancanti...
Assegnazione multipla
Fattore...
Analisi delle corrispondenze...
Scaling ottimale...
fp10 fp15 fp26 nr17 nr20 var

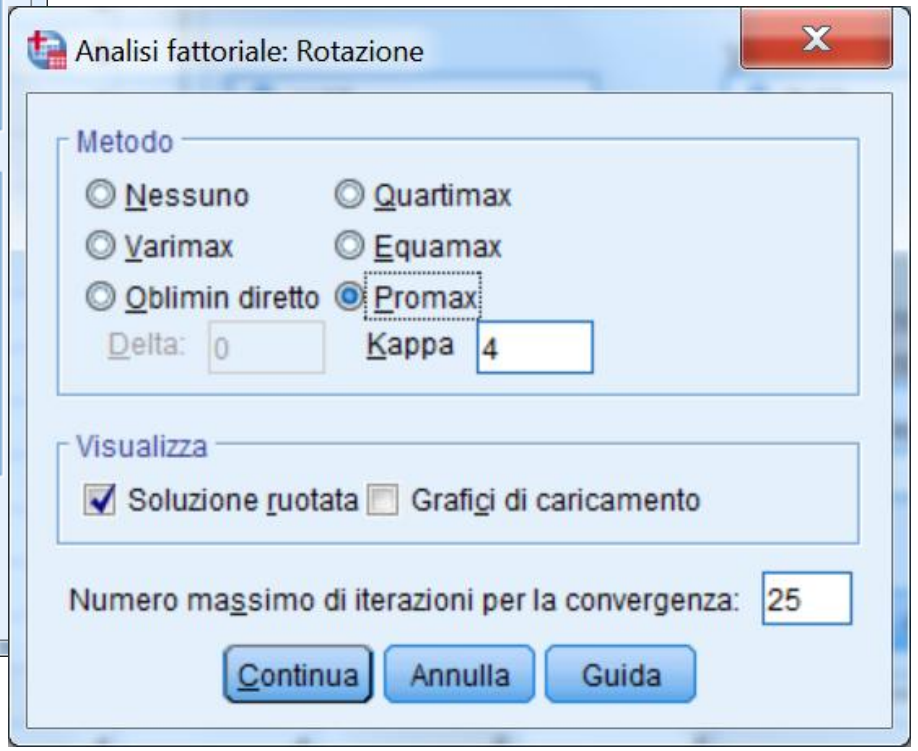
EFA IN SPSS


Analisi fattoriale
nr16
nr17
nr20
Variabili:
fp10
fp15
fp26
fp30
nr6
nr16_lg10
nr17_lg10
Variabile selezione:
Valore...
Descrittive...
Estrazione...
Rotazione...
Punteggi...
Opzioni...
OK
Incolla
Reimposta
Annulla
Guida


Analisi fattoriale: Descrittive
Statistiche
Descrittive univariate
Soluzione iniziale
Matrice di correlazione
Coefficienti
Inverso
Livelli di significatività
Riprodotto
Determinante
Anti-immagine
Test KMO e test di sfericità di Bartlett
Continua
Annulla
Guida

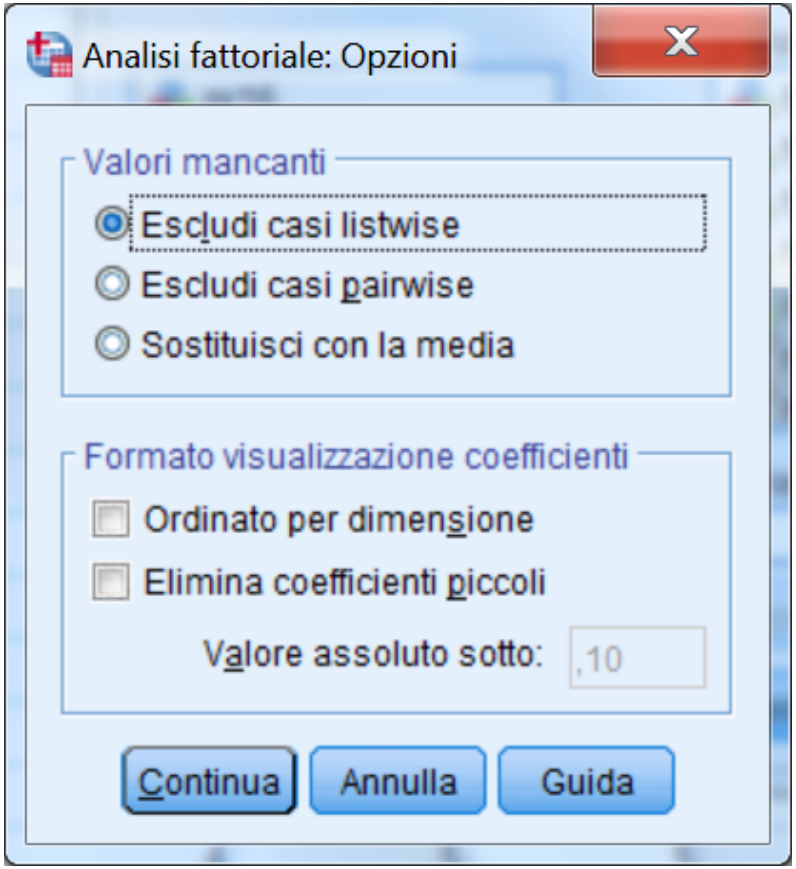
EFA IN SPSS


Analisi fattoriale: Estrazione
Metodo: Fattorizzazione dell'asse principale
Analizza
Matrice di correlazione
Matrice di covarianza
Visualizza
Soluzione fattoriale non ruotata
Grafico scree
Estrai
Basato su autovalore
Autovalori maggiori di: 1
Numero fisso di fattori
Fattori da estrarre: 2
Numero massimo di iterazioni per la convergenza: 25
Continua
Annulla
Guida


Analisi fattoriale: Rotazione
Metodo
Nessuno
Quartimax
Varimax
Equamax
Oblimin diretto
Promax
Delta: 0
Kappa 4
Visualizza
Soluzione ruotata
Grafici di caricamento
Numero massimo di iterazioni per la convergenza: 25
Continua
Annulla
Guida

EFA IN SPSS


Analisi fattoriale: Punteggi fattoriali
Salva come variabili
Metodo
Regressione
Bartlett
Anderson-Rubin
Visualizza matrice dei coefficienti di punteggio fattoriale
Continua
Annulla
Guida


Analisi fattoriale: Opzioni
Valori mancanti
Escludi casi listwise
Escludi casi pairwise
Sostituisci con la media
Formato visualizzazione coefficienti
Ordinato per dimensione
Elimina coefficienti piccoli
Valore assoluto sotto:
,10
Continua
Annulla
Guida

EFA IN SPSS

Statistiche descrittive

	Media	Deviazione std.	N analisi
fp10	2,66	1,582	819
fp15	3,17	1,492	819
fp26	3,68	1,393	819
fp30	3,15	1,553	819
nr6	4,37	1,379	819
nr16_lg10	-,4936	,15479	819
nr17_lg10	-,4891	,15665	819
nr20_lg10	-,4435	,14996	819

Matrice di correlazione[a]

		fp10	fp15	fp26	fp30	nr6	nr16_lg10	nr17_lg10	nr20_lg10
Correlazione	fp10	1,000	,368	,256	,344	,050	,039	,080	-,010
	fp15	,368	1,000	,390	,444	,155	,060	,074	,074
	fp26	,256	,390	1,000	,418	,122	,163	,130	,141
	fp30	,344	,444	,418	1,000	,120	,068	,154	,036
	nr6	,050	,155	,122	,120	1,000	,301	,310	,307
	nr16_lg10	,039	,060	,163	,068	,301	1,000	,256	,354
	nr17_lg10	,080	,074	,130	,154	,310	,256	1,000	,302
	nr20_lg10	-,010	,074	,141	,036	,307	,354	,302	1,000

a. Determinante = ,299

EFA IN SPSS

Test di KMO e Bartlett

Misura di Kaiser-Meyer-Olkin di adeguatezza del campionamento.		,737
Test della sfericità di Bartlett	Appross. Chi-quadrato	982,770
	gl	28
	Sign.	,000

Comunalità

	Iniziale	Estrazione
fp10	,184	,261
fp15	,299	,456
fp26	,253	,341
fp30	,304	,473
nr6	,189	,296
nr16_lg10	,188	,311
nr17_lg10	,172	,256
nr20_lg10	,208	,375

Metodo di estrazione: Fattorizzazione dell'asse principale.

Varianza totale spiegata

	Autovalori iniziali			Caricamenti somme dei quadrati di estrazione			Caricamenti somme dei quadrati di rotazione[a]
Fattore	Totale	% di varianza	% cumulativa	Totale	% di varianza	% cumulativa	Totale
1	2,402	30,021	30,021	1,769	22,118	22,118	1,605
2	1,655	20,688	50,709	1,001	12,508	34,625	1,354
3	,792	9,895	60,603				
4	,724	9,049	69,652				
5	,714	8,920	78,572				
6	,647	8,086	86,658				
7	,549	6,858	93,516				
8	,519	6,484	100,000				

Metodo di estrazione: Fattorizzazione dell'asse principale.

a. Quando i fattori sono correlati, i caricamenti delle somme dei quadrati non possono essere aggiunti per ottenere una varianza totale.

EFA IN SPSS

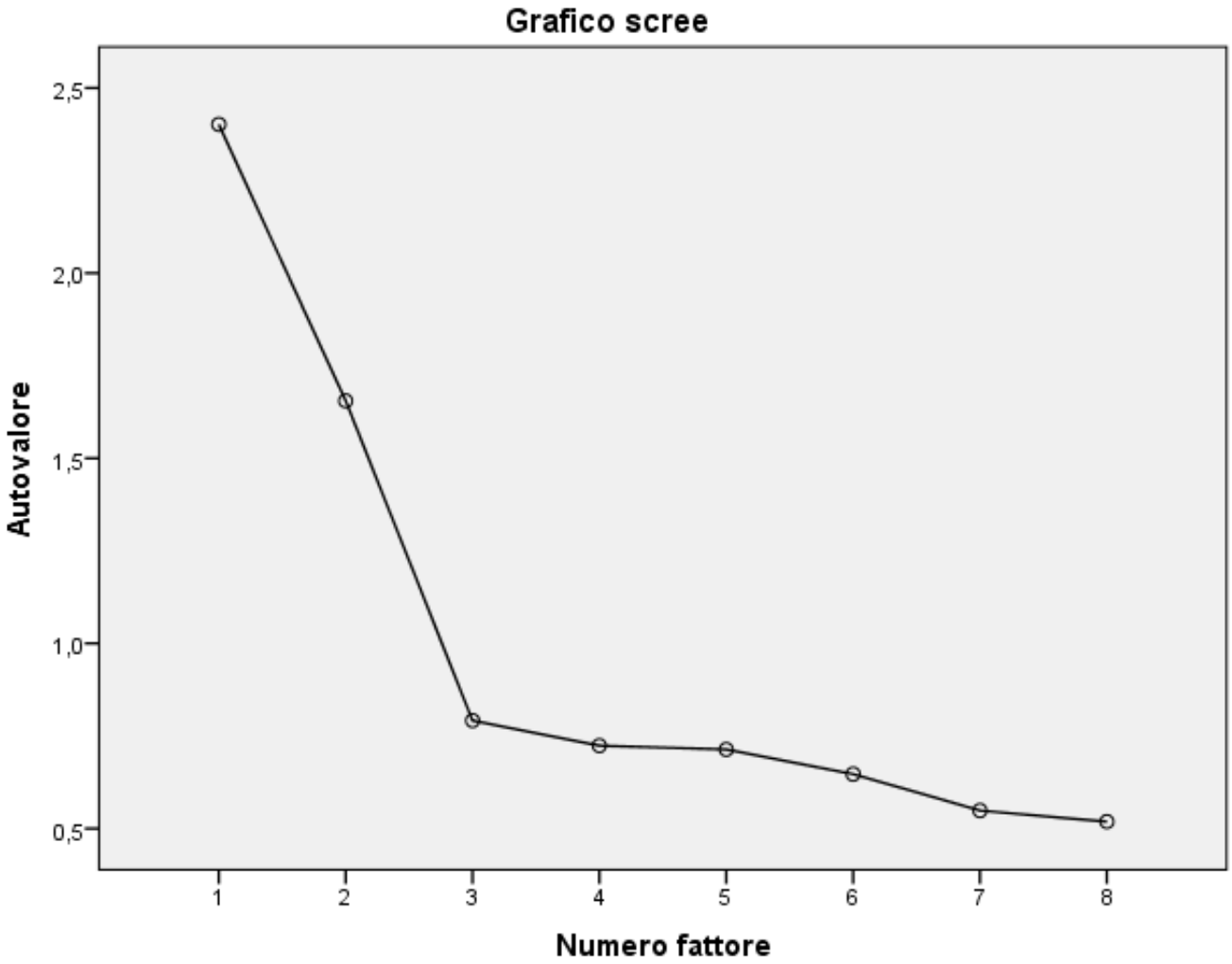


Grafico scree
Autovalore
2,5
2,0
1,5
1,0
0,5
1
2
3
4
5
6
7
8
Numero fattore

EFA IN SPSS

Correlazioni riprodotte

		fp10	fp15	fp26	fp30	nr6	nr16_lg10	nr17_lg10	nr20_lg10
Correlazione riprodotta	fp10	,261[a]	,343	,283	,349	,059	,027	,057	,006
	fp15	,343	,456[a]	,386	,464	,118	,078	,113	,055
	fp26	,283	,386	,341[a]	,394	,162	,131	,153	,119
	fp30	,349	,464	,394	,473[a]	,122	,081	,117	,057
	nr6	,059	,118	,162	,122	,296[a]	,301	,276	,327
	nr16_lg10	,027	,078	,131	,081	,301	,311[a]	,280	,340
	nr17_lg10	,057	,113	,153	,117	,276	,280	,256[a]	,303
	nr20_lg10	,006	,055	,119	,057	,327	,340	,303	,375[a]
Residuo[b]	fp10		,025	-,027	-,006	-,009	,012	,023	-,016
	fp15	,025		,004	-,020	,036	-,018	-,040	,019
	fp26	-,027	,004		,025	-,040	,031	-,023	,022
	fp30	-,006	-,020	,025		-,002	-,013	,037	-,022
	nr6	-,009	,036	-,040	-,002		,000	,034	-,020
	nr16_lg10	,012	-,018	,031	-,013	,000		-,023	,014
	nr17_lg10	,023	-,040	-,023	,037	,034	-,023		-,001
	nr20_lg10	-,016	,019	,022	-,022	-,020	,014	-,001	

Metodo di estrazione: Fattorizzazione dell'asse principale.

a. Comunalità riprodotte

b. I residui vengono calcolati tra le correlazioni osservate e riprodotte. Ci sono 0 (0,0%) residui non ridondanti con valori assoluti maggiori di 0,05.

EFA IN SPSS

Matrice dei fattori[a]

	Fattore	
	1	2
fp10	,418	-,295
fp15	,593	-,323
fp26	,559	-,169
fp30	,605	-,328
nr6	,401	,368
nr16_lg10	,363	,423
nr17_lg10	,376	,339
nr20_lg10	,362	,493

Metodo di estrazione: Fattorizzazione dell'asse principale.

a. 2 fattori estratti. 8 iterazioni richieste.

Matrice del modello[a]

	Fattore	
	1	2
fp10	,525	-,070
fp15	,678	-,013
fp26	,544	,113
fp30	,691	-,011
nr6	,046	,531
nr16_lg10	-,022	,563
nr17_lg10	,048	,492
nr20_lg10	-,071	,627

Metodo di estrazione: Fattorizzazione dell'asse principale.
Metodo di rotazione: Promax con normalizzazione Kaiser.

a. Convergenza per la rotazione eseguita in 3 iterazioni.

Matrice di struttura

	Fattore	
	1	2
fp10	,507	,067
fp15	,675	,164
fp26	,574	,256
fp30	,688	,170
nr6	,185	,543
nr16_lg10	,126	,557
nr17_lg10	,177	,504
nr20_lg10	,093	,608

Metodo di estrazione: Fattorizzazione dell'asse principale.
Metodo di rotazione: Promax con normalizzazione Kaiser.

Matrice di correlazione dei fattori

Fattore	1	2
1	1,000	,261
2	,261	1,000

Metodo di estrazione: Fattorizzazione dell'asse principale.
Metodo di rotazione: Promax con normalizzazione Kaiser.

ESERCIZIO 5:
REALIZZAZIONE DI UN MODELLO DI ANALISI FATTORIALE ESPLORATIVA

Effettuare un modello di analisi fattoriale esplorativa.

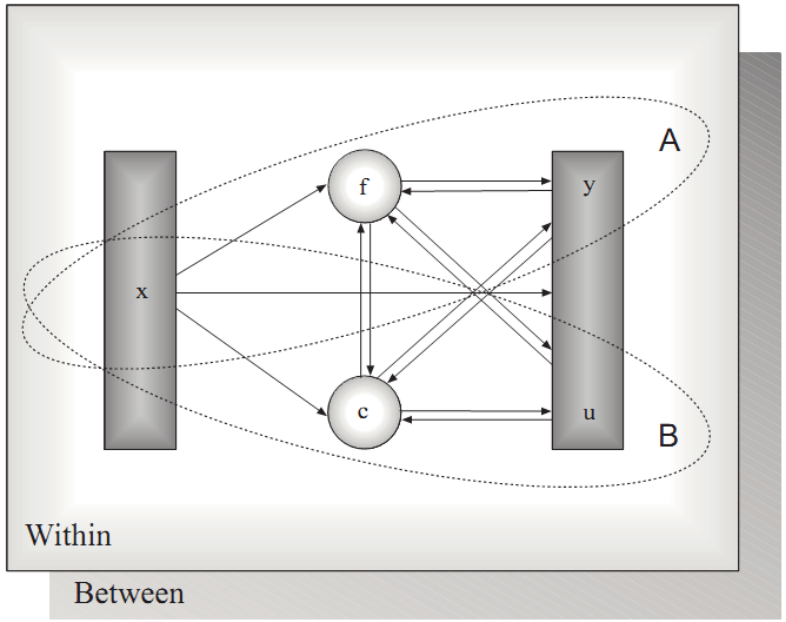
I dati sono nel file spss ESE_EFA.SAV

items:
Workload: QWI_1 QWI_2 QWI_3 QWI_4 QWI_5
Organizational Constraints: OCS_1 OCS_2 OCS_3 OCS_4 OCS_5 OCS_6 OCS_7 OCS_8 OCS_9 OCS_10 OCS_11

Effettuare l'analisi con SPSS scegliendo il metodo più adeguato per l'estrazione e la rotazione dei fattori, dopo aver esaminato le proprietà distributive degli item

MODELLI DI EQUAZIONI STRUTTURALI

Mplus



ELEMENTI DI BASE

- Cosa sono i Modelli di Equazioni Strutturali (SEM)
- Le componenti dei SEM
- Ipotesi di base e modelli matematici
- Fasi dei SEM
- Condizioni di applicabilità

I MODELLI DI EQUAZIONI STRUTTURALI

I Modelli di Equazioni Strutturali (SEM, Structural Equation Modeling) rappresentano una classe di modelli statistici che permettono di esprimere in maniera semplificata e formalizzata le relazioni tra i costrutti considerati in una determinata teoria (o in parti di essa).

I SEM consentono di esaminare se un modello in cui vengono ipotizzate determinate relazioni tra un insieme di variabili é *consistente* con i dati empirici.

I MODELLI DI EQUAZIONI STRUTTURALI

Il punto di partenza é rappresentato da una matrice di varianze/covarianze che riassume le relazioni tra le variabili osservate.

```
4.02
1.17 3.68
1.01 1.26 3.30
1.47 1.53 1.37  3.99
 .13  .55  .39   .49 3.16
 .28  .20  .46   .25  .91  2.57
 .50  .24  .49   .65  .90   .68  2.81
 .22  .31  .57   .18  .76   .81   .73  2.20
 .01 -.02  .02  -.01  .17   .07   .13   .13  .15
 .15  .09  .00   .06 -.05   .03  -.08  -.12 -.03  .16
```

I MODELLI DI EQUAZIONI STRUTTURALI

Il punto di arrivo é rappresentato da:

a) un insieme di parametri che quantificano le relazioni specificate nel modello (λ, θ, β, ψ);

b) una statistica associata ad ognuno di questi parametri che consente di esaminarne la significatività statistica (t, z);

c) una matrice delle varianze/covarianze tra le variabili osservate del modello riprodotta tramite i parametri del modello ($S\hat{}$, $\Sigma(\theta\hat{})$).

d) uno o più indici che misurano la bontà dell'adattamento del modello ai dati, cioè la corrispondenza del modello con i dati osservati (χ^2, RMSEA, SRMR, CFI, ecc.);

Rappresentazioni e Formalizzazioni nei SEM

Le relazioni tra le variabili in un modello di equazioni strutturali possono essere rappresentate in 4 modi differenti:

a) Una descrizione verbale

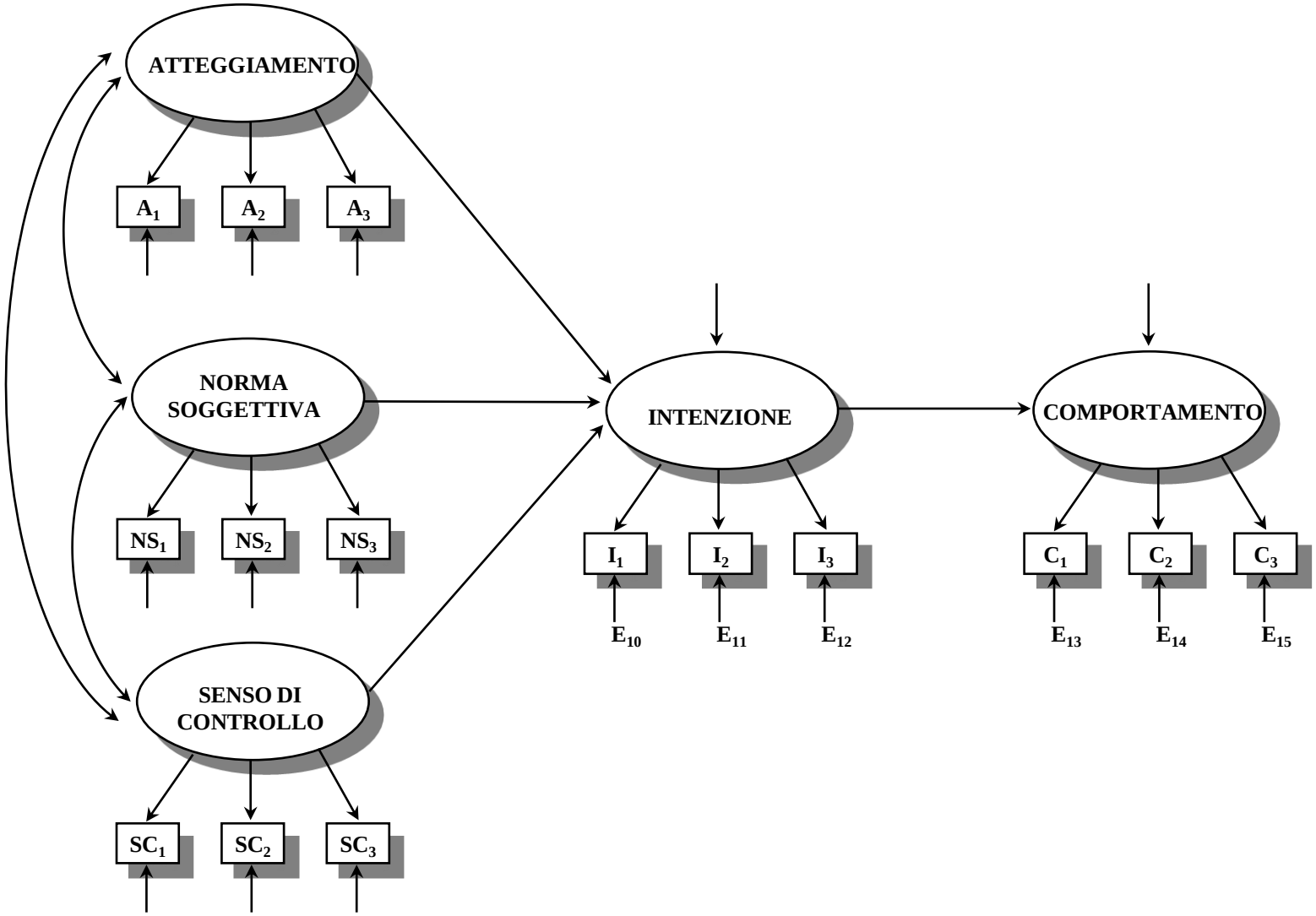
b) Una rappresentazione grafica (path diagram)

c) Una formalizzazione matematica dove le variabili *dipendenti* sono espresse come *equazioni algebriche lineari*, ovvero combinazioni lineari di altre variabili incluse nel modello, es.: $y_1=\lambda_{11}\eta_1 + \varepsilon_1$

d) Un insieme di comandi scritti nella sintassi di un linguaggio di programmazione (es., f1 by y1 y2 y3;)

Queste quattro rappresentazioni DEVONO coincidere

Formalizzazione diagrammatica di un modello strutturale



I Modelli di Equazioni Strutturali prevedono:

- una serie di equazioni in cui vengono specificate le relazioni tra i costrutti chiave della teoria (definibile modello strutturale);

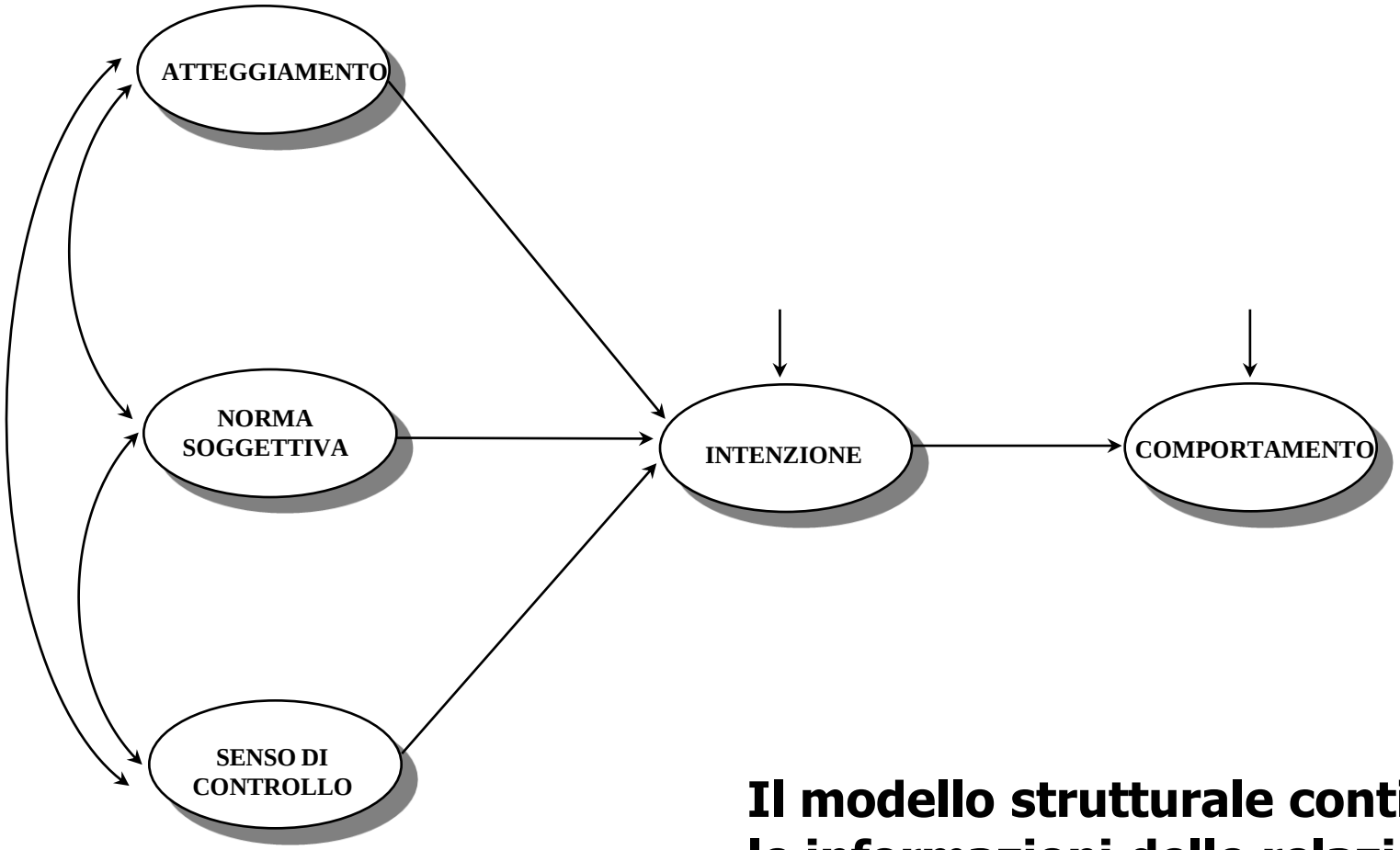
- una serie di equazioni che specificano le relazioni tra le variabili latenti e le variabili osservate (definibile modello di misura).

* la presenza soltanto del modello di misurazione definisce un modello di *analisi fattoriale confermativa* (*Confirmatory Factor Analysis*, CFA);

* la presenza soltanto del modello strutturale, definisce un modello di *path analysis* su variabili osservate.

L'analisi strutturale

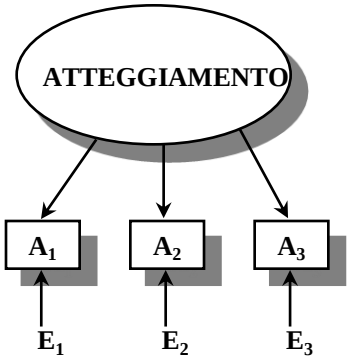
Il modello strutturale

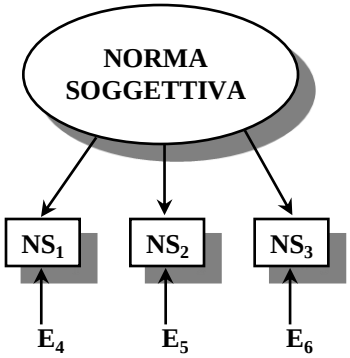


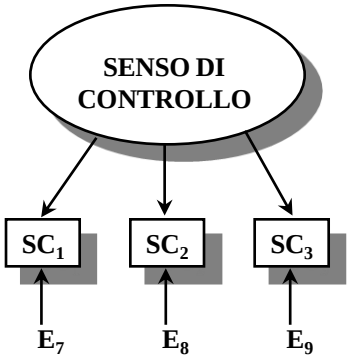
Il modello strutturale contiene le informazioni delle relazioni tra i costrutti

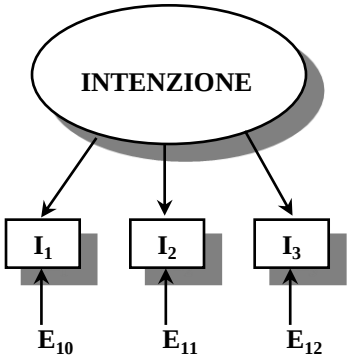
L'analisi strutturale

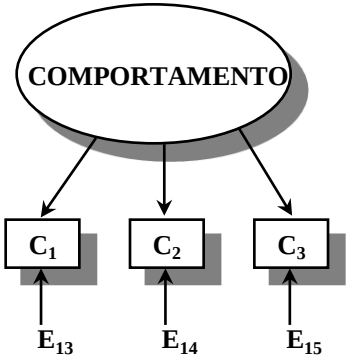
Il modello di misura











Il modello di misura descrive le relazioni tra le variabili osservate e i costrutti (o variabili latenti)

Nei modelli di Equazioni strutturali abbiamo:

* **Tipi differenti di variabili**

  Indipendenti (Esogene)/Dipendenti (Endogene)

  Latenti/Misurate

* **Tipi differenti di relazioni tra le variabili**

  Associazione (covarianza/relazione simmetrica)

  Effetto diretto (influenza diretta/relazione asimmetrica)

  Effetto indiretto (influenza indiretta)

Nei modelli di Equazioni strutturali abbiamo:

* **Tipi differenti di residui associati alle variabili dipendenti:**

  Variabili misurate: errore di misurazione

  Variabili latenti: errore di specificazione

* **Tipi differenti di parametri:**

  Liberi: parametri per cui si calcola una stima

  Fissi: parametri il cui valore è stato fissato (di solito a 0 o a 1)

  Vincolati: parametri liberi le cui stime sono vincolate ad assumere soltanto certi valori (ad esempio, 2 parametri vincolati ad assumere lo stesso valore)

I parametri di un modello di equazioni strutturali

Sono i termini delle equazioni per i quali viene prodotta una stima nella soluzione:

a) le relazioni di influenza tra le variabili (ovvero gli "effetti diretti" di una variabile su un'altra variabile)

b) le varianze e le covarianze delle variabili indipendenti

c) le varianze e le covarianze dei residui

I parametri di un modello di equazioni strutturali

*** I parametri che quantificano l'influenza diretta sono i coefficienti strutturali (pesi beta)**

*** I parametri che quantificano l'associazione non direzionale tra le variabili sono i coefficienti di covarianza**

*** Le varianze e le covarianze delle variabili dipendenti non sono parametri del modello ma vengono spiegate dal modello**

I parametri di un modello di equazioni strutturali – i coefficienti strutturali:

*** Corrispondono ai pesi beta (*beta weights*) della regressione**

*** Esprimono la quantità di variazione che ci si attende nella variabile dipendente in concomitanza di un cambiamento di una unità nelle variabili indipendenti ad essi associate, *mantenendo costanti* le altre variabili**

*** Sono i pesi che misurano l'influenza parziale di una VI su una VD, al netto delle correlazioni della VI con le altre variabili in analisi.**

I parametri di un modello di equazioni strutturali – i residui delle variabili dipendenti osservate:

Riflettono diverse componenti:

- componente stocastica: discrepanza tra dati campionari e dati della popolazione;
- errore di misurazione: le variabili non sono misurate perfettamente;
- componente di specificità della variabile: varianza unica sistematica ma non condivisa con le altre variabili;
- errore di specificazione: modello fattoriale inadeguato (troppi pochi fattori);
- errore di specificazione: forma della relazione diversa da quella lineare.

