

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 23/11: Lunedì dalle 10.30 alle 12

Dal 27/11 al 22/12: Martedì dalle 12.30 alle 14

Dall' 1/1 al 17/9/2018: 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	Martedì 10:00:12:00
	Mercoledì 11:00-14:00
	Giovedì 11:00-14:00

inizio 1° semestre	02/10/2017
fine 1° semestre	15/12/2017 possibili lezioni di recupero 18/20 dic.



ANALISI DEI DATI + LABORATORIO

CONTENUTO DEL CORSO

Il corso riguarderà i seguenti argomenti:

- Cenni di algebra matriciale**
- 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.**

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.



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

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{1,1} & \mathbf{X}_{1,2} & \mathbf{X}_{1,3} \\ \mathbf{X}_{2,1} & \mathbf{X}_{2,2} & \mathbf{X}_{2,3} \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ \mathbf{X}_{15,1} & \mathbf{X}_{15,2} & \mathbf{X}_{15,3} \end{bmatrix}$$



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' .

$$x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}; \quad x' = [2 \quad 3 \quad 4];$$

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 $c_{ij} = a_{ij} + b_{ij}$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} \mathbf{a}_{1,1} & \mathbf{a}_{1,2} \\ \mathbf{a}_{2,1} & \mathbf{a}_{2,2} \end{bmatrix} + \begin{bmatrix} \mathbf{b}_{1,1} & \mathbf{b}_{1,2} \\ \mathbf{b}_{2,1} & \mathbf{b}_{2,2} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_{1,1} + \mathbf{b}_{1,1} & \mathbf{a}_{1,2} + \mathbf{b}_{1,2} \\ \mathbf{a}_{2,1} + \mathbf{b}_{2,1} & \mathbf{a}_{2,2} + \mathbf{b}_{2,2} \end{bmatrix} = \mathbf{C}$$

Esempio numerico:

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} \mathbf{1} & \mathbf{2} \\ \mathbf{3} & \mathbf{4} \end{bmatrix} + \begin{bmatrix} \mathbf{5} & \mathbf{6} \\ \mathbf{7} & \mathbf{8} \end{bmatrix} = \begin{bmatrix} \mathbf{1+5} & \mathbf{2+6} \\ \mathbf{3+7} & \mathbf{4+8} \end{bmatrix} = \begin{bmatrix} \mathbf{6} & \mathbf{8} \\ \mathbf{10} & \mathbf{12} \end{bmatrix} = \mathbf{C}$$

Matrici e operazioni tra matrici

La matrice differenza è composta da elementi $c_{ij} = a_{ij} - b_{ij}$

$$\mathbf{A} - \mathbf{B} = \begin{bmatrix} \mathbf{a}_{1,1} & \mathbf{a}_{1,2} \\ \mathbf{a}_{2,1} & \mathbf{a}_{2,2} \end{bmatrix} - \begin{bmatrix} \mathbf{b}_{1,1} & \mathbf{b}_{1,2} \\ \mathbf{b}_{2,1} & \mathbf{b}_{2,2} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_{1,1} - \mathbf{b}_{1,1} & \mathbf{a}_{1,2} - \mathbf{b}_{1,2} \\ \mathbf{a}_{2,1} - \mathbf{b}_{2,1} & \mathbf{a}_{2,2} - \mathbf{b}_{2,2} \end{bmatrix} = \mathbf{C}$$

Esempio numerico:

$$\mathbf{A} - \mathbf{B} = \begin{bmatrix} \mathbf{1} & \mathbf{2} \\ \mathbf{3} & \mathbf{4} \end{bmatrix} - \begin{bmatrix} \mathbf{5} & \mathbf{6} \\ \mathbf{7} & \mathbf{8} \end{bmatrix} = \begin{bmatrix} \mathbf{1-5} & \mathbf{2-6} \\ \mathbf{3-7} & \mathbf{4-8} \end{bmatrix} = \begin{bmatrix} \mathbf{-4} & \mathbf{-4} \\ \mathbf{-4} & \mathbf{-4} \end{bmatrix} = \mathbf{C}$$

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, 1×1).

$$\mathbf{c} * \mathbf{A} = \mathbf{c} * \begin{bmatrix} \mathbf{a}_{1,1} & \mathbf{a}_{1,2} \\ \mathbf{a}_{2,1} & \mathbf{a}_{2,2} \end{bmatrix} = \begin{bmatrix} \mathbf{c} * \mathbf{a}_{1,1} & \mathbf{c} * \mathbf{a}_{1,2} \\ \mathbf{c} * \mathbf{a}_{2,1} & \mathbf{c} * \mathbf{a}_{2,2} \end{bmatrix}$$

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:

$$\mathbf{AB} = \begin{bmatrix} \mathbf{1} & \mathbf{2} \\ \mathbf{3} & \mathbf{4} \\ \mathbf{5} & \mathbf{6} \end{bmatrix} \begin{bmatrix} \mathbf{7} & \mathbf{8} \\ \mathbf{9} & \mathbf{1} \end{bmatrix} = \begin{bmatrix} \mathbf{1*7 + 2*9} & \mathbf{1*8 + 2*1} \\ \mathbf{3*7 + 4*9} & \mathbf{3*8 + 4*1} \\ \mathbf{5*7 + 6*9} & \mathbf{5*8 + 6*1} \end{bmatrix} =$$
$$= \begin{bmatrix} \mathbf{25} & \mathbf{10} \\ \mathbf{57} & \mathbf{28} \\ \mathbf{89} & \mathbf{46} \end{bmatrix} = \mathbf{C}$$

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:

$$a_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}; \quad b_3' = [2 \quad 3 \quad 4]; \quad a_2 * b_3' = \begin{bmatrix} 1*2 & 1*3 & 1*4 \\ 2*2 & 2*3 & 2*4 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 4 \\ 4 & 6 & 8 \end{bmatrix}$$

Prodotto 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* $c = \sum a_i b_j$, come nell'esempio seguente:

$$a_3' = [2 \quad 3 \quad 4]; \quad b_3 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}; \quad a_3' * b_3 = 2*1 + 3*2 + 4*3 = 2 + 6 + 12 = 20$$

Alcune matrici caratteristiche

Matrice trasposta. Matrice A' (o A^T) che si ottiene scambiando le righe con le colonne della matrice A .

$$\mathbf{A} = \begin{bmatrix} \mathbf{1} & \mathbf{2} & \mathbf{3} \\ \mathbf{4} & \mathbf{5} & \mathbf{6} \end{bmatrix} \longrightarrow \mathbf{A}' = \begin{bmatrix} \mathbf{1} & \mathbf{4} \\ \mathbf{2} & \mathbf{5} \\ \mathbf{3} & \mathbf{6} \end{bmatrix}$$

Alcune matrici caratteristiche

Matrice quadrata. Matrice che ha tante righe quante colonne.

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{1,1} & \mathbf{a}_{1,2} \\ \mathbf{a}_{2,1} & \mathbf{a}_{2,2} \end{bmatrix}$$

Matrice simmetrica intorno alla diagonale principale: composta da elementi $\mathbf{a}_{ij} = \mathbf{a}_{ji}$. La trasposta di una matrice simmetrica è uguale alla matrice stessa.

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{1,1} & \mathbf{a}_{1,2} \\ \mathbf{a}_{2,1} & \mathbf{a}_{2,2} \end{bmatrix} \quad \mathbf{a}_{21} = \mathbf{a}_{12}$$



Alcune matrici caratteristiche

Matrice diagonale. Ha valori diversi da zero sulla diagonale principale e valori uguali a zero al di fuori di essa.

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{1,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{a}_{2,2} \end{bmatrix}$$

Matrice identità (I): contiene soltanto valori 1 sulla diagonale principale e valori 0 al di fuori di essa.

$$\mathbf{I} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}$$



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 $n \times n$ (ovvero n righe e n colonne); la “traccia di A ” è la somma degli elementi sulla sua diagonale principale: $trA = \sum_i \sum_j a_{ij}$, 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 2×2 :

$$A = \begin{bmatrix} \mathbf{a}_{1,1} & \mathbf{a}_{1,2} \\ \mathbf{a}_{2,1} & \mathbf{a}_{2,2} \end{bmatrix} \quad \text{il determinante è:} \\ |A| = (a_{1,1} a_{2,2}) - (a_{1,2} a_{2,1}).$$

Se $|A| = 0$ la matrice non ha un'inversa, e si definisce “singolare”.



Alcuni elementi notevoli delle matrici

Combinazione lineare. considerati p vettori $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ di ordine n , e p numeri reali c_1, c_2, \dots, c_p , si definisce **combinazione lineare** dei p vettori l'espressione:

$$c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 + \dots + c_p \mathbf{x}_p$$

consideriamo tre vettori di ordine 2 e tre scalari 3, 4, 1.

$$\mathbf{x}_1 = \begin{bmatrix} \mathbf{1} \\ \mathbf{2} \end{bmatrix}; \quad \mathbf{x}_2 = \begin{bmatrix} \mathbf{3} \\ \mathbf{1} \end{bmatrix}; \quad \mathbf{x}_3 = \begin{bmatrix} \mathbf{-2} \\ \mathbf{5} \end{bmatrix}$$

Una combinazione lineare dei 3 vettori con i coefficienti 3, 4 e 1 si ottiene così:

$$3 * \begin{bmatrix} \mathbf{1} \\ \mathbf{2} \end{bmatrix} + 4 * \begin{bmatrix} \mathbf{3} \\ \mathbf{1} \end{bmatrix} + 1 * \begin{bmatrix} \mathbf{-2} \\ \mathbf{5} \end{bmatrix} = \begin{bmatrix} \mathbf{3 + 12 - 2} \\ \mathbf{6 + 4 + 5} \end{bmatrix} = \begin{bmatrix} \mathbf{13} \\ \mathbf{15} \end{bmatrix}$$

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 :

$$X = \begin{bmatrix} X_{1,1} & X_{1,2} & X_{1,3} \\ X_{2,1} & X_{2,2} & X_{2,3} \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ X_{15,1} & X_{15,2} & X_{15,3} \end{bmatrix}$$

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:

$$C = \left[\bar{X}_{.1} \quad \bar{X}_{.2} \quad \bar{X}_{.3} \right]$$

con

$$\bar{X}_{.1} = \frac{\sum_{i=1}^{15} X_{i1}}{15} \quad \bar{X}_{.2} = \frac{\sum_{i=1}^{15} X_{i2}}{15} \quad \bar{X}_{.3} = \frac{\sum_{i=1}^{15} X_{i3}}{15}$$

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:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_1^2 & \mathbf{S}_{12} & \mathbf{S}_{13} \\ \mathbf{S}_{21} & \mathbf{S}_2^2 & \mathbf{S}_{23} \\ \mathbf{S}_{31} & \mathbf{S}_{32} & \mathbf{S}_3^2 \end{bmatrix} \quad \mathbf{S}_j^2 = \frac{\sum_{i=1}^{15} (\mathbf{X}_{i1} - \bar{\mathbf{X}}_{.1})^2}{15}$$

$$\mathbf{S}_{jk} = \frac{\sum_{i=1}^{15} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{.j})(\mathbf{X}_{ik} - \bar{\mathbf{X}}_{.k})}{15}$$

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 :

$$r_{ij} = s_{ij} / (s_i s_j).$$

$$\mathbf{R} = \begin{bmatrix} \mathbf{1} & r_{12} & r_{13} \\ r_{21} & \mathbf{1} & r_{23} \\ r_{31} & r_{32} & \mathbf{1} \end{bmatrix} \quad r_{jk} = \frac{\sum_{i=1}^{15} (X_{ij} - \bar{X}_{.j})(X_{ik} - \bar{X}_{.k}) / 15}{s_j s_k}$$

R ha le stesse proprietà di S.

Calcoli matriciali di indici statistici

Matrice delle correlazioni

$$R = Z'Zn^{-1} = \begin{bmatrix} z_{1,1} & z_{2,1} & \dots & z_{15,1} \\ z_{1,2} & z_{2,2} & \dots & z_{15,2} \\ \dots & \dots & \dots & \dots \\ z_{1,3} & z_{2,3} & \dots & z_{15,3} \end{bmatrix} \begin{bmatrix} z_{1,1} & z_{1,2} & z_{1,3} \\ z_{2,1} & z_{2,2} & z_{2,3} \\ \dots & \dots & \dots \\ z_{15,1} & z_{15,2} & z_{15,3} \end{bmatrix} \begin{bmatrix} 1/n & 0 & 0 \\ 0 & 1/n & 0 \\ 0 & 0 & 1/n \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{\sum_{i=1}^n z_{i1}z_{i1}}{n} & \frac{\sum_{i=1}^n z_{i1}z_{i2}}{n} & \frac{\sum_{i=1}^n z_{i1}z_{i3}}{n} \\ \frac{\sum_{i=1}^n z_{i2}z_{i1}}{n} & \frac{\sum_{i=1}^n z_{i2}z_{i2}}{n} & \frac{\sum_{i=1}^n z_{i2}z_{i3}}{n} \\ \frac{\sum_{i=1}^n z_{i3}z_{i1}}{n} & \frac{\sum_{i=1}^n z_{i3}z_{i2}}{n} & \frac{\sum_{i=1}^n z_{i3}z_{i3}}{n} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}$$



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:

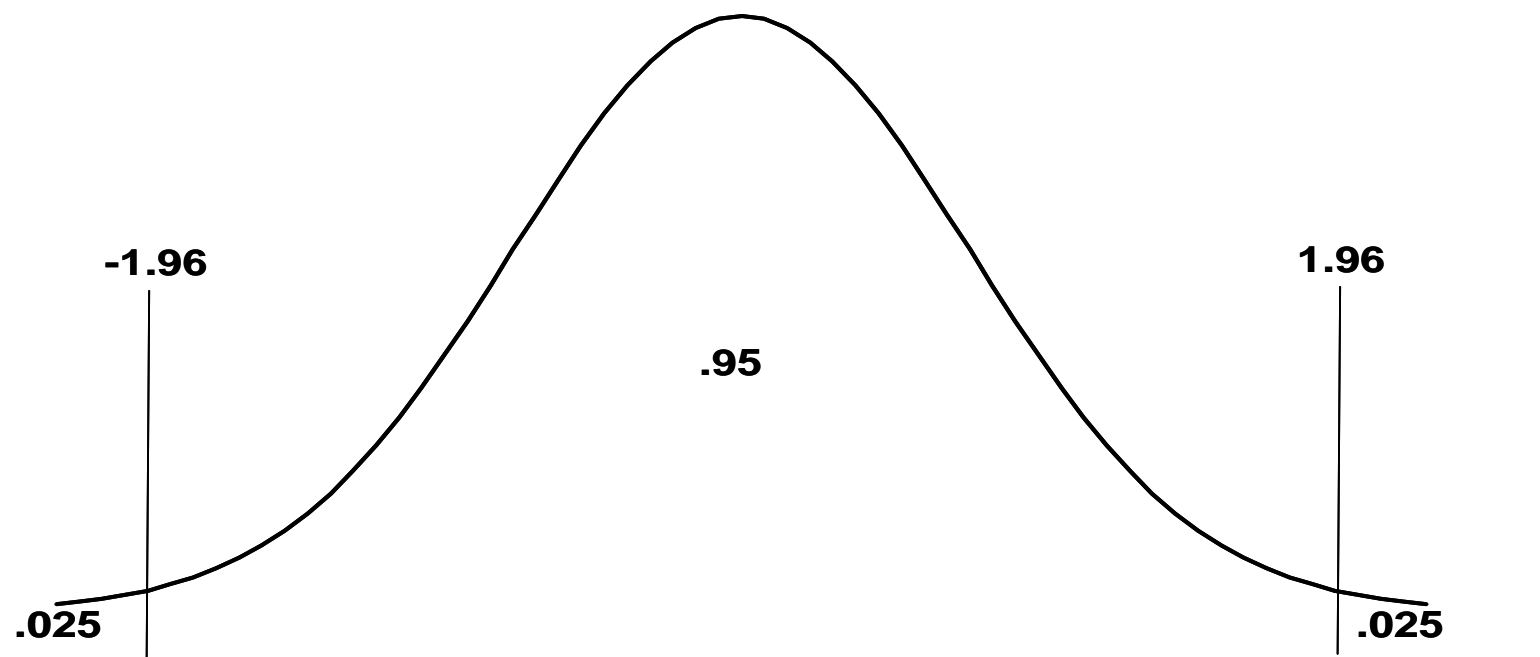
$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Forma della distribuzione

Normale standardizzata

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

$$P\left(\left[\frac{X-V}{\sigma_E}\right] \in [-1.96, 1.96]\right) = .95$$





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

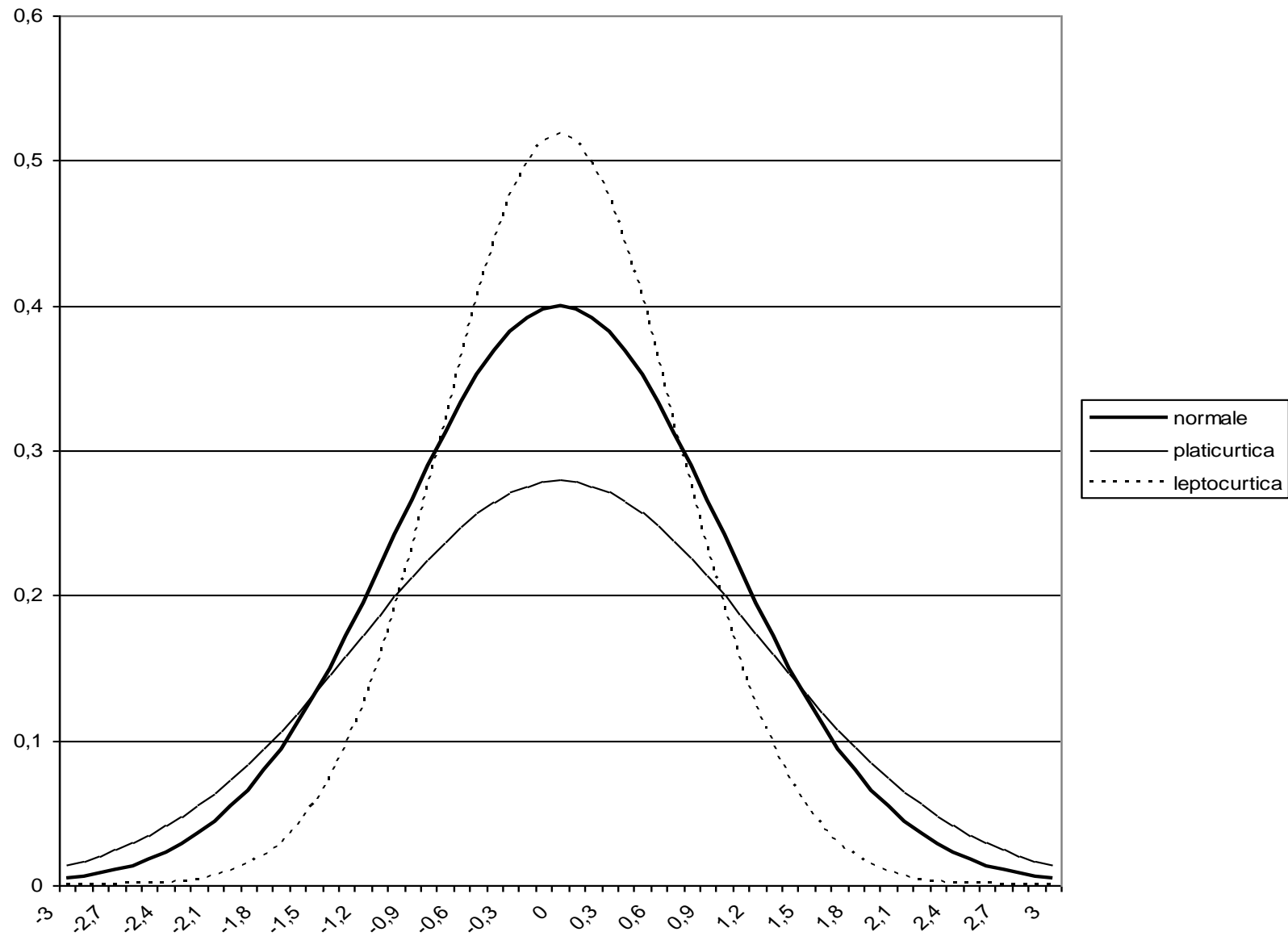
Indici che valutano la forma della distribuzione:
Curtosi e Asimmetria (o *skewness*)

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

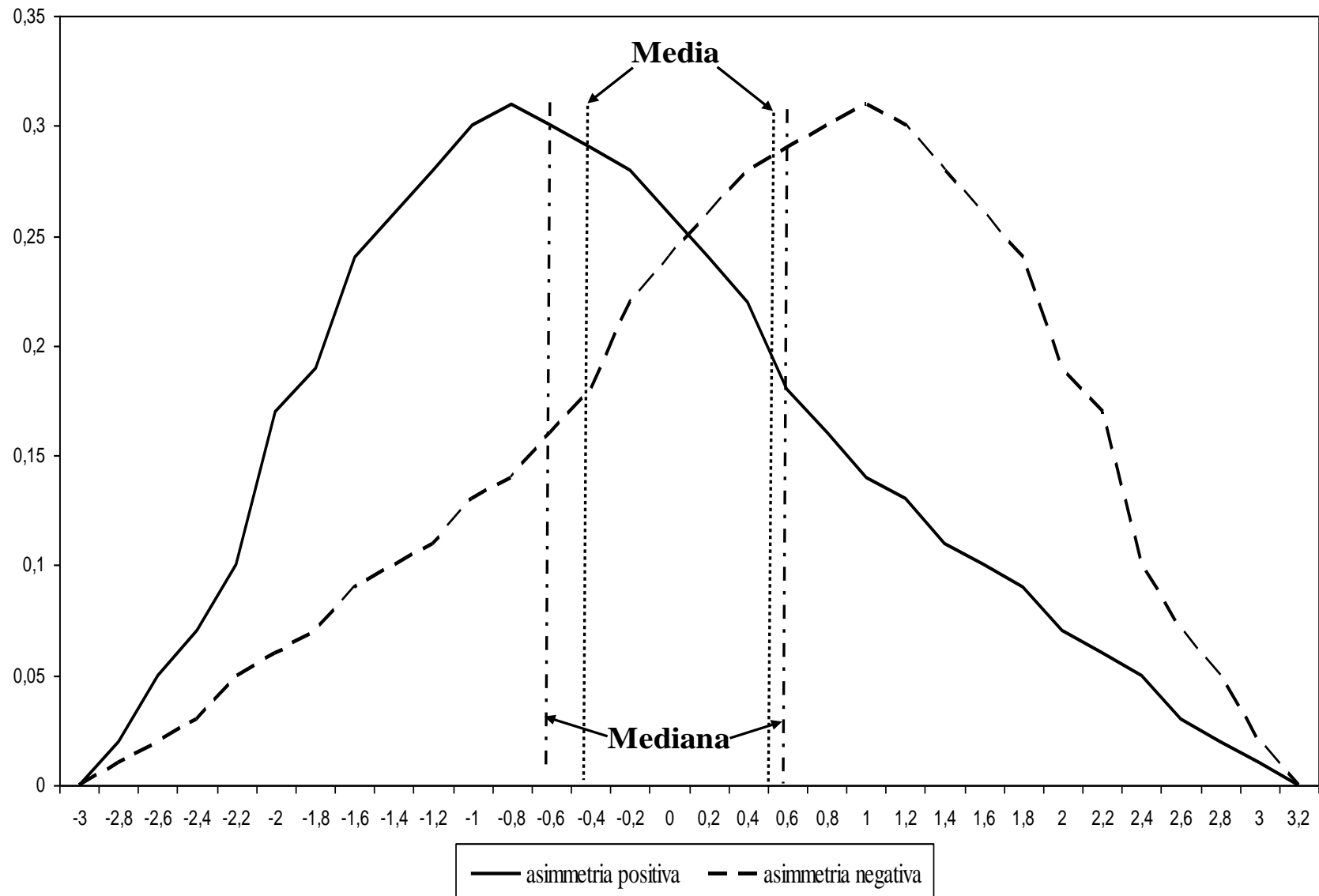
**Indici che valutano la forma della distribuzione:
Curtosi e Asimmetria (o *skewness*)**

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

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

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

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

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.