I parametri di un modello di equazioni strutturali – i residui delle variabili dipendenti latenti:

Riflettono soprattutto l'errore di specificazione:

- predittori importanti della variabile esclusi dal modello;

- predittori irrilevanti inclusi nel modello;

- forma della relazione diversa da quella lineare.

FORMALIZZAZIONE MATEMATICA: MODELLI MATEMATICI NEI SEM

Esistono diversi modelli matematici che definiscono come specificare le relazioni tra le variabili.

*** Muthén**

$$\boldsymbol{y} = \Lambda\eta + \varepsilon$$
$$\eta = B\eta + \zeta$$

*** Jöreskog, Keesling & Wiley**

$$\boldsymbol{y} = \Lambda_y\eta + \varepsilon$$
$$\boldsymbol{x} = \Lambda_x\xi + \delta$$
$$\eta = B\eta + \Gamma\xi + \zeta$$

*** Bentler & Weeks**

$$\eta = B\eta + \Gamma\xi$$

Modello di Muthén

Il modello generale di equazioni strutturali di Muthén è rappresentabile con le seguenti equazioni di base:

$$\boldsymbol{y} = \nu + \Lambda\eta + [\mathrm{K}\mathbf{x} +] \varepsilon \qquad \textbf{(a)}$$

$$\eta = \alpha + \mathrm{B}\eta + [\Gamma\mathbf{x} +] \zeta \qquad \textbf{(b)}$$

Queste due equazioni comprendono sei matrici di parametri che definiscono un modello completo MPLUS.

Modello di Muthén

Le variabili del modello

1. y (ipsilon): sono le variabili osservate che rappresentano gli indicatori delle variabili (η) nell'equazione relativa al modello di misura.

2. η (eta): sono le variabili latenti misurate dalle y nell'equazione relativa al modello di misura. Possono essere sia indipendenti che dipendenti nel modello strutturale.

3. ε (epsilon): sono i termini residuali associati alle variabili y, e non sono correlati con nessuna altra variabile del modello.

4. ζ (zeta): sono i residui o termini di disturbo associati alle variabili latenti η che risultano dipendenti.

5. x (ics): sono le variabili indipendenti osservate.

Le matrici del modello di Muthén

1. Λ (Lambda): matrice dei coefficienti di regressione per esprimere le variabili y come funzione delle variabili latenti (le η) nell'equazione relativa al modello di misura.
2. Θ_ε (Theta) = E($\varepsilon\varepsilon'$): matrice delle varianze e covarianze dei termini di errore ε (epsilon), associati alle variabili y.
3. ν (nu): vettore di intercette delle variabili osservate dipendenti y.
4. Ψ (Psi): matrice che contiene le varianze e le covarianze delle variabili latenti indipendenti, E($\eta\eta'$), e dei termini di disturbo ζ associati alle variabili latenti dipendenti, E($\zeta\zeta'$).
5. α (alfa): vettore di intercette delle variabili latenti η.

Le matrici del modello di Muthén

6. B (Beta): quando tutte le variabili osservate sono continue la matrice Beta contiene i seguenti parametri:

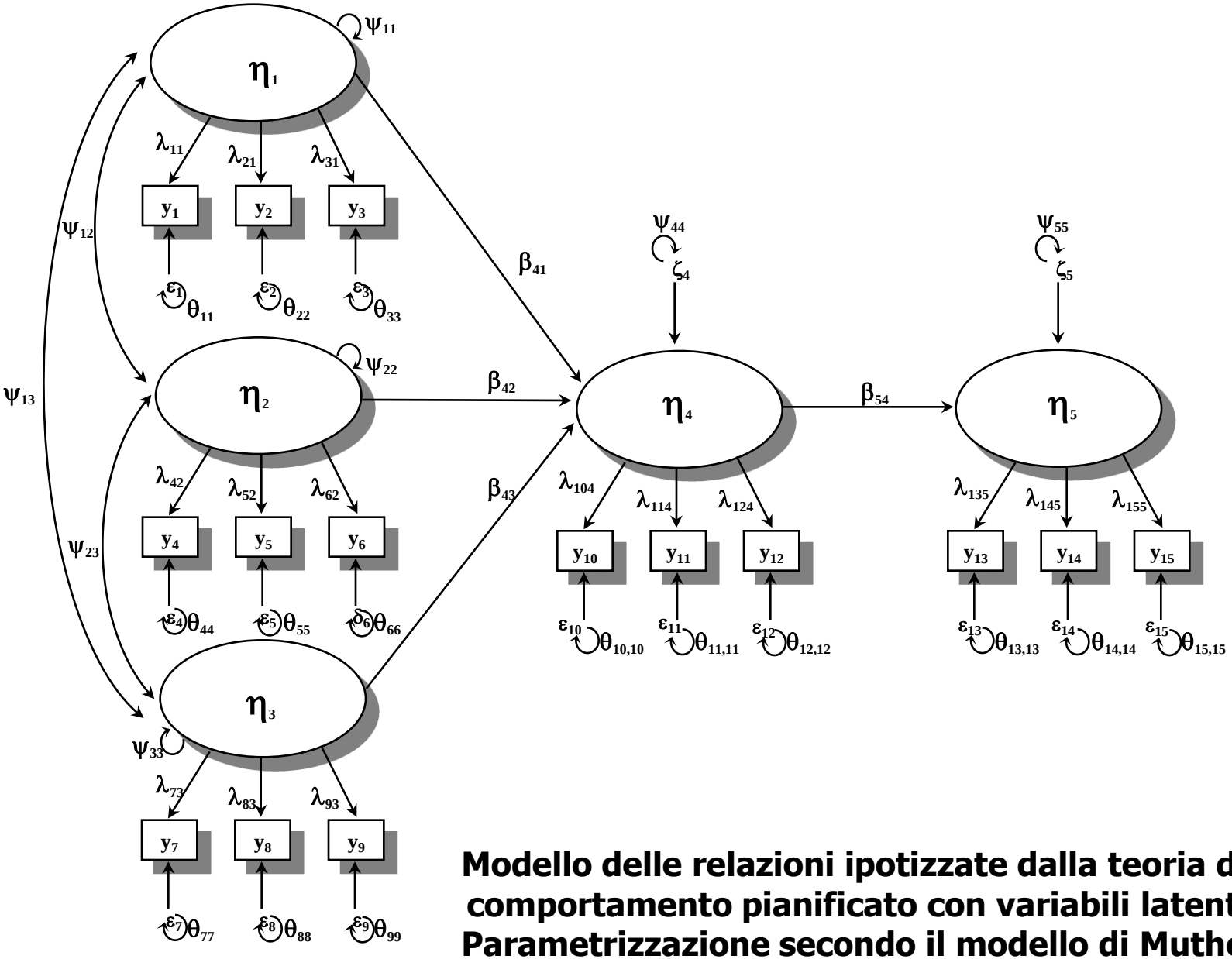
- i coefficienti di regressione (β, beta) per predire le variabili latenti (η) dalle η stesse

- i coefficienti di regressione (γ, gamma) per predire le variabili latenti *dipendenti* (η) dalle variabili *indipendenti* osservate (x)

- i coefficienti di regressione (κ, kappa) per predire le variabili osservate dipendenti y dalle variabili *indipendenti* osservate (x)

Indipendentemente dall'etichetta utilizzata, tutti questi parametri sono elementi della matrice beta, quindi sono etichettati con tale lettera nel modello di Muthén.

Rappresentazione grafica del modello di Muthén



Modello delle relazioni ipotizzate dalla teoria del comportamento pianificato con variabili latenti: Parametrizzazione secondo il modello di Muthén

$$y = \Lambda \eta + \varepsilon \qquad \Theta$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_6 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 & 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 & 0 & 0 \\ \lambda_{31} & 0 & 0 & 0 & 0 \\ 0 & \lambda_{42} & 0 & 0 & 0 \\ 0 & \lambda_{52} & 0 & 0 & 0 \\ 0 & \lambda_{62} & 0 & 0 & 0 \\ 0 & 0 & \lambda_{73} & 0 & 0 \\ 0 & 0 & \lambda_{83} & 0 & 0 \\ 0 & 0 & \lambda_{93} & 0 & 0 \\ 0 & 0 & 0 & \lambda_{104} & 0 \\ 0 & 0 & 0 & \lambda_{114} & 0 \\ 0 & 0 & 0 & \lambda_{124} & 0 \\ 0 & 0 & 0 & 0 & \lambda_{135} \\ 0 & 0 & 0 & 0 & \lambda_{145} \\ 0 & 0 & 0 & 0 & \lambda_{155} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \end{bmatrix}$$

$$\begin{bmatrix} \theta^{\varepsilon}_{11} & & & & & \\ 0 & \theta^{\varepsilon}_{22} & & & & \\ 0 & 0 & \theta^{\varepsilon}_{33} & & & \\ 0 & 0 & 0 & \theta^{\varepsilon}_{44} & & \\ 0 & 0 & 0 & 0 & \theta^{\varepsilon}_{55} & \\ 0 & 0 & 0 & 0 & 0 & \theta^{\varepsilon}_{66} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

Modello di misura

$$\eta = B \eta + \zeta$$

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \beta_{41} & \beta_{42} & \beta_{43} & 0 & 0 \\ 0 & 0 & 0 & \beta_{54} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \\ \zeta_5 \end{bmatrix}$$

$$\begin{bmatrix} \psi_{11} & & & & \\ \psi_{21} & \psi_{22} & & & \\ \psi_{31} & \psi_{32} & \psi_{33} & & \\ 0 & 0 & 0 & \psi_{44} & \\ 0 & 0 & 0 & 0 & \psi_{55} \end{bmatrix} \quad \Psi$$

Modello strutturale

Parametri del modello di Muthén

I parametri di un modello di equazioni strutturali sono:

Gli effetti diretti di una variabile su un'altra (B, Λ)

Le varianze delle variabili indipendenti e le covarianze tra le variabili indipendenti (Ψ)

Le varianze dei residui e le covarianze tra i residui (Ψ, Θ)

L'ipotesi delle Strutture di Covarianza nei SEM

Un modello di equazioni strutturali va sottoposto a verifica confrontandolo con i dati osservati.

Dai parametri del modello è possibile "ricostruire" la matrice delle varianze/covarianze tra le variabili osservate.

Questa corrispondenza consente di valutare l'adeguatezza del modello teorico, che definisce i parametri, rispetto ai dati osservati (cioè la bontà dell'adattamento del modello ai dati).

L'ipotesi delle Strutture di Covarianza nei SEM

L'adeguatezza del modello rispetto ai dati viene valutata tramite la seguente ipotesi formale:

$$\Sigma = \Sigma(\theta)$$

Secondo questa equivalenza, è possibile definire un modello che specifica le relazioni tra le variabili del modello, in modo che i parametri (θ) del modello consentano di esprimere/ricostruire la matrice di covarianze Σ tra le variabili osservate.

L'ipotesi delle Strutture di Covarianza nei SEM

Poiché gli elementi di Σ possono essere espressi come funzioni dei parametri del modello, questa ipotesi consente di valutare quanto il modello é consistente con i dati empirici.

Se il modello è corretto e i parametri sono noti la matrice viene riprodotta esattamente.

$\Sigma=\Sigma(\theta)$ rappresenta l'ipotesi nulla da verificare con i dati campionari, attraverso le stime dei parametri θ.

Poiché sotto determinate assunzioni S e $\Sigma(\theta\hat{})$ sono stimatori consistenti rispettivamente di Σ e $\Sigma(\theta)$, l'ipotesi nulla verrà accettata se S = $\Sigma(\theta\hat{})$.

Parametri del modello e matrice Σ

I parametri sono messi in relazione con la matrice di varianze e covarianze tra le variabili osservate S, tramite la seguente espressione matriciale:

$$\Sigma = \Lambda (I - B)^{-1} \Psi (I - B)^{-1'} \Lambda' + \Theta$$

che si semplifica nei casi seguenti:

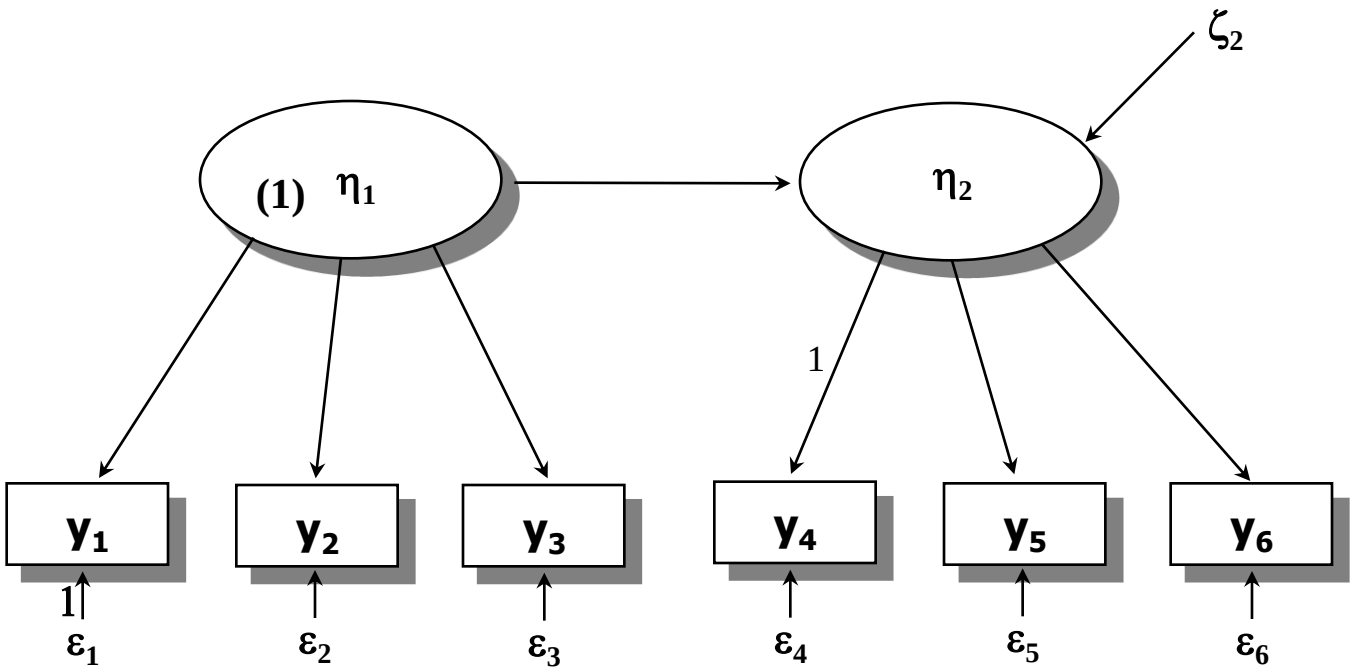
CFA con variabili osservate continue

$$\Sigma = \Lambda \Psi \Lambda' + \Theta$$

Path analysis con variabili osservate continue

$$\Sigma = (I - B)^{-1} \Psi (I - B)^{-1'}$$

ESERCIZIO 1:
TIPI DI VARIABILI/ RELAZIONI/ ERRORI - CALCOLO DEL NUMERO DEI PARAMETRI



Quante sono le VI?
Quante sono le variabili misurate?
Quanti sono gli effetti diretti?

Quante sono le VD?
Quante sono le variabili latenti?
Ci sono delle covarianze tra variabili?

Quanti sono gli errori di specificazione e quanti quelli di misurazione ?
Quanti sono i parametri del modello che vengono stimati ?

Fasi dei modelli di equazioni strutturali

a) specificazione (formulazione) del modello

b) identificazione del modello e dei suoi parametri

c) stima dei parametri del modello

d) valutazione del modello

e) modifica del modello

Specificazione del modello

a) definire le variabili latenti e osservate che compongono le diverse equazioni;

b) definire quali variabili saranno indipendenti (o "esogene") e quali dipendenti (o endogene);

c) definire le relazioni "direzionali" e "non-direzionali" che legano le variabili;

d) definire i vincoli (constraints) tra i parametri.

Identificazione dei parametri e del modello

Un modello si dice identificato se tutti i suoi parametri sono identificati, cioè se per tutti i suoi parametri esiste una soluzione numerica *unica*.

Un modello NON identificato é un modello in cui la stessa matrice di covarianza riprodotta é compatibile con più insiemi di stime numeriche per gli stessi parametri:

$$\Sigma(\theta_1\hat{}) = \Sigma(\theta_2\hat{}), \text{ ma } \theta_1\hat{} \neq \theta_2\hat{}$$

dove $\theta_1\hat{}$ e $\theta_2\hat{}$ sono due vettori che contengono valori differenti per gli stessi parametri, ovvero per uno stesso modello esaminato su uno stesso campione.

Un modello non identificato: l'EFA

Un caso noto di modello non identificato è l'Analisi Fattoriale Esplorativa. Le soluzioni fattoriali possono essere ruotate: ogni rotazione non cambia le proprietà matematiche della soluzione, ma ne cambia solo l'interpretazione *concettuale*.

	Sol. Iniziale(A)		Sol. Ruotata (B)	
	F1	F2	F1	F2
Determinato	.68	.51	.17	.83
Dinamico	.74	.48	.24	.85
Energico	.78	.33	.36	.77
Affidabile	.80	-.41	.87	.23
Responsabile	.84	-.43	.91	.24
Scrupoloso	.82	-.33	.83	.30

La % di varianza spiegata dalle due soluzioni è la stessa (circa 78%)

Un modello non identificato: l'EFA

A **A′** **R**

.68	.51	*	.68	.74	.78	.80	.84	.82	=	.72	.75	.70	.34	.35	.39
.74	.48		.51	.48	.33	-.41	-.43	-.33		.75	.78	.74	.40	.42	.45
.78	.33									.70	.74	.72	.49	.51	.53
.80	-.41									.34	.40	.49	.81	.85	.79
.84	-.43									.35	.42	.51	.85	.89	.83
.82	-.33									.39	.45	.53	.79	.83	.78

B **B′** **R**

.17	.83	*	.17	.24	.36	.87	.91	.83	=	.72	.75	.70	.34	.35	.39
.24	.85		.83	.85	.77	.23	.24	.30		.75	.78	.74	.40	.42	.45
.36	.77									.70	.74	.72	.49	.51	.53
.87	.23									.34	.40	.49	.81	.85	.79
.91	.24									.35	.42	.51	.85	.89	.83
.83	.30									.39	.45	.53	.79	.83	.78

AA′= BB′, ma A ≠ B

Un modello non identificato: l'EFA

Le due matrici sono perfettamente equivalenti da un punto di vista matematico:

- spiegano la stessa varianza delle variabili originali

- riproducono altrettanto bene la matrice delle correlazioni di partenza

La differenza dei valori delle saturazioni ha implicazioni profondamente diverse per la interpretazione del modello:

- Soluzione non ruotata:

1 fattore generale e 1 bipolare

- Soluzione ruotata:

2 fattori, Coscienziosità e Energia

Condizioni necessarie per l'identificazione

a) il numero dei coefficienti da stimare (t) deve essere inferiore al numero di elementi non ridondanti nella matrice delle covarianze (uguale a q(q+1)/2, dove q é il numero di variabili osservate) [t rule], ovvero i gradi di libertà devono essere positivi.

$$GDL = [q(q+1)/2] - t$$

b) la scala di misura delle variabili latenti deve essere fissata

Se tutti i parametri del modello possono essere espressi come funzioni delle varianze e covarianze tra le variabili osservate, allora il modello é sicuramente identificato (es: $\lambda_{21} = \sqrt{(\sigma_{32}\,\sigma_{21}/\sigma_{31})}$

Identificazione e gradi di libertà

Modello "non identificato" (underidentified): gradi di libertà negativi, il numero di parametri è maggiore del numero di var/cov.

Modello "appena identificato" (just identified), modello saturo: ha tante incognite quanti parametri noti, ovvero tanti parametri quante var/cov, per cui ha 0 gradi di libertà. Ha un fit perfetto e per questo non è interessante: non può essere disconfermato dai dati empirici.

Modello "sovra identificato" (overidentified): ha meno parametri che var/cov, quindi ha gradi di libertà maggiori di 0. Può essere disconfermato empiricamente.

Condizioni in cui é realistico considerare il modello come identificato

a) Errori di misurazione non correlati (θ è diagonale)

b) Modello ricorsivo:

- nessuna relazione di influenza reciproca (es, $\eta_1 \rightleftarrows \eta_2$)

- nessun loop (es, $\eta_1 \longrightarrow \eta_2 \longrightarrow \eta_3 \longrightarrow \eta_1$)

c) Numero di indicatori per ogni variabile latente adeguato (solitamente, almeno 3 indicatori)

Fenomeni che possono far supporre non identificazione

a) Presenza di valori non ammissibili per alcuni parametri:

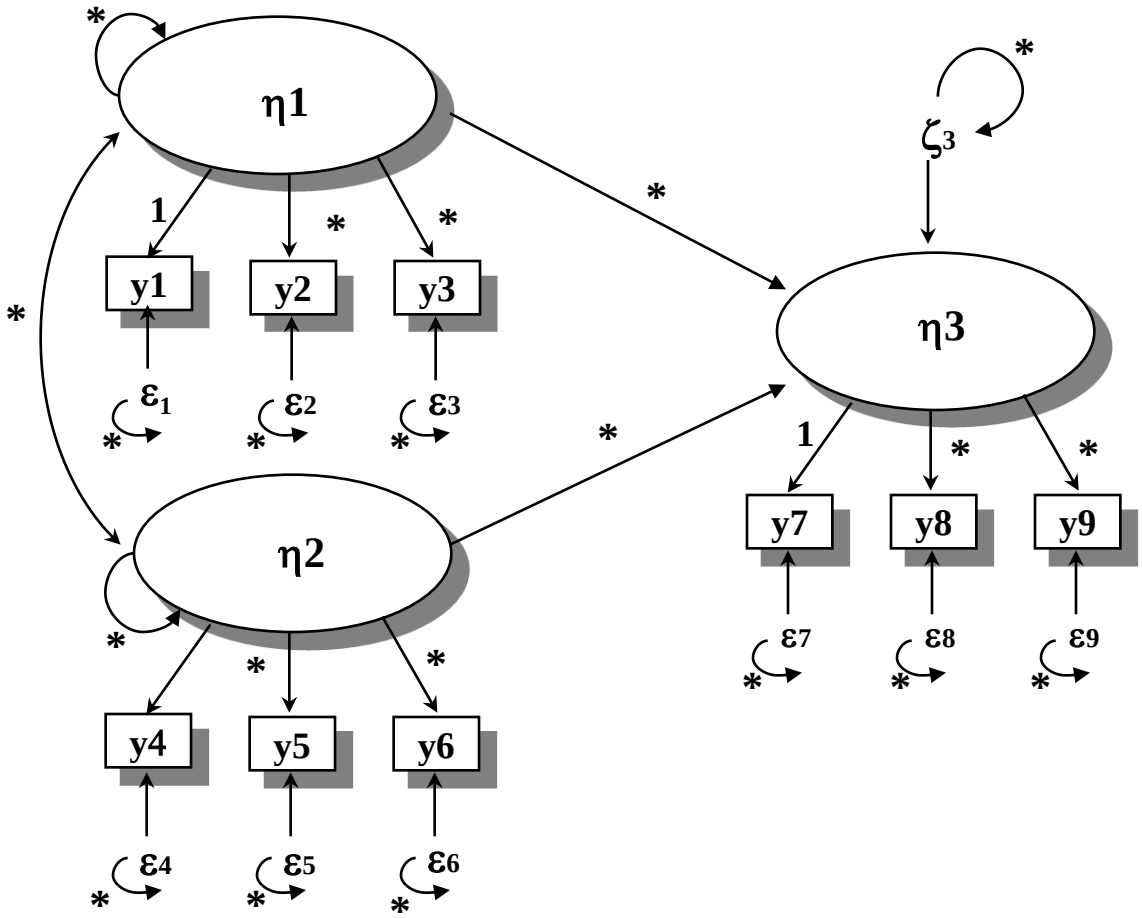
- varianze negative

- correlazioni o coefficienti strutturali maggiori di 1 in valore assoluto, nel caso di una soluzione standardizzata ("Heywood cases")

b) Valori molto elevati per gli errori standard: in un modello identificato l'errore standard dovrebbe avere un valore intorno al valore del parametro diviso √N, o di 1 diviso √N (McDonald, 1999).

c) Non convergenza del processo di iterazione

ESERCIZIO 2:
IDENTIFICAZIONE DEL MODELLO – CALCOLO DEI GRADI DI LIBERTA'

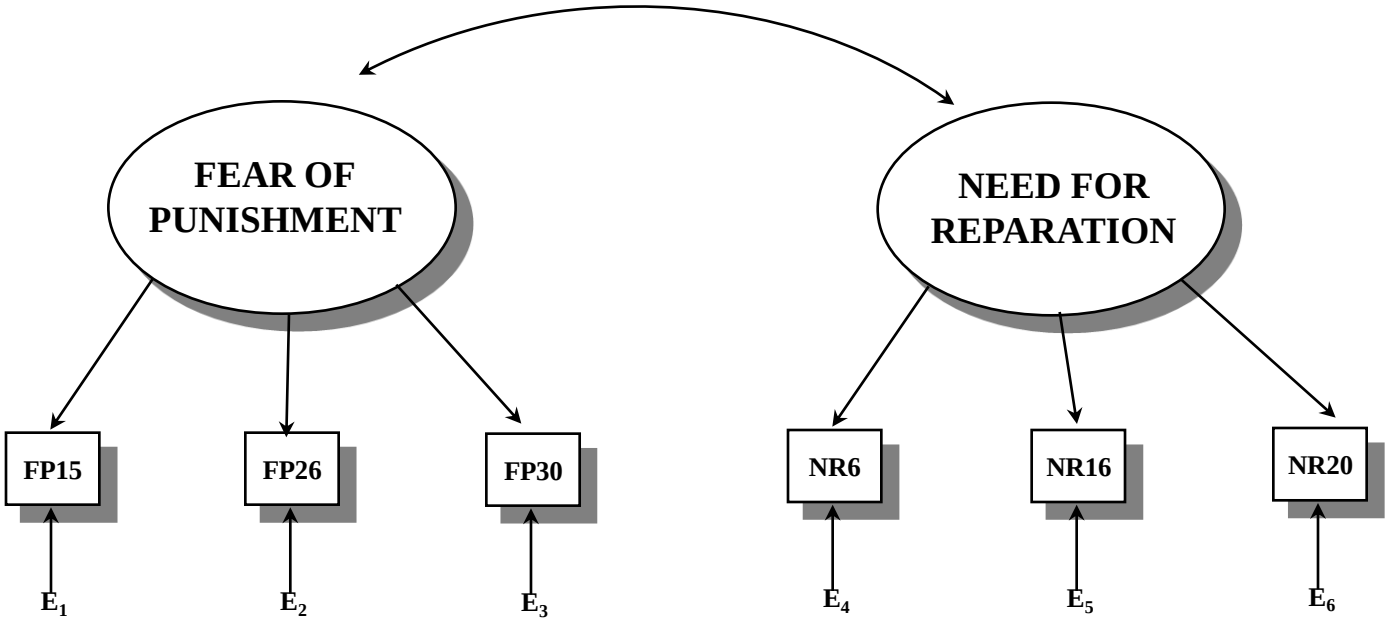


E' un modello identificato ? Perché ? Quanti sono i suoi gradi di libertà ?

NB: gli asterischi indicano dei parametri che devono essere stimati.

CFA CON MPLUS

Modello teorico della struttura fattoriale da esaminare



Item FP: FP15) Mi capita di sentirmi gravato da sentimenti di colpa; FP26) Mi capita di ripensare con timore alle conseguenze di ciò che ho fatto o detto; FP30) Mi sono sentito come se mi rimordesse la coscienza; Item NR: NR6) Sento il bisogno di riparare ai torti che posso aver procurato ad altri; NR16) Prima o poi i nodi delle proprie colpe vengono al pettine; NR20) Di fronte ai miei errori desidero riparare il prima possibile

CFA CON MPLUS

PARIS_2011_CFA.INP

```
TITLE:   CFA

DATA:
FILE IS dati_efa_cfa_grezzi.dat;


VARIABLE:
NAMES ARE
      FP10     FP15     FP26       FP30
      NR6      NR16     NR17       NR20   GENDER;
USEV ARE
      FP15     FP26       FP30     NR6      NR16       NR20    ;
 MISSING ARE ALL (9);
MODEL:
  FEARPUN BY   FP15     FP26       FP30     ;
   NEEDREP BY   NR6      NR16       NR20    ;

OUTPUT:   STANDARDIZED SAMPSTAT MODINDICES(3.84) TECH1;
```

CFA CON MPLUS

* Le righe TITLE, DATA, VARIABLE hanno lo stesso significato delle linee omonime utilizzate nel programma per la EFA.

* Le equazioni che seguono la linea "MODEL:" mostrano quali variabili devono essere messe in relazione a quali fattori.

* Il comando "**BY**" sta per "**MEASURED BY**" e indica che la variabile latente a sinistra è misurata dalle variabili osservate a destra.

CFA CON MPLUS

* Nella riga OUTPUT viene richiesto di aggiungere all'output di default :

- la soluzione completamente standardizzata [STANDARDIZED]
- le statistiche campionarie [SAMPSTAT]
- gli indici di modificazione statisticamente significativi [MODINDICES(3.84)]
- le matrici del modello di Muthén [TECH1]

CFA CON MPLUS

I settings di default nel modello CFA sono i seguenti

* La saturazione fattoriale della prima variabile dopo il "BY" viene fissata a 1

* Le saturazioni fattoriali delle altre variabili sono stimate

* Le varianze di errore sono stimate

* Le covarianze tra i residui sono fissate a 0

* Le varianze dei fattori sono stimate

* Le covarianze tra le variabili esogene sono stimate

* Il metodo di stima è ML

CFA CON MPLUS

Parametrizzazione alternativa del Modello CFA

```
MODEL:
   FEARPUN BY   FP15*     FP26       FP30     ;
   NEEDREP BY   NR6*      NR16       NR20    ;
   FEARPUN@1 NEEDREP@1;
 FEARPUN with   NEEDREP@0;
```

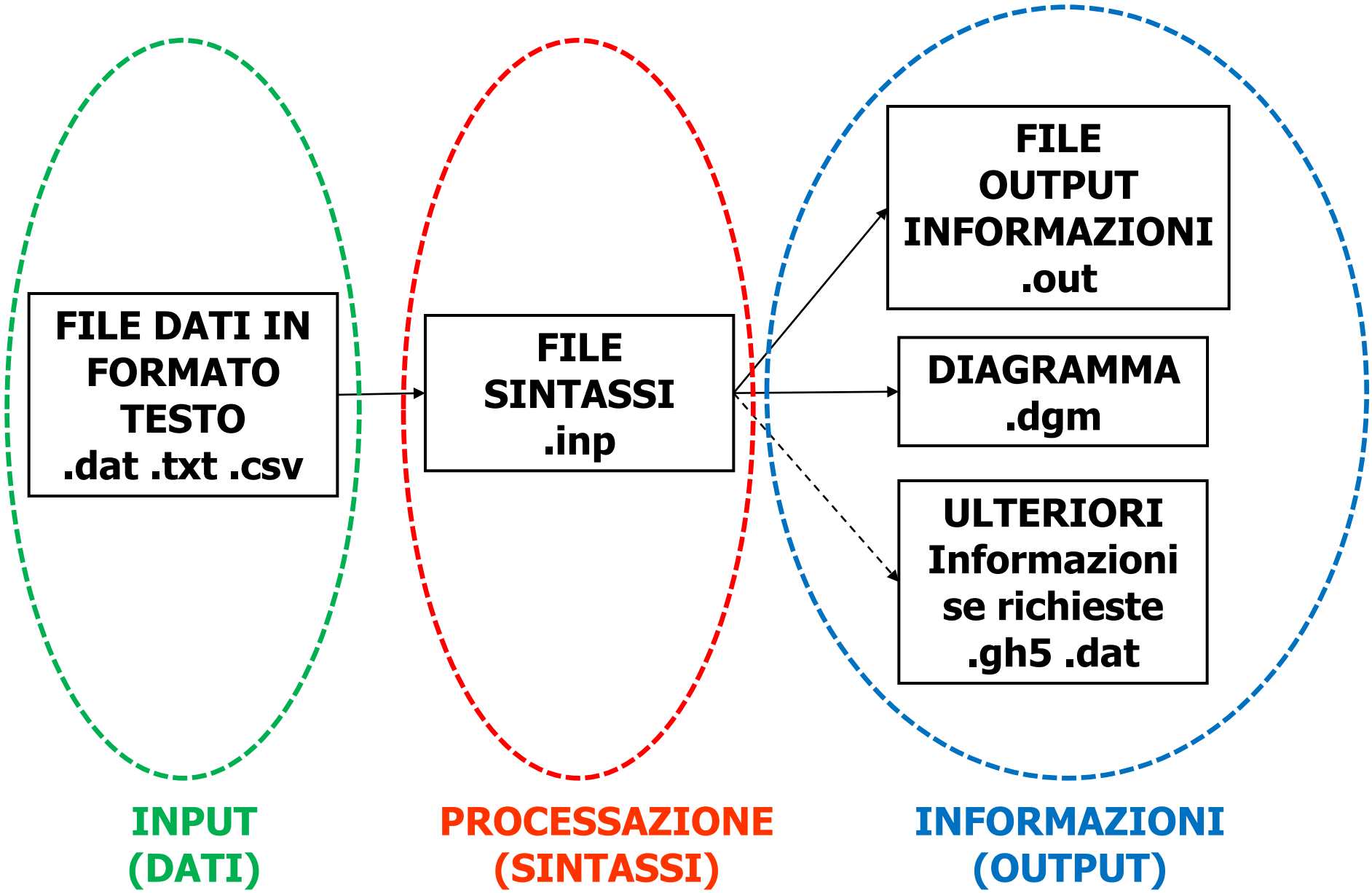
Nella sintassi di MPLUS

- L'asterisco (*) segnala che il parametro è libero.
- La chiocciolina (@) segnala che il parametro è fissato al valore che segue.

PREPARARE I DATI

DA IMPORTARE IN M*plus*

LA STRUTTURA DEL PROGRAMMA M*plus*



PREPARARE I DATI DA IMPORTARE IN M*plus*

M*plus* legge essenzialmente file dati in formato testo (con estensioni .dat. .txt. .csv. ecc).

Il file dati può contenere dati di tipo differente: *a*) può contenere i dati "grezzi", ossia una matrice Casi X Variabili; *b*) il vettore delle medie delle variabili (non obbligatorio) considerate, seguito dalla loro matrice di varianze e covarianze; *c*) il vettore delle deviazioni standard (può essere preceduto dal vettore delle medie) delle variabili, seguito dalla loro matrice di correlazioni.

L'ordine delle variabili all'interno del file dati, a differenza di altri software, è del tutto indifferente.

DATI GREZZI – TYPE IS INDIVIDUAL

File Modifica Visualizza Inserisci Formato ?

769	1.00	3.35	3.00	999.00	3.29	3.35	4.00	4.80	999.00	4.58	4.58	2.85	2.69	1.92	2.69
903	1.00	3.12	2.94	2.88	2.35	999.00	1.80	4.20	3.60	3.25	999.00	2.00	2.31	2.54	2.23
908	1.00	3.12	2.71	2.94	2.53	2.76	4.40	5.00	4.20	3.67	2.58	2.62	2.77	2.62	2.46
916	1.00	2.41	2.24	3.06	2.24	999.00	3.40	1.80	4.00	4.00	999.00	2.00	2.08	2.15	2.15
934	1.00	2.18	2.94	3.35	2.71	1.65	2.00	4.60	4.40	3.50	2.25	1.85	1.54	1.38	2.38
940	1.00	3.29	3.06	3.12	2.94	2.88	4.40	4.40	3.60	4.50	4.33	1.77	2.31	2.31	2.54
959	1.00	2.65	2.35	3.24	999.00	3.29	4.20	3.00	4.60	999.00	4.00	1.69	2.23	2.08	2.54
961	1.00	3.18	3.18	3.00	3.53	3.18	3.80	5.00	5.00	4.67	4.50	2.54	2.77	1.77	2.85
975	1.00	3.53	2.88	3.59	999.00	3.50	5.00	4.60	4.80	999.00	5.00	2.85	3.00	2.62	2.85
980	1.00	3.29	2.71	3.38	2.76	2.71	4.00	3.60	3.00	4.17	3.33	1.92	2.31	2.08	2.54
1168	1.00	3.29	3.71	2.94	3.24	999.00	5.00	4.60	4.60	4.83	999.00	2.46	2.23	2.23	2.38
2000	1.00	3.65	3.65	3.06	2.88	3.29	5.00	4.80	4.40	4.33	4.25	2.92	2.85	2.69	2.77
2001	1.00	2.71	2.53	2.88	2.47	999.00	2.40	3.20	4.00	3.00	999.00	2.15	2.00	1.92	2.62
2003	1.00	3.59	3.41	3.65	2.65	2.65	4.20	4.20	3.80	4.08	4.00	2.54	2.77	2.23	2.85
2005	1.00	3.59	2.47	3.12	2.59	2.59	4.80	1.25	3.20	3.67	4.75	1.85	2.69	2.62	2.38
2008	1.00	3.35	3.47	3.53	3.35	999.00	5.00	4.60	4.80	3.67	999.00	2.38	2.62	2.31	2.38
2012	1.00	1.50	2.88	2.94	2.94	1.82	1.60	3.80	2.80	3.64	2.83	2.46	2.85	2.31	2.17
2013	1.00	3.00	2.88	3.47	999.00	3.12	2.80	1.20	4.60	999.00	3.75	2.54	2.15	2.62	2.54
2024	1.00	2.94	2.76	2.35	2.94	3.12	3.40	3.80	5.00	4.50	4.83	2.46	2.38	2.23	2.54
2025	1.00	3.47	3.53	3.35	3.41	3.00	4.00	4.60	4.20	4.08	3.42	2.50	2.77	2.08	2.77
2028	1.00	3.53	3.65	3.76	3.88	3.94	4.60	5.00	5.00	5.00	4.92	2.54	2.31	2.08	2.54
2031	1.00	3.53	3.53	3.24	3.00	3.82	4.40	2.80	2.40	3.00	5.00	2.00	2.54	2.00	2.38
2040	1.00	3.41	3.76	3.29	3.29	3.29	4.40	5.00	4.00	3.82	4.33	2.31	2.85	2.31	2.23
2042	1.00	3.82	3.94	4.00	999.00	3.94	5.00	5.00	4.80	999.00	4.42	2.46	2.92	2.69	2.85
5005	1.00	2.41	2.29	2.24	3.00	999.00	2.80	3.00	1.40	4.25	999.00	1.85	2.08	1.69	2.00
5019	1.00	2.94	3.65	2.94	3.24	999.00	4.80	5.00	4.00	3.83	999.00	2.38	2.46	2.23	2.69
5025	1.00	3.06	3.06	2.76	2.76	2.65	4.80	4.60	4.60	3.25	3.50	2.38	2.46	2.31	2.31
5040	1.00	3.00	2.82	3.06	2.47	999.00	2.60	5.00	3.60	3.17	999.00	2.69	2.69	2.31	2.92

Per ottenere la Guida, premere F1

DATI GREZZI

1) Le righe rappresentano i soggetti (o la singola osservazione) e le colonne le variabili;

2) Tutti i valori contenuti nel file dati debbono essere preferibilmente numerici per evitare problemi nella lettura del file da parte del software;

3) È più semplice codificare tutti i valori mancanti con un unico valore comune a tutte le variabili (es. 9, 99, 999, -9, -99, -999, ecc.);

4) Vanno sostituite tutte le virgole con il punto, altrimenti M*plus* non leggerà i dati.

DATI GREZZI – Creare il file da SPSS

1) Il suggerimento è salvare un file .sav includendo solo le variabili cui siamo interessati (procedura estremamente utile quando abbiamo un dataset d'origine molto ampio) e, qualora se ne disponga, del codice identificativo del soggetto;

2) Ricodificare tutti i valori mancanti con un unico valore, ricordandosi di specificare anche nel file .sav il valore che abbiamo scelto;

3) Salvare il file nel suo formato originale ed esportarlo come un file dati formato testo, attraverso il percorso: FILE – SALVA CON NOME – SALVA COME TIPO: Ascii fisso (*.dat).

DATI GREZZI – Creare il file da SPSS

4) Aprire il file dati in comune editor di testo (es. Blocco note o WordPad) e sostituire (qualora ce ne fossero) tutte le virgole con il punto (nel blocco note, il percorso è MODIFICA – SOSTITUISCI – TROVA(.) – SOSTITUISCI CON(.) – SOSTITUISCI TUTTO);

5) Ricordarsi di salvare il file dopo aver effettuato tali modifiche.

Il file così creato avrà una estensione .dat

DATI GREZZI – Creare il file da SPSS

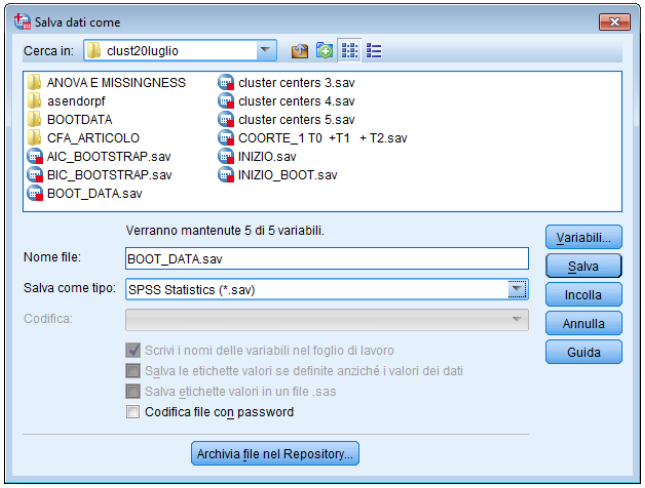
1)

	CASENUM	GIUDIZIO	AIUT	EMON	SRL
1	1	2,00	3,33	2,33	3,00
2	2	2,00	3,67	2,67	5,00
3	3	2,25	3,00	3,33	3,00
4	3	2,25	3,00	3,33	3,00
5	5	3,25	3,33	3,33	4,00
6	6	3,25	2,67	3,33	4,00
7	7	2,25	3,00	2,00	2,67
8	7	2,25	3,00	2,00	2,67
9	8	3,50	3,67	4,33	4,67
10	9	2,75	4,00	3,00	4,00
11	9	2,75	4,00	3,00	4,00
12	9	2,75	4,00	3,00	4,00

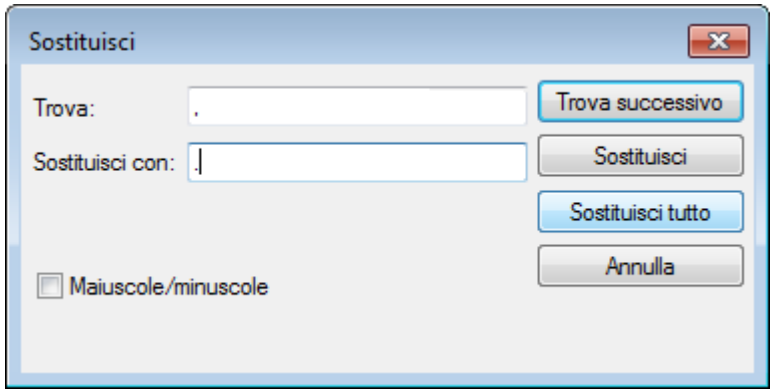
2)

Nome	Tipo	Larghezza	Decimali	Etichetta	Valori	Mancante/i
CASENUM	Numerico	8	0		Nessuno	Nessuno
GIUDIZIO	Numerico	8	2		Nessuno	-999,00
AIUT	Numerico	8	2		Nessuno	-999,00
EMON	Numerico	8	2		Nessuno	-999,00
SRL	Numerico	8	2		Nessuno	-999,00

3)



4)



DATI GREZZI – Creare il file da EXCEL

1) Anche qui, ricodificare tutti i valori mancanti con un solo valore che li rappresenti all'interno di tutte le variabili;

2) Esportare il file in formato testo, attraverso il percorso: FILE – SALVA CON NOME – SALVA COME TIPO: TEXT Tab delimited (*.txt);

3) Aprire il file in un editor di testo e si noterà che se la prima riga del file excel era occupata dalle etichette delle variabili queste verranno trasportate anche nel file testo (va cancellata!);

4) Sostituire tutte le virgole con il punto nell'editor di testo.

PREPARARE I DATI DA IMPORTARE IN M*plus*

RIASSUMENDO…

L'ordine delle variabili all'interno del file dati, a differenza di altri software, è del tutto indifferente.

I file che Mplus può leggere sono sempre file di testo.

Si possono importare dati grezzi (TYPE IS INDIVIDUAL, consigliata la codifica con un unico valore per i missing all'interno di variabili differenti) oppure file contenenti *summary data* (TYPE IS COVARIANCE, TYPE IS MEANS COVARIANCE, TYPE IS CORRELATIONS, ecc., molto utile per riprodurre modelli con i dati reperibili normalmente all'interno degli articoli).

Stima dei parametri

Bisogna stimare i parametri θ in modo che la matrice riprodotta $\Sigma(\theta)$ sia più vicina possibile a quella osservata Σ.

Esistono diversi metodi di stima che danno origine ad altrettante "funzioni di adattamento" (fit function) utilizzabili per questo compito.

Le funzioni di adattamento sono rappresentate dall'espressione "F(S,$\Sigma(\theta\hat{})$)".

I metodi di stima più comunemente utilizzati sono:

*** la massima verosimiglianza (Maximum Likelihood)**

*** i minimi quadrati (Least Squares)**

Il metodo della Massima Verosimiglianza (ML)

In generale scopo del metodo ML è quello di stimare il parametro della popolazione che sia più vicino al valore campionario osservato, ovvero tale che la probabilità di osservare quel valore campionario data quella stima sia massima, ovvero:

$$P[(\hat{\theta} - \bar{x}) < \varepsilon] = \max$$

Tra tutte le stime possibili di θ la stima ML è quella che massimizza la probabilità o la verosimiglianza (likelihood) che le differenze tra stima e valore osservato siano dovute solo al caso.

Il metodo della Massima Verosimiglianza (ML) per le stime dei parametri di un modello Mplus

Nel modello Mplus abbiamo visto che ci sono diverse matrici di parametri. Tramite queste matrici è possibile riprodurre la matrice di covarianze campionaria S nel modo seguente:

$$\Sigma(\theta) = \Lambda\,(I - B)^{-1}\,\Psi\,(I - B)^{-1\prime}\Lambda' + \Theta$$

Indichiamo la matrice riprodotta in questo modo $\Sigma(\theta)$. Se il modello funziona perfettamente allora $S = \Sigma(\theta)$

Come facciamo a trovare le migliori stime possibili per i parametri del modello affinché S e $\Sigma(\theta)$ siano più simili possibile ?

Il metodo della Massima Verosimiglianza (ML) per le stime dei parametri di un modello Mplus

Il miglior insieme di stime per i parametri θ è quello che rende minima la differenza tra S e $\Sigma(\theta)$.

Questo vuol dire che le migliori stime sono quelle che forniscono la più elevata probabilità (cioè, verosimiglianza) di osservare S da un campione casuale di una popolazione che ha matrice di covarianza Σ, la cui stima è rappresentata da $\Sigma(\theta)$.

Le stime θ che determinano tale matrice $\Sigma(\theta)$ sono le stime di Massima Verosimiglianza.

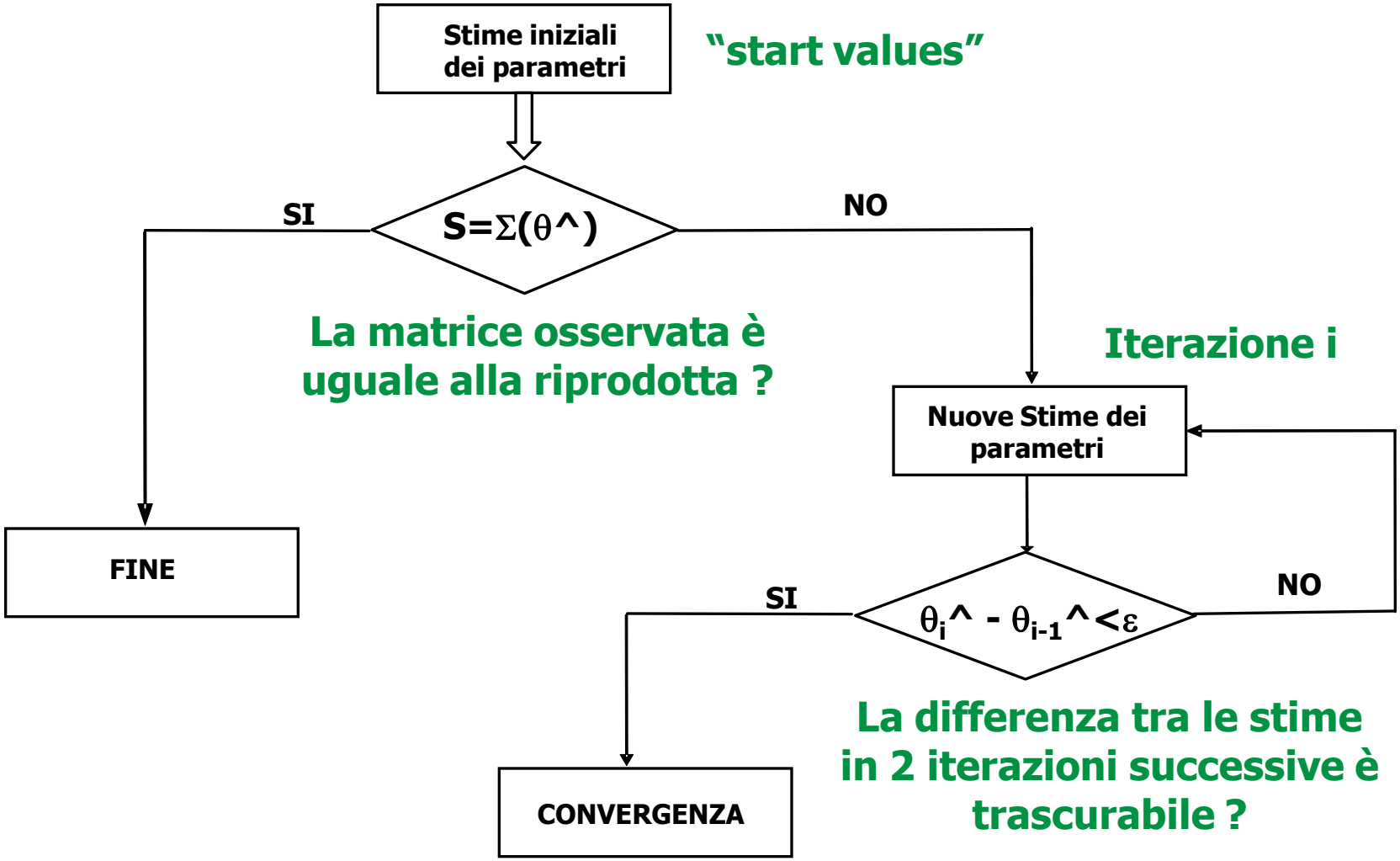
$$F_{ML}(S,\Sigma(\theta\hat{}\,)) = tr(S\Sigma(\theta)^{-1}) + \ln|\Sigma(\theta)| - \ln|S| - q$$

Il metodo della Massima Verosimiglianza (ML) per le stime dei parametri di un modello Mplus

Minimizzare la funzione F_{ML} vuol dire trovare dei valori per le stime θ che MINIMIZZANO la differenza tra S e $\Sigma(\theta)$. Questi valori vengono ricavati attraverso una procedura iterativa, basata su alcuni consolidati algoritmi di calcolo numerico.

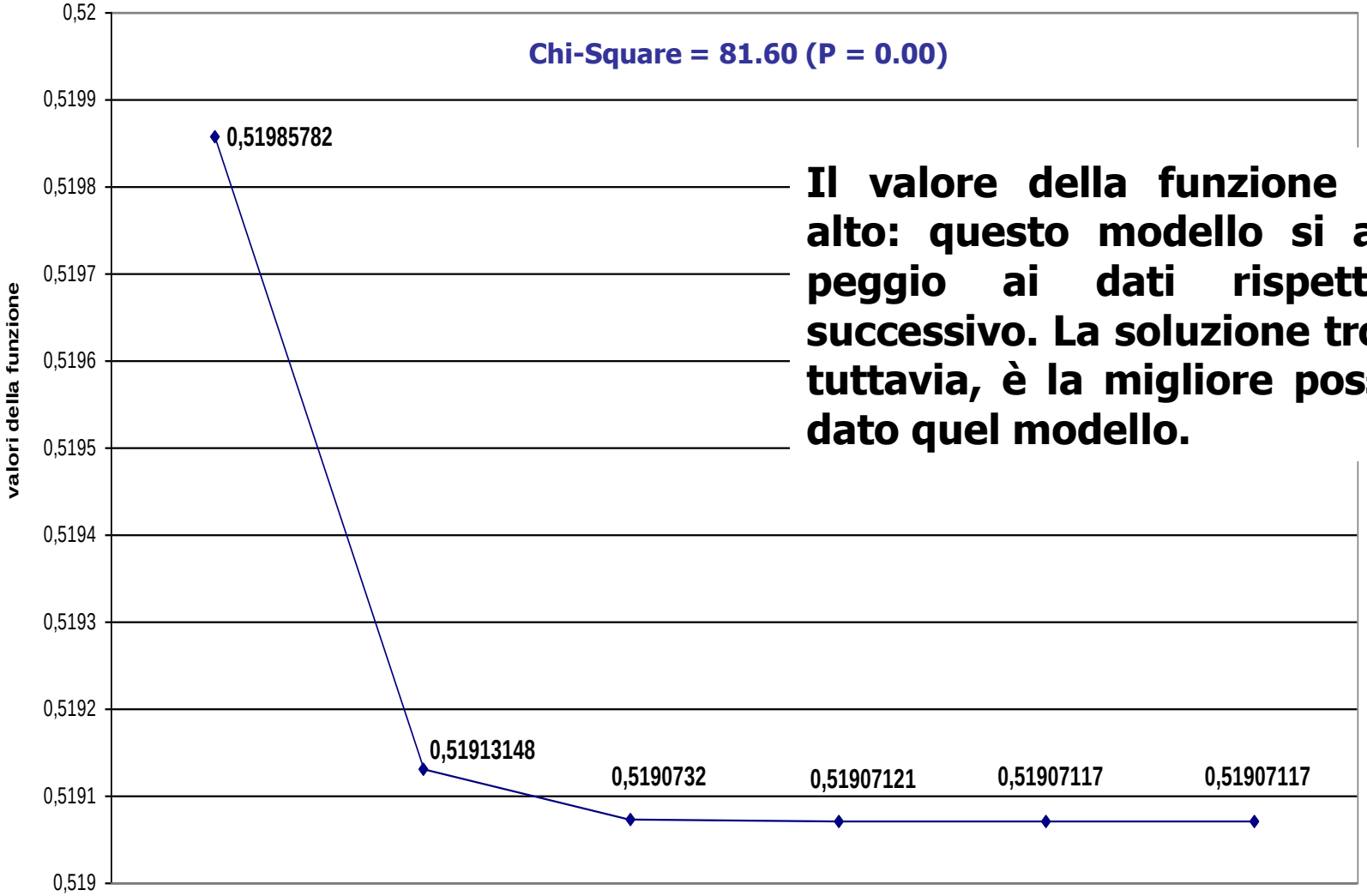
E' necessario utilizzare un metodo iterativo perché la funzione F_{ML} è una funzione non lineare complessa, per la quale non è possibile arrivare facilmente ad una soluzione esplicita.

Le stime di Massima Verosimiglianza (ML): Il processo iterativo di stima dei parametri in Mplus



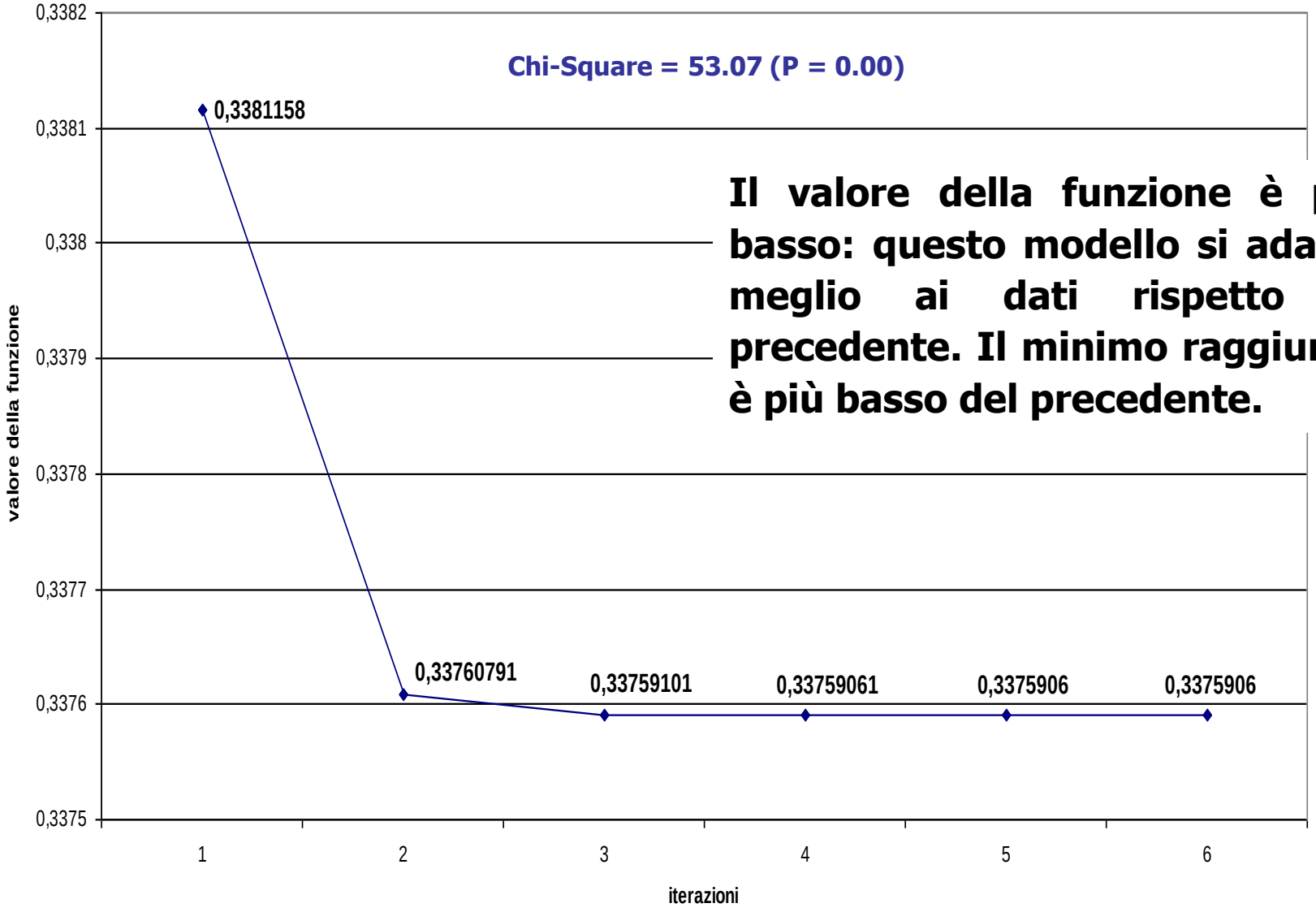
Le stime ottenute sono quelle che rendono minima la differenza tra Σ(θ^) e S

Il processo di stima dei parametri



Il valore della funzione è più alto: questo modello si adatta peggio ai dati rispetto al successivo. La soluzione trovata, tuttavia, è la migliore possibile, dato quel modello.

Il processo di stima dei parametri



Il valore della funzione è più basso: questo modello si adatta meglio ai dati rispetto al precedente. Il minimo raggiunto è più basso del precedente.

CFA IN MPLUS CON VARIABILI CONTINUE

Metodi di stima dei parametri

* GLS – minimi quadrati generalizzati
* WLS – minimi quadrati ponderati (ADF)
* ML – massima verosimiglianza (DEFAULT)

*MLM, MLMV, MLR, MLF – metodi robusti che forniscono risultati corretti anche in presenza di violazioni della normalità (solo con dati grezzi)
MLM = Correzione di Satorra-Bentler
MLMV = Correzione rispetto alle varianze-covarianze e alle medie
MLR = Correzione di Yuan-Bentler, robusta anche alla non indipendenza delle osservazioni
MLF = Standard error approssimati, Chi-quadrato tradizionale

Alcune funzioni di adattamento "speciali" in Mplus

I metodi robusti che forniscono risultati corretti anche in presenza di violazioni della normalità (solo con dati grezzi)

Massima Verosimiglianza Robusta (ML-M):
Il chi-quadrato e gli errori standard vengono opportunamente ponderati per fornire stime corrette in presenza di non normalità (Satorra e Bentler).

***Minimi Quadrati Ponderati Robusti* (*WLS-MV*):**
Fornisce stime corrette dei parametri e valori corretti del chi-quadrato e degli errori standard in presenza di non normalità e di variabili categoriali (Muthén).

Alcune funzioni di adattamento "speciali" in Mplus

```
TITLE:  CFA
DATA:
LISTWISE=ON ;
FILE IS dati_efa_cfa_grezzi.dat;
VARIABLE:
NAMES ARE
    FP10   FP15   FP26    FP30
    NR6    NR16   NR17    NR20  GENDER;
USEV ARE
    FP15   FP26    FP30   NR6    NR16    NR20  ;
 MISSING ARE ALL (9);
ANALYSIS:  ESTIMATOR = MLM;
MODEL:
  FEARPUN BY  FP15   FP26    FP30   ;
   NEEDREP BY  NR6    NR16    NR20  ;
OUTPUT:  STANDARDIZED SAMPSTAT MODINDICES(3.84)
TECH1;
```

PARIS_2011_CFA_MLM.INP

CFA CON MPLUS

Lettura dell'*Output : Elementi più importanti*

SAMPLE STATISTICS

Covariances

	FP15	FP26	FP30	NR6	NR16
FP15	2.232				
FP26	0.821	1.952			
FP30	1.037	0.910	2.426		
NR6	0.323	0.241	0.264	1.905	
NR16	0.115	0.290	0.155	0.537	1.589
NR20	0.156	0.245	0.087	0.485	0.510

Covariances

	NR20
NR20	1.353

CFA CON MPLUS

Lettura dell'*Output : Elementi più importanti*

Correlations

	FP15	FP26	FP30	NR6	NR16
FP15	1.000				
FP26	0.393	1.000			
FP30	0.446	0.418	1.000		
NR6	0.156	0.125	0.123	1.000	
NR16	0.061	0.165	0.079	0.308	1.000
NR20	0.090	0.151	0.048	0.302	0.348

Correlations

	NR20
NR20	1.000

CFA CON MPLUS

Lettura dell'*Output : Elementi più importanti*

UNIVARIATE SAMPLE STATISTICS

UNIVARIATE HIGHER-ORDER MOMENT DESCRIPTIVE STATISTICS

Variable/ Sample Size	Mean/ Variance	Skewness/ Kurtosis	Minimum/ Maximum	% with Min/Max	Percentiles 20%/60%	40%/80%	Median
FP15	3.176	0.063	1.000	18.50%	2.000	3.000	3.000
816.000	2.234	-0.956	6.000	6.74%	4.000	4.000	
FP26	3.686	-0.321	1.000	9.48%	2.000	4.000	4.000
812.000	1.952	-0.608	6.000	9.11%	4.000	5.000	
FP30	3.148	0.057	1.000	21.89%	1.000	3.000	3.000
813.000	2.426	-1.117	6.000	6.40%	4.000	5.000	
NR6	4.376	-0.856	1.000	6.24%	3.000	4.000	5.000
817.000	1.904	0.177	6.000	22.40%	5.000	6.000	
NR16	4.676	-1.096	1.000	4.02%	4.000	5.000	5.000
820.000	1.587	1.048	6.000	29.76%	5.000	6.000	
NR20	5.045	-1.510	1.000	2.57%	4.000	5.000	5.000
816.000	1.352	2.343	6.000	44.85%	6.000	6.000	

CFA CON MPLUS

```
 Chi-Square Test of Model Fit

          Value                             21.622
          Degrees of Freedom                    8
          P-Value                           0.0057

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.046
          90 Percent C.I.                    0.023  0.069
          Probability RMSEA <= .05           0.584

CFI/TLI

          CFI                                0.979
          TLI                                0.961
```

CFA CON MPLUS

E' il modello in cui non viene ipotizzata nessuna relazione tra le variabili.

```
Chi-Square Test of Model Fit for the Baseline Model

          Value                            663.540
          Degrees of Freedom                15
          P-Value                          0.0000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.026
```

Valutazione globale del modello

Il valore minimo della funzione di fit moltiplicato per il numero di soggetti meno 1 segue approssimativamente la distribuzione del χ^2 con gradi di libertà (df) uguali a (q(q+1)/2)-t, dove q è il numero di variabili osservate e t il numero di parametri stimati, e con valore atteso $E(\chi^2)=df$.