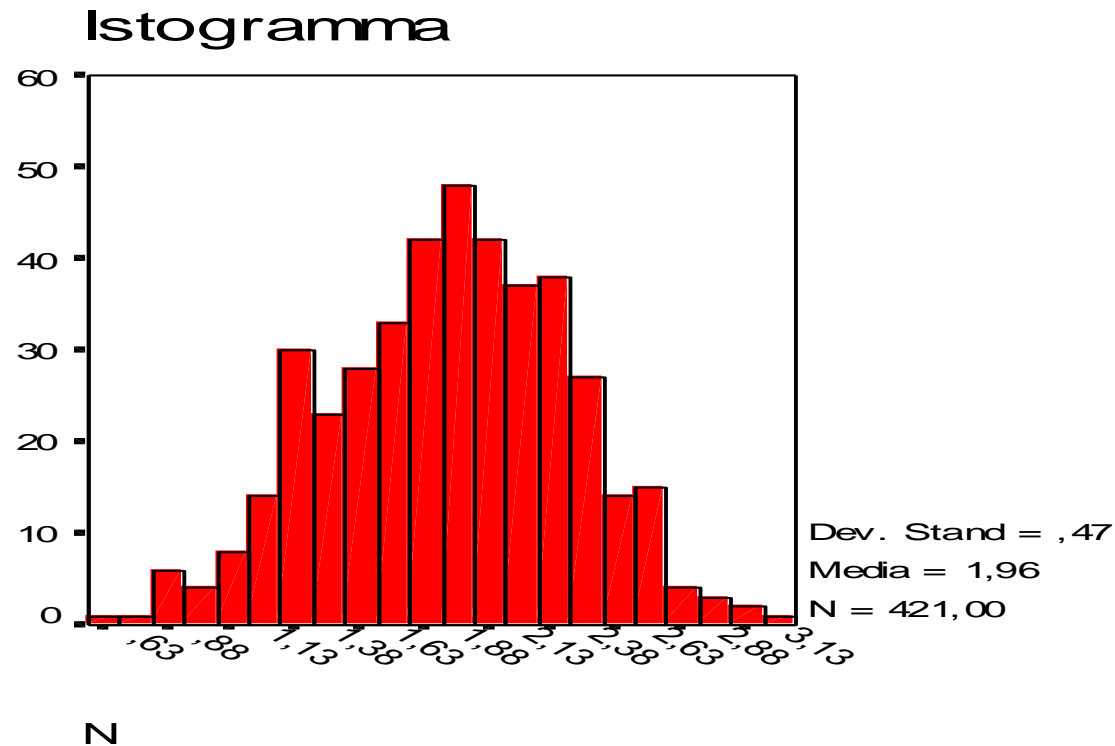
Criterio empirico: accettabili valori compresi tra -1 e 1
Nella verifica delle ipotesi su asimmetria e curtosi utilizzare un livello di alpha più basso ($.01$ o $.001$).

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



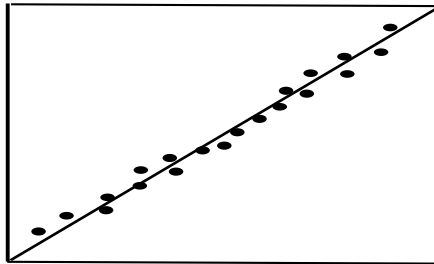


Forma della distribuzione

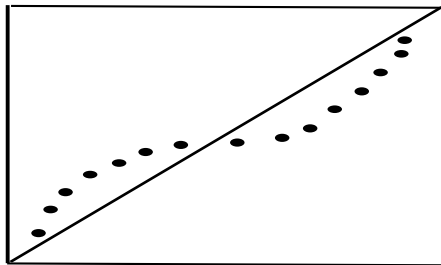
Grafici per l'esame della normalità

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.

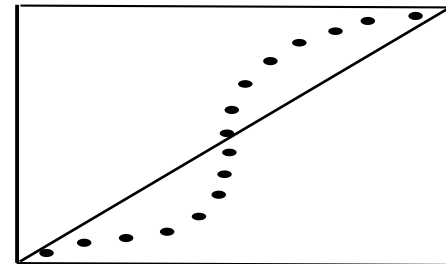
Forma della distribuzione



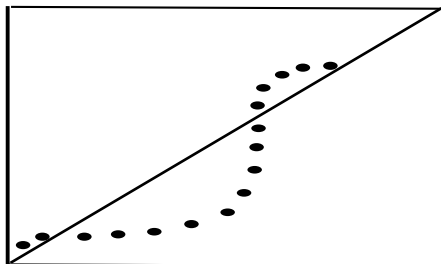
1. Normale



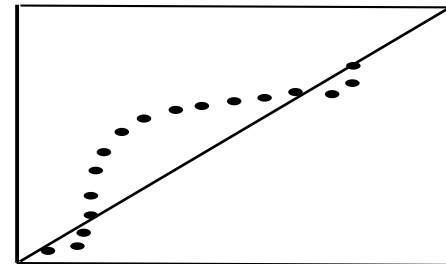
2. Platicurtica



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.

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.

Valori anomali (outliers) univariati

Influenza degli outliers univariati

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;

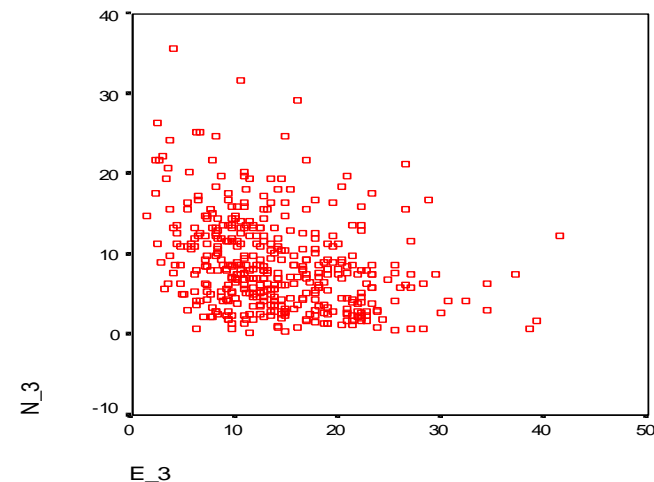
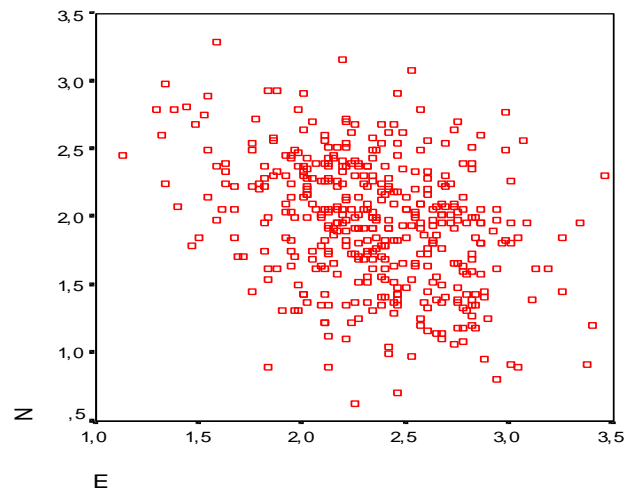
statistiche "robuste" (es., media "*trimmed*" calcolata eliminando il 5% dei casi con punteggi più elevati e più bassi).

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

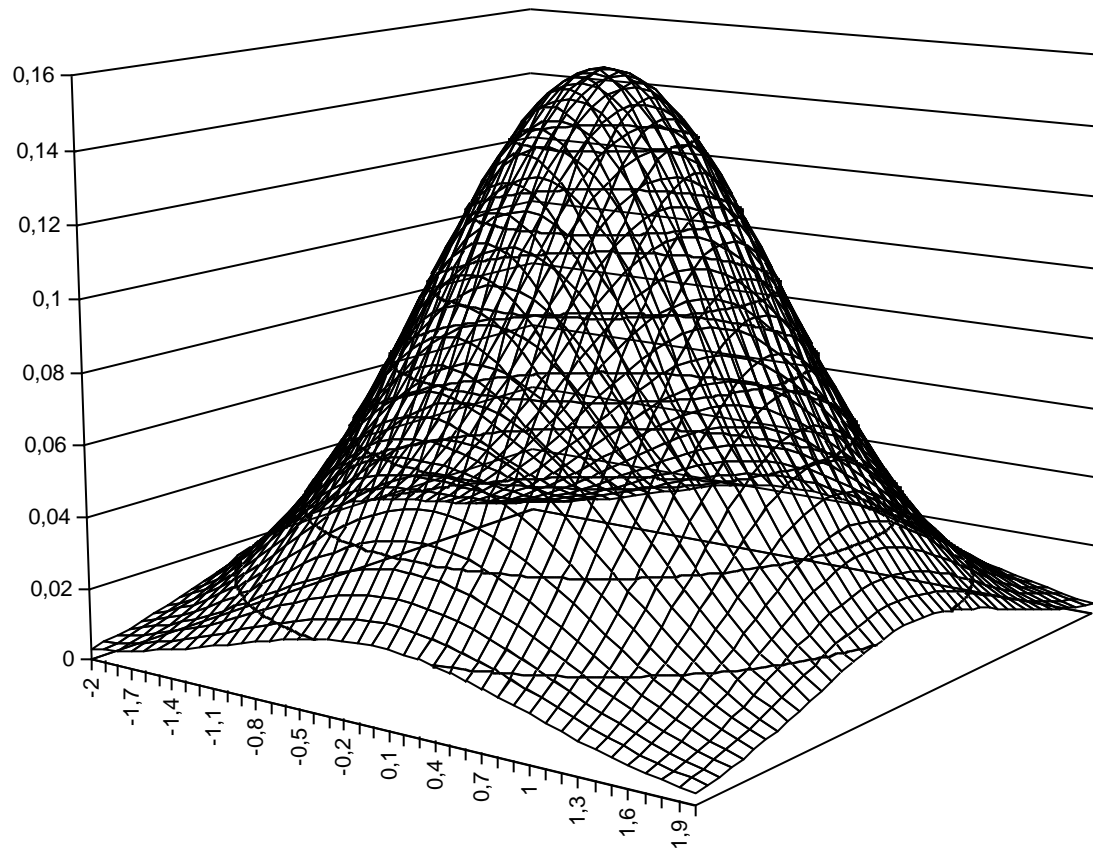
**Non-linearità e non-normalità: fenomeni collegati.
Tecniche per rendere "normale" la distribuzione.**

Problema	Trasformazione [$X^* = f(X)$]
Asimmetria Positiva Estrema (valori >2)	Reciproco: $X^* = 1/X$
Asimmetria Positiva Sostanziale (valori tra 1 e 2)	Logaritmo: $X^* = \text{Log}_{10}(X)$
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 = $X^* = \text{Log}_{10}(K - X)$
Asimmetria Negativa Estrema (valori <-2)	Reciproco = $X^* = 1/(K - X)$

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

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.:

$$f(\mathbf{x}, \mathbf{y}; \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy}) =$$

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

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:

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

dove p è il numero di variabili nel vettore \mathbf{y} , $\boldsymbol{\Sigma}$ è la matrice di varianze e covarianze tra le p variabili, $|\boldsymbol{\Sigma}|$ è il suo determinante, $\boldsymbol{\mu}$ 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:**

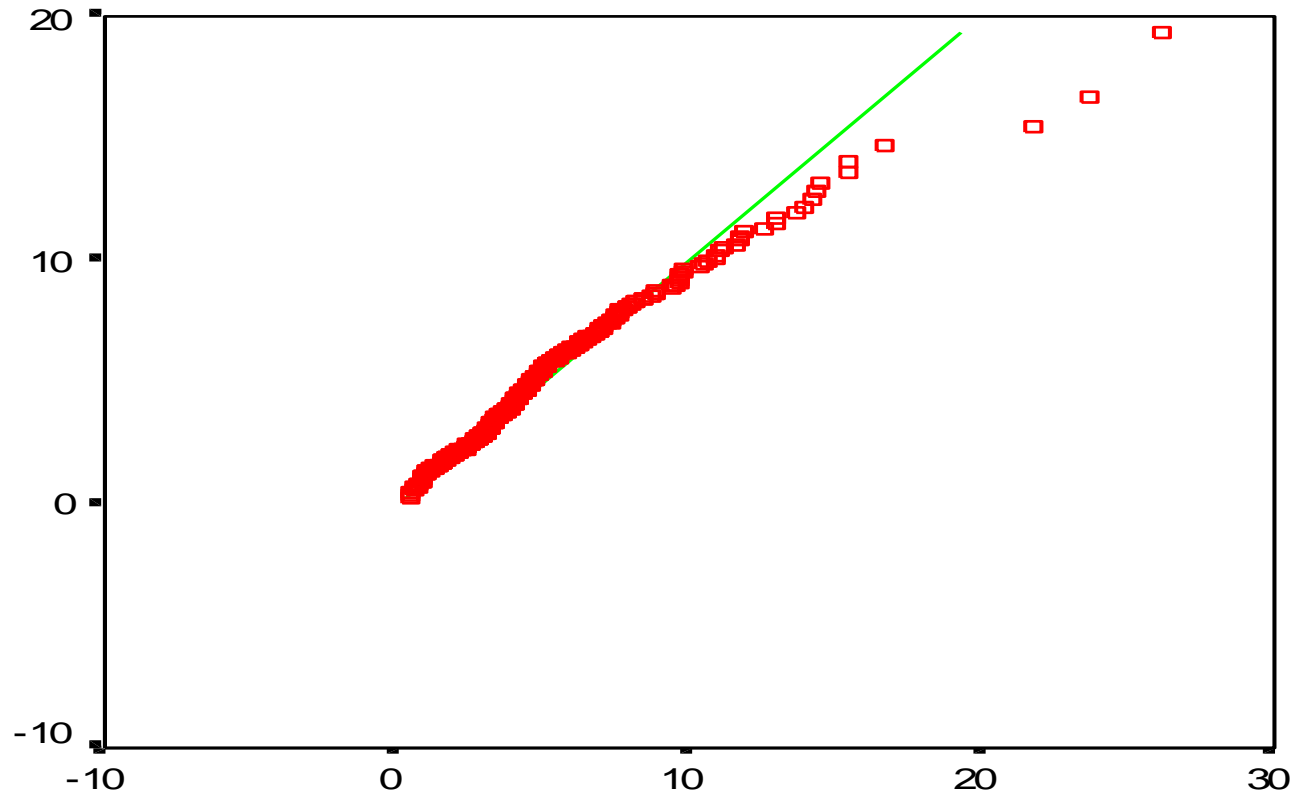
$$D_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}})$$

Rappresenta la distanza del vettore di punteggi di un soggetto (\mathbf{X}_i) dal centroide del campione $\bar{\mathbf{X}}$, pesata per le var/covarianze (\mathbf{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

Grafico Q-Q Chi-quadrato di Mahalan



Valore osservato

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

$$k = \frac{\sum_{i=1}^N (D_i^2)^2}{N}$$

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

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

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

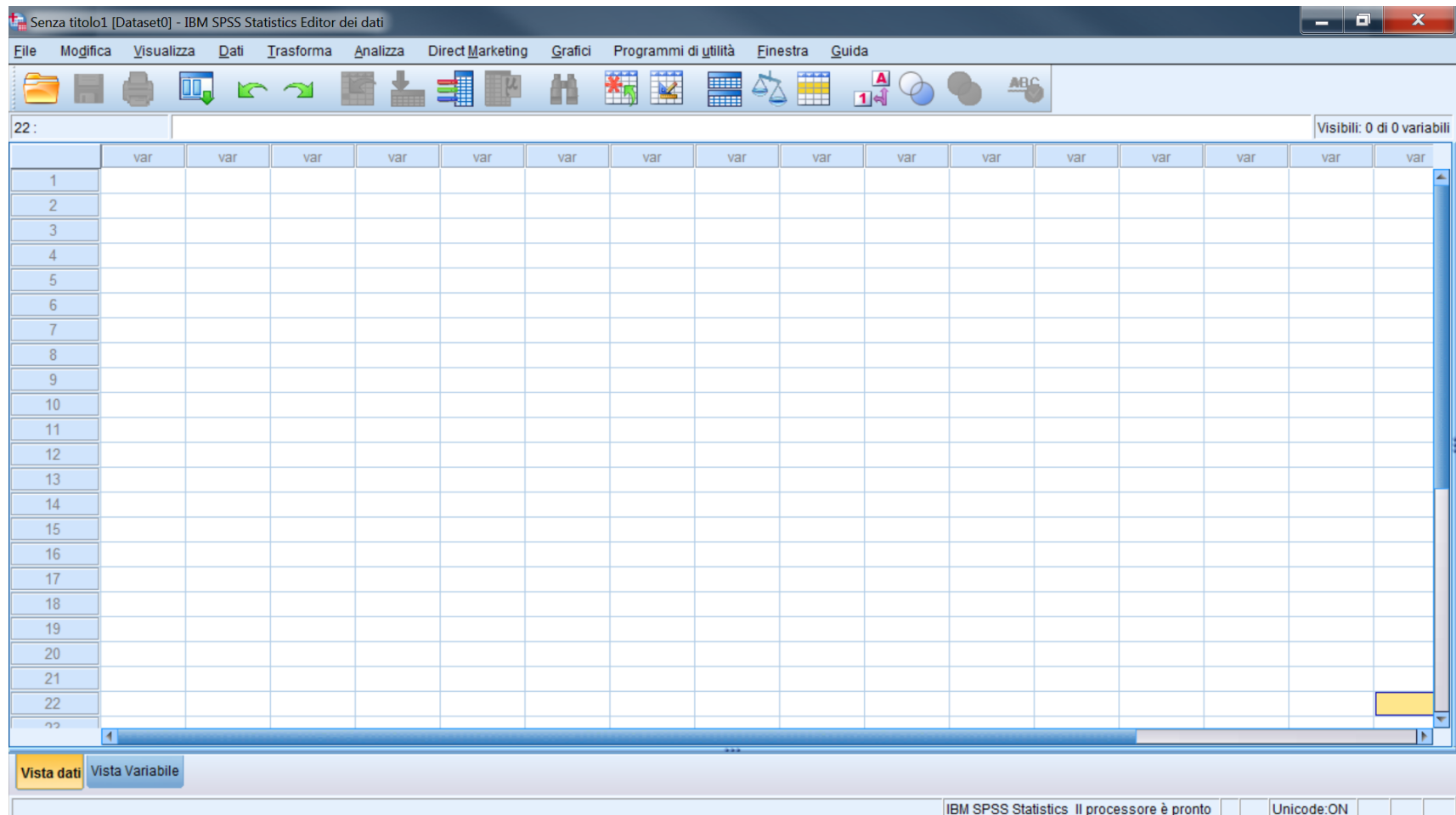
LE FINESTRE DI SPSS

- **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

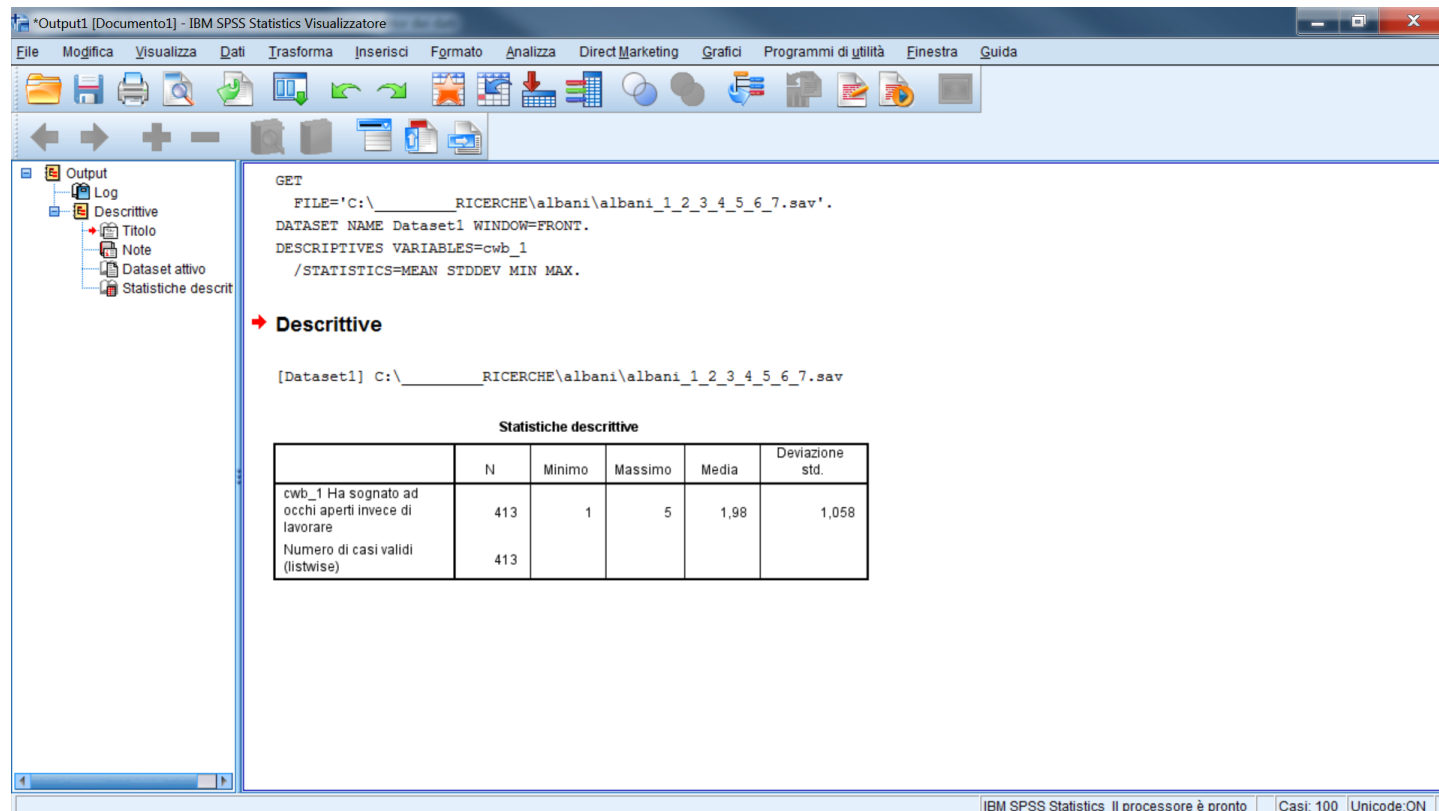
Le celle vuote all'interno dei confini del data file vengono considerate come un valore mancante, ovvero vengono convertite in valori mancanti "di sistema"