Questo consente di esaminare statisticamente l'adattamento del modello ai dati tramite l'ipotesi nulla $\Sigma=\Sigma(\theta)$.

Lo standard per questo confronto a livello di campione é che $\Sigma\hat{}$ e S siano uguali.

Valutazione globale del modello

Se il chi-quadrato risulta statisticamente non significativo, l'ipotesi nulla H_0: $\Sigma=\Sigma(\theta)$ non può essere rifiutata, dunque il modello teorico risulta compatibile con i dati empirici.

Se invece il valore del χ^2 risulta statisticamente significativo l'ipotesi nulla va rifiutata, dunque il modello teorico non risulta compatibile con i dati empirici.

Valutazione globale del modello

Indici alternativi di fit

Il chi quadrato è fortemente dipendente dalla numerosità del campione, e questo lo rende conservativo se il campione è ampio (rifiuta troppo spesso H_0) ma liberale se il campione è esiguo (rifiuta troppo di rado H_0).

Per questo sono stati sviluppati indici alternativi per valutare globalmente la bontà dell'adattamento.

* **Misure di fit nel campione (SRMR, RMR)**
* **Indici incrementali o comparativi (TLI, CFI)**
* **Indici di approssimazione (RMSEA)**

Valutazione globale del modello

Misure di fit nel campione

RMR = ROOT MEAN-SQUARE RESIDUAL =
$= [2\Sigma_i\Sigma_j(s_{ij} - \hat{\sigma}_{ij})^2 /(q(q+1))]^{1/2}$
SRMR = STANDARDIZED RMR

Indicano la media della varianza e covarianza residua, cioè non spiegata dal modello.

L'RMR può essere utilizzato per confrontare l'adattamento di due differenti modelli specificati sugli stessi dati, oppure di uno stesso modello specificato su dati differenti. Allora, il modello da privilegiare è quello che presenta un RMR più basso.

L'SRMR può essere interpretato in assoluto. Valori bassi indicano buon fit.

Valutazione globale del modello

Indici incrementali o comparativi

Valutano l'adeguatezza del modello rispetto ad un modello nullo in cui si ipotizza che non ci siano relazioni tra le variabili.

TUCKER AND LEWIS INDEX (TLI) - (NNFI, NON NORMED FIT INDEX)

$$\frac{\left(\chi^2_{nullo}/df_{nullo}\right)-\left(\chi^2_{target}/df_{target}\right)}{\left(\chi^2_{nullo}/df_{nullo}\right)-1}$$

Valutazione globale del modello

Indici incrementali o comparativi

Valutano l'adeguatezza del modello rispetto ad un modello nullo in cui si ipotizza che non ci siano relazioni tra le variabili.

COMPARATIVE FIT INDEX (CFI)

$$1 - \frac{\max\left[\left(\chi^2_{target} - df_{target}\right), 0\right]}{\max\left[\left(\chi^2_{nullo} - df_{nullo}\right), \left(\chi^2_{target} - df_{target}\right), 0\right]}$$

Il CFI varia da 0 a 1, il TLI può risultare anche maggiore di 1. Valori intorno a 1 indicano buon fit.

Valutazione globale del modello

Indice di approssimazione
RMSEA (Root Means Square Error of Approximation)

Il chi quadrato valuta l'ipotesi nulla che $\Sigma=\Sigma(\theta)$ e presuppone che esista un modello vero nella popolazione, ovvero un modello che è perfettamente consistente con i dati empirici, ovvero che rappresenta una "fotografia" perfetta della realtà.

Spesso questo è irrealistico. I modelli al più possono fornire un'immagine approssimativa della realtà.

Valutazione globale del modello

Indice di approssimazione

RMSEA (Root Means Square Error of Approximation)

Allora, l'RMSEA valuta quanto errore commettiamo nell'approssimare la realtà con il nostro modello.

Se l'errore è contenuto, il nostro modello non sarà una fotografia perfetta della realtà ma almeno la approssima sufficientemente bene.

Se l'errore è grande, il nostro modello non può considerarsi nemmeno un'approssimazione della realtà.

L'RMSEA valuta l'ipotesi che $\Sigma \approx \Sigma(\theta)$.

Valutazione globale del modello

Indice di approssimazione

RMSEA (Root Means Square Error of Approximation)

E' una misura della discrepanza tra Σ e $\Sigma(\theta)$ dovuta all'approssimazione, ponderata per i gradi di libertà del modello, e quindi é una misura del fit che tiene in considerazione la parsimonia del modello.

$$RMSEA = \frac{\hat{F}_0}{df}; \quad \hat{F}_0 = \max\left[\frac{(\min F - df)}{n}, 0\right]$$

RMSEA≤.05: errore di approssimazione minimo,
.05 <RMSEA≤ .08: errore di appross. accettabile,
RMSEA>.08: il modello non tiene nella popolazione.

Valutazione globale del modello

Indice di approssimazione
RMSEA (Root Means Square Error of Approximation)

Test of Close fit

Esamina l'ipotesi nulla che l'RMSEA sia inferiore a .05 e quindi la seguente ipotesi nulla: H_0: $\Sigma \approx \Sigma(\theta)$.

Se p(RMSEA ≤.05) > .05 non possiamo rifiutare l'ipotesi nulla che il modello sia almeno approssimativamente adeguato.

Valutazione globale del modello

Indice di approssimazione
RMSEA (Root Means Square Error of Approximation)

L'RMSEA viene fornito anche nella versione di stima intervallare (con un intervallo di confidenza del 10%).

Il limite inferiore dell'intervallo di confidenza dell'RMSEA deve essere minore di .05, il limite superiore minore di .08: Allora il test of close fit risulta non significativo.

Valutazione globale del modello

Le indicazioni di Hu e Bentler (1998, 1999) sui valori degli indici alternativi

Presentare nei risultati dei SEM almeno due indici:

SRMR: è il più sensibile a segnalare modelli errati

un indice tra: TLI CFI RMSEA

Valori di cut-off per gli indici "migliori"

TLI e CFI maggiore/uguale a .95

SRMR minore di .08

RMSEA minore di .06 (o di .08)

Fattori che influenzano il fit

Ampiezza del Campione: Modelli testati su campioni più grandi hanno un fit peggiore (soprattutto il chi-quadrato).

Numero di variabili nel modello/ Gradi di libertà: Un modello con più variabili generalmente tende ad avere un fit peggiore di un modello con meno variabili/parametri. Più è grande la matrice delle covarianza più è difficile ottenere un buon fit. Modelli con pochi gradi di libertà (a parità di altre condizioni) ottengono più facilmente un buon fit (Marsh, Hau, Balla, & Grayson, 1998).

Fattori che influenzano il fit

Relazioni deboli tra le variabili osservate (grandezza delle covarianze): **covarianze piccole aumentano la probabilità di ottenere un basso chi-quadrato, e quindi un buon fit (Dillon & Goldstein, 1985; Fornell, 1983).**

Dati non-normali (specialmente la curtosi): **aumentano il chi quadrato e le misure assolute. Gli indici incrementali e comparativi sono meno influenzati (Muthén & Kaplan, 1985).**

Attendibilità delle variabili osservate: **variabili più attendibili determinano un migliore fit (Jackson, 2001, 2003).**

CFA CON MPLUS

MODEL RESULTS

MODEL RESULTS

Stima non standardizzata — t = Stima/errore — Two-Tailed

	Estimate	S.E.	Est./S.E.	P-Value
		Errore standard		probabilità
FEARPUN BY				
FP15	1.000	0.000	999.000	999.000
FP26	0.903	0.081	11.117	0.000
FP30	1.086	0.094	11.575	0.000
NEEDREP BY				
NR6	1.000	0.000	999.000	999.000
NR16	1.002	0.126	7.925	0.000
NR20	0.903	0.113	8.026	0.000
NEEDREP WITH				
FEARPUN	0.209	0.046	4.529	0.000
Variances				
FEARPUN	0.932	0.117	7.949	0.000
NEEDREP	0.547	0.096	5.697	0.000

CFA CON MPLUS

Stima non standardizzata

t = Stima/errore

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
		Errore standard		probabilità
Residual Variances				
FP15	1.300	0.099	13.094	0.000
FP26	1.191	0.087	13.677	0.000
FP30	1.327	0.110	12.017	0.000
NR6	1.358	0.097	14.051	0.000
NR16	1.041	0.084	12.364	0.000
NR20	0.907	0.070	12.889	0.000

CFA CON MPLUS

STDYX Standardization

Stima non standardizzata | t = Stima/errore | Errore standard | probabilità

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
FEARPUN BY				
FP15	0.646	0.033	19.482	0.000
FP26	0.624	0.034	18.436	0.000
FP30	0.673	0.033	20.261	0.000
NEEDREP BY				
NR6	0.536	0.042	12.691	0.000
NR16	0.588	0.042	13.841	0.000
NR20	0.574	0.042	13.654	0.000
NEEDREP WITH				
FEARPUN	0.292	0.053	5.466	0.000
Variances				
FEARPUN	1.000	0.000	999.000	999.000
NEEDREP	1.000	0.000	999.000	999.000
Residual Variances				
FP15	0.582	0.043	13.581	0.000
FP26	0.610	0.042	14.429	0.000
FP30	0.547	0.045	12.221	0.000
NR6	0.713	0.045	15.745	0.000
NR16	0.655	0.050	13.120	0.000
NR20	0.670	0.048	13.868	0.000

CFA CON MPLUS

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
FP15	0.418	0.043	9.741	0.000
FP26	0.390	0.042	9.218	0.000
FP30	0.453	0.045	10.130	0.000
NR6	0.287	0.045	6.345	0.000
NR16	0.345	0.050	6.921	0.000
NR20	0.330	0.048	6.827	0.000

Valutazione dei singoli parametri

* I valori dei parametri devono essere ammissibili

* Significatività statistica delle stime dei parametri:

$$t = (\theta_i\text{^} - \theta_0\text{^})/SE$$

dove θ_0^=0, e SE = errore standard del parametro

Valori maggiori o uguali a |1.96| indicano una significatività al livello di probabilità di .05.

* Varianza spiegata (per le variabili dipendenti):

R quadrato

Modifica del modello

Per "migliorare" un modello si possono:

*** Fissare a 0 i valori dei parametri che non sono risultati significativamente differenti da 0 [Il test di Wald in EQS calcola il cambiamento determinato nel chi-quadrato dall'aver fissato a 0 uno o più parametri].**

*** Liberare parametri che erano stati fissati a zero: esistono dei "diagnostici" che permettono di individuare quali parametri modificare [Modification Indexes, Moltiplicatori di Lagrange]. Questi valutano il cambiamento nel χ^2 determinato dalla "liberazione" dei parametri precedentemente fissati a zero.**

Modifica del modello

* Anche l'esame delle covarianze residue standardizzate può evidenziare difficoltà del modello nel ricostruire alcune specifiche covarianze osservate.

* E' fondamentale validare i "nuovi" modelli che risultano da modifiche su un campione diverso. In assenza di campioni diversi si può applicare una strategia di cross-validation considerando un sotto-campione per generare le modifiche e un altro validare e generalizzare tali modifiche.

* La strategia migliore è comunque quella di specificare modelli alternativi prima dell'analisi dei dati stessa. Questo evita problemi di mancata generalizzabilità e di capitalizzazione sulle caratteristiche di uno specifico campione.

Modifica del modello

Spesso gli indici di modificazione suggeriscono di liberare alcune covarianze tra gli errori di misurazione. Secondo Fornell (1983) e Bagozzi (1983) queste possono essere liberate a patto che:

a) Siano sensate da un punto di vista teorico o metodologico

b) Non alterino le stime dei parametri strutturali

c) Non alterino le stime dei parametri del modello di misurazione

Anche se queste tre condizioni sono rispettate, l'inclusione delle covarianze tra gli errori di misurazione indebolisce l'interpretazione del modello, perché il modello di misurazione non è corretto (il numero di fattori non basta a rendere ragione delle covarianze tra le variabili osservate).

Modifica del modello

E' diverso il caso della covarianza tra residui ζ delle variabili latenti (Kline, 2005).

Il residuo di una variabile latente (ζ) riflette l'errore di specificazione: non tutte le "cause" della variabile latente (η) sono state considerate.

Se due ζ covariano questo dipende dal fatto che probabilmente c'è almeno una variabile, che rappresenta una "causa" comune per le due η, che è stata omessa dal modello.

Questo fenomeno è frequentissimo nella ricerca psicologico-sociale.

CFA CON MPLUS

ATTENZIONE !! SONO INDICI UNIVARIATI ! I PARAMETRI FISSI VANNO LIBERATI UNO PER VOLTA !!

Sono da considerare significativi gli indici di modifica che risultano maggiori di 3.84 (per α=.05) o di 6.63 (per α=.01).

		M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.
BY Statements					
NEEDREP	BY FP26	7.744	0.246	0.182	0.130
NEEDREP	BY FP30	4.583	-0.215	-0.159	-0.102
WITH Statements					
FP26	WITH FP15	4.613	-0.401	-0.401	-0.323
FP30	WITH FP15	7.791	0.663	0.663	0.505
NR6	WITH FP15	4.450	0.124	0.124	0.093
NR16	WITH FP15	5.077	-0.120	-0.120	-0.104
NR16	WITH FP26	5.222	0.115	0.115	0.103
NR20	WITH FP30	5.263	-0.118	-0.118	-0.107

ESERCIZIO: REALIZZAZIONE DI UN MODELLO DI CFA

Effettuare un modello di analisi fattoriale confermativa.

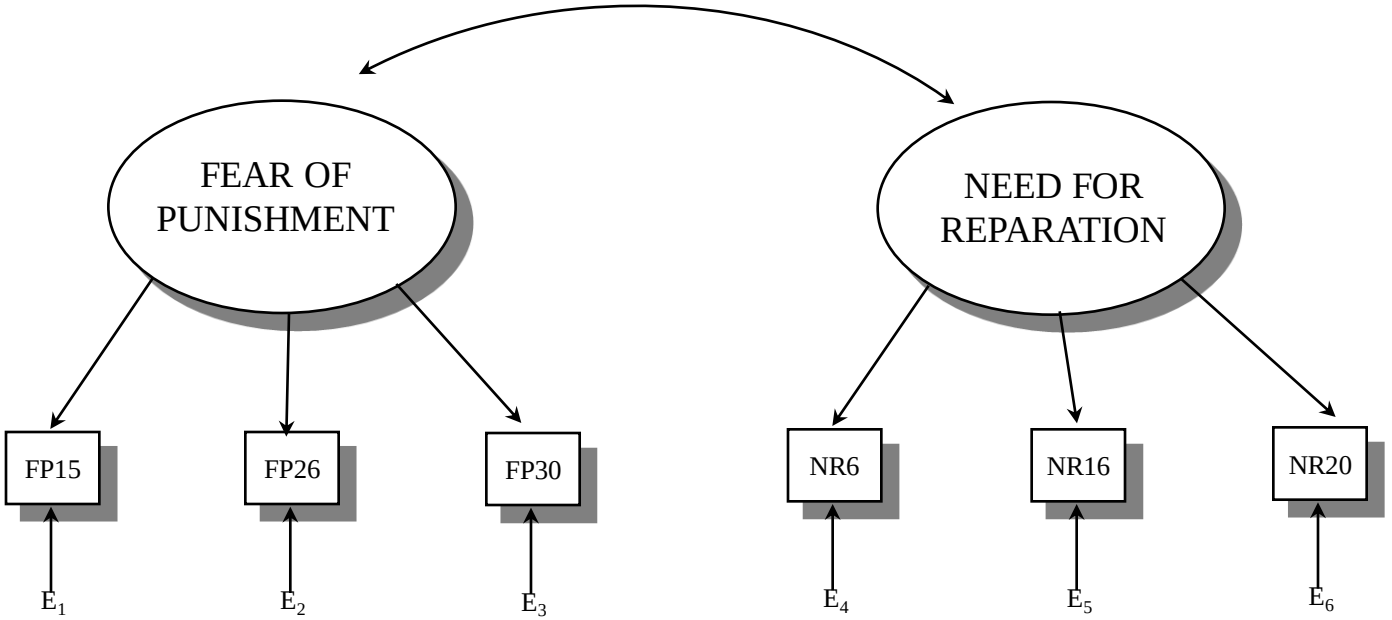
6 item – 2 fattori
Y1-Y3 → F1
Y4-Y6 → F2

es4.dat

CONFRONTARE IL FIT DEL MODELLO IN CUI I FATTORI SONO ORTOGONALI CON IL FIT DEL MODELLO IN CUI I FATTORI SONO CORRELATI.

Modelli *concatenati* (*nested*)

Due modelli sono concatenati quando l'uno é ottenibile dall'altro applicando determinate restrizioni. In questo caso i parametri di uno sono un sottoinsieme dei parametri dell'altro.



Modelli *concatenati* (*nested*)

Se i modelli sono concatenati é possibile esaminare statisticamente quale modello presenta un indice di adattamento migliore.

Infatti la differenza dei loro chi quadrati:

$$\chi^2_{diff} = \chi^2_{p\text{-}m} - \chi^2_{p}$$

si distribuisce ancora come un chi quadrato con gradi di libertà pari a $df_{p\text{-}m} - df_{p}$.

Se la differenza tra i due chi quadrato (χ^2_{diff}) é significativa, le restrizioni apportate peggiorano significativamente il modello.

Condizioni di applicabilità

a. Indipendenza delle osservazioni

Violazione: Le osservazioni non sono indipendenti

Test statistici e stime dei parametri non esatti

Rimedio: modelli multilivello

b. Forma lineare della relazione tra le variabili
c. Distribuzione normale multivariata delle variabili

Non linearità, non normalità

Sono fenomeni spesso associati. Se presenti i metodi di stima classici (ML) non possono essere applicati perché danno risultati inessatti

Rimedio: metodi di stima "robusti" (correzione di Satorra-Bentler, MLM).

Condizioni di applicabilità

d. Livello di misura delle variabili: almeno intervalli

Livello di misura "basso"

I metodi sviluppati per variabili continue possono essere utilizzati quando una variabile ha 4 o più categorie "ordinabili" e la distribuzione è sostanzialmente normale. Se le variabili hanno poche categorie, i metodi per variabili continue (es., ML) danno risultati non corretti.

Rimedio: Se il numero di categorie é inferiore bisogna ricorrere a metodi di stima alternativi (Mplus costruisce una matrice di correlazioni speciali per il livello di misura).

Condizioni di applicabilità

e. La matrice da analizzare é quella delle Covarianze

Si analizza la matrice di correlazione

Test del chi quadrato ed errori standard non esatti.

Rimedio: convertire la matrice in matrice di covarianza; utilizzare programmi che forniscono stime corrette (es., SEPATH); interpretare con cautela i risultati in assenza di correzioni.

Condizioni di applicabilità

f. Il modello testato deve essere specificato a priori rispetto ai dati sui quali viene esaminato

Il modello testato viene specificato in funzione dei dati sui quali viene esaminato

I risultati non sono generalizzabili e sono soggetti a capitalizzazione sul caso.

Rimedio: mettere in atto procedure di cross-validazione dei risultati.

Condizioni di applicabilità

g. I metodi di stima hanno proprietà asintotiche: richiedono campioni molto numerosi

Il campione utilizzato è di numerosità modesta

Le stime sono inaccurate. Il test di fit è poco potente.

Numerosità del campione: come regolarsi ?

Condizioni di applicabilità

Numerosità del campione: come regolarsi ?

Le raccomandazioni relative alla numerosità del campione proposte in letteratura sono varie.

Boomsa (1983) ha concluso che campioni di almeno 100 unità rappresentano un requisito minimo ma campioni di più di 200 unità sono preferibili.

Geweke e Singleton (1980), al contrario, hanno mostrato che possono essere utilizzati campioni anche più piccoli.

Se il campione è inferiore a 100 SS il test di bontà dell'adattamento non segue la distribuzione del chi-2. In generale campioni piccoli tendono a generare risultati instabili.

Condizioni di applicabilità

Numerosità del campione: come regolarsi ?

Tanaka (1987) ha evidenziato che il numero assoluto di soggetti non é un problema rilevante, ed ha focalizzato l'attenzione su fattori come la grandezza del modello (e quindi il numero di parametri da stimare), il numero di variabili osservate e quello di variabili latenti.

Bentler e Chou (1987) suggeriscono un rapporto minimo di 5 soggetti per ogni parametro libero, per metodi di stima basati sulla distribuzione normale multivariata, e un rapporto minimo di 10 soggetti per ogni parametro libero, per metodi di stima "distribution free".

Condizioni di applicabilità

Numerosità del campione: come regolarsi ?

Kline (2005) indica che da 5 a 10 o 20 soggetti per ogni parametro stimato dovrebbe rappresentare un campione sufficiente.

A parità di altre condizioni, un numero maggiore di osservazioni determina una potenza statistica maggiore.

I ricercatori incoraggiano ad utilizzare campioni grandi quando si devono esaminare modelli più complessi (Kim, 2000; McCallum et al., 2006).

Condizioni di applicabilità

Numerosità del campione: come regolarsi ?

Se non si anticipano problemi nei dati (es., valori mancanti, distribuzioni non normali) si raccomanda un campione minimo di 200 SS per ogni SEM.

Tuttavia, se le variabili sono molto attendibili, e molto correlate, gli effetti sono forti e il modello non è troppo complesso, campioni più piccoli possono risultare sufficienti (Bearden, Sharma & Teel 1982; Bollen, 1990).

E' stato dimostrato che in questi casi i modelli SEM possono funzionare bene anche con campioni molto piccoli (50 – 100 SS).

I MODELLI DI EQUAZIONI STRUTTURALI:

APPLICAZIONI CON MPLUS

- Analisi Confermativa Vincolata
- Indici di attendibilità
- Modelli con variabili osservate
- SEM completi ("full SEM")
- L'analisi della mediazione statistica

ANALISI FATTORIALE CONFERMATIVA VINCOLATA

È possibile formulare diverse ipotesi relativamente ai valori assunti dai termini λ_i e θ_i.

Se si ipotizza che tutti i λ_i abbiano lo stesso valore (cioè, $\lambda_1=\lambda_2=\lambda_3$) ma che i θ_i possano avere valori differenti, il modello fattoriale specificato viene definito "tau equivalente": si ipotizza cioè che ogni indicatore del fattore η abbia soltanto la stessa varianza comune mentre la varianza unica può essere differente per i diversi indicatori.

ANALISI FATTORIALE CONFERMATIVA VINCOLATA

Se si ipotizza che tutti i λ_i abbiano lo stesso valore (cioè, $\lambda_1=\lambda_2=\lambda_3$) e che anche tutti i θ_i abbiano lo stesso valore (cioè, $\theta_1=\theta_2=\theta_3$) il modello fattoriale specificato viene definito modello delle "forme parallele": si ipotizza cioè che ogni indicatore del fattore η abbia la stessa varianza comune e la stessa varianza unica.

Il modello fattoriale in cui sia i λ_i che i θ_i possano avere valori differenti, il modello fattoriale specificato viene definito "congenerico": non viene fatta nessuna assunzione relativamente al valore della varianza comune e della varianza unica, mentre si ipotizza solamente che ogni indicatore del fattore η sia saturo in η.

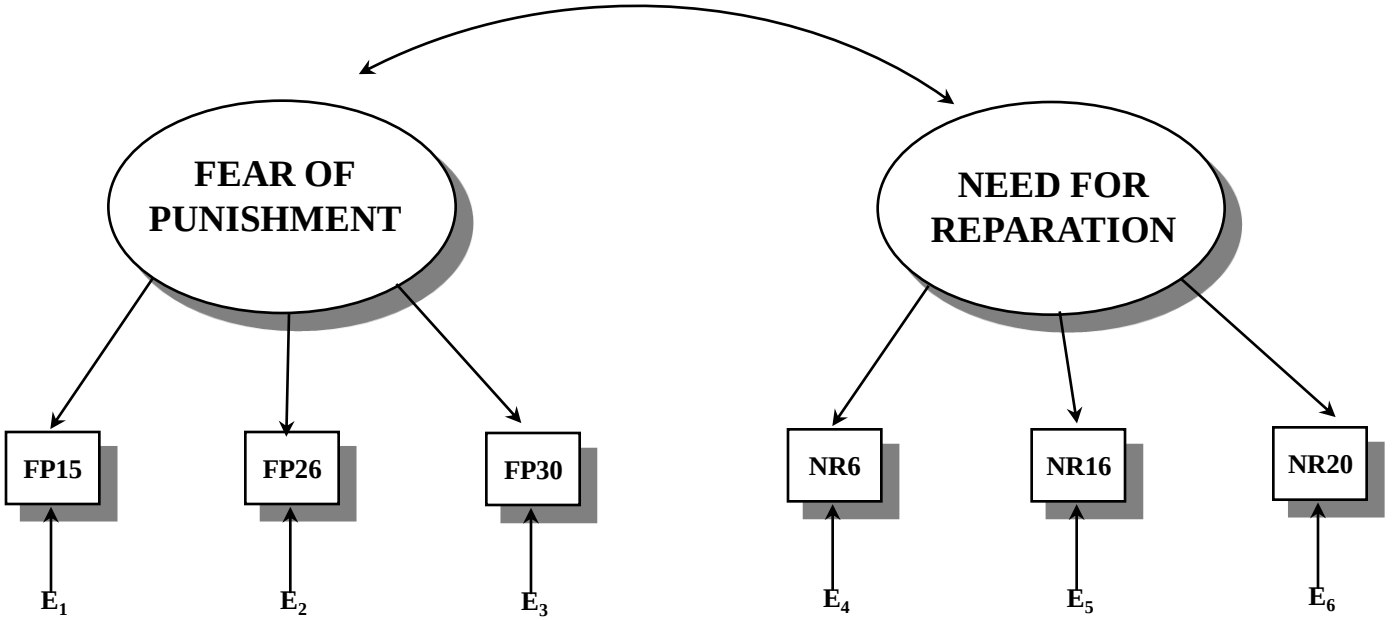
ANALISI FATTORIALE CONFERMATIVA VINCOLATA

Questi tre modelli fanno riferimento a tre specifici modelli identificati nella letteratura psicometrica.

È possibile sottoporre a verifica empirica ognuna delle tre differenti ipotesi, e vedere quale sia la più consistente rispetto ai dati empirici.

Programmi come EQS, LISREL, MPLUS consentono di specificare i diversi vincoli necessari per operazionalizzare i diversi modelli, e di ricavare gli indici di bontà dell'adattamento che consentono di optare per un modello piuttosto che per un altro.

ANALISI FATTORIALE CONFERMATIVA VINCOLATA



Modello "tau equivalente"

Gli indicatori di ciascun fattore comune hanno la stessa saturazione cioè la *stessa varianza comune.*

VINCOLI SUI PARAMETRI IN MPLUS

Come imporre le restrizioni nel modello tau equivalente:

FEARPUN BY FP15* FP26 FP30 (1);
NEEDREP BY NR6*.5 NR16 NR20 (2);

Il numero (1) sta a indicare che le stime delle 3 saturazioni delle variabili FP15 FP26 FP30 vengono vincolate ad essere uguali.

Il numero (2) sta a indicare che le stime delle 3 saturazioni delle variabili NR6 NR16 NR20 vengono vincolate ad essere uguali.

Vengono specificati due numeri diversi (1) e (2) in modo che le stime siano vincolate all'interno dello stesso fattore ma non attraverso i due fattori.

MODELLO "TAU EQUIVALENTE"

CFA_TAU

```
TITLE:  CFA
 DATA:
 FILE IS dati_efa_cfa_grezzi.dat;

 VARIABLE:
 NAMES ARE
    FP10   FP15   FP26    FP30
    NR6    NR16   NR17    NR20  GENDER;
 USEV ARE
    FP15   FP26    FP30   NR6    NR16    NR20  ;
  MISSING ARE ALL (9);

 MODEL:
   FEARPUN BY  FP15*   FP26    FP30   (1);
   NEEDREP BY  NR6*.5   NR16   NR20   (2);
   FEARPUN @1;
   NEEDREP @1;
 OUTPUT:  STANDARDIZED SAMPSTAT MODINDICES(3.84) TECH1;
```

MODELLO "TAU EQUIVALENTE"

```
Chi-Square Test of Model Fit

          Value                             26.739
          Degrees of Freedom                    12
          P-Value                           0.0084

RMSEA (Root Mean Square Error Of Approximation)
          Estimate                           0.039
          90 Percent C.I.                    0.019  0.059
          Probability RMSEA <= .05           0.811

CFI/TLI
          CFI                                0.977
          TLI                                0.972

SRMR (Standardized Root Mean Square Residual)
          Value                              0.031
```

MODELLO "TAU EQUIVALENTE"

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
FEARPUN BY				
FP15	0.954	0.036	26.846	0.000
FP26	0.954	0.036	26.846	0.000
FP30	0.954	0.036	26.846	0.000
NEEDREP BY				
NR6	0.713	0.031	22.726	0.000
NR16	0.713	0.031	22.726	0.000
NR20	0.713	0.031	22.726	0.000
NEEDREP WITH				
FEARPUN	0.301	0.053	5.710	0.000

NB. I factor loadings sono uguali all'interno dello stesso fattore solo nella soluzione non standardizzata (Estimate) e in quella standardizzata (Std), ma non in quella completamente standardizzata (StdYX), perché le varianze di errore non sono vincolate.

MODELLO "TAU EQUIVALENTE"

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
FEARPUN BY				
FP15	0.639	0.020	32.440	0.000
FP26	0.672	0.019	34.548	0.000
FP30	0.622	0.020	31.374	0.000
NEEDREP BY				
NR6	0.518	0.021	24.630	0.000
NR16	0.568	0.022	25.396	0.000
NR20	0.608	0.023	26.377	0.000
NEEDREP WITH				
FEARPUN	0.301	0.053	5.710	0.000

MODELLO "TAU EQUIVALENTE"

Residual Variances

FP15	0.592	0.025	23.526	0.000
FP26	0.548	0.026	20.944	0.000
FP30	0.613	0.025	24.799	0.000
NR6	0.731	0.022	33.536	0.000
NR16	0.678	0.025	26.689	0.000
NR20	0.630	0.028	22.434	0.000

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
FP15	0.408	0.025	16.220	0.000
FP26	0.452	0.026	17.274	0.000
FP30	0.387	0.025	15.687	0.000
NR6	0.269	0.022	12.315	0.000
NR16	0.322	0.025	12.698	0.000
NR20	0.370	0.028	13.188	0.000

MODELLO "TAU EQUIVALENTE"

			M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.