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)



The screenshot shows the IBM SPSS Statistics Visualizzatore window. The main area displays the following output:

```

GET
  FILE='C:\_____RICERCHE\albani\albani_1_2_3_4_5_6_7.sav'.
DATASET NAME Dataset1 WINDOW=FRONT.
DESCRIPTIVES VARIABLES=cwb_1
  /STATISTICS=MEAN STDDEV MIN MAX.
  
```

→ **Descrittive**

[Dataset1] C:_____RICERCHE\albani\albani_1_2_3_4_5_6_7.sav

Statistiche descrittive

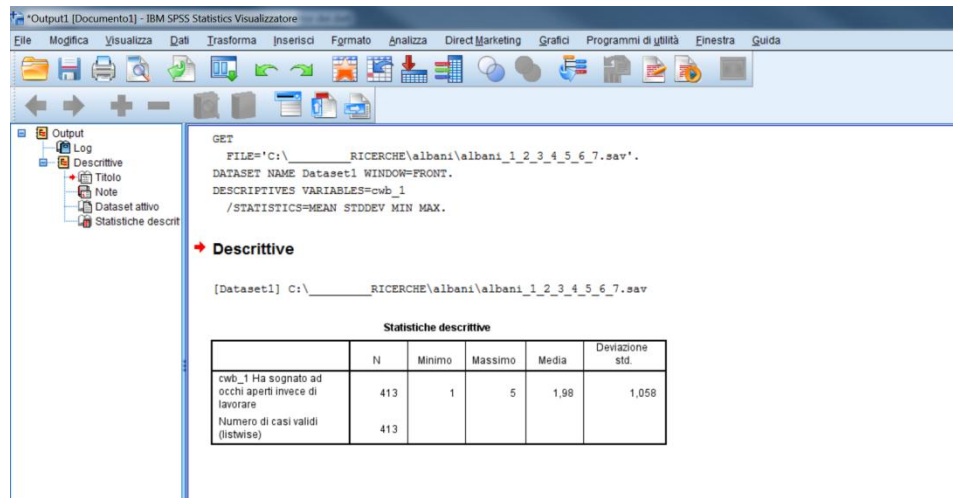
	N	Minimo	Massimo	Media	Deviazione std.
cwb_1 Ha sognato ad occhi aperti invece di lavorare	413	1	5	1,98	1,058
Numero di casi validi (listwise)	413				

At the bottom of the window, the status bar shows: IBM SPSS Statistics Il processore è pronto Cast: 100 Unicode: ON

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.

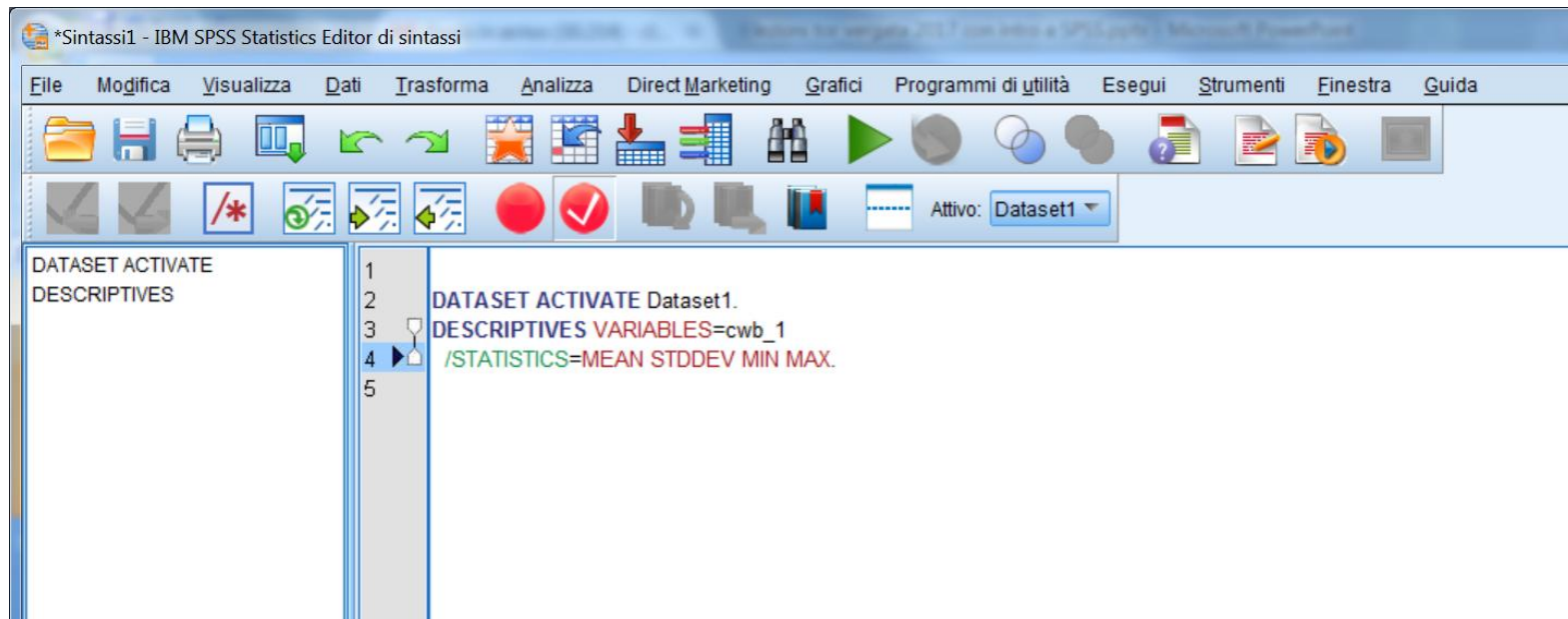
-Il quadro di destra contiene gli elementi veri e propri dell'output (tabelle, grafici e testo).

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



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.

 albani_1_2_3_4_5_6_7.sav [Dataset1] - IBM SPSS Statistics Editor dei dati

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



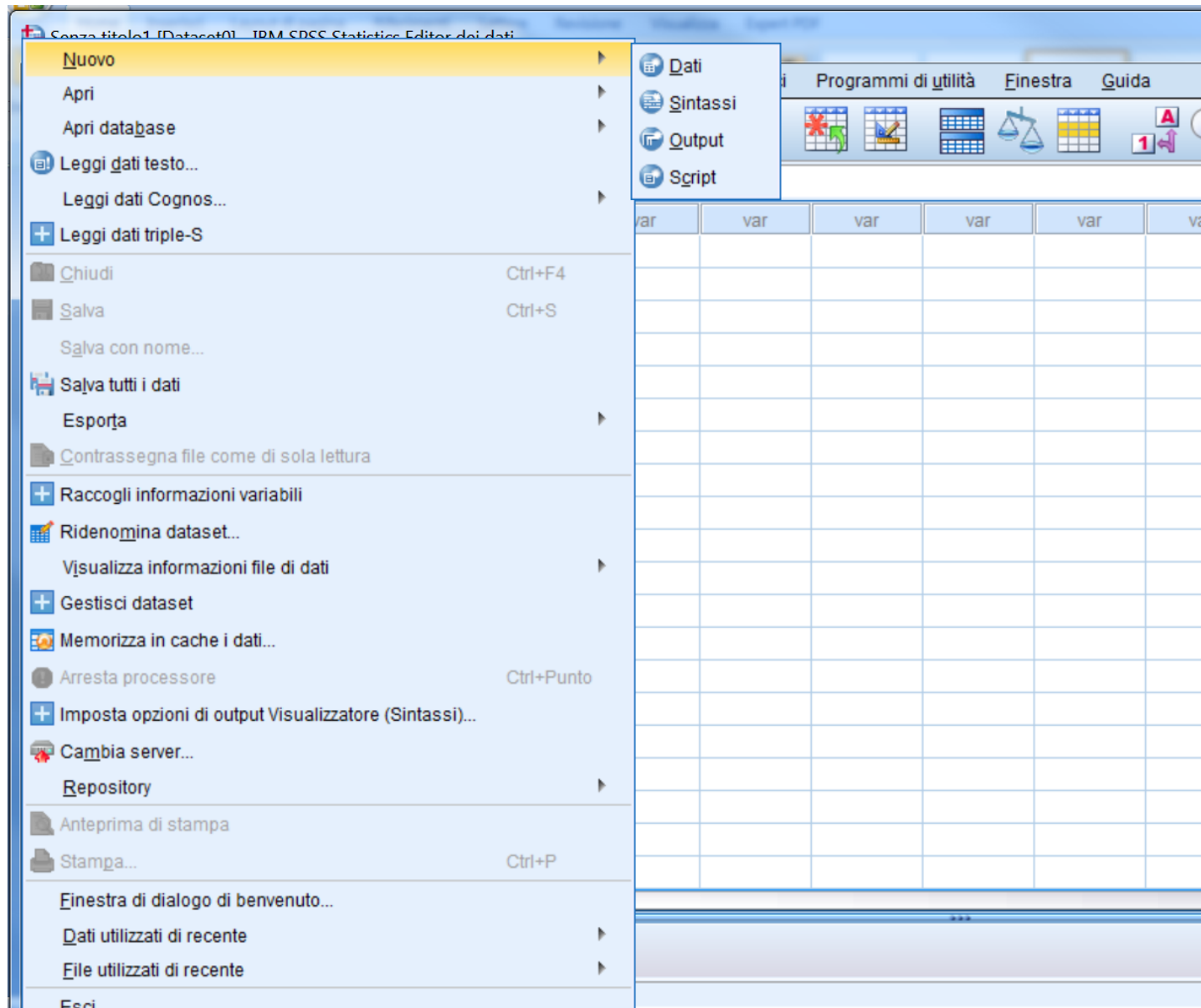
SPSS

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).

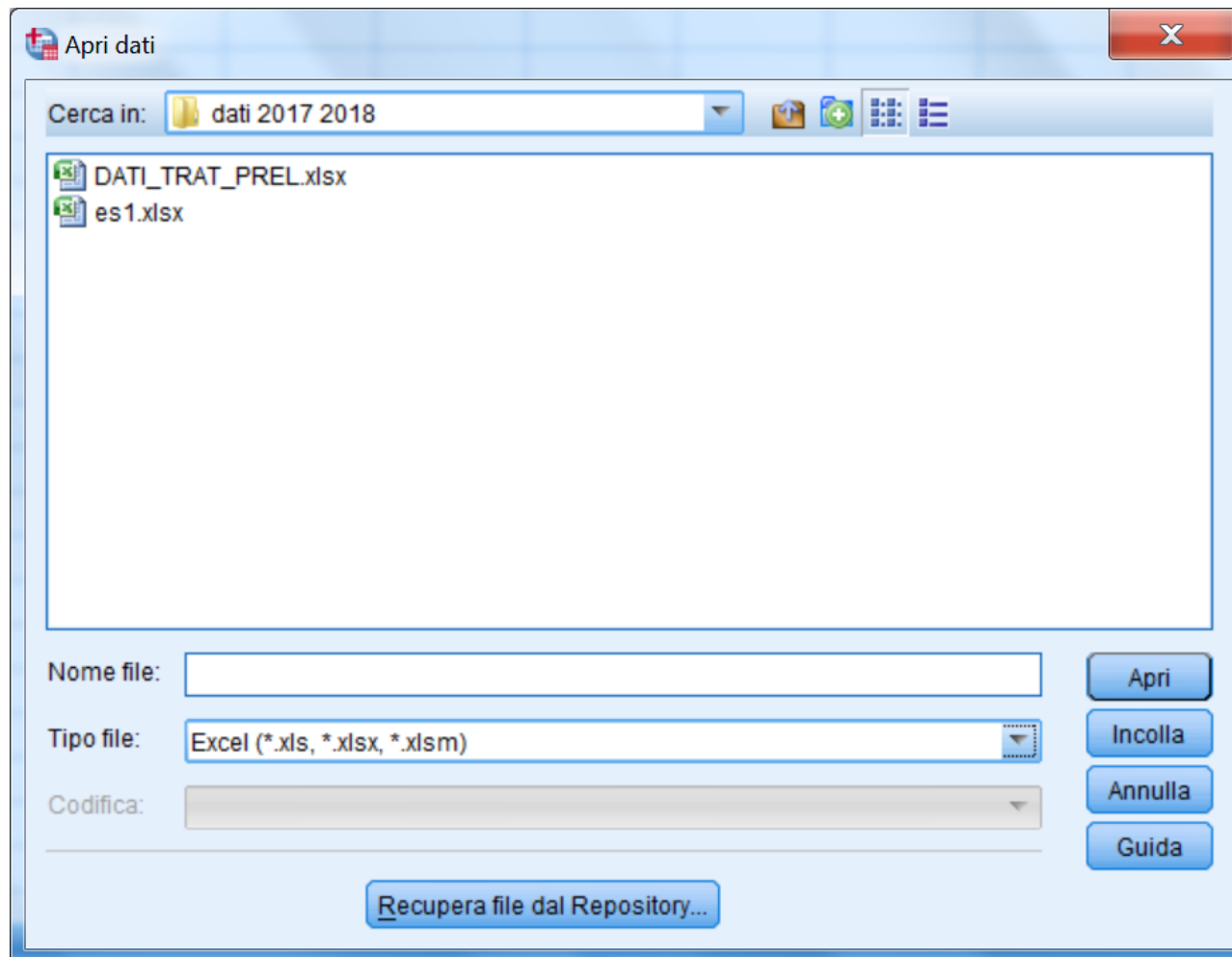


Aprire un file di dati in formato excel

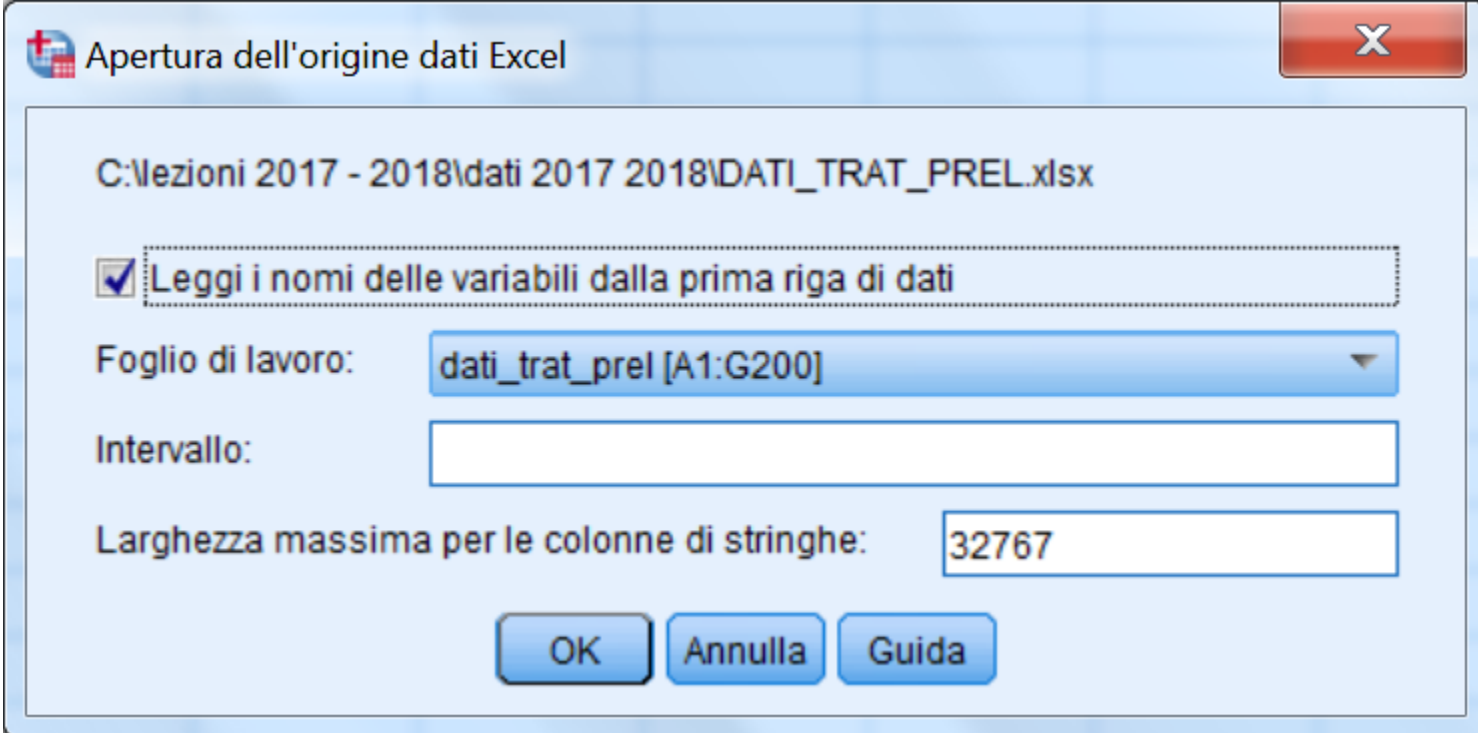
	A	B	C	D	E	F	G	H
1	sex	age	att	ns	contco	compas	int	
2	1	43	16	9	10	2	7	
3	1	30	54	6	3	0	3	
4	1	45	29	4	2	1	4	
5	1	34	30	8	2	0	2	
6	9	99	37	4	2	0	2	
7	2	51	32	2	8	0	6	
8	9	99	31	10	4	2	4	
9	2	28	30	6	2	0	4	
10	2	26	30	8	8	4	2	
11	1	30	42	8	10	0	2	
12	1	51	43	10	8	8	8	
13	1	50	22	8	9	0	4	
14	1	29	34	6	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	
21	2	25	24	4	4	0	2	

DATI_TRAT_PREL.xlsx

Aprire i dati



Leggere i dati



Apertura dell'origine dati Excel

C:\lezioni 2017 - 2018\dati 2017 2018\DATI_TRAT_PREL.xlsx

Leggi i nomi delle variabili dalla prima riga di dati

Foglio di lavoro: dati_trat_prel [A1:G200]

Intervallo:

Larghezza massima per le colonne di stringhe: 32767

OK Annulla Guida

Aprire i dati

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

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

	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

Vista dati Vista Variabile



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

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

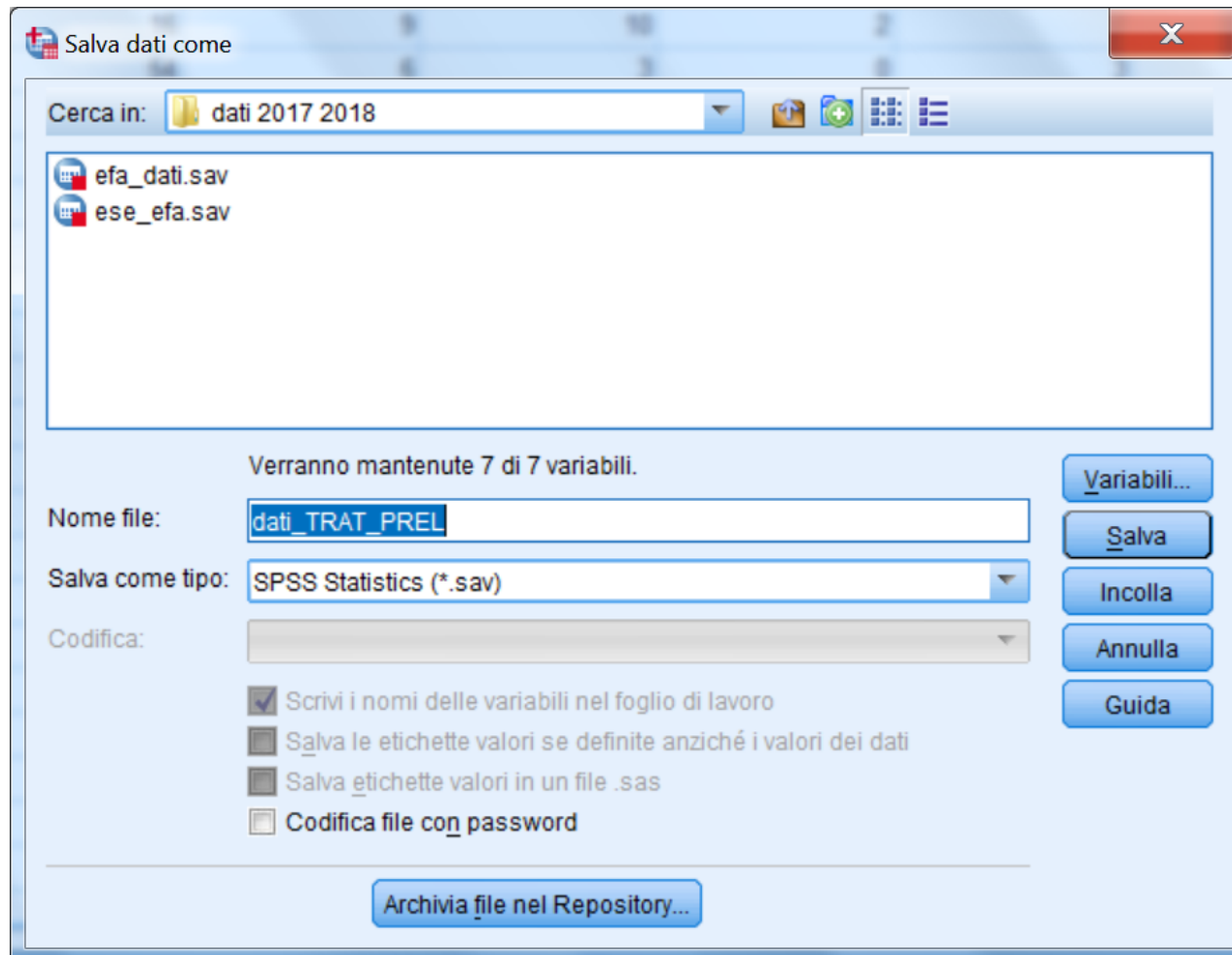
File Modifica Visualizza Dati Trasforma Analizza Direct Marketing Grafici Programmi d

Nuovo ▶
 Apri ▶
 Apri database ▶
 Leggi dati testo...
 Leggi dati Cognos... ▶
 Leggi dati triple-S
 Chiudi Ctrl+F4
 Salva Ctrl+S
 Salva con nome...
 Salva tutti i dati
 Esporta ▶
 Contrassegna file come di sola lettura
 Raccogli informazioni variabili
 Ridenomina dataset...
 Visualizza informazioni file di dati ▶
 Gestisci dataset
 Memorizza in cache i dati...
 Arresta processore Ctrl+Punto
 Imposta opzioni di output Visualizzatore (Sintassi)...
 Cambia server...
 Repository ▶
 Anteprima di stampa

	compas	
10	2	
3	0	
2	1	
2	0	
2	0	
8	0	
4	2	
2	0	
8	4	
10	0	
8	8	
9	0	
2	0	
10	4	
8	1	
10	0	
8	0	
3	1	
10	0	
4	0	

Salvare i dati

Salviamo il nostro file importato da excel per usarlo come file .sav nei prossimi esempi (altrimenti alla chiusura 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".

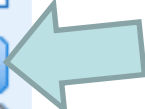
SPSS

Il menu file

Verranno mantenute 7 di 7 variabili.

Nome file:

Salva come tipo:



Salva dati come: Variabili

Solo le variabili selezionate verranno salvate sul file di dati specificato.

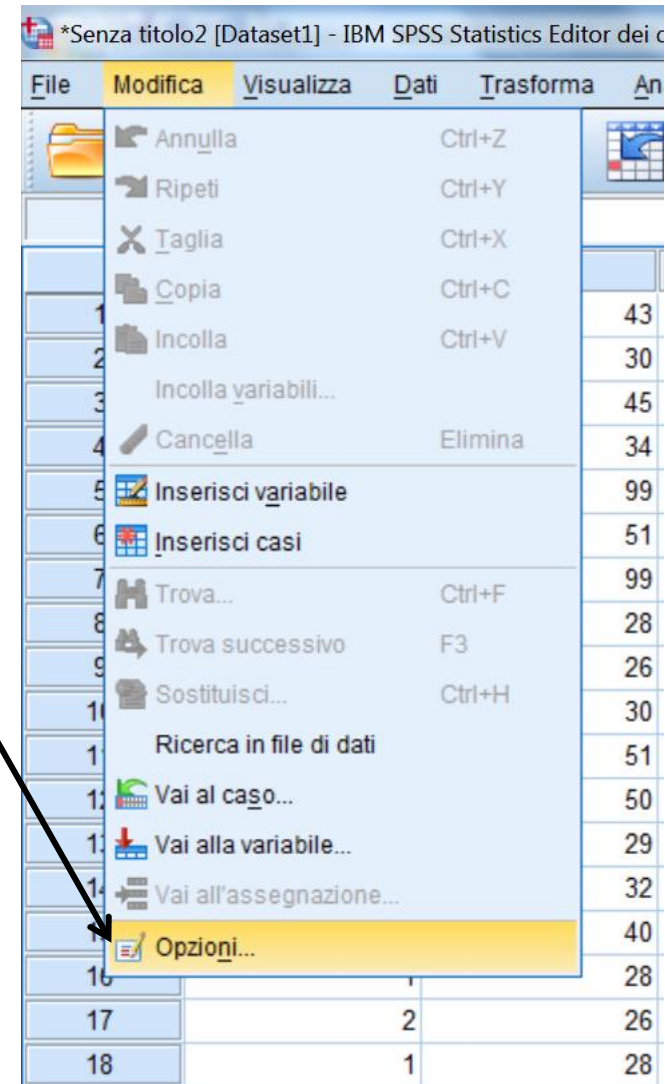
Mantieni	Nome	Etichetta	Ordine
<input checked="" type="checkbox"/>	sex		1
<input checked="" type="checkbox"/>	age		2
<input checked="" type="checkbox"/>	att		3
<input checked="" type="checkbox"/>	ns		4
<input checked="" type="checkbox"/>	contco		5
<input checked="" type="checkbox"/>	compas		6
<input checked="" type="checkbox"/>	int		7

Selezionate: 7 di 7 variabili.

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

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

File Modifica **Visualizza** Dati Trasforma Analizza Direct marketing

- Barra di stato
- Barre degli strumenti
- Editor del menu...
- Caratteri...
- Linee della griglia
- Etichette valori
- Contrassegna dati assegnati
- Personalizza vista Variabile...
- Variabili Ctrl+T

1				16
2				54
3				29
4				30
5				37
6				32
7	9	99		31
8	2	28		30
9	2	26		30
10	1	30		42

Vista dati Vista Variabile

IBM SPSS Statistics Il processore è pronto Unicode:ON

SPSS

Vista dati



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

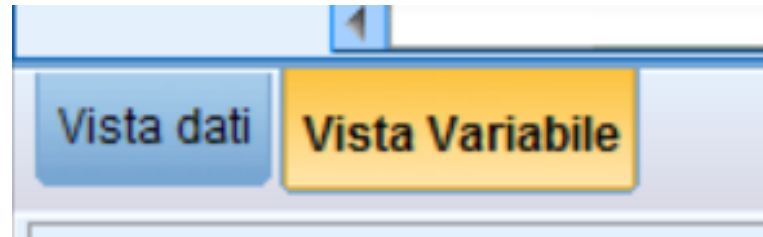
File Modifica Visualizza Dati Trasforma Analizza Direct marketing Grafici Programmi

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

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

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

	Nome	Tipo	Larghezza	Decimali	Etichetta	Valori	Mancante/i	Colonne
1	sex	Numerico	12	0		Nessuno	Nessuno ...	12
2	age	Numerico	12	0		Nessuno	Nessuno	12
3	att	Numerico	12	0		Nessuno	Nessuno	12
4							Nessuno	12
5							Nessuno	12

Valori mancanti

Nessun valore mancante
 Valori mancanti discreti

 Intervallo più un valore mancante discreto facoltativo
 Basso: Alto:
 Valore discreto:

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
1	sex	Numerico	12	0		Nessuno	Nessuno	12	Destra	Nominale
2	age	Numerico	12	0		Nessuno	Nessuno	12	Destra	Scala
3	att	Numerico	12	0		Nessuno	Nessuno	12	Destra	Scala
4	ns	Numerico	12	0						
5	contco	Numerico	12	0						
6	compas	Numerico	12	0						
7	int	Numerico	12	0						
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										

Etichette valori

Etichette valori

Valore: Controllo ortografico...

Etichetta:

1 = "MASCHIO"

SPSS

Il menu dati

Effettua operazioni sulle variabili e sui casi.

The screenshot shows the SPSS 'Data' menu open over a dataset named 'teams.sav'. The menu options include:

- Definisci proprietà variabili...
- Imposta livello di misurazione per sconosciuto...
- Crea etichette valore dai dati
- Copia proprietà dei dati...
- Nuovo attributo personalizzato...
- Definisci data e ora...
- Definisci data dai dati...
- Definisci insiemi a risposta multipla...
- Convalida
- Identifica casi duplicati...
- Identifica casi insoliti...
- Confronta dataset...
- Ordina casi...
- Variabili di ordinamento...
- Trasponi...
- Adatta larghezze stringa tra i file
- Unisci file
- Prodotto cartesiano
- Ristruttura...
- Esegui raking dei pesi...
- Messa in corrispondenza punteggi propensione...
- Corrispondenza controllo casi...
- Aggrega...
- Disegno ortogonale
- Suddividi in file
- Copia dataset
- File suddiviso...
- Seleziona casi...
- Pesa casi...

The data table in the foreground shows the following values for the 'dysfunc' variable:

	dysfunc
1	-,23
2	-,13
3	,00
4	-,33
5	,39
6	1,02
7	-,35
8	-,23
9	,39
10	-,08
11	-,23
12	,09
13	-,29
14	-,06
15	,27
16	,18
17	,38
18	,43
19	,03
20	-,50

The background shows a Microsoft Word document titled 'e.tabelle.doc [Modalità di compatibilità] - Microsoft Word' with a grid of cells, some containing the text 'var'.

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, insieme 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:

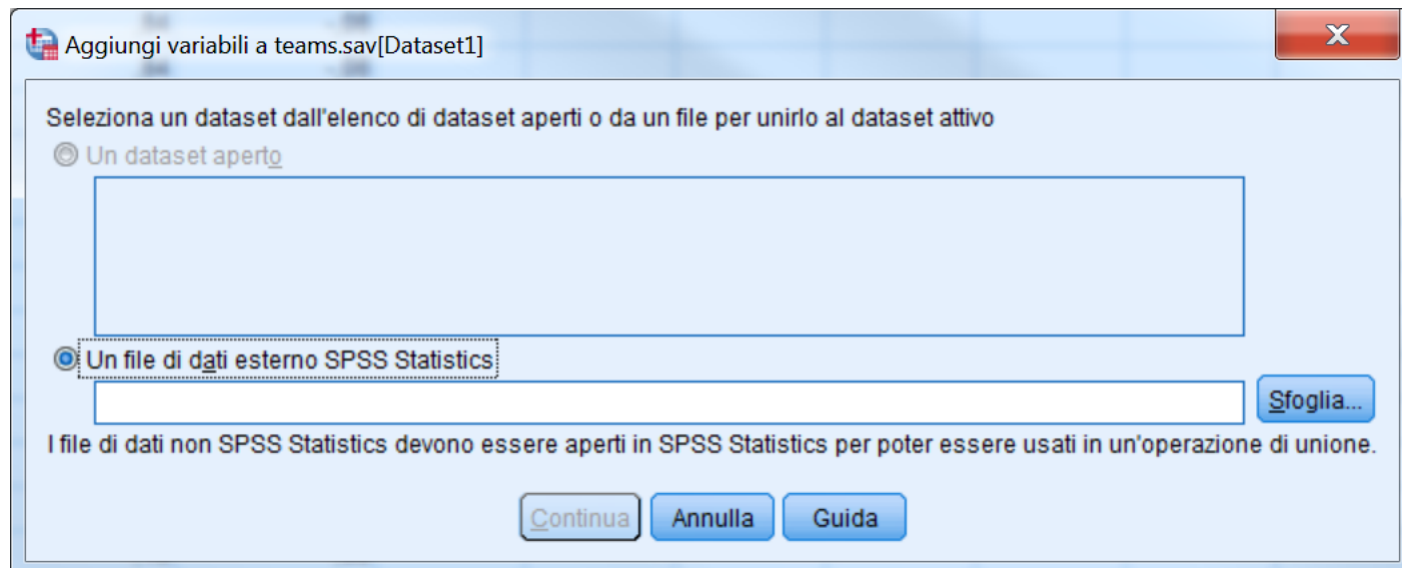
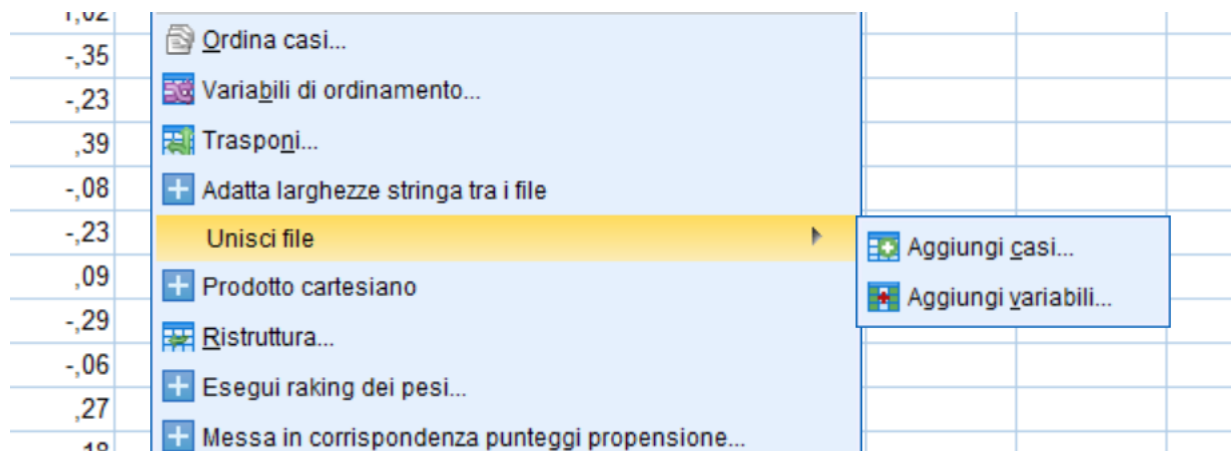
Aggiungi casi e Aggiungi variabili.

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:

“***Selezione 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

The image shows two overlapping dialog boxes in SPSS. The background dialog is 'Seleziona casi' (Select Cases) and the foreground dialog is 'Seleziona casi: Se...' (Select Cases: If...).

Seleziona casi (Background Dialog):

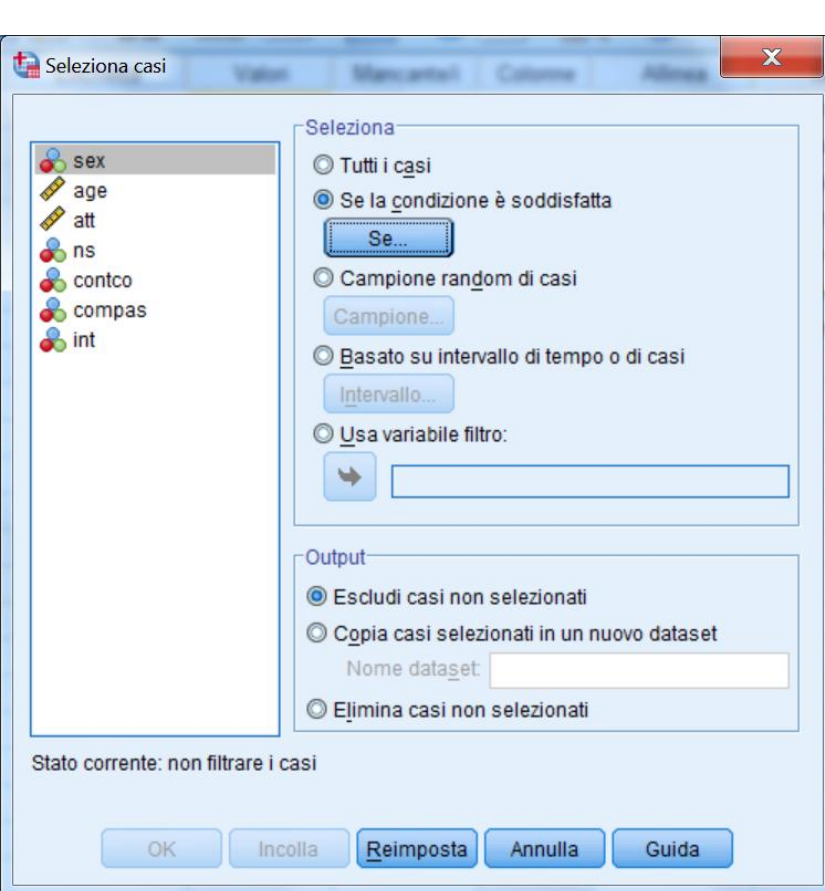
- Seleziona:** Tutti i casi, Se la condizione è soddisfatta, Campione random di..., Basato su intervallo di..., Usa variabile filtro: []
- Output:** Escludi casi non selezionati, Copia casi selezionati in un nuovo dataset (Nome dataset: []), Elimina i casi non selezionati
- Stato corrente: non filtrare i casi
- Buttons: OK, Incolla, Reimposta, Annulla

Seleziona casi: Se... (Foreground Dialog):

- Condition: `sex = 1`
- Buttons: +, <, >, 7, 8, 9, -, <=, >=, 4, 5, 6, *, =, ~=, 1, 2, 3, /, &, |, 0, ., **, ~, (), Elimina
- Gruppo di funzioni: Tutto, Aritmetico, CDF e CDF noncentrale, Conversione, Data/Ora corrente, Aritmetica data, Creazione data
- Funzioni e variabili speciali: []
- Buttons: Continua, Annulla, Guida

SPSS

Il menu dati



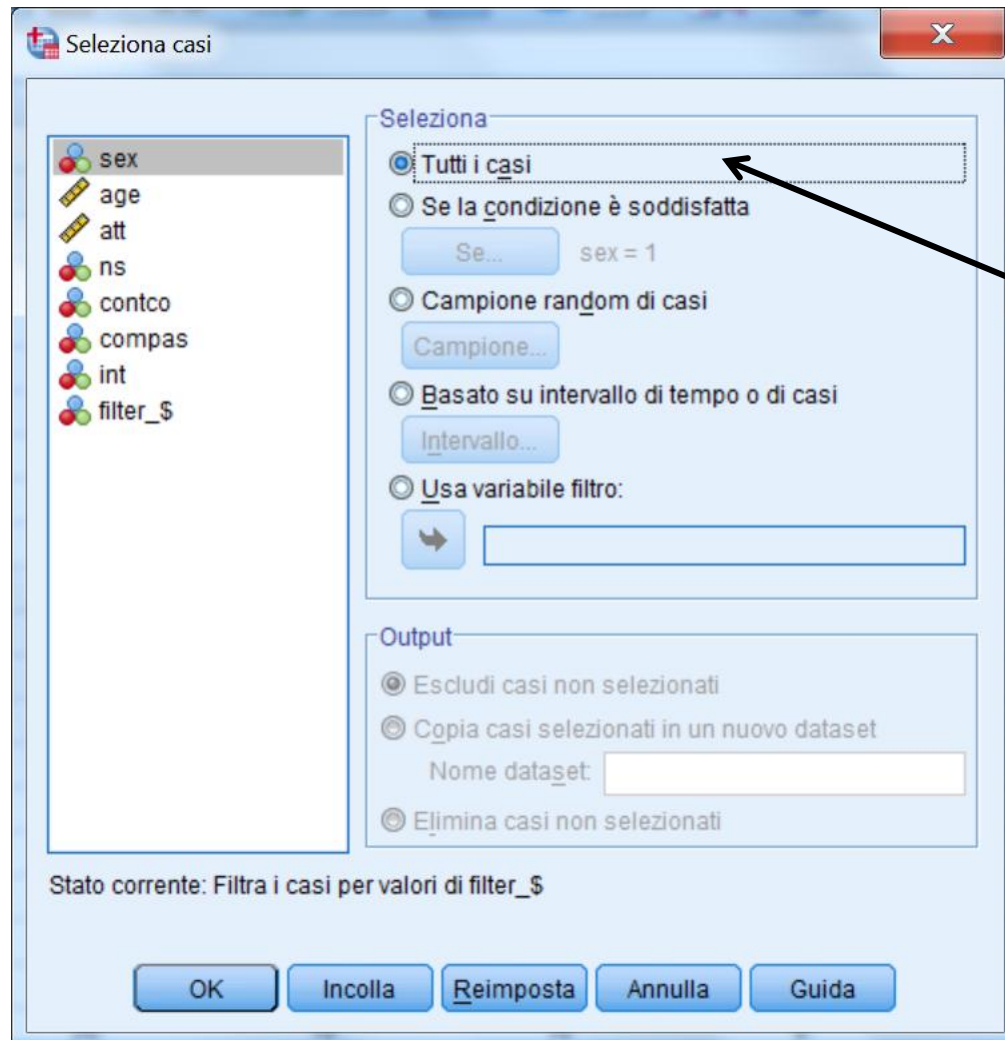
*Senza titolo2 [Dataset1] - IBM SPSS Statistics Editor dei dati

File Modifica Visualizza Dati Trasforma Analizza Direct marketing Grafici

	sex	age	att	ns
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
11	1	51	43	10
12	1	50	22	8
13	1	29	34	6
14	2	32	27	8
15	1	40	50	4
16	1	28	28	6
17	2	26	50	10
18	1	28	32	6
19	2	18	42	4
20	2	25	24	4
21	2	22	50	10

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:

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

The screenshot shows the IBM SPSS Statistics Editor interface. The 'Trasforma' menu is open, displaying various transformation options. The background shows a dataset with a 'sex' variable and a grid of data points.