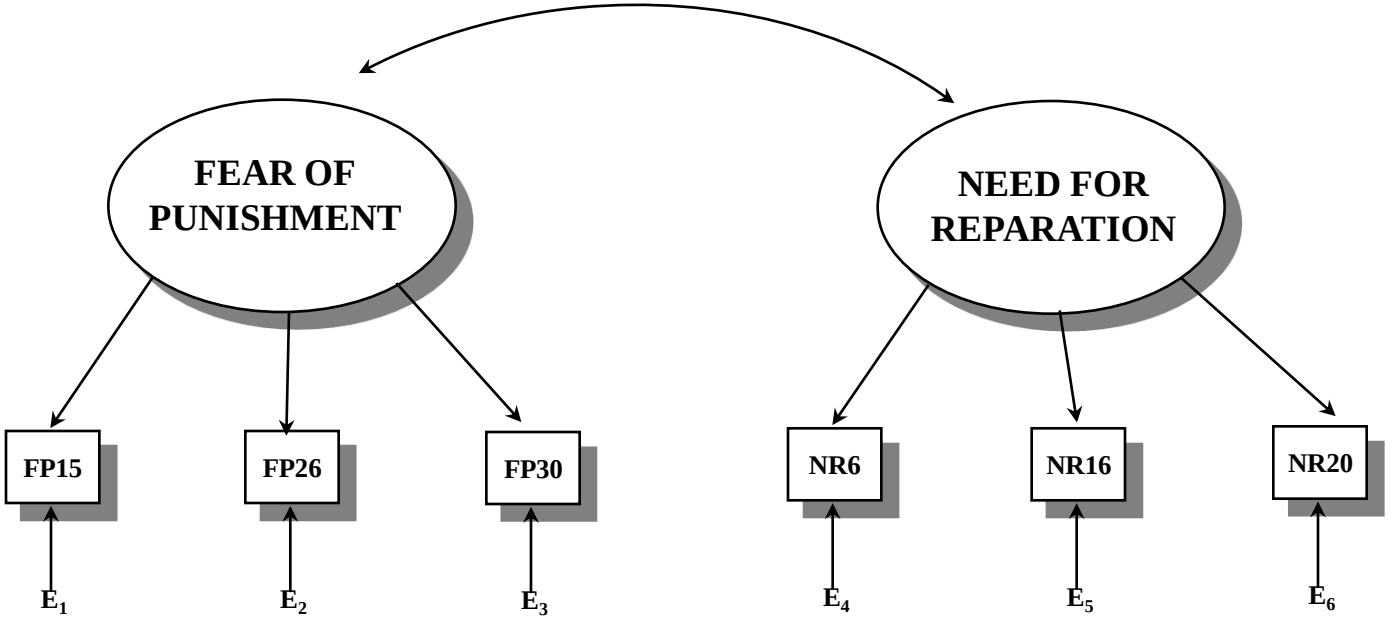
Cambiamento nel chi-quadrato determinato dalla stima di parametri fissati a zero

WITH Statements

FP26	WITH	FP15	5.583	-0.168	-0.168	-0.139
FP30	WITH	FP15	7.262	0.194	0.194	0.141
NR6	WITH	FP15	4.669	0.127	0.127	0.094
NR16	WITH	FP15	4.939	-0.118	-0.118	-0.099
NR16	WITH	FP26	4.822	0.110	0.110	0.101
NR20	WITH	FP30	5.219	-0.116	-0.116	-0.104

Nessuno dei MI relativi al cambiamento nel chi-quadrato determinato dal rilasciamento dei vincoli risulta significativo

ANALISI FATTORIALE CONFERMATIVA VINCOLATA



Modello "forme parallele"

Gli indicatori di ciascun fattore comune hanno la stessa saturazione cioè la *stessa varianza comune* e la stessa varianza residua (errore di misurazione).
Il risultato di questo doppio vincolo è che gli indicatori hanno la stessa attendibilità.

VINCOLI SUI PARAMETRI IN MPLUS

Come imporre le restrizioni nel modello forme parallele:

```
FEARPUN BY   FP15*   FP26    FP30   (1);
NEEDREP BY  NR6*.5    NR16   NR20   (2);
FP15   FP26    FP30  (3);
NR6    NR16    NR20  (4);
```

I vincoli indicati con i numeri (1) e (2) sono quelli già imposti nel modello tau equivalente.

Il numero (3) sta a indicare che le stime delle 3 varianze residue delle variabili FP15 FP26 FP30 vengono vincolate ad essere uguali.

Il numero (4) sta a indicare che le stime delle 3 varianze residue delle variabili NR6 NR16 NR20 vengono vincolate ad essere uguali.

MODELLO "FORME PARALLELE"

CFA_PARALL

```
TITLE:  CFA
DATA:
FILE IS dati_efa_cfa_grezzi.dat;

VARIABLE:
NAMES ARE
    FP10   FP15   FP26    FP30
    NR6    NR16   NR17    NR20  GENDER;
USEV ARE
    FP15   FP26    FP30   NR6    NR16    NR20   ;
 MISSING ARE ALL (9);

MODEL:
  FEARPUN BY  FP15*   FP26    FP30   (1);
  NEEDREP BY  NR6*.5   NR16   NR20   (2);
  FEARPUN @1;
  NEEDREP @1;
FP15   FP26    FP30  (3);
NR6    NR16    NR20  (4);

OUTPUT:  STANDARDIZED SAMPSTAT MODINDICES(3.84) TECH1;
```

MODELLO "FORME PARALLELE"

```
Chi-Square Test of Model Fit

          Value                             62.331
          Degrees of Freedom                16
          P-Value                           0.0000

RMSEA (Root Mean Square Error Of Approximation)
          Estimate                          0.059
          90 Percent C.I.                   0.044  0.075
          Probability RMSEA <= .05          0.146

CFI/TLI
          CFI                               0.929
          TLI                               0.933

SRMR (Standardized Root Mean Square Residual)
          Value                             0.114
```

MODELLO "FORME PARALLELE"

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
FEARPUN BY				
FP15	0.961	0.036	26.921	0.000
FP26	0.961	0.036	26.921	0.000
FP30	0.961	0.036	26.921	0.000
NEEDREP BY				
NR6	0.714	0.032	22.469	0.000
NR16	0.714	0.032	22.469	0.000
NR20	0.714	0.032	22.469	0.000
NEEDREP WITH				
FEARPUN	0.304	0.053	5.744	0.000
Residual Variances				
FP15	1.281	0.045	28.486	0.000
FP26	1.281	0.045	28.486	0.000
FP30	1.281	0.045	28.486	0.000
NR6	1.105	0.039	28.559	0.000
NR16	1.105	0.039	28.559	0.000
NR20	1.105	0.039	28.559	0.000

MODELLO "FORME PARALLELE"

I factor loadings sono uguali all'interno dello stesso fattore anche nella soluzione completamente standardizzata, perché le varianze di errore sono vincolate.

STDYX Standardization	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
FEARPUN BY				
FP15	0.647	0.017	38.715	0.000
FP26	0.647	0.017	38.715	0.000
FP30	0.647	0.017	38.715	0.000
NEEDREP BY				
NR6	0.562	0.020	27.954	0.000
NR16	0.562	0.020	27.954	0.000
NR20	0.562	0.020	27.954	0.000
NEEDREP WITH				
FEARPUN	0.304	0.053	5.744	0.000
Residual Variances				
FP15	0.581	0.022	26.867	0.000
FP26	0.581	0.022	26.867	0.000
FP30	0.581	0.022	26.867	0.000
NR6	0.684	0.023	30.277	0.000
NR16	0.684	0.023	30.277	0.000
NR20	0.684	0.023	30.277	0.000

MODELLO "FORME PARALLELE"

MODEL MODIFICATION INDICES

Minimum M.I. value for printing the modification index 3.840

	M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.
BY Statements				
FEARPUN BY FP26	9.113	-0.118	-0.118	-0.080
FEARPUN BY FP30	7.249	0.105	0.105	0.071
FEARPUN BY NR6	4.919	0.121	0.121	0.095
NEEDREP BY NR6	12.841	0.140	0.140	0.110
NEEDREP BY NR20	12.077	-0.136	-0.136	-0.107
Variances/Residual Variances				
FP26	6.134	-0.173	-0.173	-0.079
FP30	4.710	0.152	0.152	0.069
NR6	23.478	0.274	0.274	0.170
NR20	17.490	-0.237	-0.237	-0.147

Cambiamento nel chi-quadrato determinato dal rilasciamento dei vincoli

MODELLO "FORME PARALLELE"

MODEL MODIFICATION INDICES

Minimum M.I. value for printing the modification index 3.840

		M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.
WITH Statements					
NR6	WITH FP15	4.182	0.109	0.109	0.092
NR16	WITH FP15	5.133	-0.121	-0.121	-0.101
NR16	WITH FP26	5.388	0.124	0.124	0.104
NR20	WITH FP30	4.303	-0.111	-0.111	-0.093
NR20	WITH NR16	4.207	0.108	0.108	0.098

Cambiamento nel chi-quadrato determinato dal rilasciamento dei parametri fissati a zero

CONFRONTO TRA I CHI² DEI MODELLI

MODELLO CONGENERICO: $\chi^2(8) = 21.62, p<.01$

MODELLO TAU-EQUIVALENTE: $\chi^2(12) = 26.74, p<.01$

MODELLO FORME PARALLELE: $\chi^2(16) = 62.33, p<.001$

TAU-EQUIVALENTE – CONGENERICO:
$\chi^2_{diff}(4)=5.06, p=.28$

FORME PARALLELE - TAU-EQUIVALENTE:
$\chi^2_{diff}(4)=35.59, p<.001$

CHI_2_DIFF.xls

ESERCIZIO 4: REALIZZAZIONE DI UN MODELLO DI UN MODELLO CFA TAU-EQUIVALENTE E DELLE FORME PARALLELE

Realizzare il modello CFA con i dati del file es4.dat

Verificare sugli stessi dati i modelli congenerico, tau-equivalente e delle forme parallele, applicando le opportune restrizioni.

Confrontare poi i modelli congenerico, tau equivalente e delle forme parallele, sulla base dei loro chi^2.

ANALISI FATTORIALE CONFERMATIVA: INDICI DI ATTENDIBILITA' (Fornell e Larcker, 1981)

Sia x_i un indicatore almeno congenerico di una variabile latente ξ, ovvero $x_i = \lambda_i\xi + \delta_i$

Sia θ_i la varianza di errore, ovvero $E(\delta_i\delta_i)=\theta_i$

Si supponga che i termini ε_i dei diversi indicatori almeno congenerici di ξ abbiano media zero e non siano correlati tra di loro né con la variabile latente.

Per comodità si assuma che la varianza della variabile latente sia uguale a 1.

Allora si possono definire i seguenti indici:

ANALISI FATTORIALE CONFERMATIVA: INDICI DI ATTENDIBILITA' (Fornell e Larcker, 1981)

a) Attendibilità del singolo indicatore: è data dal rapporto tra la varianza spiegata dal fattore comune e la varianza totale della variabile:

$$\rho_i = \frac{\lambda_i^2}{\lambda_i^2 + \theta_i}$$

Questo indice varia da 0 a 1: vanno considerati come adeguati valori maggiori/uguali a .3.

ANALISI FATTORIALE CONFERMATIVA: INDICI DI ATTENDIBILITA' (Fornell e Larcker, 1981)

a) Attendibilità del singolo indicatore: se il singolo indicatore è influenzato da più di un fattore la formula è un po' più complessa. Ad esempio nel caso di due fattori esse diventa:

$$\rho_i = \frac{\lambda_{i1}^2 + \lambda_{i2}^2 + 2\lambda_{i1}\lambda_{i2}\phi_{21}}{\lambda_{i1}^2 + \lambda_{i2}^2 + 2\lambda_{i1}\lambda_{i2}\phi_{21} + \theta_i}$$

Si interpreta sempre come rapporto tra la varianza spiegata dai fattori comuni e la varianza totale della variabile

ANALISI FATTORIALE CONFERMATIVA: INDICI DI ATTENDIBILITA'

b) Attendibilità composita del fattore (simile al coefficiente omega): è data dal rapporto tra il quadrato della somma delle saturazioni nel fattore comune e questo stesso numero più la proporzione di varianza totale attribuibile alla unicità delle variabili:

$$\rho_c = \frac{(\sum_i \lambda_i)^2}{(\sum_i \lambda_i)^2 + \sum_i \theta_i}$$

Questo indice varia da 0 a 1: vanno considerati come adeguati valori maggiori/uguali a.6.

ANALISI FATTORIALE CONFERMATIVA: INDICI DI ATTENDIBILITA'

c) Varianza media spiegata dal fattore (AVE, Average Extracted Variance): è data dal rapporto tra la somma della varianza di ogni variabile spiegata dal fattore comune e la varianza totale delle variabili, cioè:

$$\rho_v = \frac{(\sum_i \lambda_i^2)}{(\sum_i \lambda_i^2) + \sum_i \theta_i}$$

Questo indice varia da 0 a 1. Anche se alcuni ricercatori indicano un valore maggiore/uguale a .50 come ideale, altri considerano il valore di tale indice come "accessorio" rispetto agli indici di fit (globali e dei singoli parametri).

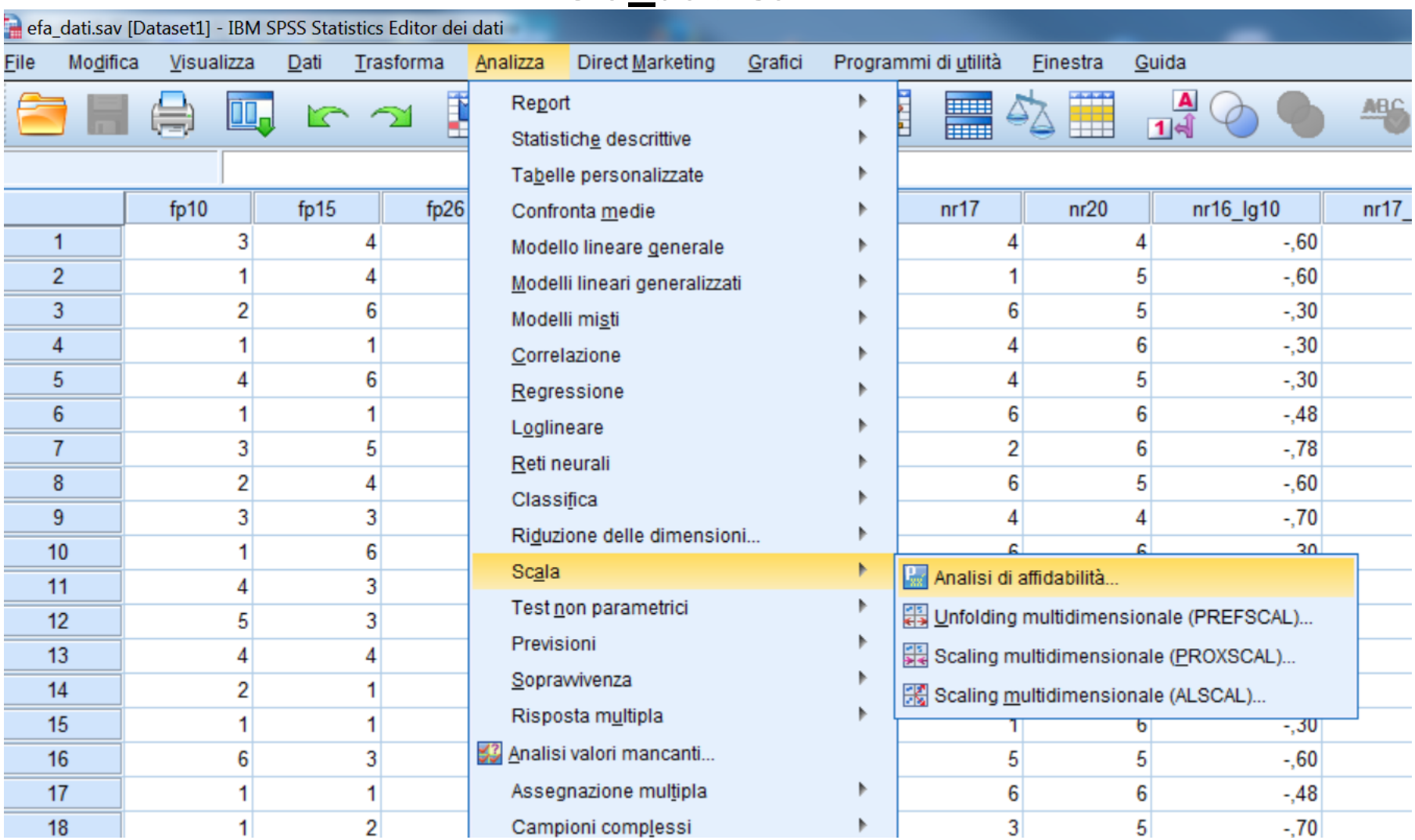
ANALISI FATTORIALE CONFERMATIVA: INDICI DI ATTENDIBILITA'

	Attendibilità Indicatore	Attendibilità Composita	Varianza Spiegata Media (AVE)
FP15	0,418	0,685	0,42
FP26	0,391		
FP30	0,453		
NR6	0,287	0,586	0,32
NR16	0,345		
NR20	0,330		

Reliability.xls

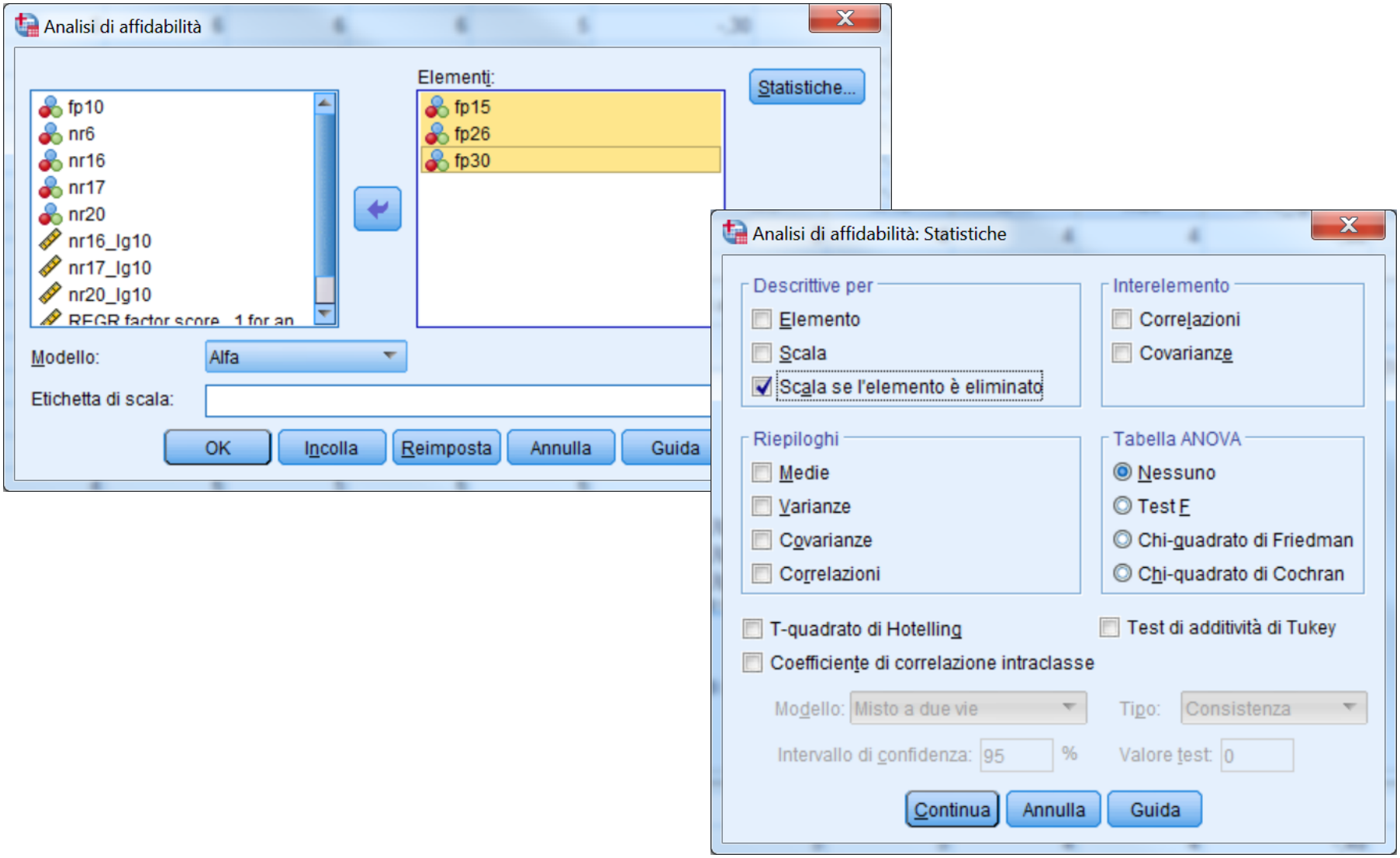
ANALISI DI ATTENDIBILITA' CON SPSS

efa_dati.sav



	fp10	fp15	fp26	nr17	nr20	nr16_lg10	nr17_
1	3	4		4	4	-,60	
2	1	4		1	5	-,60	
3	2	6		6	5	-,30	
4	1	1		4	6	-,30	
5	4	6		4	5	-,30	
6	1	1		6	6	-,48	
7	3	5		2	6	-,78	
8	2	4		6	5	-,60	
9	3	3		4	4	-,70	
10	1	6					
11	4	3					
12	5	3					
13	4	4					
14	2	1					
15	1	1		1	6	-,30	
16	6	3		5	5	-,60	
17	1	1		6	6	-,48	
18	1	2		3	5	-,70	

ANALISI DI ATTENDIBILITA' CON SPSS



ANALISI DI ATTENDIBILITA' CON SPSS

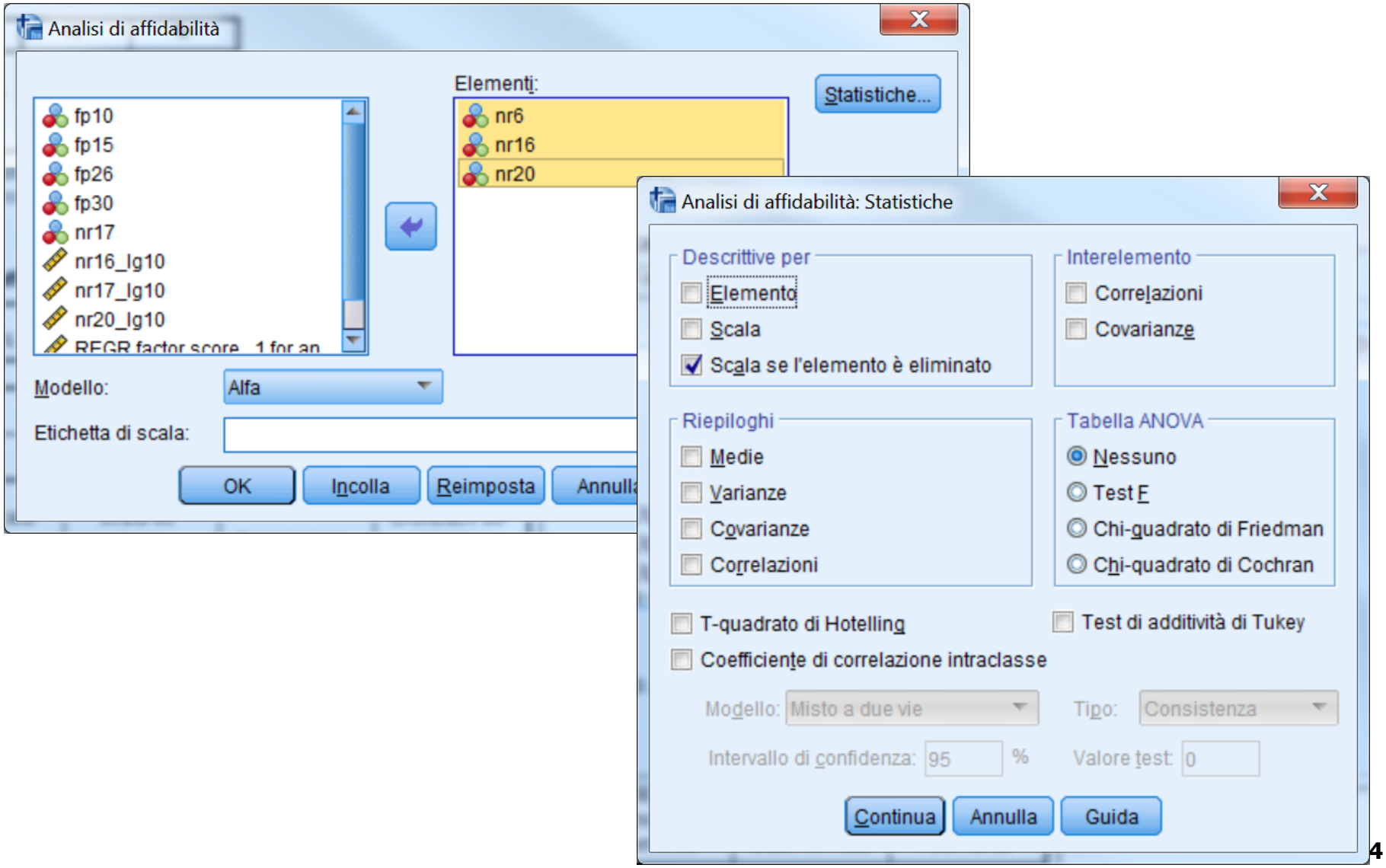
Statistiche di affidabilità

Alpha di Cronbach	N. di elementi
,682	3

Statistiche elemento-totale

	Media scala se viene eliminato l'elemento	Varianza scala se viene eliminato l'elemento	Correlazione elemento-totale corretta	Alpha di Cronbach se viene eliminato l'elemento
fp15	6,83	6,163	,497	,587
fp26	6,32	6,697	,476	,615
fp30	6,85	5,789	,518	,560

ANALISI DI ATTENDIBILITA' CON SPSS



ANALISI DI ATTENDIBILITA' CON SPSS

Statistiche di affidabilità

Alpha di Cronbach	N. di elementi
,579	3

Statistiche elemento-totale

	Media scala se viene eliminato l'elemento	Varianza scala se viene eliminato l'elemento	Correlazione elemento-totale corretta	Alpha di Cronbach se viene eliminato l'elemento
nr6	9,71	3,953	,371	,511
nr16	9,41	4,213	,403	,455
nr20	9,05	4,565	,396	,471

ESERCIZIO 5: INDICI DI ATTENDIBILITA' RICAVATI DALLA SOLUZIONE CFA

Utilizzando i risultati del modello realizzato nell'esercizio 4:

Calcolare gli indici a, b e c per la soluzione congenerica.

Effettuare l'analisi di attendibilità con SPSS e confrontare la soluzione con gli indici ricavati tramite Mplus.

MODELLI CON SOLE VARIABILI OSSERVATE

Il modello matematico si semplifica in:

$$\eta = \alpha + B\eta + \zeta$$

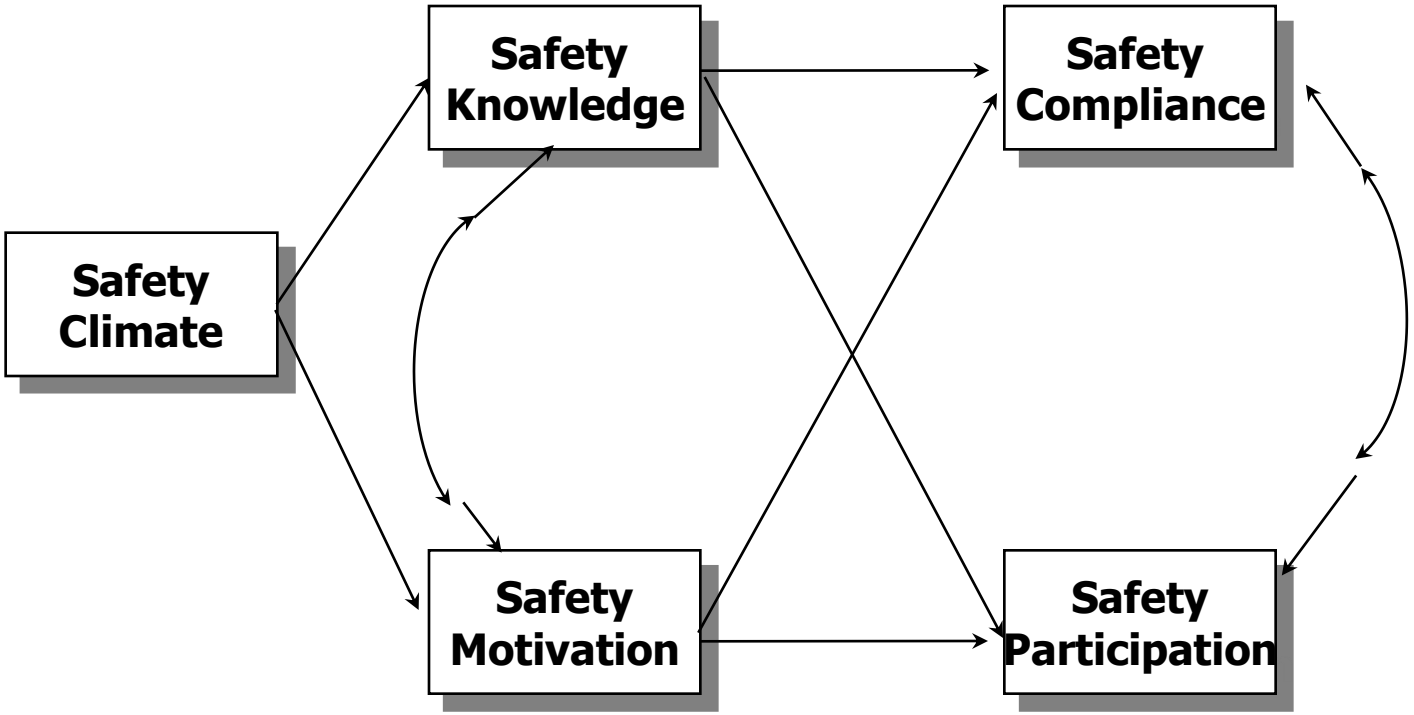
η è il vettore delle variabili dipendenti. Ogni η è misurata da una sola y la cui saturazione λ è fissata a 1 e la cui varianza d'errore θ è fissata a 0: $y=1\eta + \varepsilon$; $\theta=0$

ζ sono gli errori di specificazione dell'equazione le cui varianze sono nella matrice Ψ

α sono le intercette delle y (solitamente uguali a 0)

x sono le variabili osservate indipendenti

MODELLI CON VARIABILI OSSERVATE



MODELLI CON VARIABILI OSSERVATE

TITLE: neal 2011

neal_2011_item_path

DATA:
FILE IS ITALY_2012.DAT;
TYPE IS INDIVIDUAL;

VARIABLE:
NAMES ARE .....

usev are
saf_cli saf_knw ind_mot SAF_COMP SAF_PART ;

define:
saf_cli = SAFCLI1 +SAFCLI2 +SAFCLI3 +SAFCLI4+
SAFCLI5 +SAFCLI6 +SAFCLI7 +SAFCLI8 +SAFCLI9+
SAFCLI10 +SAFCLI11 +SAFCLI12 +SAFCLI13+
SAFCLI14 +SAFCLI15 +SAFCLI16 ;
saf_knw =KN_21 +KN_22 +KN_23 +KN_24 ;
ind_mot =MOT_25 +MOT_26 +MOT_27 +MOT_28;
SAF_COMP =COMP_29+ COMP_30 +COMP_31 +COMP_32;
SAF_PART =PART_33 +PART_34 +PART_35 +PART_36 ;

MODELLI CON VARIABILI OSSERVATE

```
 TITLE:  neal 2011

.....

analysis:
estimator=MLMV;

MODEL:

saf_knw on saf_cli;
ind_mot on saf_cli;
SAF_COMP ON saf_knw ind_mot;
SAF_PART ON saf_knw ind_mot;

saf_knw WITH ind_mot;
SAF_COMP WITH SAF_PART;

OUTPUT:  STANDARDIZED SAMPSTAT   MODINDICES(3.84);
```

MODELLI CON SOLE VARIABILI OSSERVATE E LIVELLI DI AGGREGAZIONE

Se un costrutto è misurato con una scala composta da più item, si definiscono diversi livelli di aggregazione degli indicatori della variabile latente.

Disaggregazione Totale. **Ogni item è usato come separato indicatore del costrutto. Rappresenta il livello di analisi più dettagliato per l'analisi dei SEM (Bagozzi & Heatherington 1994). Si possono esaminare le proprietà psicometriche di ciascun singolo item. Se il numero di item è molto elevato può risultare una eccessiva potenza della verifica, per cui l'ipotesi nulla viene rifiutata quando invece non dovrebbe esserlo (Bagozzi & Heatherington 1994).**

MODELLI CON SOLE VARIABILI OSSERVATE E LIVELLI DI AGGREGAZIONE

Disaggregazione parziale. **Per ogni variabile latente vengono costruiti diversi compositi di item aggregati. Questi compositi vengono chiamati "parcels": i parcels rappresentano il risultato della somma/aggregazione di più item (Bentler &Wu 1995; Dabholkar, Thorpe et al. 1996).**
Il costrutto viene modellato come una variabile latente misurata dai parcels e non dagli item.
I parcels possono essere individuati in modo indipendente dal loro contenuto, e quindi non vanno interpretati direttamente ma solo come indicatori del costrutto. Viceversa i parcels possono essere creati individuando delle "sottodimensioni" (o "facets") del costrutto più globale.