Case	sex
1	1
2	1
3	1
4	1
5	9
6	2
7	9
8	2
9	2
10	1
11	1
12	1
13	1
14	2
15	1
16	1
17	2
18	1
19	2
20	2
21	2

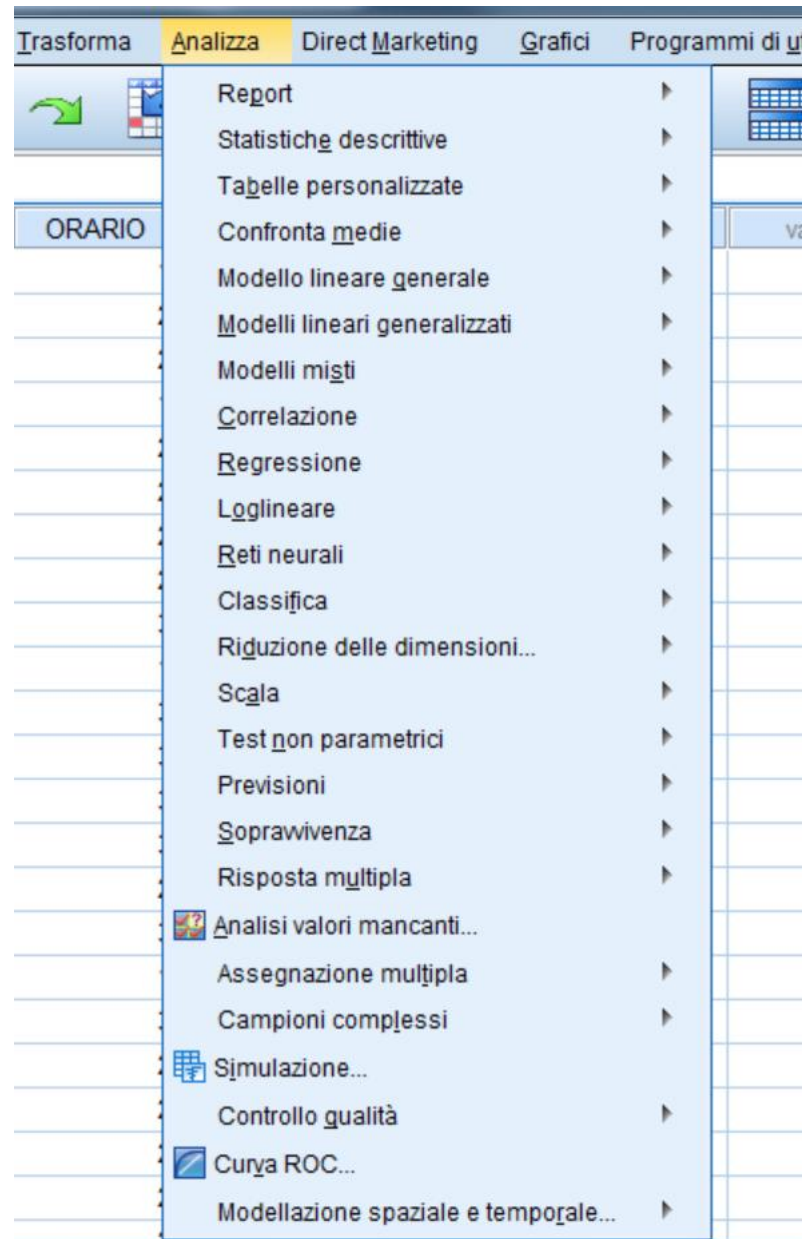
The 'Trasforma' menu options include:

- Calcola variabile...
- Trasformazione programmabilità...
- Conta valori all'interno dei casi...
- Sposta valori...
- Ricodifica nelle stesse variabili...
- Ricodifica in variabili differenti...
- Ricodifica automatica...
- Crea variabili dummy
- Raccolta visuale...
- Raccolta ottimale...
- Prepara dati per modellazione
- Classificazione casi...
- Procedura guidata Data e ora...
- Crea serie storiche...
- Sostituisci valori mancanti...
- Generatori numeri random...
- Esegui trasformazioni in sospeso (Ctrl+G)

SPSS

Il menu Analizza

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

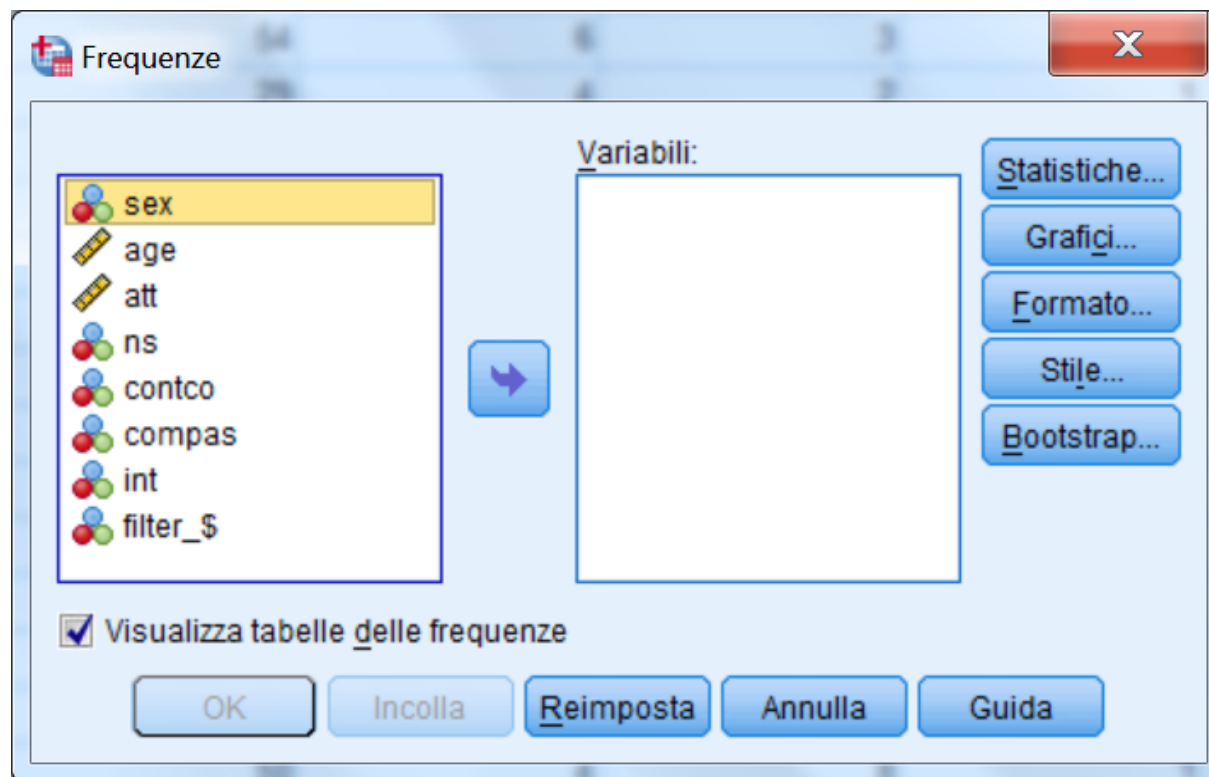


Il menù **ANALIZZA** si trova in tutte le finestre di SPSS

SPSS

Le finestre di Dialogo

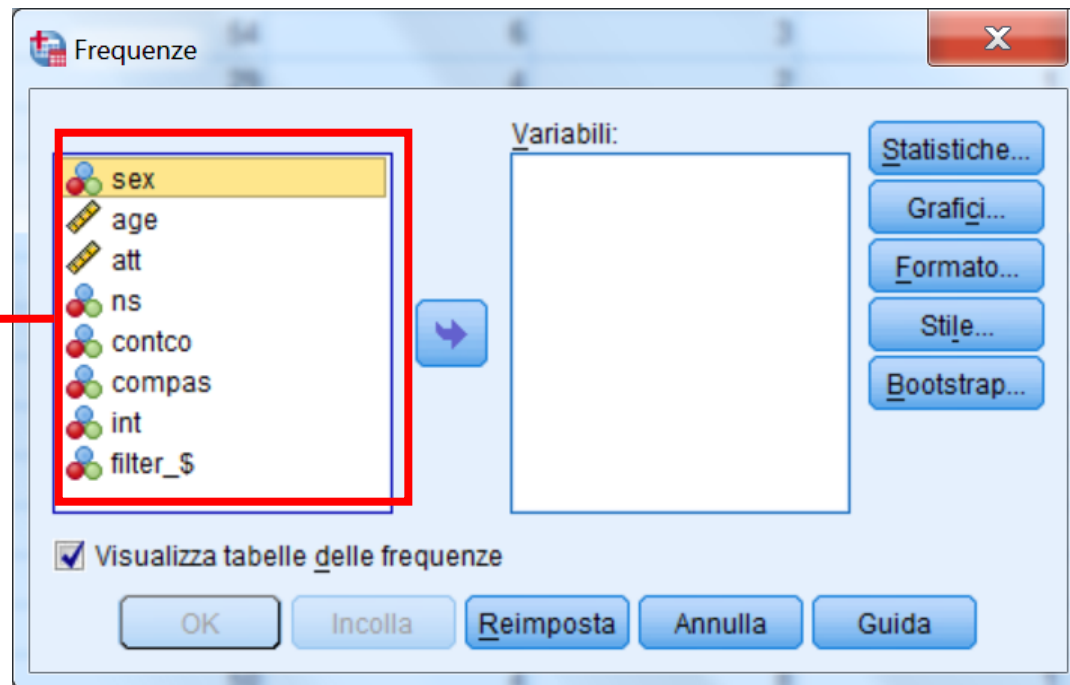
La maggior parte delle opzioni nel menu "Analizza" consentono di aprire delle "finestre di dialogo"



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



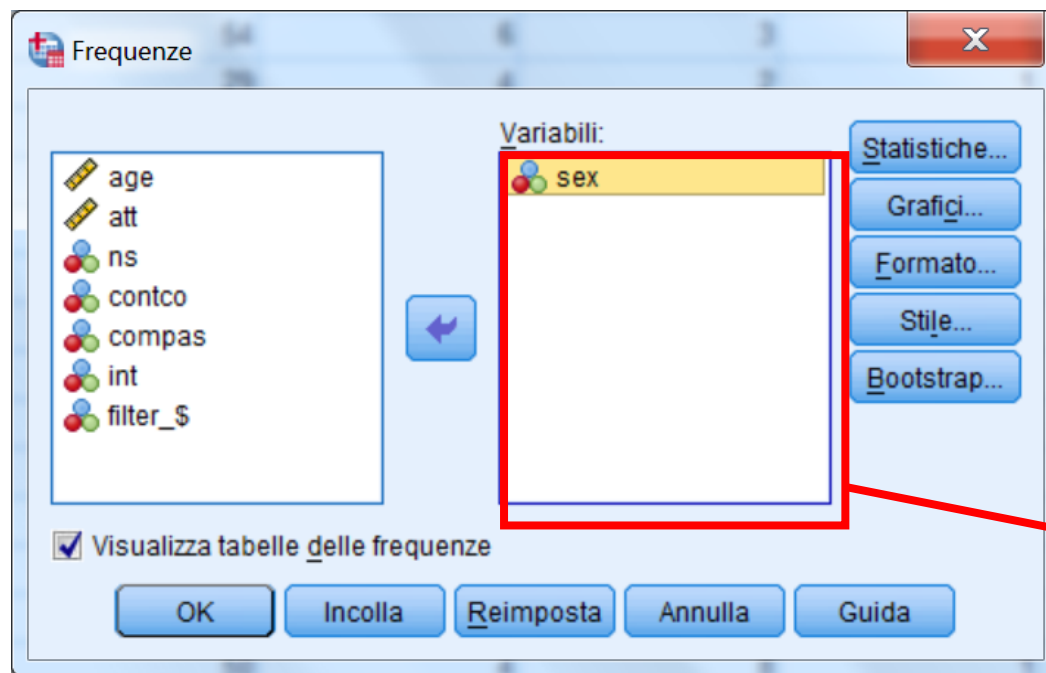
Lista delle variabili origine

- **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

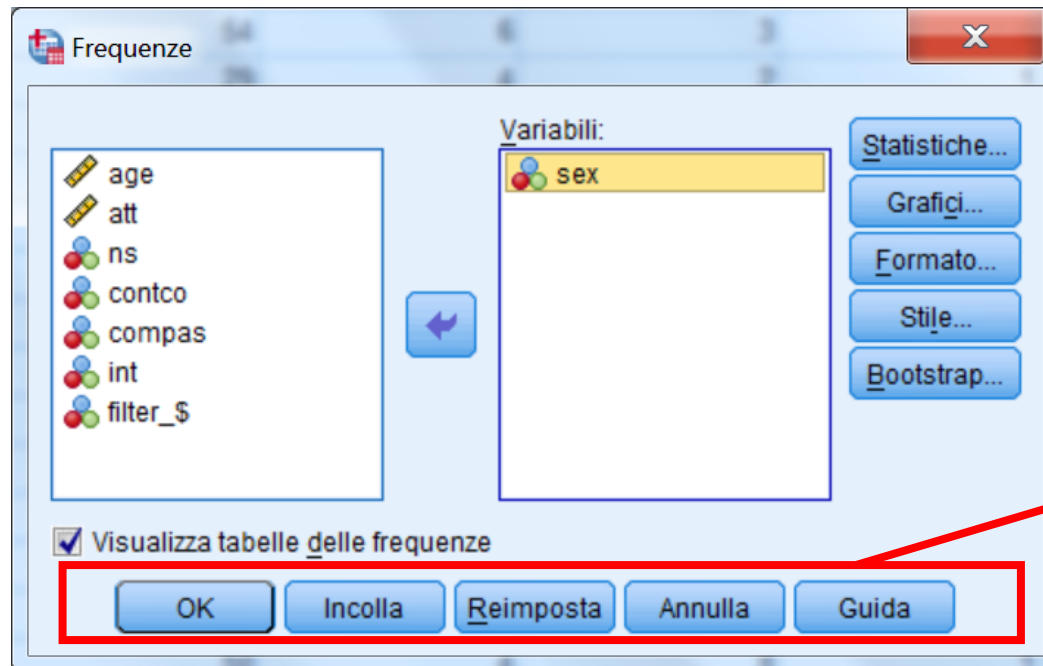


Lista delle
variabili
bersaglio

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



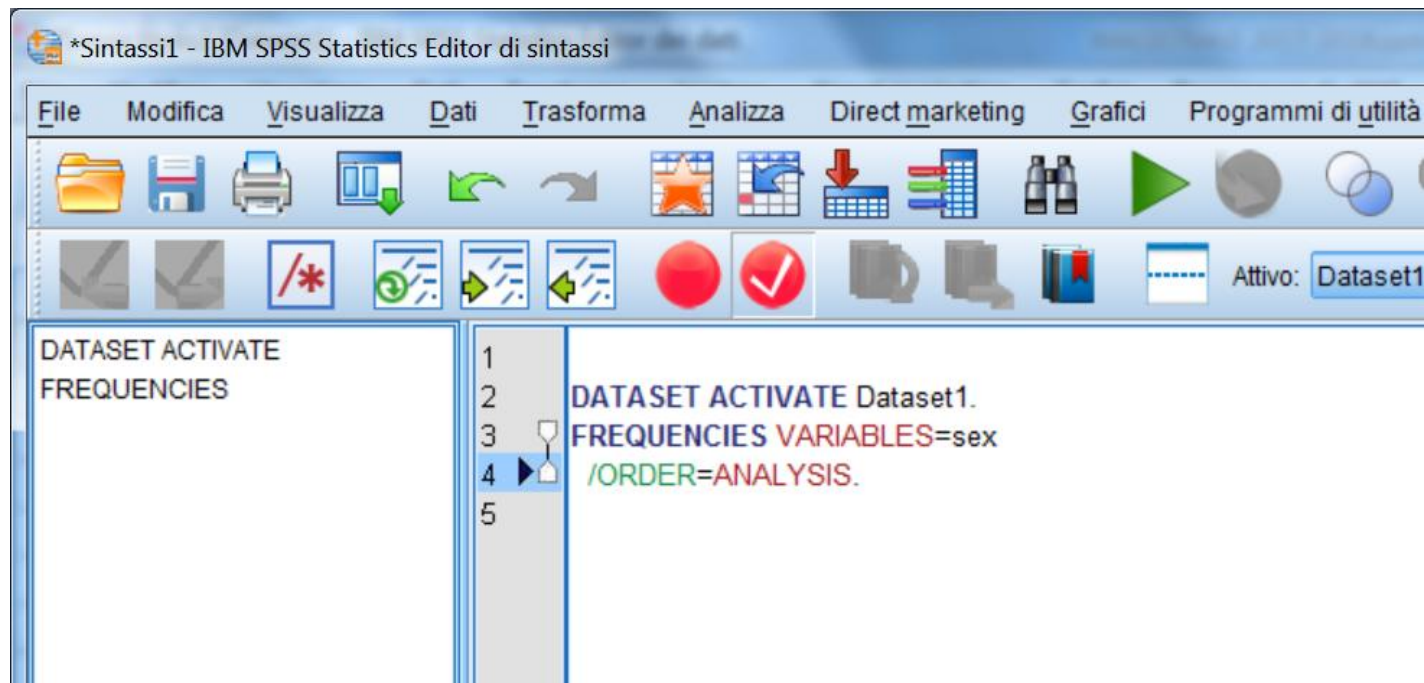
**I Bottoni
dei
comandi**

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

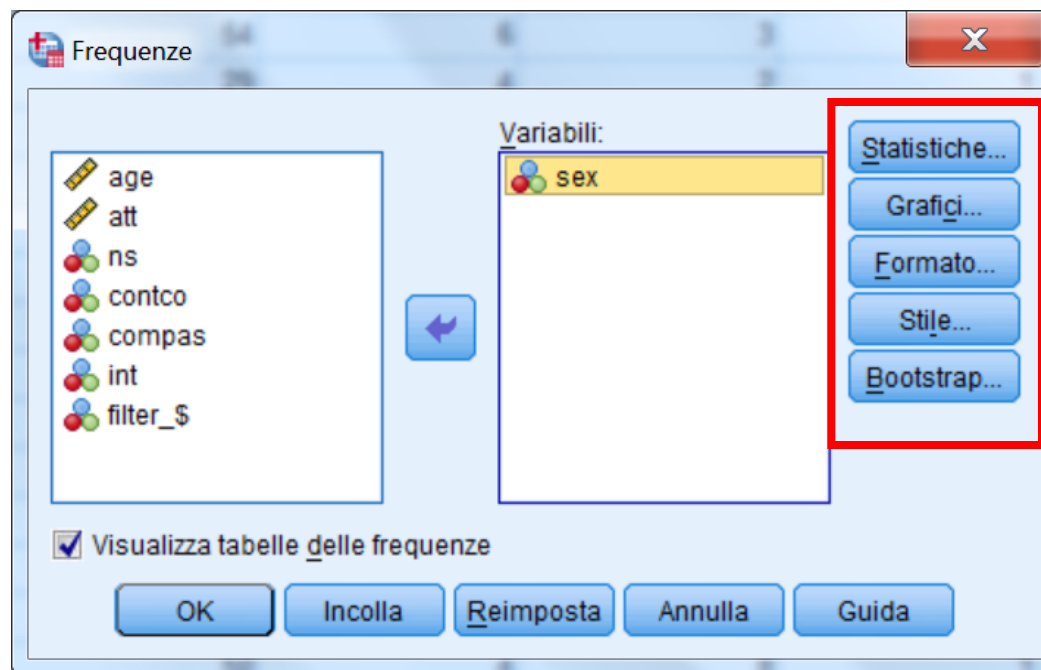
INCOLLA consente di “tradurre” i comandi e le opzioni selezionate nella finestra di dialogo nel linguaggio di programmazione di SPSS.