MODELLI CON SOLE VARIABILI OSSERVATE E LIVELLI DI AGGREGAZIONE

La disaggregazione parziale può risultare particolarmente utile per la definizione di modelli complessi, poiché:

- **riduce l'errore di misurazione**
- **consente di ottenere stime più stabili riducendo il numero di parametri da stimare**
- **migliora l'approssimazione alla distribuzione normale**

(Bagozzi & Heatherington 1994; Baumgartner &Homberg 1996; Dabholkar, Thorpe et al. 1996).

MODELLI CON SOLE VARIABILI OSSERVATE E LIVELLI DI AGGREGAZIONE

Aggregazione Totale. **Utilizza singole misure per ciascun costrutto, ognuna risultato della aggregazione/somma degli item che compongono la scala di misura del costrutto (Bagozzi & Heatherington 1994).**

Il principale vantaggio di questo approccio è nella semplicità e nella capacità di catturare gli elementi essenziali del modello concettuale (Bagozzi & Heatherington 1994; Baumgartner & Homberg 1996).

Il principale svantaggio è nel fatto che non viene modellato l'errore di misurazione delle variabili.

MODELLI CON SOLE VARIABILI OSSERVATE E LIVELLI DI AGGREGAZIONE

Esistono due principali tipi di modelli di aggregazione totale:

a) Modelli con sole variabili osservate, in cui l'errore di misurazione delle variabili osservate (somma di item) viene fissato a zero.

b) Modelli in cui l'errore di misurazione delle variabili osservate (somma di item) viene fissato ad un valore maggiore di 0. Si tratta dei modelli con "single indicator latent variable" , o con "reliability correction".

MODELLI CON SOLE VARIABILI OSSERVATE E LIVELLI DI AGGREGAZIONE

Reliability correction

In questi modelli si utilizza una stima dell'attendibilità della scala composta dai diversi item che vengono aggregati per misurare il costrutto. Ad esempio si può considerare il valore dell'alfa di Cronbach come stima dell'attendibilità della scala. Quindi si utilizza il valore (1-alfa) come stima della inattendibilità della scala. Questo valore viene moltiplicato per la varianza della scala S^2 in modo tale da ottenere una stima della varianza unica della scala. Nel modello il valore della varianza residua della scala viene dunque fissato a $(1\text{-alfa})*S^2$.

MODELLI CON SOLE VARIABILI OSSERVATE E LIVELLI DI AGGREGAZIONE

Reliability correction

Questa tecnica non stima la varianza unica della variabile come parte del modello, ma non assume neanche che essa sia uguale a zero (come nella "path analysis"). Pertanto, con questa tecnica è possibile ottenere delle stime dei parametri (beta, ecc.) del modello più corretta (Stephenson & Holbert, 2003). Rispetto ai modelli della disaggregazione (totale o parziale) tuttavia il modello della reliablity correction è meno preciso perché la stima della varianza unica sulla quale si basa contiene sia la varianza dovuta all'errore di misurazione sia la varianza specifica della variabile.

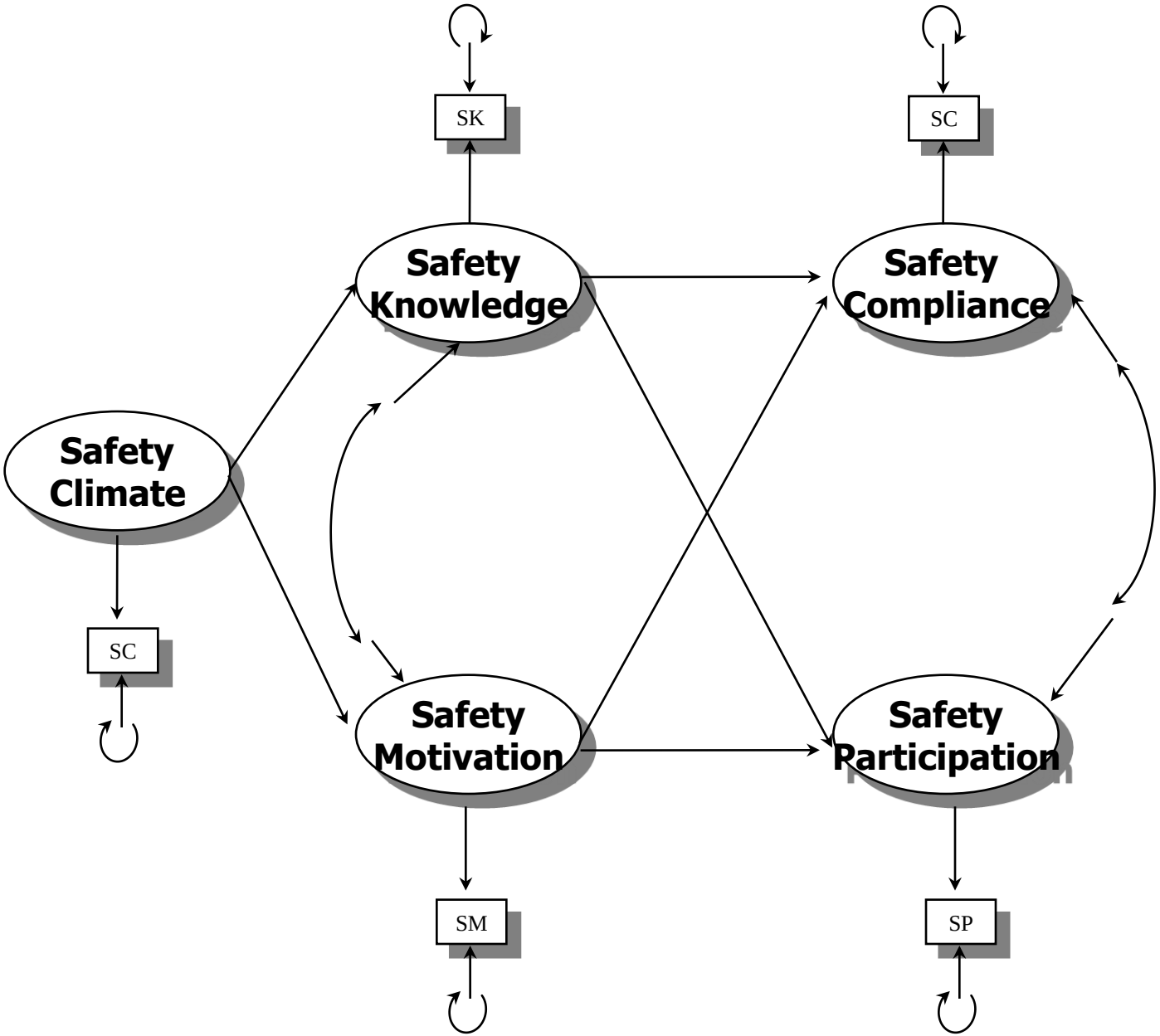
MODELLI CON SOLE VARIABILI OSSERVATE E LIVELLI DI AGGREGAZIONE

Ovunque possibile è bene specificare modelli con variabili latenti (disaggregazione totale o parziale) piuttosto che modelli con variabili osservate (aggregazione totale), poiché in questi ultimi si assume che le variabili siano misurate senza errore o si fissa l'errore ad un valore derivante da una stima (Kline, 1998). Questa assunzione non viene fatta quando invece si usano variabili latenti.

MODELLI CON SOLE VARIABILI OSSERVATE E LIVELLI DI AGGREGAZIONE

Nei modelli con variabili latenti la varianza unica viene stimata per ogni indicatore di ogni variabile latenti, e quindi la varianza di errore viene stimata come parte del modello. Utilizzare un modello che assume che le variabili siano prive di errore quando in realtà non lo sono può compromettere le stime dei parametri del modello. Nella maggior parte delle situazioni questa compromissione porta a una sottostima delle relazioni tra le variabili. L'uso di variabili latenti in cui l'errore di misura viene esplicitamente modellato porta invece a stime corrette di tali parametri.

MODELLO "RELIABILITY CORRECTION"



MODELLO "RELIABILITY CORRECTION"

Stime per la correzione della varianza residua

	alfa	s^2	(1-alfa)s^2
CLIMATE	0,97	458,174	13,745
KNOWL	0,909	23,124	2,104
MOTIV	0,925	21,440	1,608
COMPLI	0,921	25,153	1,987
PARTICI2	0,836	24,893	4,082

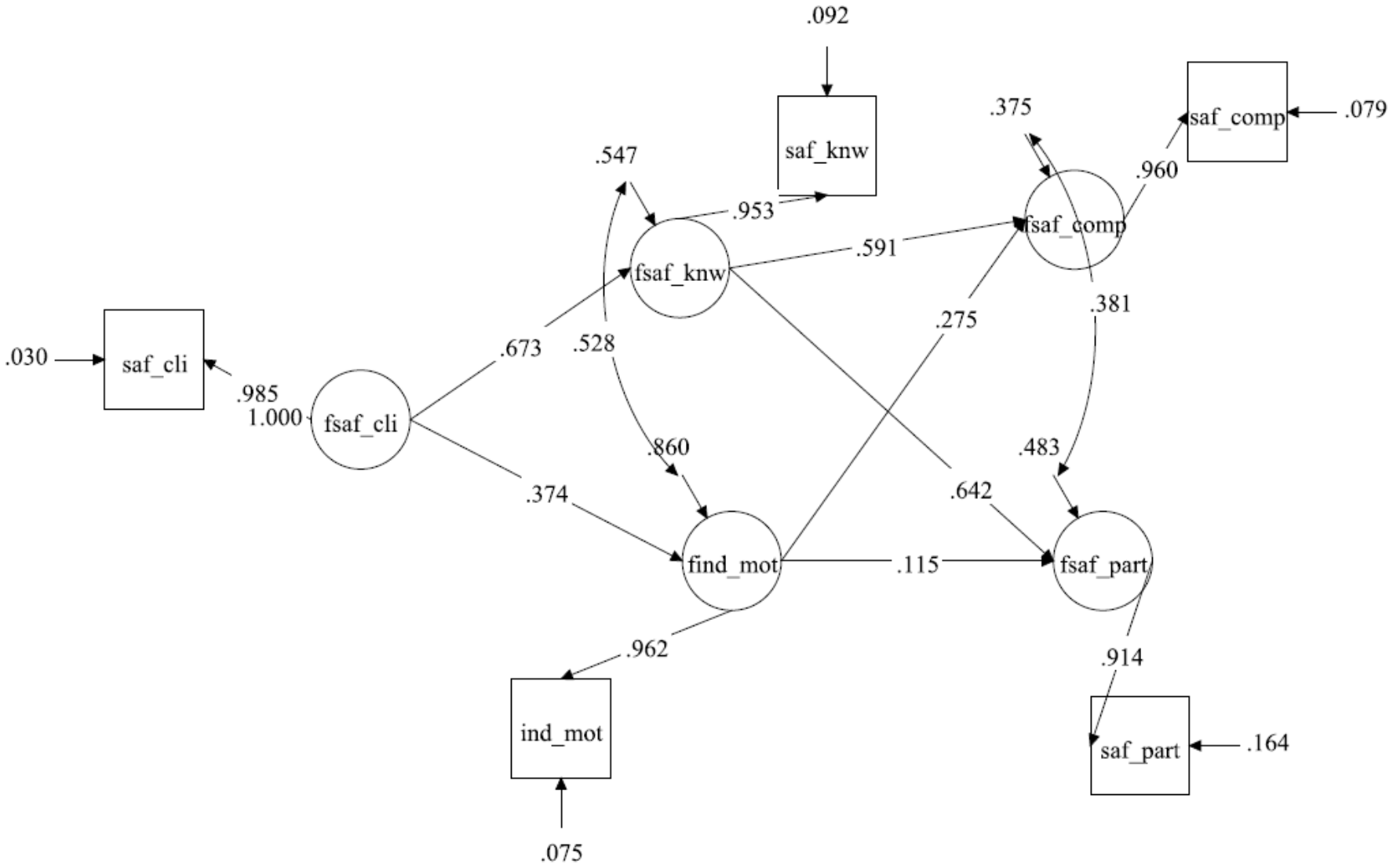
MODELLO "RELIABILITY CORRECTION"

MODEL:

neal_2011_item_rel corr

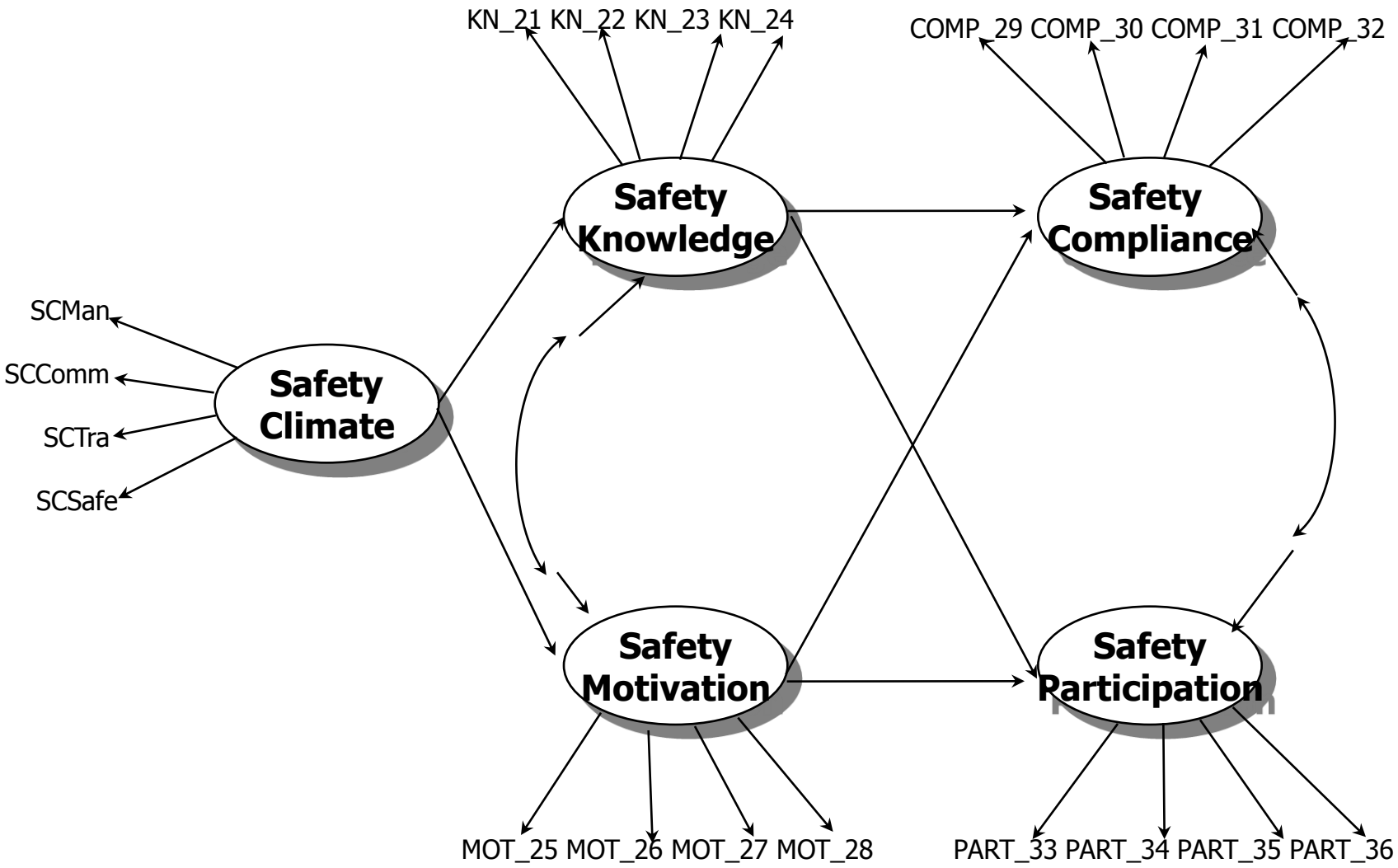
```
Fsaf_cli BY saf_cli;
saf_cli @13.745;
Fsaf_knw BY saf_knw ;
saf_knw @2.104;
Find_mot BY ind_mot ;
ind_mot @1.608;
FSAF_COMP BY SAF_COMP ;
SAF_COMP @1.987;
FSAF_PART BY SAF_PART ;
SAF_PART @4.082;

Fsaf_knw on Fsaf_cli;
Find_mot on Fsaf_cli;
FSAF_COMP ON Fsaf_knw Find_mot;
FSAF_PART ON Fsaf_knw Find_mot;
Fsaf_knw WITH Find_mot;
FSAF_COMP WITH FSAF_PART;
```

MODELLO "RELIABILITY CORRECTION"



neal_2011_item_rel corr

MODELLO "PARTIAL AGGREGATION"



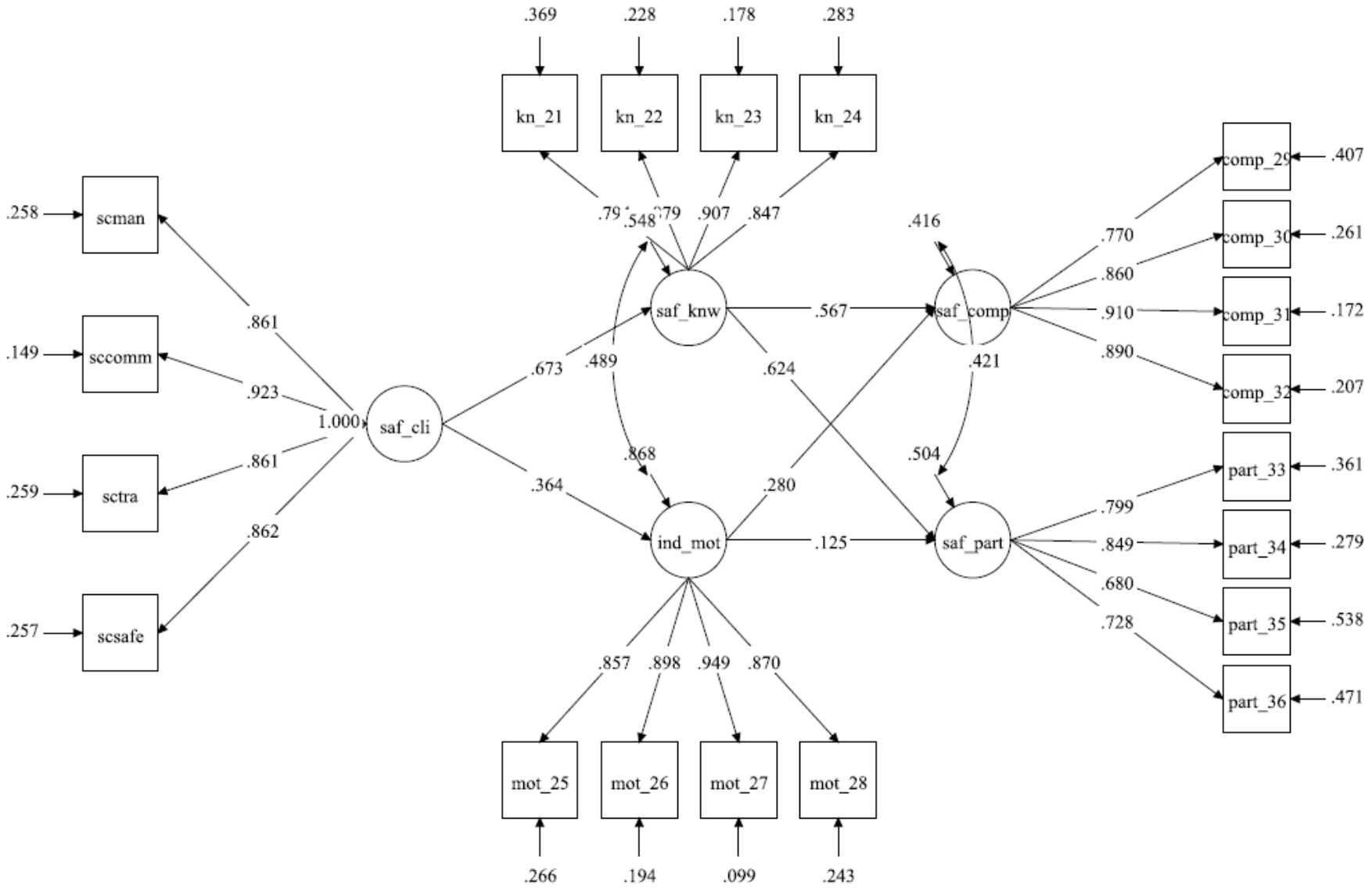
MODELLO "PARTIAL AGGREGATION"

```
usev are
  KN_21 KN_22 KN_23 KN_24    MOT_25 MOT_26 MOT_27 MOT_28
  COMP_29 COMP_30 COMP_31 COMP_32    PART_33 PART_34
PART_35 PART_36   SCMan  SCComm SCTra SCSafe;
define:
 SCMan = SAFCLI1  +SAFCLI2  +SAFCLI3   +SAFCLI4;
 SCComm =SAFCLI5  +SAFCLI6   +SAFCLI7  +SAFCLI8  +SAFCLI9;
 SCTra =SAFCLI10 +SAFCLI11 +SAFCLI12 +SAFCLI13;
 SCSafe =SAFCLI14 +SAFCLI15 +SAFCLI16 ;
analysis: estimator=MLMV;
```

neal_2011_part_aggreg

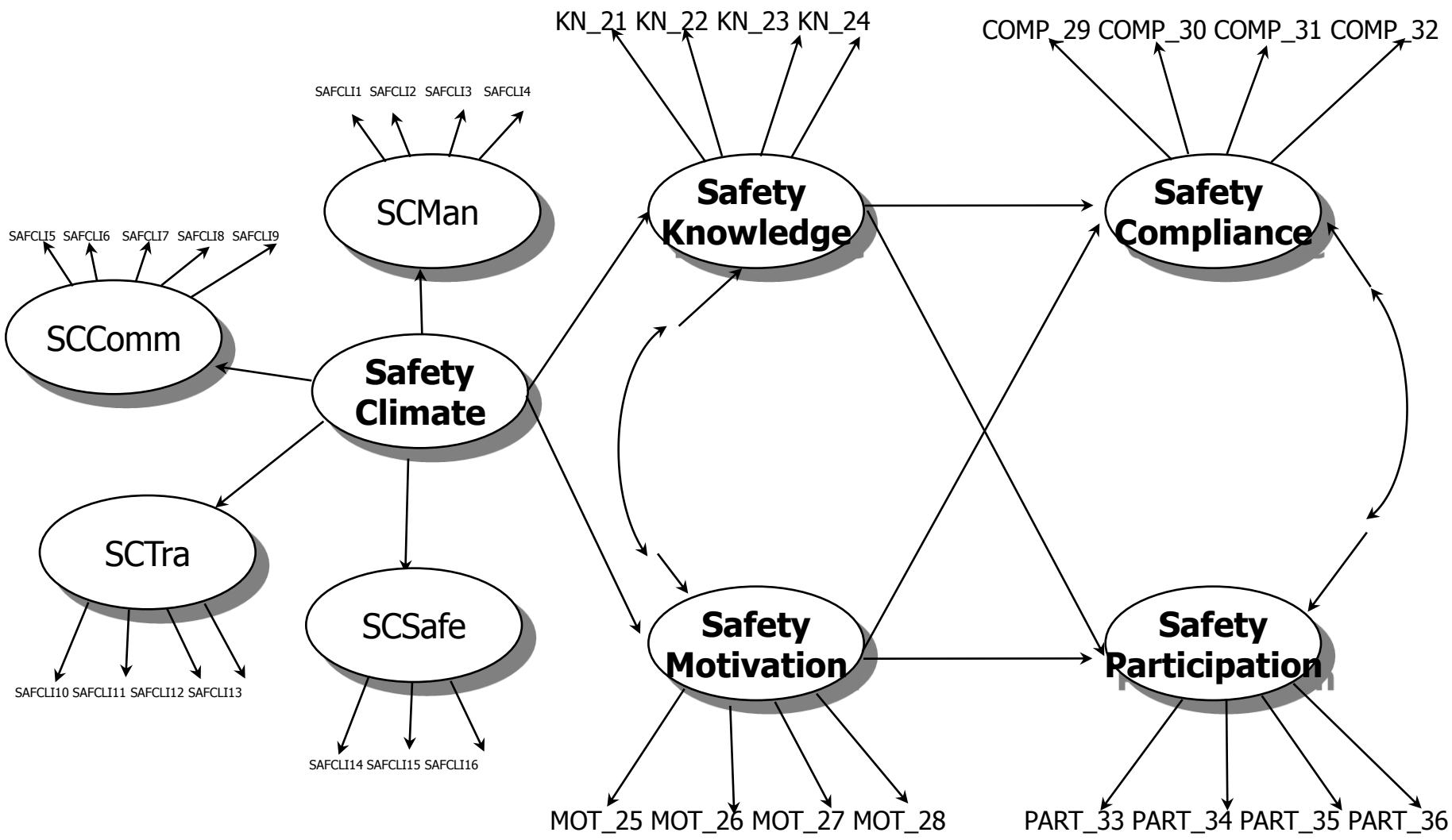
```
MODEL:
saf_cli by SCMan SCComm SCTra SCSafe;
saf_knw by  KN_21 KN_22 KN_23 KN_24 ;
ind_mot by  MOT_25 MOT_26 MOT_27 MOT_28;
SAF_COMP by COMP_29 COMP_30 COMP_31 COMP_32;
SAF_PART by PART_33 PART_34 PART_35 PART_36 ;
saf_knw on saf_cli; ind_mot on saf_cli;
SAF_COMP ON saf_knw ind_mot;
SAF_PART ON saf_knw ind_mot;
saf_knw WITH ind_mot; SAF_COMP WITH SAF_PART;
```

MODELLO "PARTIAL AGGREGATION"



neal_2011_item_rel corr

MODELLO "FULL"



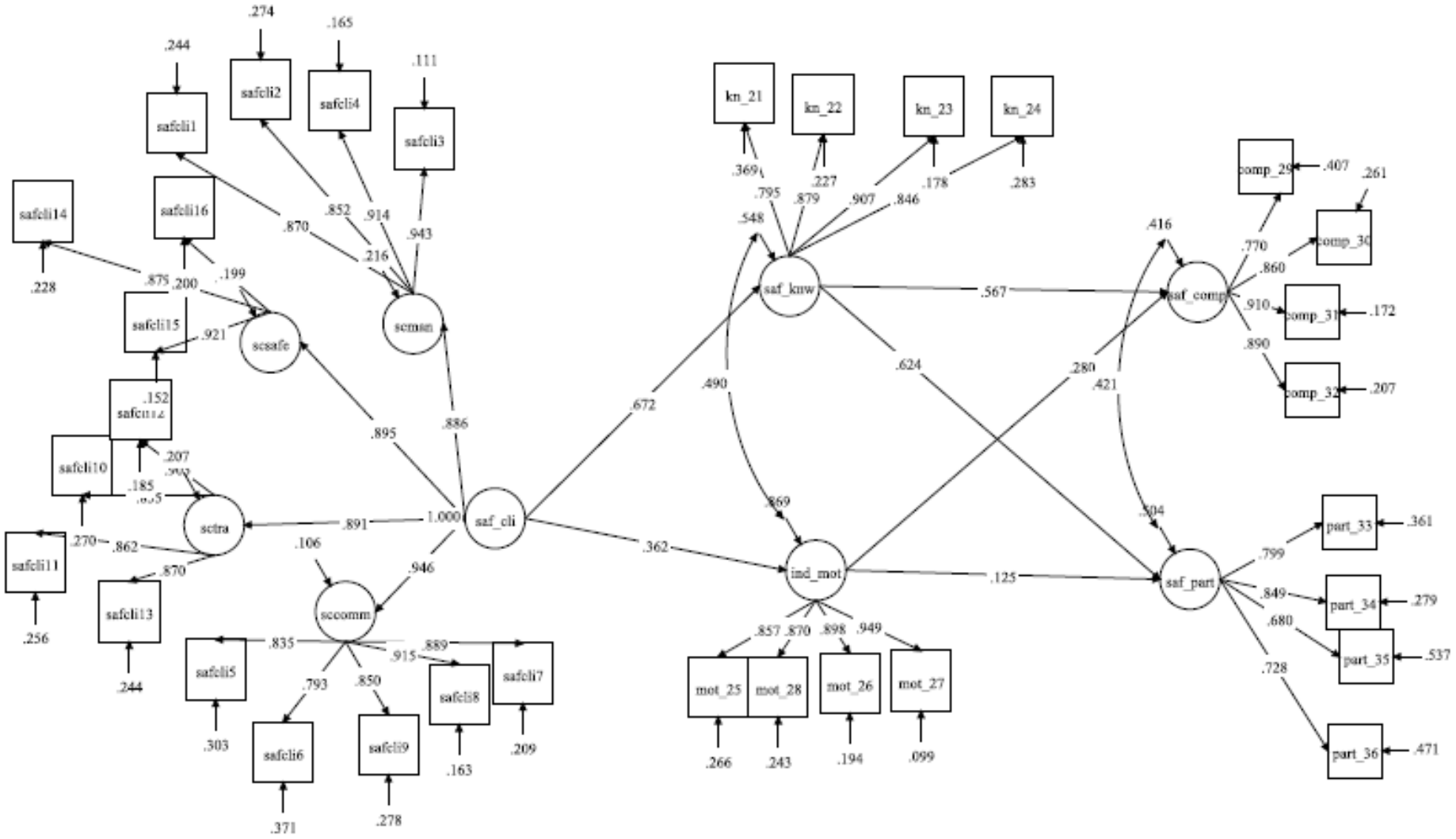
MODELLO FULL

```
MODEL:
 SCMan by   SAFCLI1  SAFCLI2  SAFCLI3   SAFCLI4;
 SCComm by  SAFCLI5  SAFCLI6   SAFCLI7  SAFCLI8  SAFCLI9;
 SCTra by   SAFCLI10 SAFCLI11 SAFCLI12 SAFCLI13;
 SCSafe by  SAFCLI14 SAFCLI15 SAFCLI16 ;
saf_cli by SCMan SCComm SCTra SCSafe;
saf_knw by  KN_21 KN_22 KN_23 KN_24 ;
ind_mot by  MOT_25 MOT_26 MOT_27 MOT_28;
SAF_COMP by COMP_29 COMP_30 COMP_31 COMP_32;
SAF_PART by PART_33 PART_34 PART_35 PART_36 ;
```

neal_2011_item_SEM.inp

```
saf_knw on saf_cli;
ind_mot on saf_cli;
SAF_COMP ON saf_knw ind_mot;
SAF_PART ON saf_knw ind_mot;
saf_knw WITH ind_mot;
SAF_COMP WITH SAF_PART;

SAF_COMP ON saf_cli ;
SAF_PART ON saf_cli ;
```

MODELLO FULL



neal_2011_item_SEM.inp

MODELLI A CONFRONTO

	chi_2	df	TLI	CFI	Rmsea	p(RMSEA)	SRMR
Var. oss.	170	2	0,67	0,93	0,24	0,001	0,059
Rel. Corr.	117	2	0,78	0,96	0,2	0,001	0,041
Partial agg.	802	162	0,94	0,95	0,052	ns	0,06
Full SEM	1400	452	0,95	0,95	0,038	ns	0,065

	Var. oss.	Rel. Corr.	Partial agg.	full SEM
Climate ->Knowledge	0.607	0.673	0.673	0.672
Climate ->Motivation	0.347	0.374	0.364	0.362
Knowledge->Compl.	0.503	0.591	0.567	0.567
Motivation->Compl.	0.307	0.275	0.280	0.280
Knowledge->Particip.	0.519	0.642	0.624	0.624
Motivation->Particip.	0.157	0.115	0.125	0.125
R^2 Knowledge	0.369	0.453	0.452	0.452
R^2 Motivation	0.120	0.140	0.132	0.131
R^2 Compliance	0.521	0.625	0.584	0.584
R^2 Participation	0.386	0.517	0.496	0.496