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:

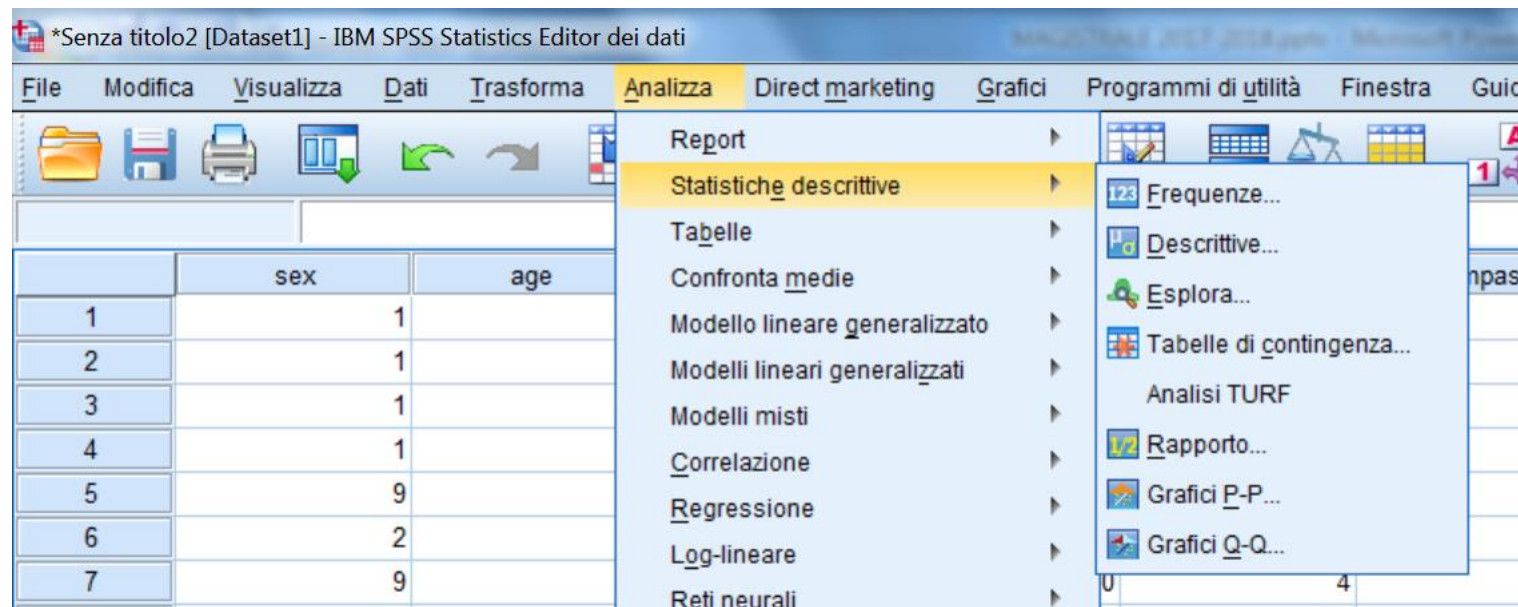
- ci sono errori nei dati, e quindi le variabili assumono valori fuori scala (ad esempio, un item che varia da 1 a 5 ha un punteggio di 8)**
- ci sono "valori anomali" (outliers) ovvero soggetti che presentano valori estremamente elevati in una variabile**
- ci sono casi con valori mancanti**

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**



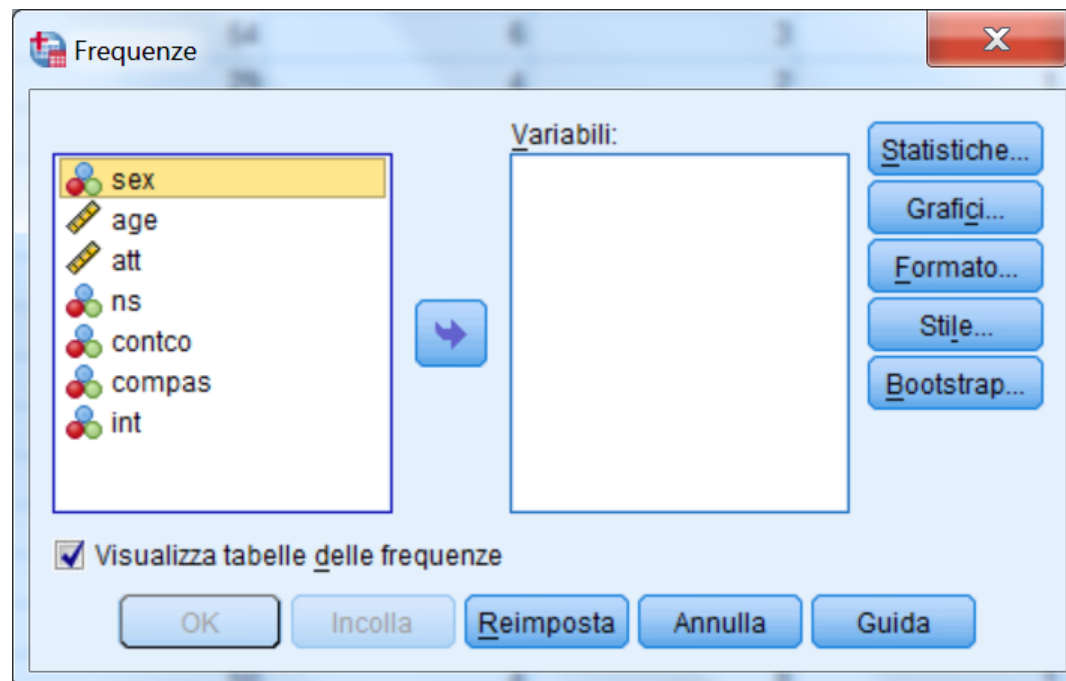
The screenshot shows the IBM SPSS Statistics Editor dei dati interface. The 'Analizza' menu is open, and the 'Statistiche descrittive' option is selected. A sub-menu is open, showing the following options: 'Frequenze...', 'Descrittive...', 'Esplora...', 'Tabelle di contingenza...', 'Analisi TURF', 'Rapporto...', 'Grafici P-P...', and 'Grafici Q-Q...'. The background shows a data table with columns 'sex' and 'age' and rows 1 through 7.

	sex	age
1	1	
2	1	
3	1	
4	1	
5	9	
6	2	
7	9	

SPSS

La procedura Frequenze

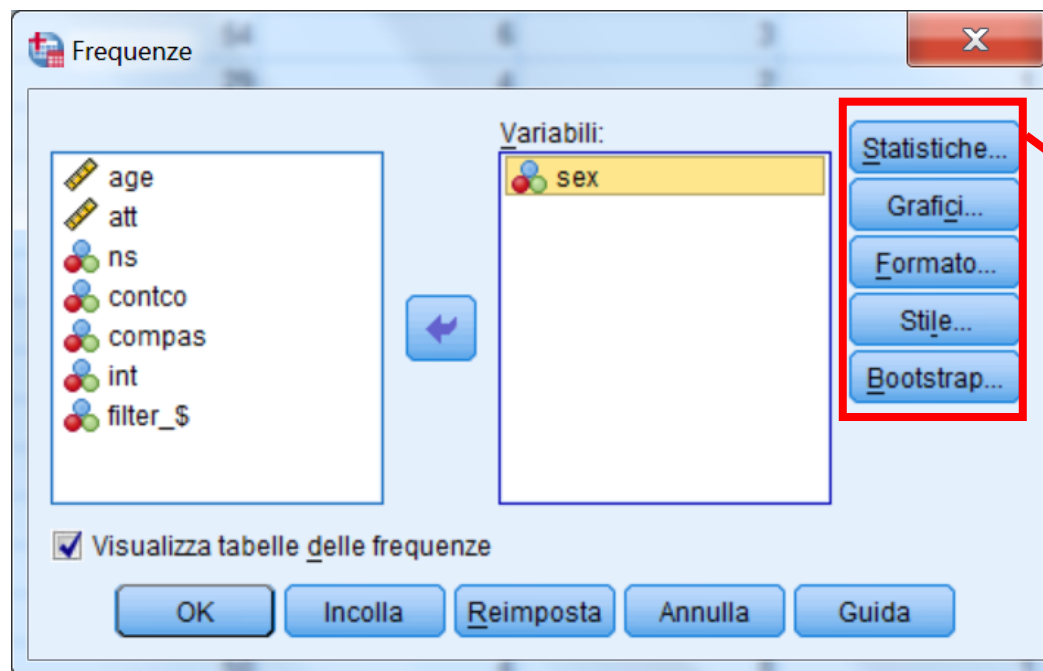
La procedura "Frequenze" consente di effettuare una serie di analisi preliminari, tramite statistiche descrittive e grafici. Selezionando la procedura frequenze si aprirà questa finestra di dialogo:



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).

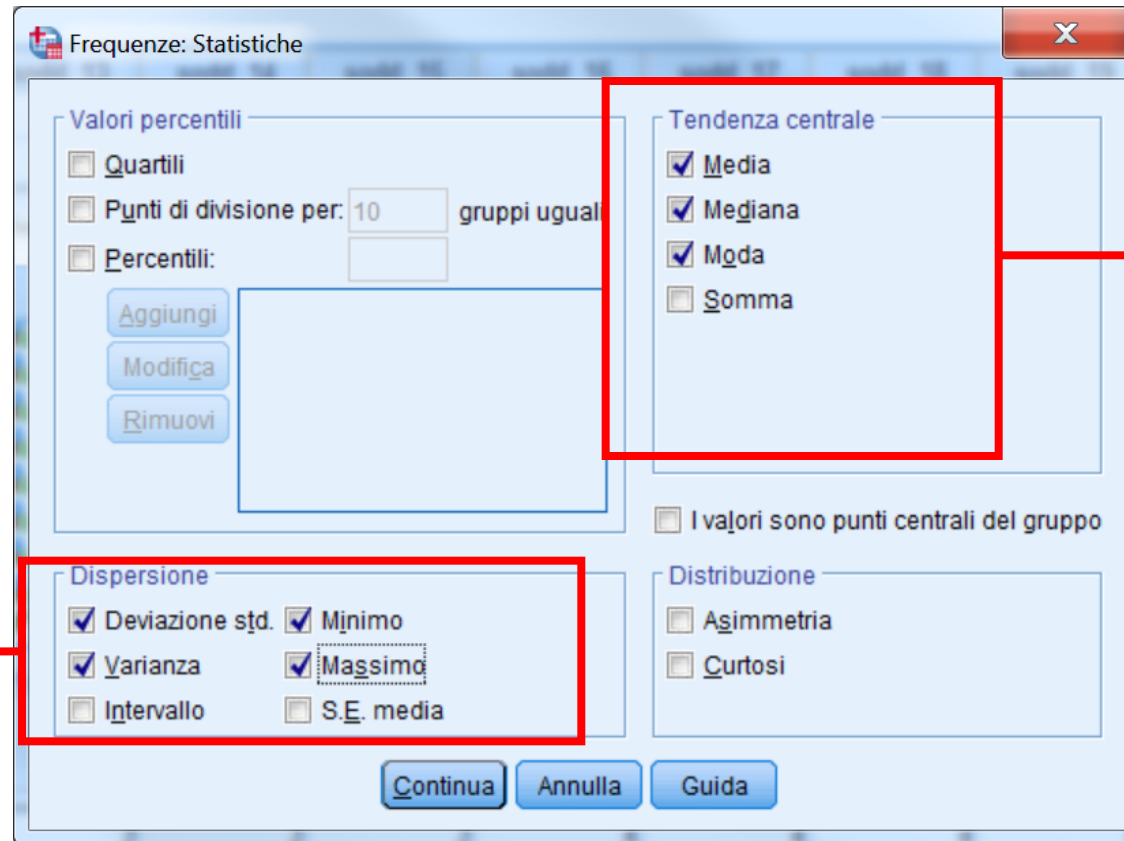


Il pulsante "Statistiche" consente di richiedere una serie di statistiche descrittive

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:



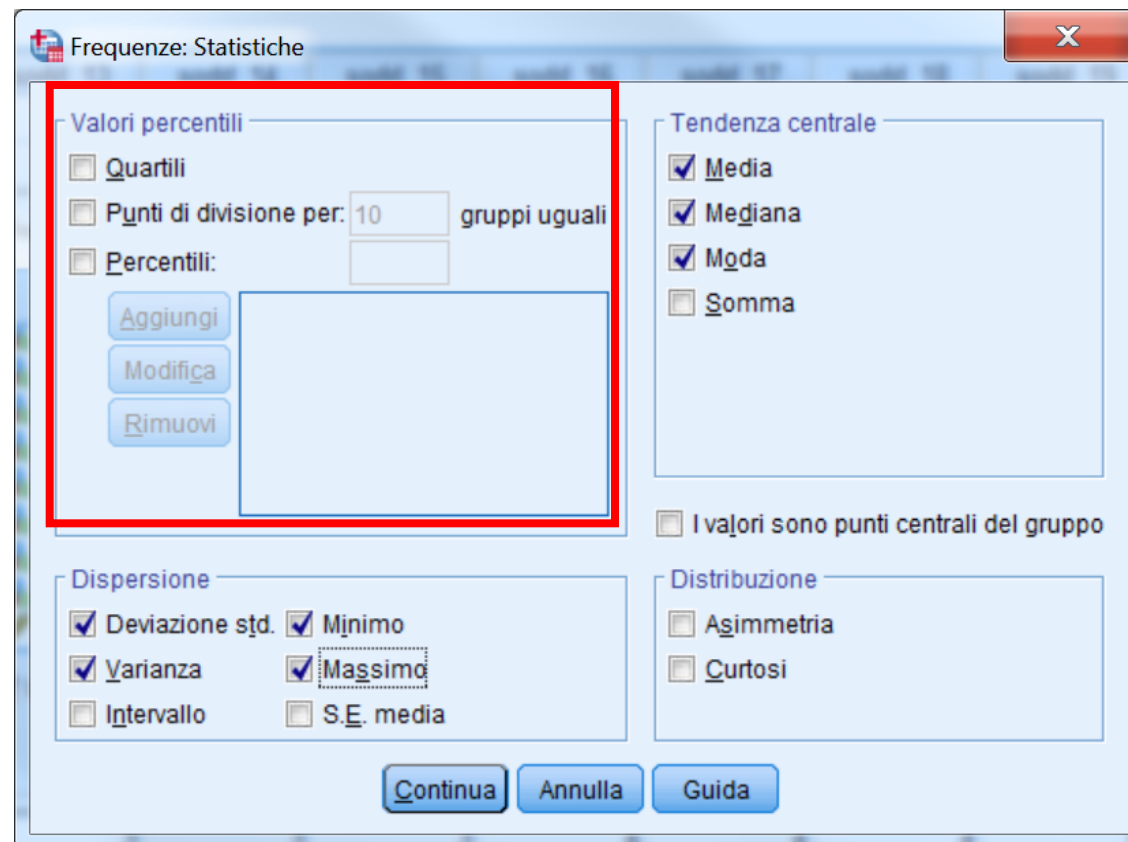
Indici di dispersione

Indici di tendenza centrale

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.

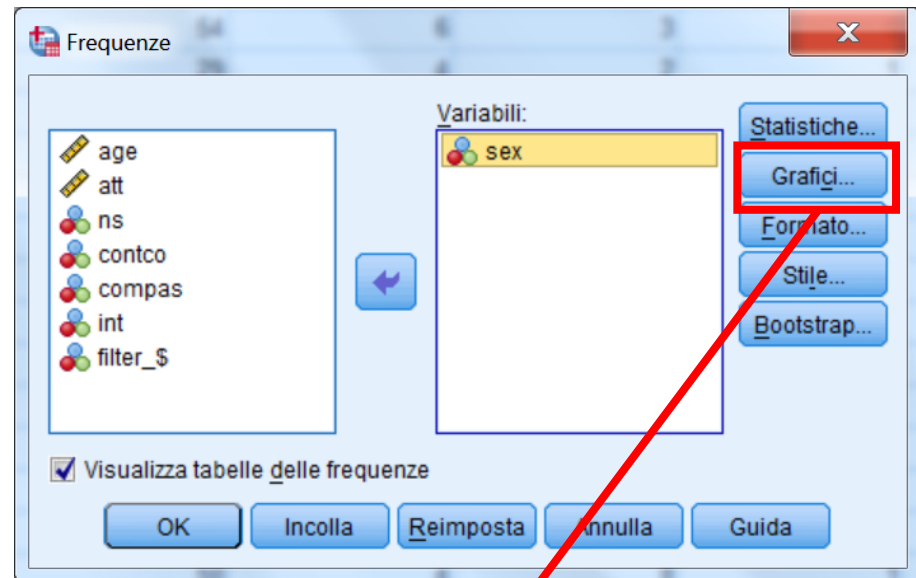


RICORDA: i Quartili indicano quei valori che dividono la distribuzione in quattro parti uguali.

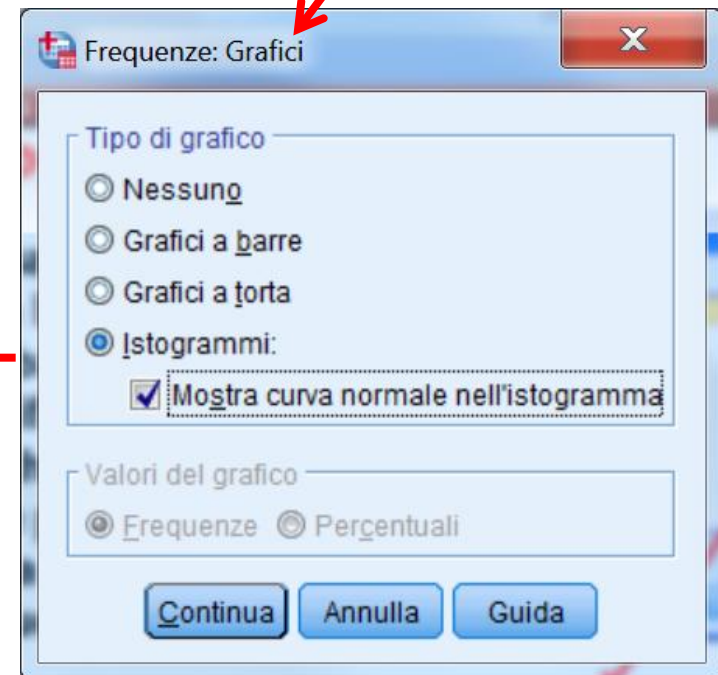
I Percentili indicano quei valori che dividono la distribuzione in 100 parti uguali.

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

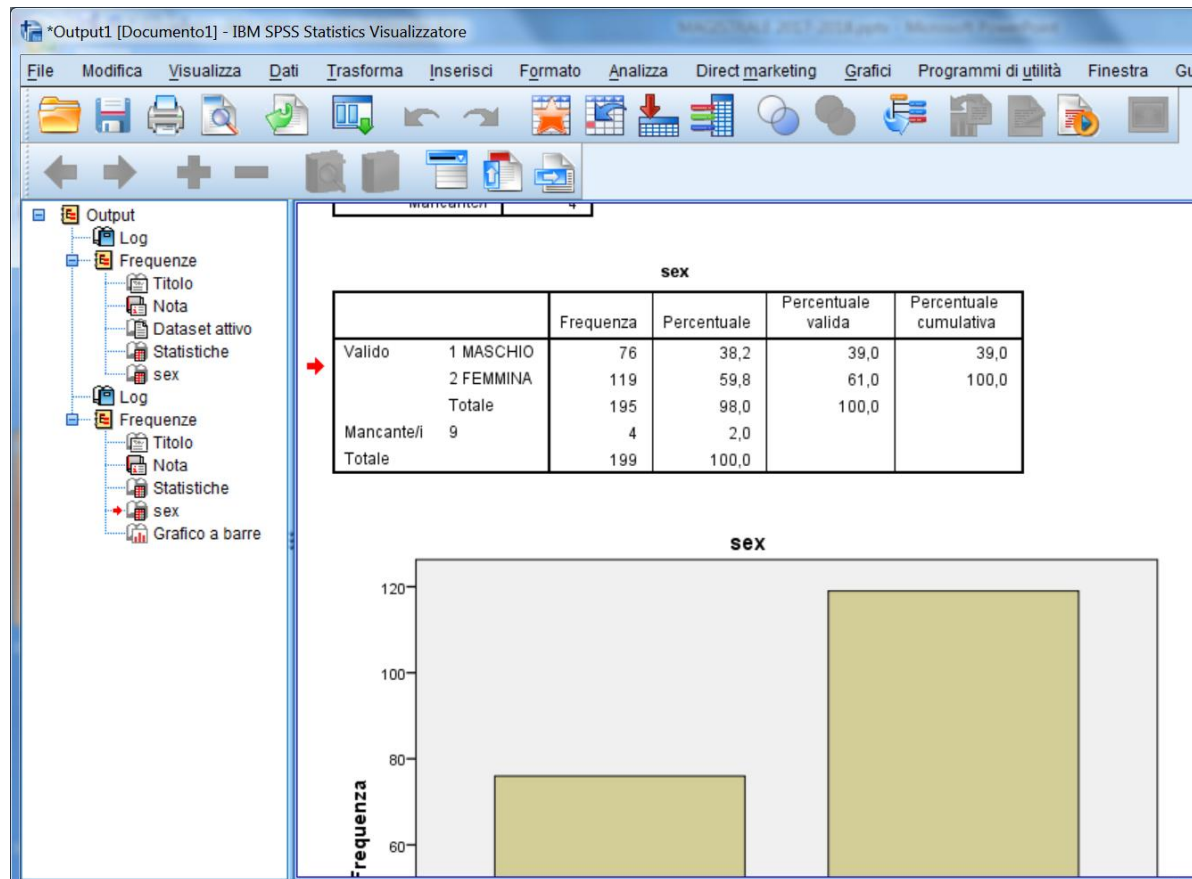


SPSS

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

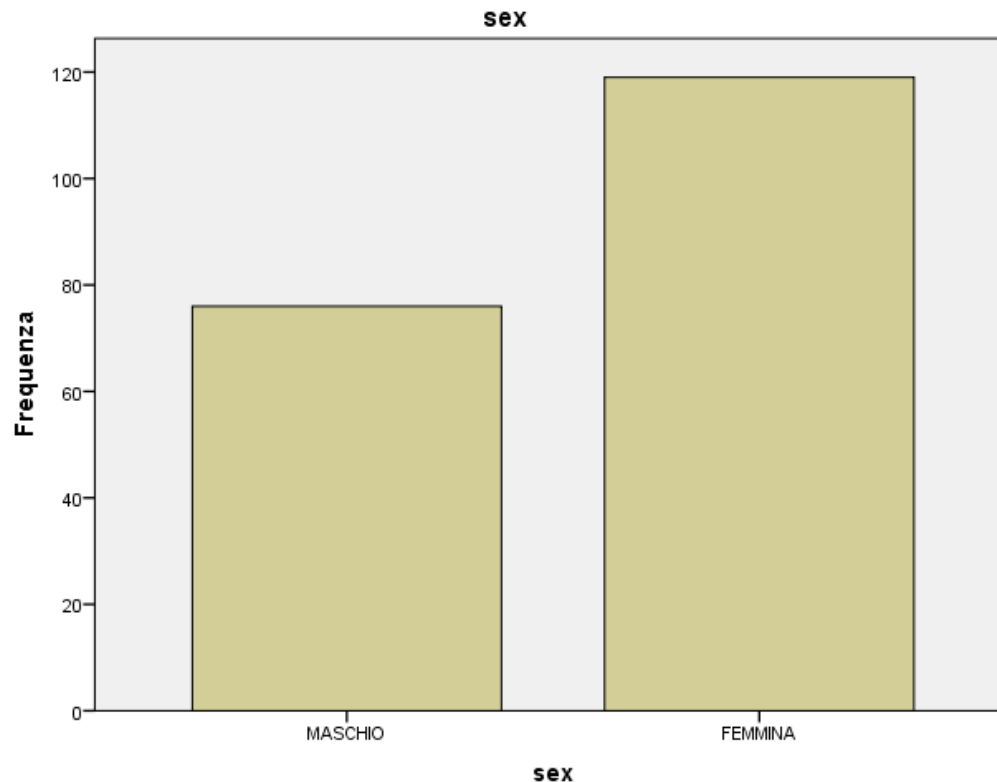
		Frequenza	Percentuale	Percentuale valida	Percentuale cumulativa
Valido	1 MASCHIO	76	38,2	39,0	39,0
	2 FEMMINA	119	59,8	61,0	100,0
	Totale	195	98,0	100,0	
Mancanteli	9	4	2,0		
Totale		199	100,0		