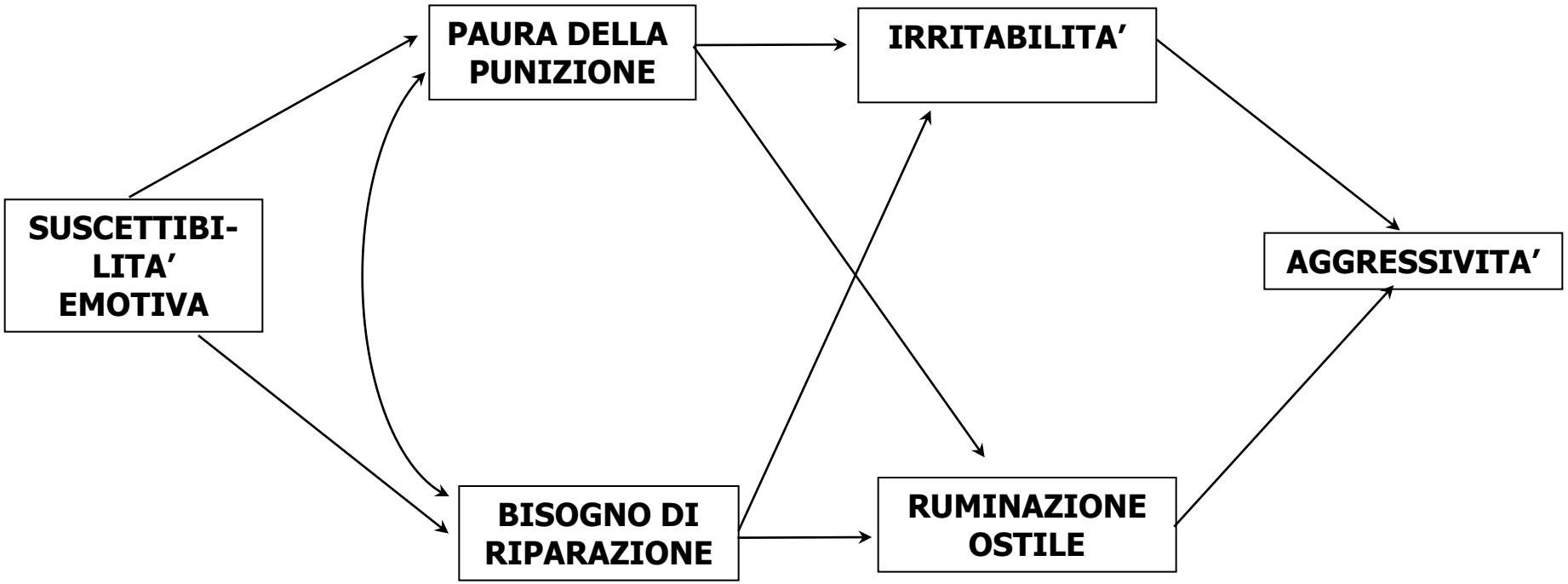
MODELLI A CONFRONTO

- I modelli di aggregazione totale hanno molti meno gradi di libertà. Se si utilizzano tali modelli, le stime fornite dall'approccio "reliability correction" sono meno distorte di quelle ottenute fissando a 0 la varianza residua

- I modelli di disaggregazione parziale e totale hanno molti più gradi di libertà. Essi correggono le stime dei parametri per l'errore di misurazione. Vanno sempre preferiti ai modelli di aggregazione totale

- Nel nostro esempio modelli di disaggregazione parziale e totale danno risultati sovrapponibili. Il modello "reliability correction" fornisce stime non troppo distanti da quelli "parcels" e "full". Il modello con variabili osservate non corrette è quello più critico

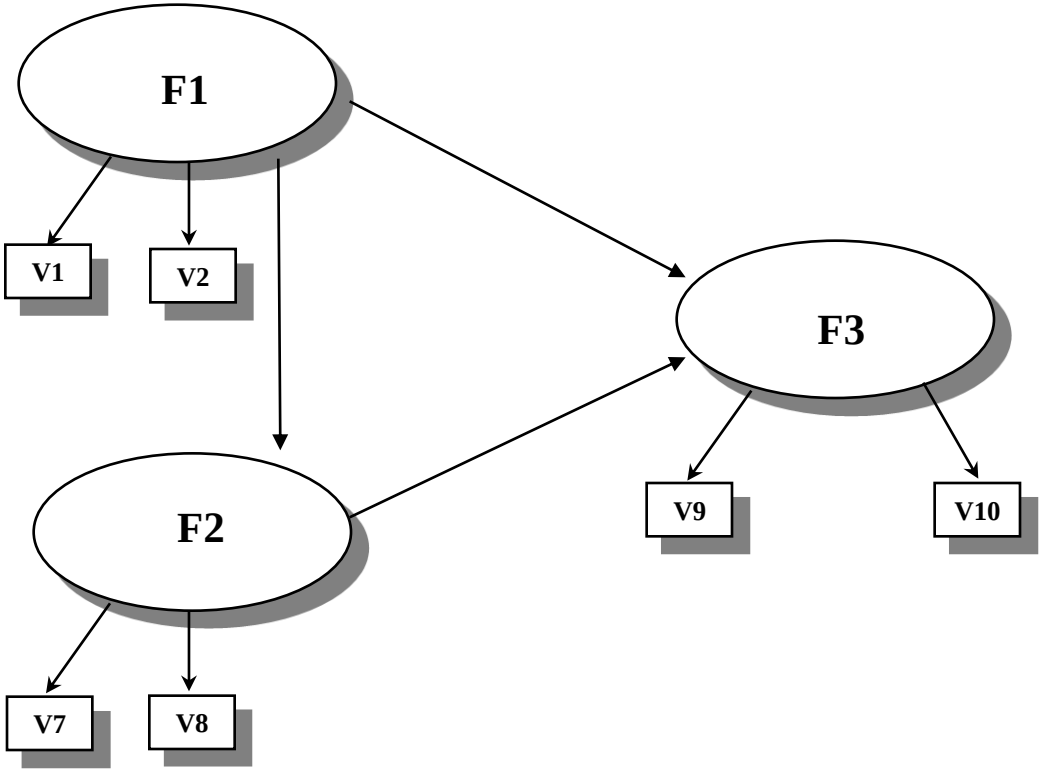
ESERCIZIO: REALIZZAZIONE DI UN MODELLO DI EQUAZIONI STRUTTURALI CON VARIABILI OSSERVATE



FILE IS OBSERVED_PARIS_2011.DAT;

ESERCIZIO: REALIZZAZIONE DI UN MODELLO DI EQUAZIONI STRUTTURALI COMPLETO

Analizzare il seguente modello di equazioni strutturali completo:



ESERCIZIO: REALIZZAZIONE DI UN MODELLO DI EQUAZIONI STRUTTURALI COMPLETO

V1=ISEE6; V2=ISEE2; V3=ISEE13; V4=ISEE11; V5=ISEE12; V6=ISEE3; V7=COMMUN; V8=MONIT; V9=SELF_EST; V10=SODD; V11=OTTIM;

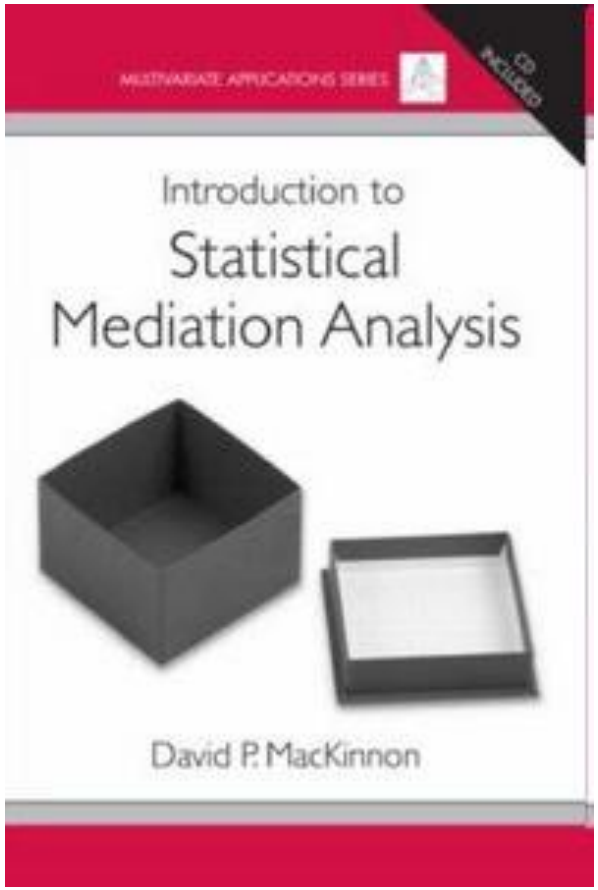
F1: autoefficacia nell'espressione delle emozioni positive
F2: relazione con i genitori
F3: benessere

NUMERO DI SOGGETTI=412

La matrice di varianza/covarianza osservata si trova nel file ES2.dat

L'analisi della mediazione statistica

Riferimenti concettuali e metodologici:



http://www.public.asu.edu/~davidpm/ripl/mediate.htm

L'analisi della mediazione statistica

Definizione di Mediatore: Una variabile intermedia nel processo causale che lega una variabile indipendente a una dipendente

Alcuni esempi dalla letteratura psicologica

Attitudes cause intentions which then cause behavior (Azjen & Fishbein, 1980)

Prevention programs change norms which promote healthy behavior (Judd & Kenny, 1981)

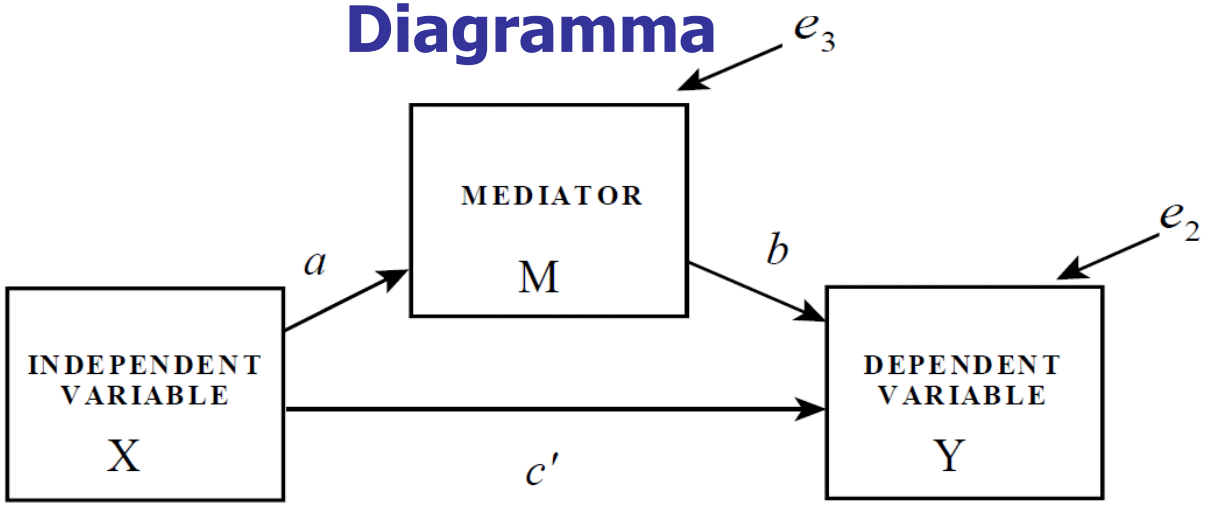
Increasing exercise skills increases self-efficacy which increases physical activity (Bandura, 1977)

L'analisi della mediazione statistica

Tre modi per specificare un modello di mediazione

Descrizione Verbale: la variabile M è intermedia nella sequenza causale che lega X a Y.

Diagramma



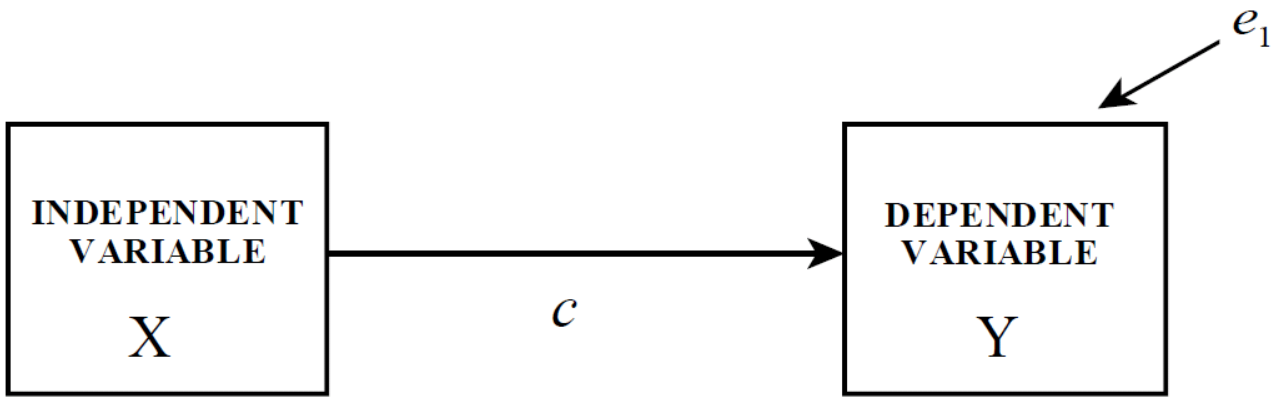
Equazioni

$$Y = i_2 + c'X + bM + e_2$$

$$M = i_3 + aX + e_3$$

L'analisi della mediazione statistica

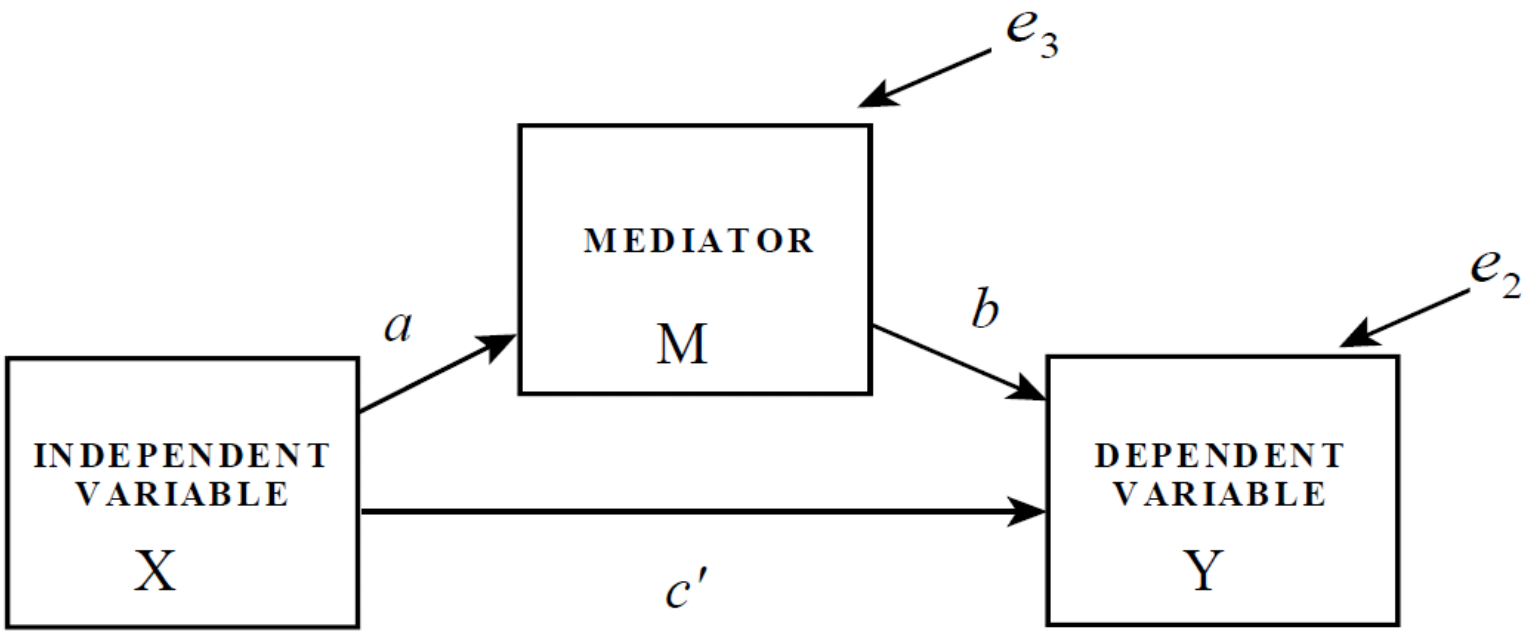
Nella sua forma più semplice la mediazione rappresenta l'aggiunta di una terza variabile nella relazione tra una variabile dipendente Y e una variabile indipendente X, dove la X "causa" la terza variabile di mediazione "M".



Modello senza mediatore "M"

L'analisi della mediazione statistica

Modello con il mediatore "M"



$$Y = i_2 + c'X + bM + e_2$$

$$M = i_3 + aX + e_3$$

L'analisi della mediazione statistica

Ci sono tre approcci principali all'analisi della mediazione statistica:
(*a*) l'approccio "causal steps"
(*b*) l'approccio della differenza dei coefficienti
(*c*) l'approccio del prodotto dei coefficienti

Tutti questi metodi utilizzano l'informazione proveniente dalle seguenti equazioni di base:

$$Y = i_1 + cX + e_1 \qquad \text{(eq. 1)}$$

$$Y = i_2 + c'X + bM + e_2 \qquad \text{(eq. 2)}$$

$$M = i_3 + aX + e_3 \qquad \text{(eq. 3)}$$

L'analisi della mediazione statistica

L'approccio "causal steps" è quello proposto da Judd e Kenny (1981) e da Baron e Kenny (1986) e consiste nella verifica delle seguenti condizioni:

a) È necessaria una relazione significativa tra X e Y nell'eq. 1

b) È necessaria una relazione significativa tra X e M nell'eq. 3

c) È necessaria una relazione significativa tra M e Y nell'eq. 2, al netto della relazione tra X e Y nella stessa equazione

L'analisi della mediazione statistica

L'approccio "causal steps" è quello proposto da Judd e Kenny (1981) e da Baron e Kenny (1986) e consiste nella verifica delle seguenti condizioni:

d) Il coefficiente c che lega X a Y nell'eq. 1 deve essere maggiore del coefficiente c' che lega X a Y nell'eq. 2 (ovvero in presenza del mediatore M)

d') Nella formulazione originale di Judd e Kenny l'enfasi era sulla mediazione totale: per cui si è in presenza di un effetti di mediazione quando c è significativo e c' non lo è

Baron and Kenny's (1986) article had been cited by 12,688 journal articles as of September 2009, according to Social Sciences Citation Index.

L'analisi della mediazione statistica

La condizione d', mediatzione totale, è irrealistica nelle scienze sociali (Baron & Kenny, 1986)

Si parla di mediazione parziale, invece, quando *c'* è minore di *c*, ma non è necessariamente non-significativo. E' una condizione più realistica e in quanto tale è diventata parte del "causal steps approach".

Una volta accertate le condizioni per la presenza della mediazione, di solito viene utilizzato il prodotto dei coefficienti (ab) per stimare l'effetto di mediazione.

L'analisi della mediazione statistica

Problemi nell'approccio "causal steps" di Judd e Kenny (1981) e Baron e Kenny (1986):

a) La potenza della verifica di questo approccio è molto bassa. Sono necessari tanti test di ipotesi nulle, e questo abbassa la potenza e aumenta l'errore ti tipo I.

b) L'approccio mira a definire quali sono le condizioni per cui è presente una mediazione: il test esplicito dell'effetto indiretto di X su Y è accessorio.

c) E' difficile estendere questo approccio per incorporare mediatori multipli e valutarne l'effetto sulla variabile dipendente.

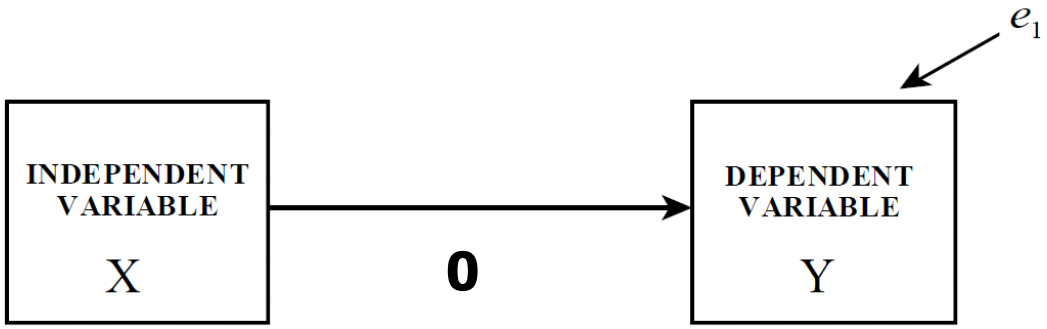
L'analisi della mediazione statistica

Problemi nell'approccio "causal steps" di Judd e Kenny (1981) e Baron e Kenny (1986):

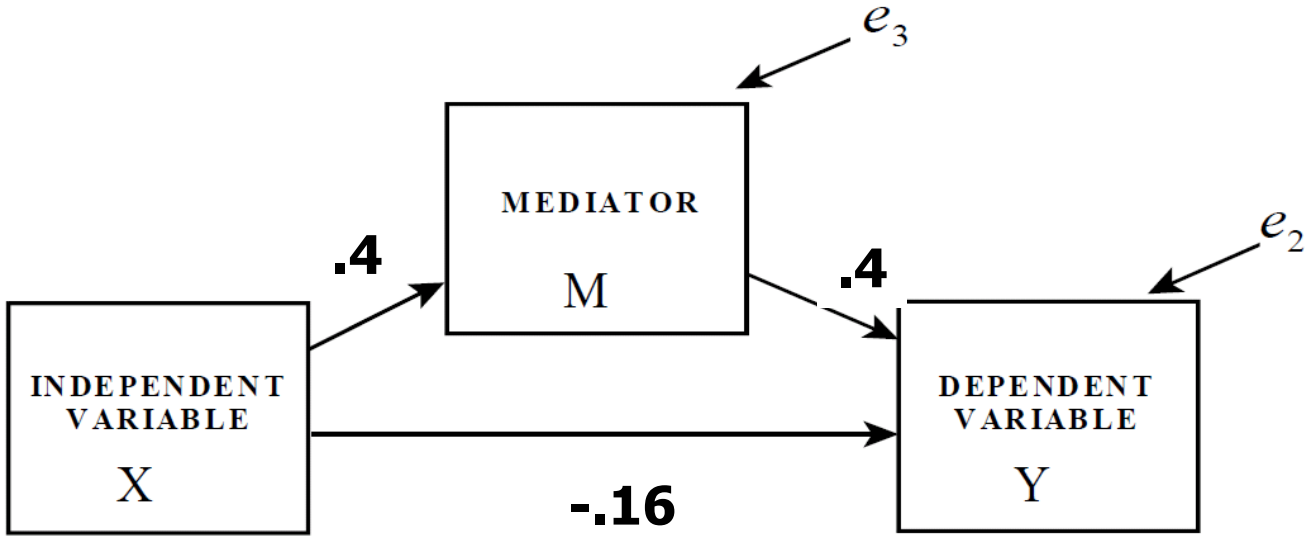
d) Può esserci mediazione anche in presenza di un effetto non significativo di X su Y: si tratta dei cosiddetti modelli di mediazione inconsistente, ovvero modelli in cui la variabile di mediazione M agisce come soppressore della relazione tra X e Y per cui in presenza di M la relazione tra X e Y, prima non significativa, ora lo diventa.

In questi modelli solitamente l'effetto diretto e l'effetto indiretto sono di segno opposto e si cancellano, come è illustrato nella figura successiva.

L'analisi della mediazione statistica



Modello senza il mediatore



Modello con il mediatore

L'analisi della mediazione statistica

Calcolo dell'effetto di mediazione (o "indiretto"): differenza e prodotto dei coefficienti.

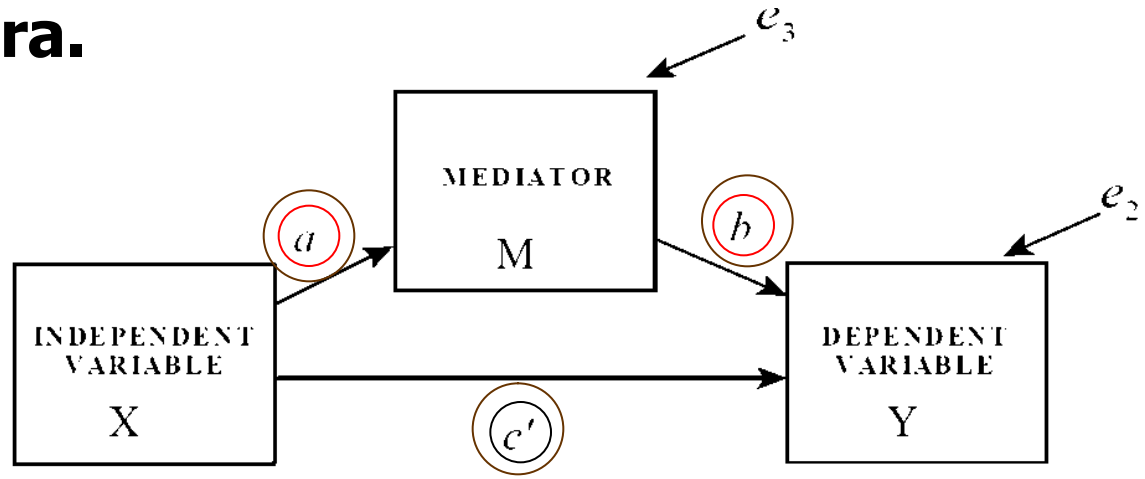
La **differenza** dei coefficienti (*c-c'*) nelle eq. 1 e 2 corrisponde alla riduzione dell'effetto di X su Y dovuto all'aggiunta di M.

Il **prodotto** dei coefficienti (*ab*) nelle eq. 2 e 3 corrisponde al razionale (Alwin & Hauser 1975) per cui la mediazione dipende da quanto la variabile indipendente X cambia il mediatore M (coefficiente a), e da quanto il mediatore M influenza la variabile dipendente Y al netto dell'impatto di X su Y (b).

MacKinnon et al. (1995) hanno dimostrato l'equivalenza algebrica di questi due differenti metodi di stima dell'effetto di mediazione

L'analisi della mediazione statistica

L'approccio del prodotto dei coefficienti è migliore perché può essere facilmente applicato a modelli con: mediatori multipli, variabili indipendenti e dipendenti multiple, variabili osservate e latenti. In questo approccio la scomposizione degli effetti è molto chiara.



Effetto Indiretto di X su Y = *ab*

Effetto Diretto di X su Y = *c'*

Effetto Totale di X su Y = *ab*+*c'*

L'analisi della mediazione statistica

Inferenza statistica ed effetti di mediazione

L'errore standard più comunemente utilizzato per ab, s_{ab}, è dato dalla seguente formula derivata da Sobel (1982), che si basa sulle derivate prime ed utilizza il metodo multivariato delta (Folmer, 1981).

$$s_{ab} = \sqrt{a^2 s_b^2 + b^2 s_a^2 + 2ab\,cov_{ab}}$$

In questa formula il termine cov_{ab} è generalmente uguale a zero. Si tratta della formula utilizzata in molti programmi per i SEM, EQS (Bentler, 1997), Mplus (Muthén & Muthén, 2009) e LISREL (Jöreskog & Sörbom, 2001).

L'analisi della mediazione statistica

Inferenza statistica ed effetti di mediazione

La stima dell'effetto di mediazione e del suo errore standard può essere utilizzata per costruire intervalli di confidenza per l'effetto di mediazione. Gli intervalli di confidenza sono utilizzati perché incorporano l'errore nella stima del parametro, fornendo un insieme di possibili valori per un effetto piuttosto che un singolo valore stimato.

Limiti dell'intervallo di confidenza: $ab \pm z_{crit}\, s_{ab}$

$LSup = ab + z_{crit}\, s_{ab}$, $LInf = ab - z_{crit}\, s_{ab}$

z_{crit} è il valore critico. Si può utilizzare anche la t.

Limiti per l'intervallo al 95% : $ab \pm 1.96\, s_{ab}$

$LSup = ab + 1.96\, s_{ab}$, $LInf = ab - 1.96\, s_{ab}$

L'analisi della mediazione statistica

Inferenza statistica ed effetti di mediazione

Se lo zero è incluso nell'intervallo di confidenza l'effetto di mediazione non è statisticamente significativo. Invece, se zero è fuori dell'intervallo, l'effetto di mediazione è statisticamente significativo.

Si può esaminare la significatività dell'effetto di mediazione dividendo la stima "ab" per il suo errore standard (s_{ab}) e confrontando questo valore con quello critico nella distribuzione normale. Se il valore assoluto di questo rapporto è maggiore/uguale a 1.96 allora l'effetto di mediazione è significativamente diverso da zero al livello di significatività dello 0.05.

L'analisi della mediazione statistica

Problemi nella verifica delle ipotesi sull'effetto di mediazione

I metodi descritti sopra, basati sul coefficiente sviluppato da Sobel possono dare risultati inaccurati in diverse circostanze.

Il problema più ricorrente è quello della ridotta potenza della verifica derivante dal fatto che la distribuzione campionaria del prodotto ab è asimmetrica.

Mackinnon et al. (2006) hanno sviluppato un programma per il calcolo degli intervalli "asimmetrici" di confidenza, PRODCLIN: http://www.public.asu.edu/~davidpm/ripl/Prodclin/

Il metodo di Mackinnon et al. (2006) può però sottostimare l'incidenza dell'errore del I tipo quando a=0 e b≠0, o a ≠0 e b=0. In questi casi è più prudente utilizzare il metodo di Sobel.

Inferenza Statistica: Il metodo Bootstrap

Il Bootstrap è una tecnica non parametrica basata sul campionamento con re-immissione. Essa consente di generare una rappresentazione empirica della distribuzione campionaria dell'effetto indiretto trattando il campione ottenuto di ampiezza n come una rappresentazione della popolazione in miniatura.

Il campionamento viene condotto con re-immissione, per cui un nuovo campione di ampiezza n è costruito campionando i casi dalla popolazione originale ma consentendo che ogni caso una volta estratto possa essere re-immesso nel campione in modo che possa essere estratto di nuovo per costruire un nuovo campione di ampiezza n.

Inferenza Statistica: Il metodo Bootstrap

Una volta che il campione è costruito, i coefficienti a b sono stimati su questo data-set ri-campionato, e il prodotto dei coefficienti calcolato.

Questo processo viene ripetuto per un totale di k volte (almeno 5000 volte).

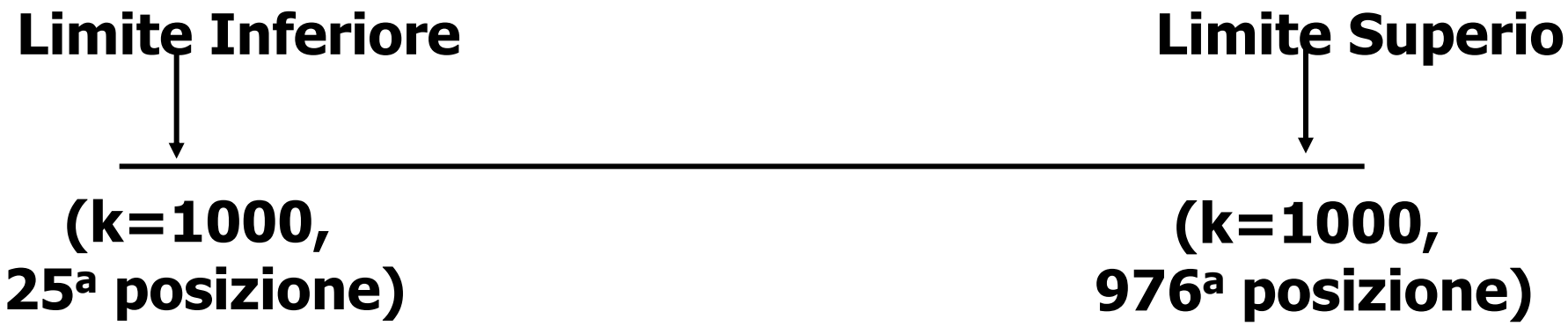
Alla fine, il ricercatore avrà a disposizione k stime diverse dell'effetto indiretto, e la distribuzione di tali stime rappresenta una approssimazione empirica della distribuzione campionaria dell'effetto indiretto.

Inferenza Statistica: Il metodo Bootstrap

Questa procedura consente di ottenere un intervallo di confidenza per l'effetto indiretto basato sulle stime boostrap. Se il valore zero non è compreso tra il limite inferiore e quello superiore dell'intervallo, allora l'effetto indiretto non è zero con una confidenza del ci%.

Questo procedimento è concettualmente equivalente a rifiutare l'ipotesi nulla che l'effetto indiretto è zero al livello di significatività ci 100-ci%.

Inferenza Statistica: Il metodo Bootstrap



Se k=1000, per un intervallo di fiducia al 95%, mettendo in ordine crescente i k valori di ab, il limite inferiore dell'intervallo è definito come il valore di ab nella 25ª posizione, e il limite superiore è il valore di ab nella 976ª posizione.

Inferenza Statistica: Il metodo Bootstrap

Il bootstrap è l'approccio più sensibile per individuare gli effetti indiretti quando sono presenti.

Il bootstrap non fa assunzioni sulla distribuzione campionaria.

Non si basa sul calcolo di errori standardi, e quindi è immune al problema che può verificarsi quando le assunzioni non sono rispettate.