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

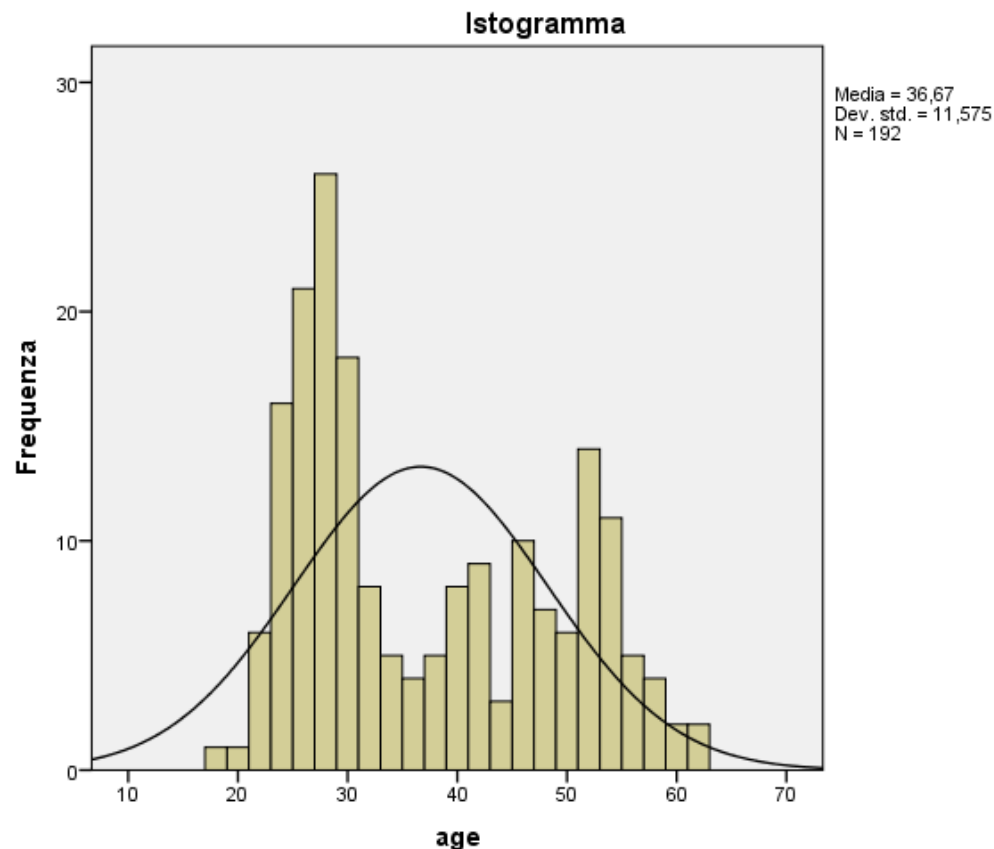
age

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

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

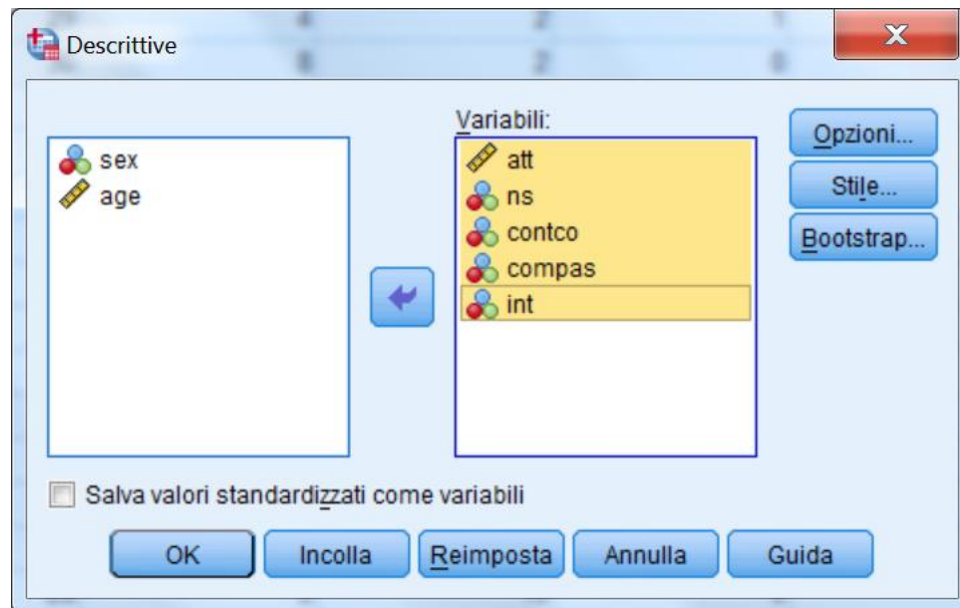
File Modifica Visualizza Dati Trasforma **Analizza** Direct marketing Grafici Programmi di utilità Finestra Guida

Report
Statistiche descrittive
 Tabelle
 Confronta medie
 Modello lineare generalizzato
 Modelli lineari generalizzati
 Modelli misti
 Correlazione
 Regressione
 Log-lineare
 Reti neurali
 Classifica
 Riduzione delle dimensioni
 Scala

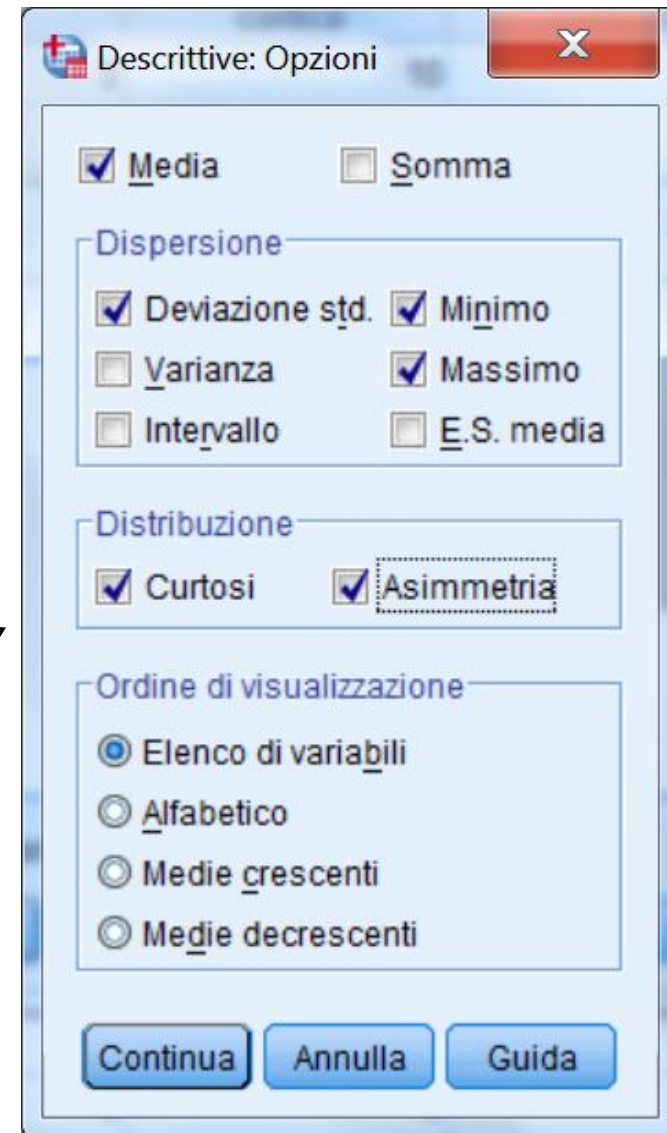
123 Frequenze...
Descrittive...
 Esplora...
 Tabelle di contingenza...
 Analisi TURF
 Rapporto...
 Grafici P-P...
 Grafici Q-Q...

	sex	age
1	1	
2	1	
3	1	
4	1	
5	9	
6	2	
7	9	
8	2	
9	2	
10	1	
11	1	

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 !

Ricodificare la variabile "contco"

Asimmetria Negativa Sostanziale
(valori tra -1 e -2)

Logaritmo = $X^* = \text{Log}_{10}(K-X)$
(qui, $K = 10+1=11$)

The screenshot shows the IBM SPSS Statistics Editor interface. The 'Trasforma' menu is open, and the 'Calcola variabile...' option is selected, indicated by a black arrow. The background shows a data table with columns 'sex' and 'contco'.

	sex	contco
1	1	9
2	1	6
3	1	4
4	1	8
5	9	4
6	2	2
7	9	10
8	2	6
9	2	8
10	1	8
11	1	10
12	1	8

Ricodificare la variabile "contco"

Asimmetria Negativa Sostanziale
(valori tra -1 e -2)

Logaritmo = $X^* = \text{Log}_{10}(K-X)$
($K = \max(X)+1=10+1=11$)

Calcola variabile

Variabile di destinazione: contco_2

Espressione numerica: LG10(11-contco)

Tipo ed etichetta...

- sex
- age
- att
- ns
- contco
- compas
- int

Gruppo di funzioni:

- Tutto
- Aritmetico
- CDF e CDF noncentrale
- Conversione
- Data/Ora corrente
- Aritmetica data
- Creazione data

Funzioni e variabili speciali:

- Idf.Strange
- Idf.T
- Idf.Uniform
- Idf.Weibull
- Lag(1)
- Lag(2)
- Length
- Lg10
- Ln
- Lngamma
- Lower

LG10(esprnum). Numerica. Fornisce il logaritmo in base 10 di esprnum, che deve essere numerico e maggiore di 0.

Se... (condizione di selezione dei casi facoltativa)

OK Incolla Reimposta Annulla Guida



Ricodificare la variabile "contco"

Asimmetria Negativa Sostanziale
(valori tra -1 e -2)

Logaritmo = $X^* = \text{Log}_{10}(K-X)$
(qui, $K = 10+1=11$)

Comando di Sintassi:

COMPUTE contco_2=LG10(11-contco).



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

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

1 : contco_2 ,0

	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

1

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.**

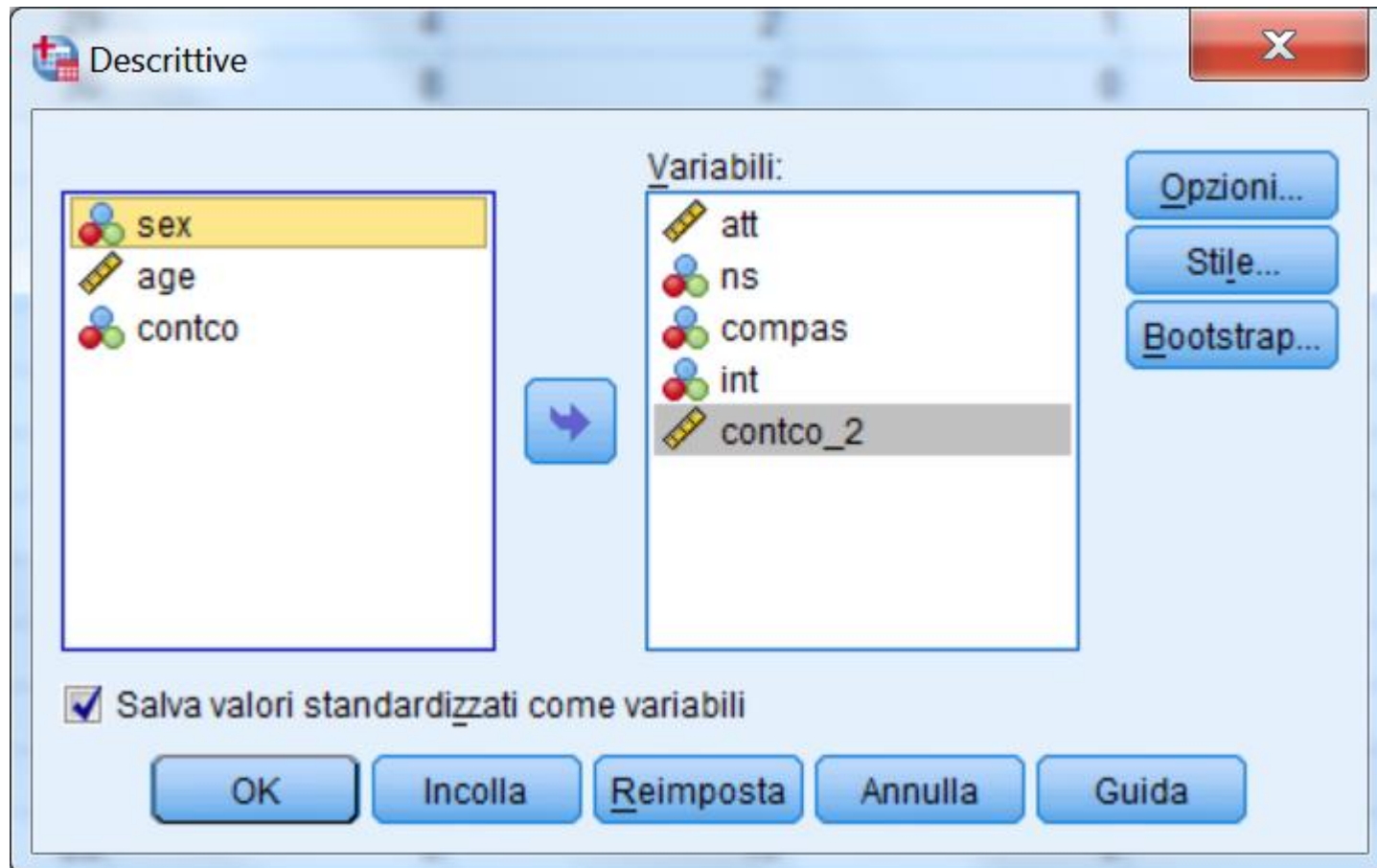
Per individuare gli outliers univariati è possibile **standardizzare i punteggi relativi alla variabile in esame e chiedere una distribuzione delle frequenze.**

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.

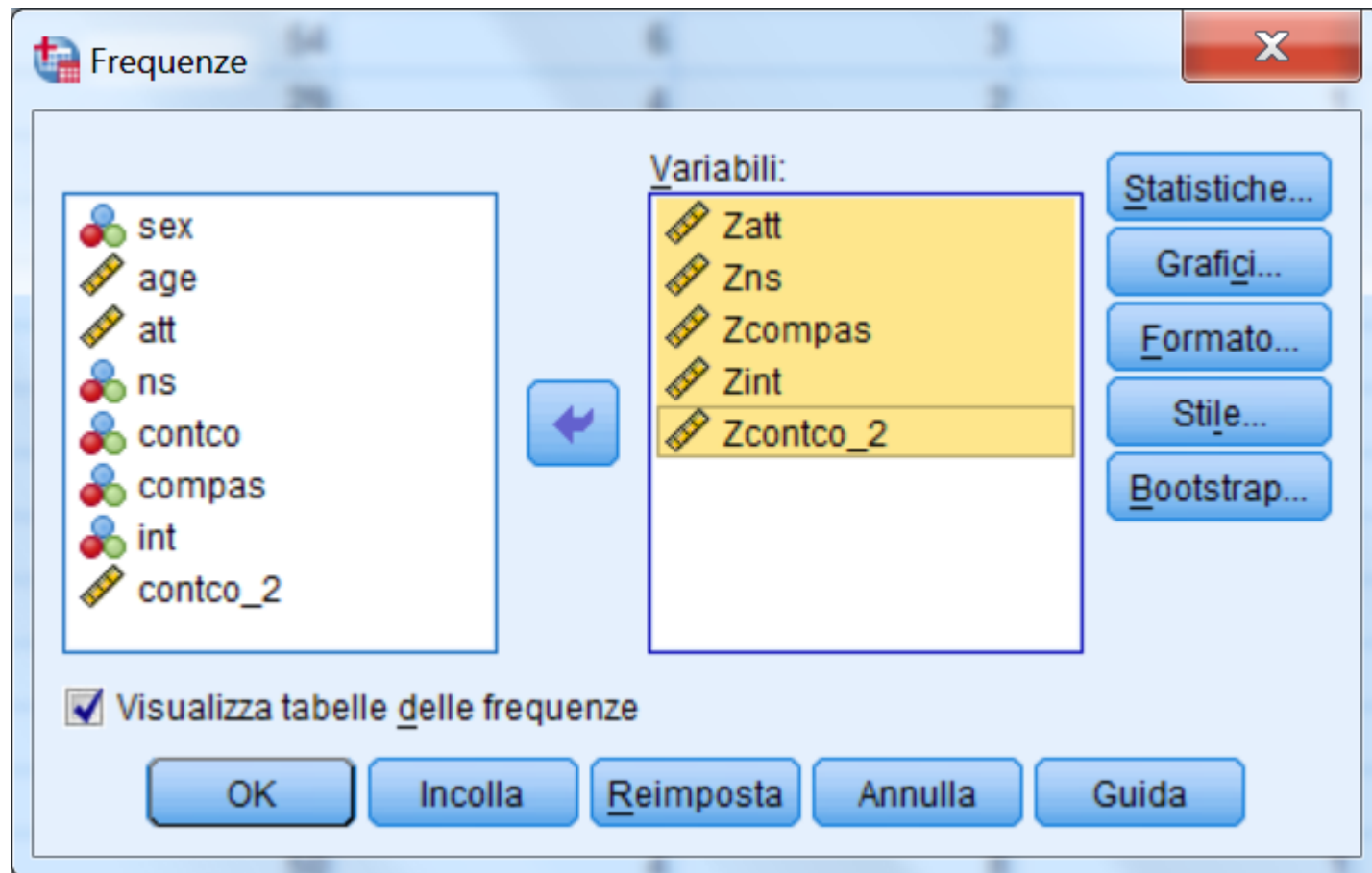
Esplorazione dei dati: gli outliers (valori anomali)

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

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

File Modifica Visualizza **Dati** Trasforma Analizza Direct marketing Grafici

1 : Zatt -3,665

	int
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	

- Definisci proprietà variabili...
- Imposta livello di misurazione per sconosciuto...
- Copia proprietà dei dati...
- Nuovo attributo personalizzato...
- Definisci date...
- Definisci insiemi a risposta multipla...
- Convalida
- Identifica casi duplicati...
- Identifica casi insoliti...
- Confronta dataset...
- Ordina casi...
- Ordina le variabili...
- Trasponi...
- Unisci file
- Ristruttura...
- Esegui raking dei pesi...
- Messa in corrispondenza punteggi propensione...
- Corrispondenza caso-controllo...
- Aggrega...
- Suddividi in file
- Progettazione ortogonale
- Copia dataset
- File di suddivisione...
- Seleziona casi...**
- Pesa casi...

Vista dati Vista Variabile

Seleziona casi...

**Filtrare i soggetti
escludendo i due
outliers**

Filtrare i soggetti escludendo i due outliers

The image shows a screenshot of the SPSS 'Seleziona casi' (Select Cases) dialog box. The 'Seleziona' section is set to 'Se la condizione è soddisfatta' (If the condition is satisfied). The variable 'Zns' is selected in the list of variables. The filter condition 'Zatt > -3 & Zns > -3' is entered in the text box. A calculator interface is visible, showing the input of the condition. The 'Gruppo di funzioni' (Group of functions) list includes 'Tutto', 'Aritmetico', 'CDF e CDF noncentrale', 'Conversione', 'Data/Ora corrente', 'Aritmetica data', and 'Creazione data'. The 'Funzioni e variabili speciali' (Special functions and variables) section is empty. The 'Stato cc' (Case status) section is also empty. The 'Continua' (Continue), 'Annulla' (Cancel), and 'Guida' (Help) buttons are visible at the bottom.