Inferenza Statistica: Il metodo Bootstrap

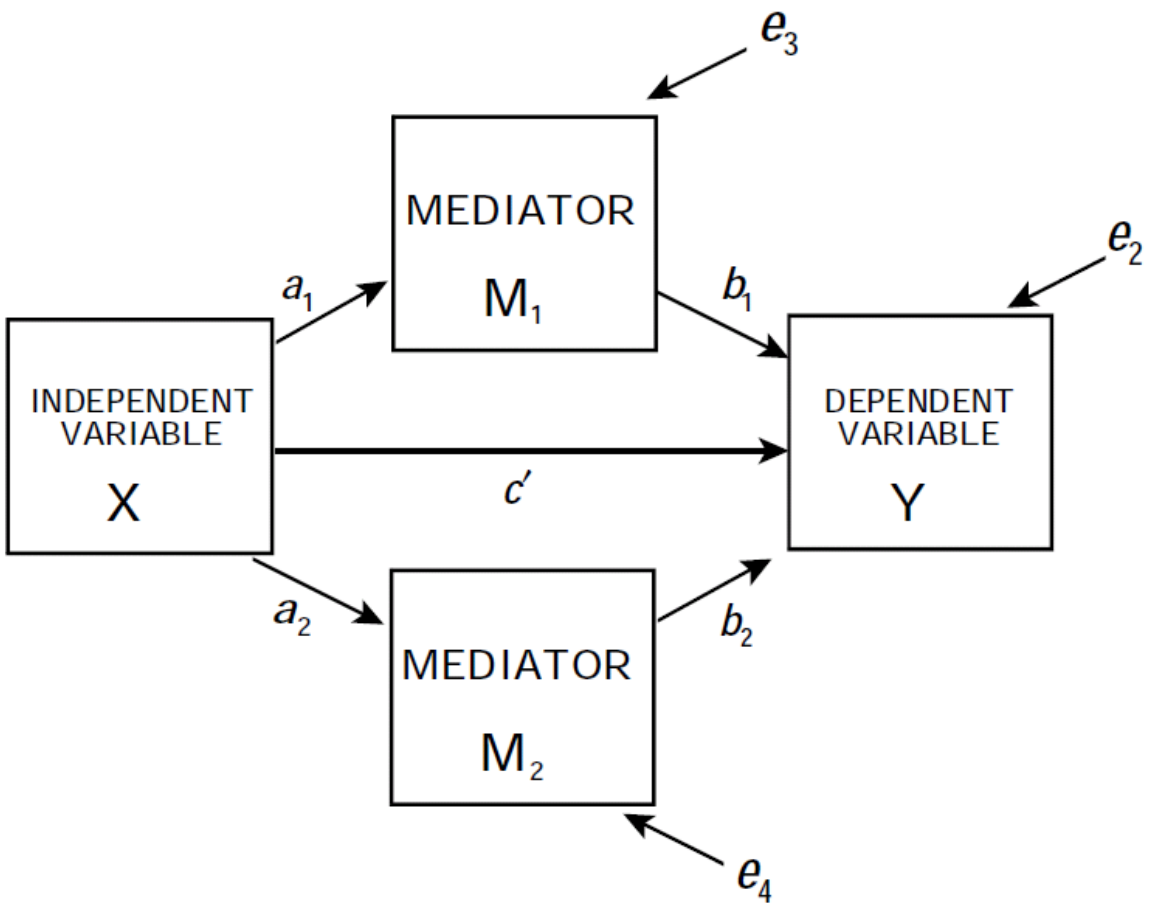
Per queste proprietà il bootstrap ha maggiore potenza statistica di qualsiasi altra strategia di verifica di ipotesi utilizzabile per esaminare la significatività statistica degli effetti indiretti.

Si tratta di un approccio molto generale, che può essere utilizzato in ogni analisi di mediazione. Hayes e Preacher hanno sviluppato delle macro in SPSS e in SAS che possono essere scaricate per poter effettuare i test boostrap degli effetti indiretti.

Mplus implementa la procedura bootstrap in modo estremamente semplice.

L'analisi della mediazione statistica

Modelli con mediatori multipli



L'analisi della mediazione statistica

Modello con Mediatori Multipli

Equazioni di regressione usate per stimare il modello di mediazione con 2 mediatori

$$Y = i_2 + c'X + b_1M_1 + b_2M_2 + e_2$$

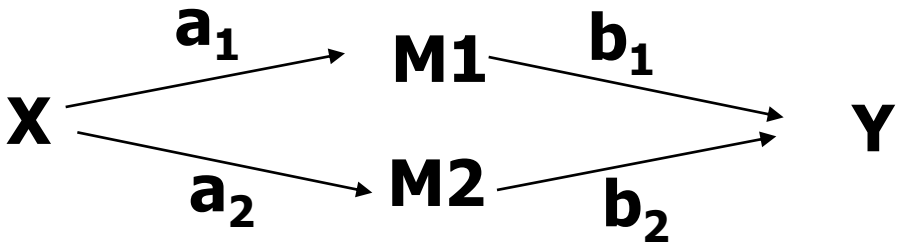
$$M_1 = i_3 + a_1X + e_3$$

$$M_2 = i_4 + a_2X + e_4$$

L'analisi della mediazione statistica

Scomposizione degli effetti nei modelli con mediatori multipli

Nei modelli con mediatori multipli è possibile definire due tipi di effetti di mediazione:



La mediazione specifica è l'effetto dovuto a uno specifico mediatore, a_1b_1 (per M1), a_2b_2 (per M2).

Mediazione totale: è l'effetto dovuto a tutti i possibili percorsi attraverso i quali X influenza indirettamente Y, ovvero $a_1b_1+a_2b_2$.

L'analisi della mediazione statistica

Scomposizione degli effetti nei modelli con mediatori multipli

L'effetto diretto è definibile sempre come:

L'effetto totale: è la somma di tutti gli effetti che agiscono su Y = $c' + (a_1b_1 + a_2b_2)$

Effetto diretto: c'

Effetto indiretto (attraverso M1): a_1b_1

Effetto indiretto (attraverso M2): a_2b_2

L'analisi della mediazione statistica

Modelli di mediazione in MPLUS

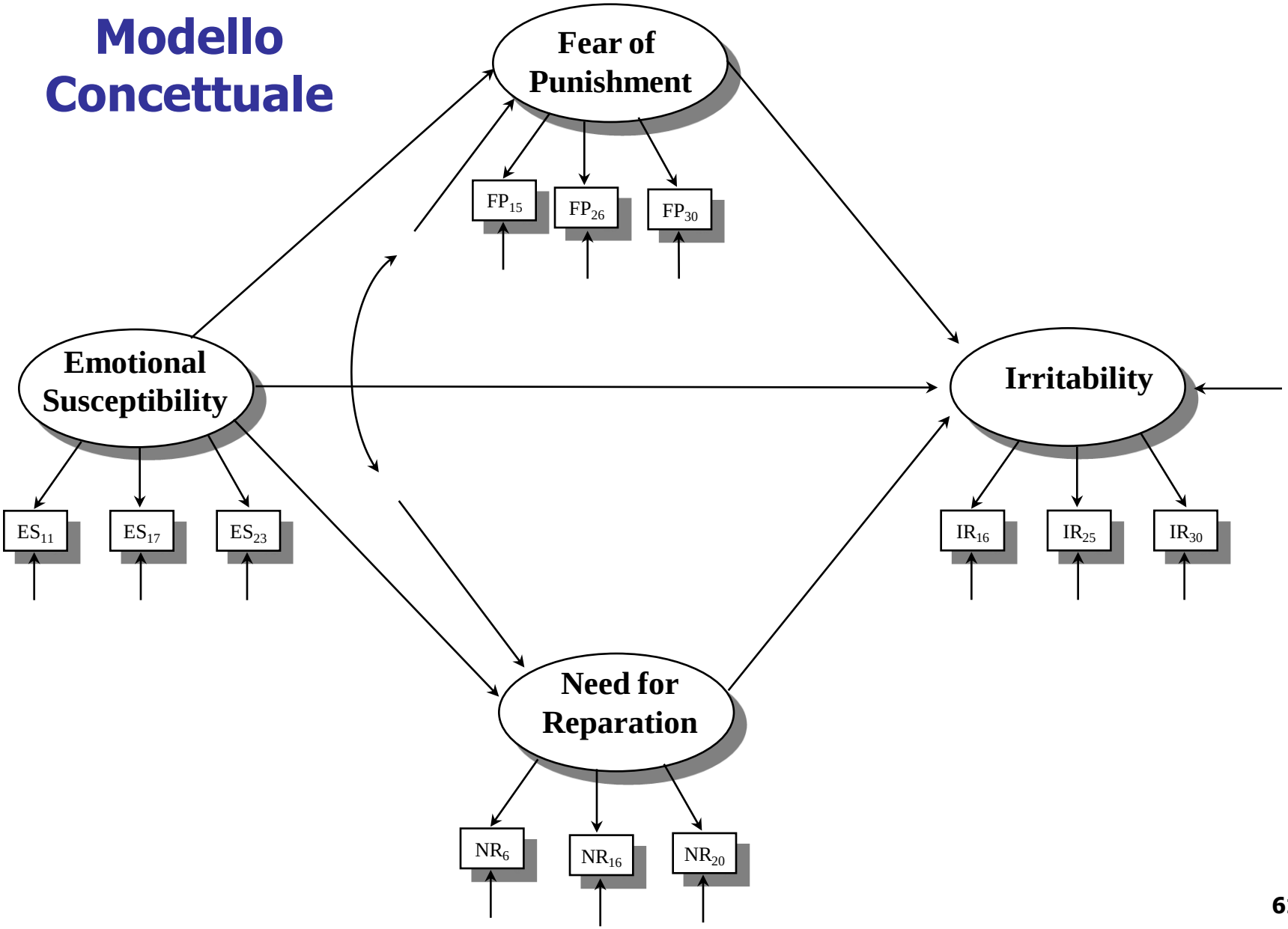
I modelli di mediazione rappresentano un caso specifico dei SEM che può essere implementato in MPLUS.

Il modello generale consente di testare modelli più complicati con variabili indipendenti e dipendenti multiple, latenti e osservate.

Bisogna utilizzare i metodi di stima della massima verosimiglianza o dei minimi quadrati, perché i residui delle variabili dipendenti correlano.

L'analisi della mediazione statistica

Modello Concettuale



L'analisi della mediazione statistica

Modelli di mediazione in MPLUS

Gli effetti indiretti possono essere ottenuti tramite i comandi IND e VIA nella sezione della sintassi MPLUS "MODEL INDIRECT".

IND è utilizzato per ottenere tutti gli effetti indiretti (totali e specifici).

MODEL INDIRECT:

Y IND X;

Y è la variabile dipendente, X la indipendente.

L'analisi della mediazione statistica

Modelli di mediazione in MPLUS

VIA è utilizzata per ottenere solo l'effetto indiretto specifico definito dalle variabili nel comando.

MODEL INDIRECT:

Y VIA M1 X;

Y è la variabile dipendente, X la indipendente, M1 il mediatore. Solo l'effetto indiretto specifico X->M1->Y viene calcolato.

L'analisi della mediazione statistica

Modelli di mediazione in MPLUS

PARIS_2011_SEM

```
TITLE: SEM
DATA:
 FILE IS SEM_PARIS.DAT;

VARIABLE:
  NAMES ARE
 es11 es17 es23  ir16 ir25 ir30
 FP15 FP26 FP30 nr6 nr16  nr20  ;
  USEV ARE
 es11 es17 es23
 ir16 ir25 ir30
 FP15 FP26 FP30
 nr6 nr16  nr20  ;
```

L'analisi della mediazione statistica

Modelli di mediazione in MPLUS

```
MODEL:
 EMSUS BY  es11 es17 es23;
 FEARPUN BY  FP15    FP26      FP30    ;
 NEEDREP BY  NR6     NR16    NR20    ;
 IRRIT BY  ir16 ir25 ir30;
FEARPUN ON EMSUS;
NEEDREP ON EMSUS;
IRRIT   ON EMSUS FEARPUN NEEDREP;
NEEDREP WITH  FEARPUN;

MODEL INDIRECT:
IRRIT IND EMSUS;

OUTPUT:  STANDARDIZED SAMPSTAT TECH1 TECH4 CINTERVAL
MODINDICES(3.84);
```

Calcola gli intervalli di confidenza

Calcola la scomposizione degli effetti

Calcola la covarianza tra le stime

L'analisi della mediazione statistica

Modelli di mediazione in MPLUS

TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Effects from EMSUS to IRRIT				
Total	0.321	0.059	5.426	0.000
Total indirect	0.151	0.036	4.224	0.000
Specific indirect				
IRRIT FEARPUN EMSUS	0.180	0.037	4.855	0.000
IRRIT NEEDREP EMSUS	-0.029	0.016	-1.853	0.064
Direct				
IRRIT EMSUS	0.170	0.063	2.708	0.007

L'analisi della mediazione statistica

Modelli di mediazione in MPLUS

CONFIDENCE INTERVALS OF TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

	Lower .5%	Lower 2.5%	Lower 5%	Estimate	Upper 5%	Upper 2.5%	Upper .5%
Effects from EMSUS to IRRIT							
Total	0.169	0.205	0.224	0.321	0.418	0.437	0.473
Total indirect	0.059	0.081	0.092	0.151	0.210	0.221	0.243
Specific indirect							
IRRIT FEARPUN EMSUS	0.085	0.108	0.119	0.180	0.242	0.253	0.276
IRRIT NEEDREP EMSUS	-0.070	-0.060	-0.055	-0.029	-0.003	0.002	0.011
Direct							
IRRIT EMSUS	0.008	0.047	0.067	0.170	0.273	0.292	0.331

Lower/Upper .5%: p=.01; Lower/Upper 2.5%: p=.05; Lower/Upper 5%: p=.10

L'analisi della mediazione statistica

Modelli di mediazione in MPLUS

Stime Boostrap

PARIS_2011_SEM

```
 ANALYSIS:
        Bootstrap = 5000;
MODEL:
 EMSUS BY  es11 es17 es23;
 FEARPUN BY  FP15    FP26      FP30    ;
 NEEDREP BY  NR6     NR16    NR20    ;
 IRRIT BY  ir16 ir25 ir30;
FEARPUN ON EMSUS;
NEEDREP ON EMSUS;
IRRIT   ON EMSUS FEARPUN NEEDREP;
NEEDREP WITH  FEARPUN;
MODEL INDIRECT:
IRRIT IND EMSUS;
OUTPUT:  STANDARDIZED SAMPSTAT TECH1 TECH4
CINTERVAL(BOOTSTRAP) MODINDICES(3.84);
```

L'analisi della mediazione statistica

Modelli di mediazione in MPLUS

CONFIDENCE INTERVALS OF TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

	Lower .5%	Lower 2.5%	Lower 5%	Estimate	Upper 5%	Upper 2.5%	Upper .5%
Effects from EMSUS to IRRIT							
Total	0.170	0.204	0.223	0.321	0.418	0.438	0.480
Total indirect	0.061	0.081	0.091	0.151	0.220	0.238	0.260
Specific indirect							
IRRIT							
FEARPUN							
EMSUS	0.094	0.110	0.120	0.180	0.253	0.268	0.294
IRRIT							
NEEDREP							
EMSUS	-0.087	-0.069	-0.061	-0.029	-0.006	-0.002	0.005
Direct							
IRRIT							
EMSUS	-0.010	0.036	0.058	0.170	0.277	0.298	0.338

La stima boostrap per i limiti dell'IC dell'effetto specifico di IRRIT attraverso NEEDREP ora è significativa (p<.05).

L'analisi della mediazione statistica

- L'approccio di Baron e Kenny è problematico e va sostituito con approcci più moderni, basati sulla stima dell'effetto indiretto

- Il metodo più utile per calcolare l'effetto di mediazione è quello che si basa sul prodotto dei coefficienti

- Soprattutto su piccoli campioni è importante utilizzare corrette procedure per la stima dei parametri (es, quelle basate sugli intervalli asimmetrici)

I SEM per l'analisi di Mediazione

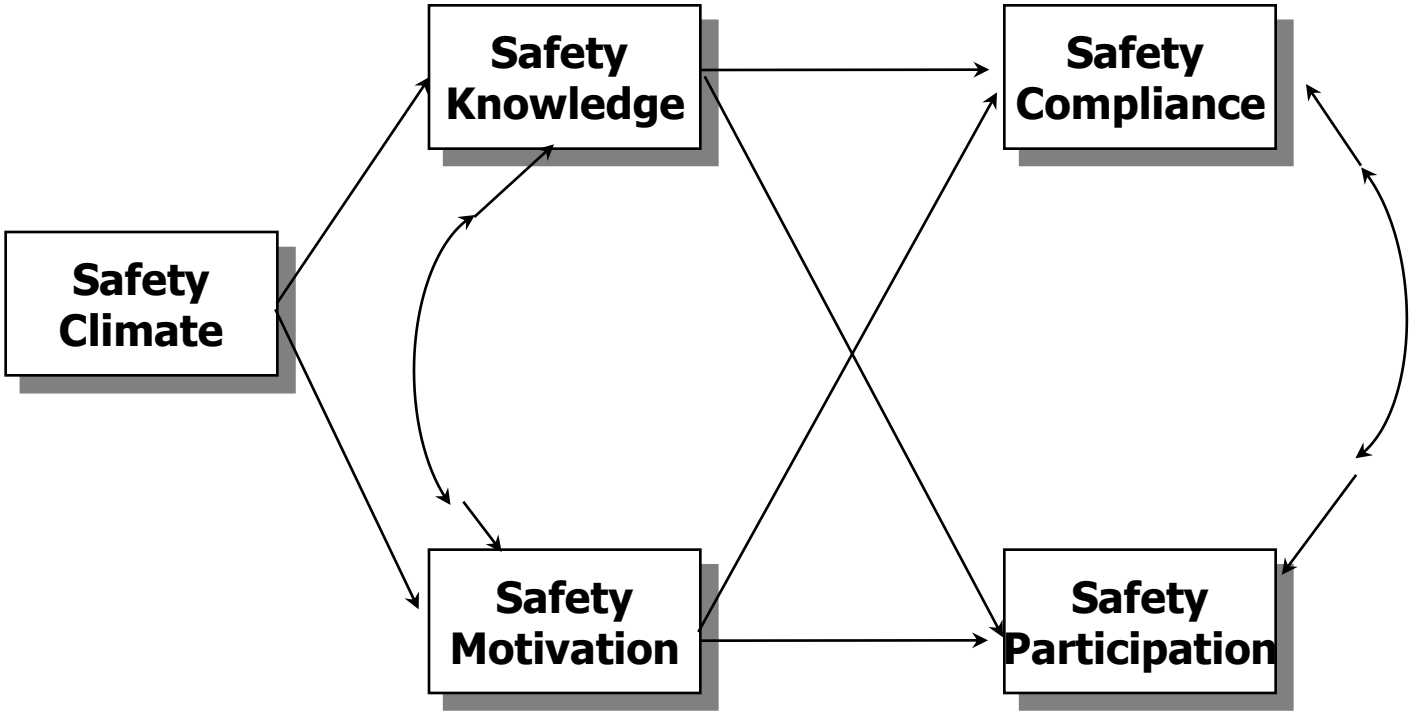
I SEM offrono la massima flessibilità per esaminare modelli con: variabili indipendenti multiple, mediatori multipli (in parallelo o concatenati), variabili dipendenti multiple, relazioni ricorsive o non ricorsive, variabili osservate e/o latenti

I SEM garantiscono flessibilità nel calcolo dei parametri utilizzando stimatori full information. Le tecniche "limited information" (come i minimi quadrati ordinari) stimano i parametri di ciascuna equazione separatamente. Le tecniche "Full information" (FIML) stimano i parametri di un modello simultaneamente per tutte le equazioni.

I SEM esaminano sia i singoli parametri sia il modello nella sua globalità, grazie all'utilizzo di molteplici indici di fit. I SEM consentono di esaminare modelli molto complessi, come quelli multilivello o quelli su gruppi multipli.

ESERCIZIO 7: ANALISI DELLA MEDIAZIONE STATISTICA

Utilizzando il modello dell'esempio



- Esaminare gli effetti diretti, indiretti e totali

- Calcolare gli intervalli di fiducia al 95% per gli effetti indiretti specifici applicando la procedura basata sul metodo bootstrap

POTENZIALITÀ DEI SEM

* Analisi simultanea delle relazioni tra costrutti (osservati e latenti) specificati da una teoria

* Incorporazione dell'errore di misurazione nel modello

* Esame probabilistico della consistenza tra teoria e dati

* Possibilità di considerare simultaneamente:

- variabili osservate e variabili latenti
- strutture di covarianza e medie strutturali
- strutture di covarianza e/o medie su più campioni
- variabili continue e categoriali (MPLUS)
- data set complessi (modelli multilivello)

LIMITI DEI SEM

* Necessità di ipotesi teoriche trasferibili in un modello
* Necessità di dati di qualità elevata (analisi preliminari)
* Condizioni stringenti di applicazione (es., ML)
* Condizioni stringenti per l'identificabilità dei modelli
* Uso "esplorativo" - non guidato dalla teoria (es., MI)
* Interpretazione causale dei nessi
* Direzionalità dei nessi di influenza

Cenni di algebra matriciale

Matrici e operazioni tra matrici

Matrice: tabella di numeri tra due parentesi o linee.
Indicate con *lettere maiuscole in carattere grassetto* (A).
Costituita da un certo numero di righe e da un certo numero di colonne.
Variabili sulle colonne, soggetti o casi sulle righe ("*matrici Casi X Variabili*")
Elemento generico (soggetto i-esimo sulla variabile j-esima): indicato come x_{ij}: per i=2 e j=3 quindi $x_{2,3}$.

Es., con 15 righe e 3 colonne abbiamo la seguente matrice:

Matrici e operazioni tra matrici

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ \ldots\ldots & & \\ \ldots\ldots & & \\ x_{15,1} & x_{15,2} & x_{15,3} \end{bmatrix}$$

Matrici e operazioni tra matrici

Vettore: particolare matrice che ha una sola riga o una sola colonna.

Vettore composto da una sola colonna: x.
Vettore composto da una sola riga: x'.

$$x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}; \quad x' = \begin{bmatrix} 2 & 3 & 4 \end{bmatrix};$$

Matrici e operazioni tra matrici

***Matrice somma e differenza*. Per sommare e sottrarre due matrici è necessario che esse abbiano lo stesso numero di righe e lo stesso numero di colonne. La matrice somma è composta da elementi $c_{ij} = a_{ij} + b_{ij}$**

$$\mathbf{A}+\mathbf{B}=\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}+\begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix}=\begin{bmatrix} a_{1,1}+b_{1,1} & a_{1,2}+b_{1,2} \\ a_{2,1}+b_{2,1} & a_{2,2}+b_{2,2} \end{bmatrix}=\mathbf{C}$$

Esempio numerico:

$$\mathbf{A}+\mathbf{B}=\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}+\begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}=\begin{bmatrix} 1+5 & 2+6 \\ 3+7 & 4+8 \end{bmatrix}=\begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix}=\mathbf{C}$$

Matrici e operazioni tra matrici

La matrice differenza è composta da elementi $c_{ij} = a_{ij} - b_{ij}$

$$\mathbf{A}-\mathbf{B}=\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}-\begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix}=\begin{bmatrix} a_{1,1}-b_{1,1} & a_{1,2}-b_{1,2} \\ a_{2,1}-b_{2,1} & a_{2,2}-b_{2,2} \end{bmatrix}=\mathbf{C}$$

Esempio numerico:

$$\mathbf{A}-\mathbf{B}=\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}-\begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}=\begin{bmatrix} 1-5 & 2-6 \\ 3-7 & 4-8 \end{bmatrix}=\begin{bmatrix} -4 & -4 \\ -4 & -4 \end{bmatrix}=\mathbf{C}$$

Matrici e operazioni tra matrici

***Prodotto di una matrice per uno scalare*: Il risultato è una matrice in cui ogni elemento viene moltiplicato per lo scalare c (uno scalare è un singolo numero, ovvero una matrice composta da una sola riga e una sola colonna, 1x1).**

$$c * A = c * \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} = \begin{bmatrix} c * a_{1,1} & c * a_{1,2} \\ c * a_{2,1} & c * a_{2,2} \end{bmatrix}$$

Esempio numerico:

$$2 * \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix}$$

Matrici e operazioni tra matrici

Prodotto fra due matrici: **è possibile se il numero di colonne della prima matrice è uguale al numero delle righe della seconda. La matrice prodotto risultante ha tante righe quante ne ha la prima matrice (A) e tante colonne quante ne ha la seconda matrice (B).**

Esempio numerico:

$$AB = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \begin{bmatrix} 7 & 8 \\ 9 & 1 \end{bmatrix} = \begin{bmatrix} 1*7+2*9 & 1*8+2*1 \\ 3*7+4*9 & 3*8+4*1 \\ 5*7+6*9 & 5*8+6*1 \end{bmatrix} =$$

$$= \begin{bmatrix} 25 & 10 \\ 57 & 28 \\ 89 & 46 \end{bmatrix} = C$$

Matrici e operazioni tra matrici

Prodotto matriciale (o esterno): **prodotto fra un vettore *colonna* a_j ed un vettore *riga* b_k'. Dà luogo ad una matrice $C_{j,k}$ con j righe e k colonne, come nell'esempio seguente:**

$$a_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}; \quad b_3' = \begin{bmatrix} 2 & 3 & 4 \end{bmatrix}; \quad a_2 * b_3' = \begin{bmatrix} 1*2 & 1*3 & 1*4 \\ 2*2 & 2*3 & 2*4 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 4 \\ 4 & 6 & 8 \end{bmatrix}$$

P*rodotto scalare (o interno):* **prodotto fra un vettore riga a_i' ed un vettore colonna b_j. Il numero di righe del primo vettore deve essere uguale al numero di colonne del secondo. Il risultato sarà uno *scalare* $c = \Sigma\, a_i b_j$, come nell'esempio seguente:**

$$a_3' = \begin{bmatrix} 2 & 3 & 4 \end{bmatrix}; \quad b_3 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}; a_3' * b_3 = 2*1 + 3*2 + 4*3 = 2 + 6 + 12 = 20$$

Alcune matrici caratteristiche

Matrice trasposta. **Matrice A′ (o A^T) che si ottiene scambiando le righe con le colonne della matrice A.**

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \longrightarrow A' = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Alcune matrici caratteristiche

***Matrice quadrata*. Matrice che ha tante righe quante colonne.**

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}$$

Matrice simmetrica **intorno alla diagonale principale: composta da elementi $a_{ij}=a_{ji}$. La trasposta di una matrice simmetrica è uguale alla matrice stessa.**

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \qquad a_{21}=a_{12}$$

Alcune matrici caratteristiche

Matrice diagonale. **Ha valori diversi da zero sulla diagonale principale e valori uguali a zero al di fuori di essa.**

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & 0 \\ 0 & a_{2,2} \end{bmatrix}$$

Matrice identità **(I): contiene soltanto valori 1 sulla diagonale principale e valori 0 al di fuori di essa.**

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Alcune matrici caratteristiche

***Matrice inversa*. Definita per le matrici quadrate. Data una matrice A, la sua inversa, indicata con la notazione A^{-1}, è tale che:**

$$AA^{-1} = A^{-1}A = I$$

Calcolo di una matrice inversa: piuttosto complesso.

Alcuni elementi notevoli delle matrici

Traccia. **Sia A una matrice quadrata di ordine nXn (ovvero n righe e n colonne); la "traccia di A" è la somma degli elementi sulla sua diagonale principale:** $trA = \Sigma_i \Sigma_j a_{ij}$, **con i=j.**

Determinante. **E' un numero che si ottiene effettuando la somma algebrica dei prodotti ognuno costituito da elementi appartenenti a righe e colonne diverse della matrice. In una matrice 2x2 :**

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}$$

il determinante è:

$$|A| = (a_{1,1}\, a_{2,2}) - (a_{1,2}\, a_{2,1}).$$

Se $|A| = 0$ la matrice non ha un'inversa, e si definisce "singolare".

Alcuni elementi notevoli delle matrici

Combinazione lineare. considerati p vettori $x_1, x_2, .., x_p$ di ordine n, e p numeri reali $c_1, c_2, .. c_p$, si definisce combinazione lineare dei p vettori l'espressione:

$$c_1x_1 + c_2x_2 + \ldots + c_px_p$$

consideriamo tre vettori di ordine 2 e tre scalari 3, 4, 1.

$$x_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}; \quad x_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}; \quad x_3 = \begin{bmatrix} -2 \\ 5 \end{bmatrix}$$

Una combinazione lineare dei 3 vettori con i coefficienti 3, 4 e 1 si ottiene così:

$$3*\begin{bmatrix} 1 \\ 2 \end{bmatrix} + 4*\begin{bmatrix} 3 \\ 1 \end{bmatrix} + 1*\begin{bmatrix} -2 \\ 5 \end{bmatrix} = \begin{bmatrix} 3+12-2 \\ 6+4+5 \end{bmatrix} = \begin{bmatrix} 13 \\ 15 \end{bmatrix}$$

Un vettore che è combinazione lineare di altri vettori viene detto "linearmente dipendente".

Espressioni matriciali di indici statistici

Consideriamo la seguente matrice di dati X:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ \ldots\ldots & & \\ \ldots\ldots & & \\ x_{15,1} & x_{15,2} & x_{15,3} \end{bmatrix}$$

Si tratta di una matrice "casiXvariabili":
i casi sono le righe, le variabili sono le colonne.

Espressioni matriciali di indici statistici

Da X possiamo ricavare diverse altre matrici.

Centroide: **Vettore delle medie delle variabili, avrà una riga e tre colonne:**

$$\mathbf{c} = \left[\bar{X}_{.1} \quad \bar{X}_{.2} \quad \bar{X}_{.3}\right]$$

con

$$\bar{X}_{.1} = \frac{\sum_{i=1}^{15} X_{i1}}{15} \qquad \bar{X}_{.2} = \frac{\sum_{i=1}^{15} X_{i2}}{15} \qquad \bar{X}_{.3} = \frac{\sum_{i=1}^{15} X_{i3}}{15}$$

medie calcolate su tutti i soggetti, per ogni variabile

Espressioni matriciali di indici statistici

La matrice di covarianza S (o matrice delle varianze e delle covarianze) per le tre variabili considerate ha il seguente aspetto:

$$S = \begin{bmatrix} s_1^2 & s_{12} & s_{13} \\ s_{21} & s_2^2 & s_{23} \\ s_{31} & s_{32} & s_3^2 \end{bmatrix}$$

$$s_j^2 = \frac{\sum_{i=1}^{15}(X_{i1} - \bar{X}_{.1})^2}{15}$$

$$s_{jk} = \frac{\sum_{i=1}^{15}(X_{ij} - \bar{X}_{.j})(X_{ik} - \bar{X}_{.k})}{15}$$

La matrice S è *simmetrica* intorno alla *diagonale principale*. Sulla diagonale principale: varianze (s_j^2) delle singole variabili. Fuori della diagonale: covarianze (s_{jk}) tra le variabili.

Espressioni matriciali di indici statistici

Dividendo le covarianze per le deviazioni standard delle singole variabili (come nel caso univariato) si trasforma la matrice di covarianza nella matrice di correlazione R. Infatti, per due variabili i e j:

$$r_{ij}=s_{ij}/(s_i s_j).$$

$$\mathbf{R}=\begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}$$

$$r_{jk} = \frac{\sum_{i=1}^{15}(X_{ij}-\bar{X}_{.j})(X_{ik}-\bar{X}_{.k})/15}{s_j s_k}$$

R ha le stesse proprietà di S.

Calcoli matriciali di indici statistici

Matrice delle correlazioni

$$\mathbf{R} = \mathbf{Z}'\mathbf{Z}n^{-1} = \begin{bmatrix} z_{1,1} & z_{2,1} & \ldots & z_{15,1} \\ z_{1,2} & z_{2,2} & \ldots & z_{15,2} \\ z_{1,3} & z_{2,3} & \ldots & z_{15,3} \end{bmatrix} \begin{bmatrix} z_{1,1} & z_{1,2} & z_{1,3} \\ z_{2,1} & z_{2,2} & z_{2,3} \\ \ldots & \ldots & \ldots \\ z_{15,1} & z_{15,2} & z_{15,3} \end{bmatrix} \begin{bmatrix} 1/n & 0 & 0 \\ 0 & 1/n & 0 \\ 0 & 0 & 1/n \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{\sum_{i=1}^{n} z_{i1}z_{i1}}{n} & \frac{\sum_{i=1}^{n} z_{i1}z_{i2}}{n} & \frac{\sum_{i=1}^{n} z_{i1}z_{i3}}{n} \\ \frac{\sum_{i=1}^{n} z_{i2}z_{i1}}{n} & \frac{\sum_{i=1}^{n} z_{i2}z_{i2}}{n} & \frac{\sum_{i=1}^{n} z_{i2}z_{i3}}{n} \\ \frac{\sum_{i=1}^{n} z_{i3}z_{i1}}{n} & \frac{\sum_{i=1}^{n} z_{i3}z_{i2}}{n} & \frac{\sum_{i=1}^{n} z_{i3}z_{i3}}{n} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}$$