Seleziona casi

Seleziona

Tutti i casi

Se la condizione è soddisfatta

Seleziona casi: Se

Zatt > -3 & Zns > -3

Gruppo di funzioni:

- Tutto
- Aritmetico
- CDF e CDF noncentrale
- Conversione
- Data/Ora corrente
- Aritmetica data
- Creazione data

Funzioni e variabili speciali:

Stato cc

Zatt > -3 & Zns > -3

Continua Annulla Guida

Filtrare i soggetti escludendo i due outliers

*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 : sex 1

	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

IBM SPSS Statistics Il processore è pronto

Unicode.ON

Filtro attivo



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 !

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 così 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"

Calcola variabile

Variabile di destinazione: nord = Espressione numerica: \$CASENUM

Tipo ed etichetta...

- sex
- age
- att
- ns
- contco
- compas
- int
- contco_2
- Zatt
- Zns
- Zcompas
- Zint
- Zcontco_2
- filter_\$

Numero di sequenza del caso corrente. Per ogni caso, \$CASENUM rappresenta il numero di casi letti incluso il caso corrente. Il formato è F8.0. Il valore di \$CASENUM non è necessariamente il numero di riga in una finestra dell'Editor dati (disponibile negli ambienti a finestre) e il valore cambia se il file viene ordinato o se vengono inseriti nuovi casi prima della fine del file.

Gruppo di funzioni:

- Tutto
- Aritmetico
- CDF e CDF noncentrale
- Conversione
- Data/Ora corrente
- Aritmetica data
- Crescimento data

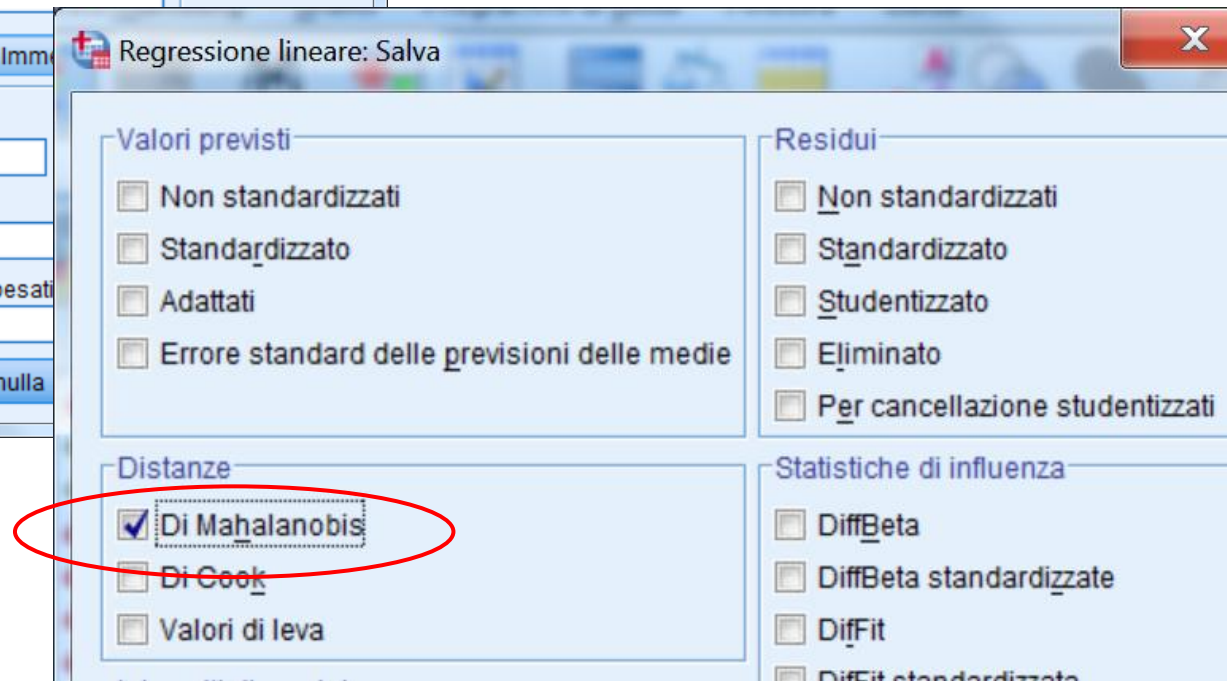
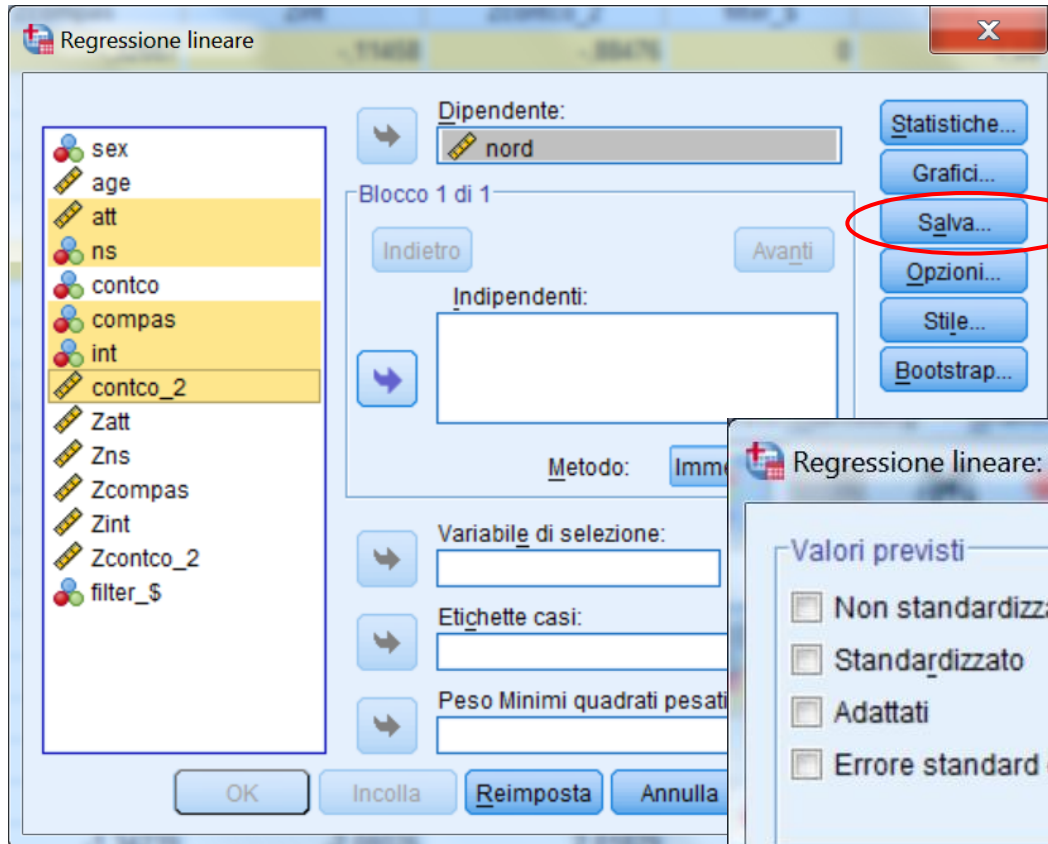
Funzioni e variabili speciali:

- \$Casenum
- \$Date
- \$Date11
- \$JDate
- \$Systemis
- \$Time
- Abs
- Any
- Applymodel
- Arsin
- Artan

Se... (condizione di selezione dei casi facoltativa)

OK Incolla Reimposta Annulla Guida

Calcolo della distanza di Mahalanobis tramite regressione



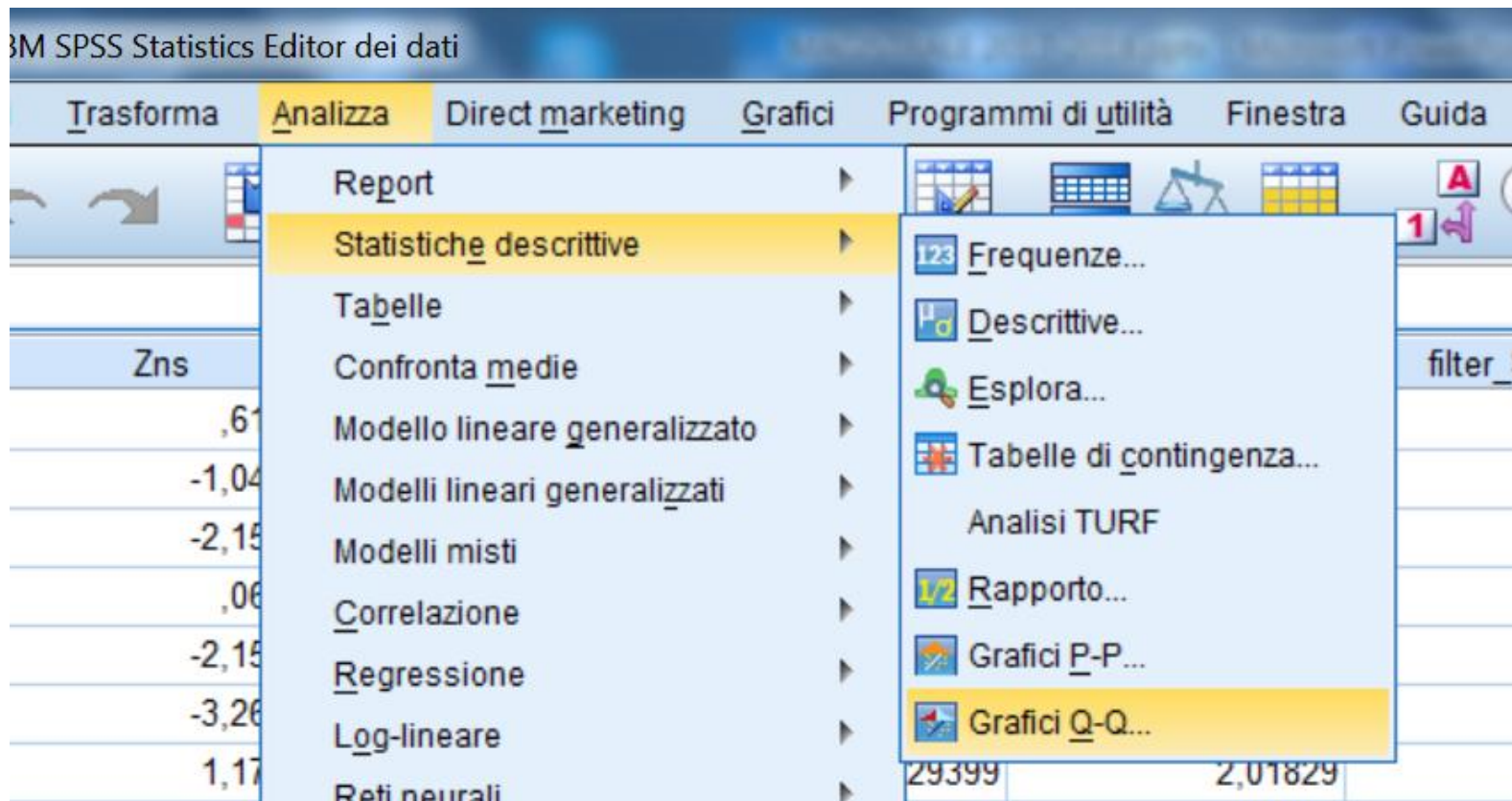
Esplorazione dei dati: la normalità multivariata e



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

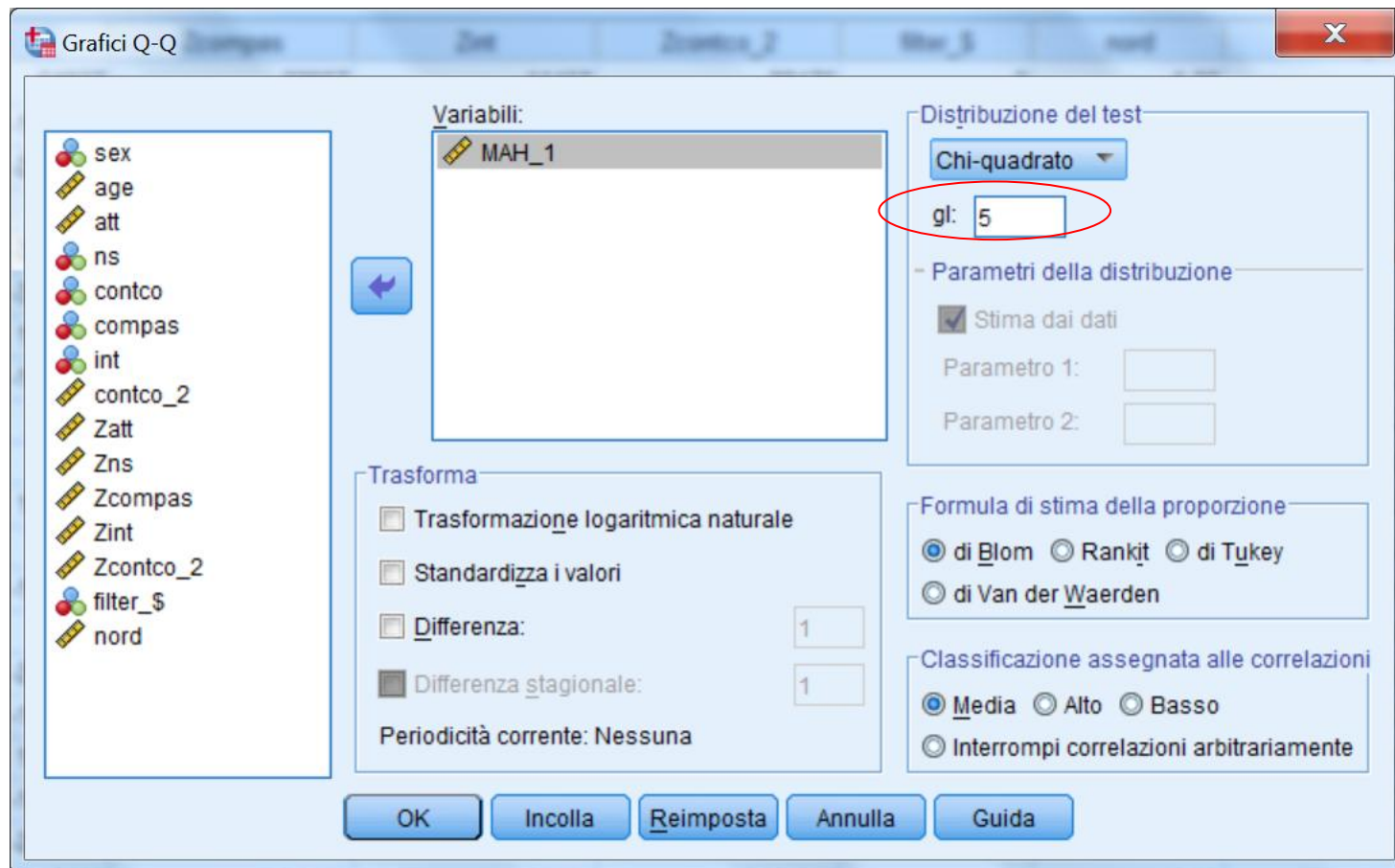
Viene creata la variabile MAH_1 nel datafile

Test grafico Q-Q Plot



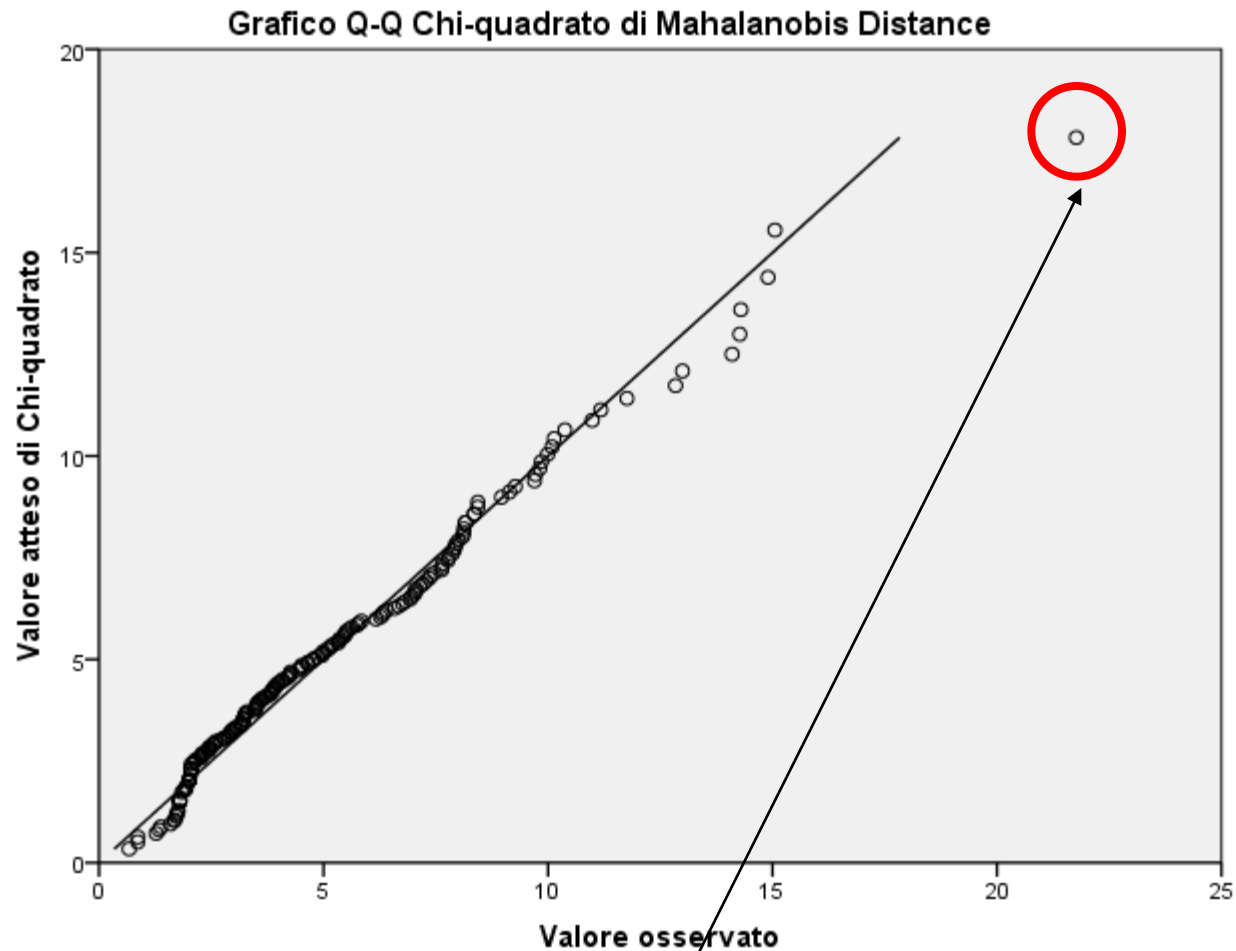
Si chiede tramite Statistiche descrittive...

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:

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

Calcoliamo il denominatore

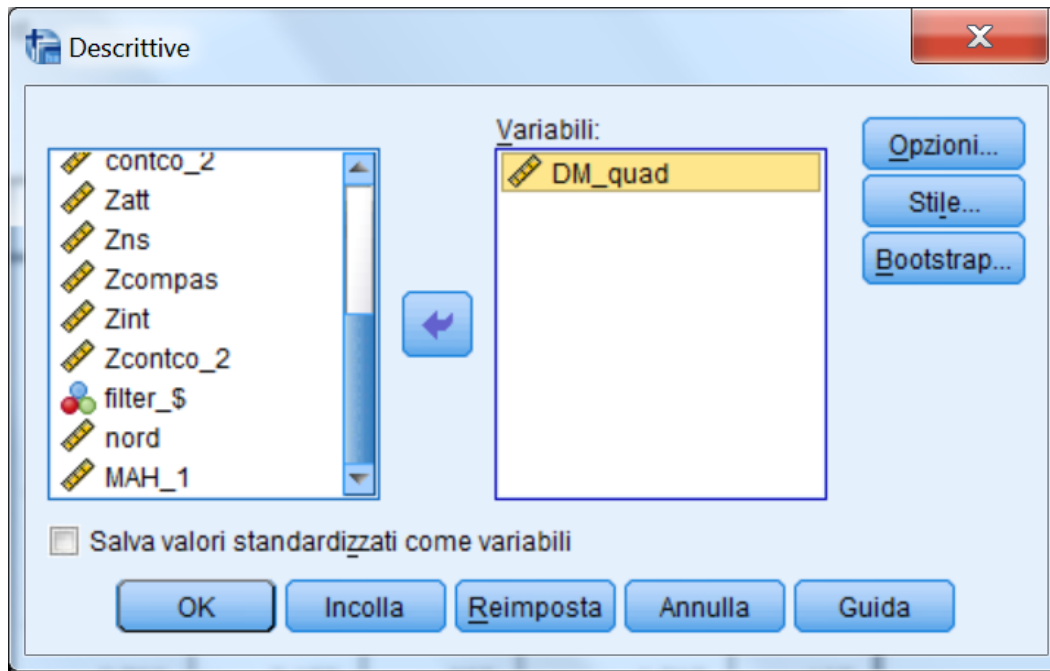
$$D_i^2 = MAH_1$$

Per cui:

$$(D_i^2)^2 = (MAH_1)^2$$

The screenshot shows the 'Calcola variabile' (Calculate Variable) dialog box in SPSS. The 'Variabile di destinazione' (Destination Variable) is 'DM_quad'. The 'Espressione numerica' (Numerical Expression) is 'MAH_1**2'. The '**' operator is highlighted with a red circle. The variable list on the left includes 'MAH_1'. The 'Gruppo di funzioni' (Function Group) is set to 'Tutto' (All). The 'Funzioni e variabili speciali' (Special Functions and Variables) list includes '\$Casenum', '\$Date', '\$Date11', '\$JDate', '\$Systemis', '\$Time', 'Abs', 'Any', 'Applymodel', 'Arasin', and 'Artan'.

Calcolo del coefficiente di curtosi multivariata



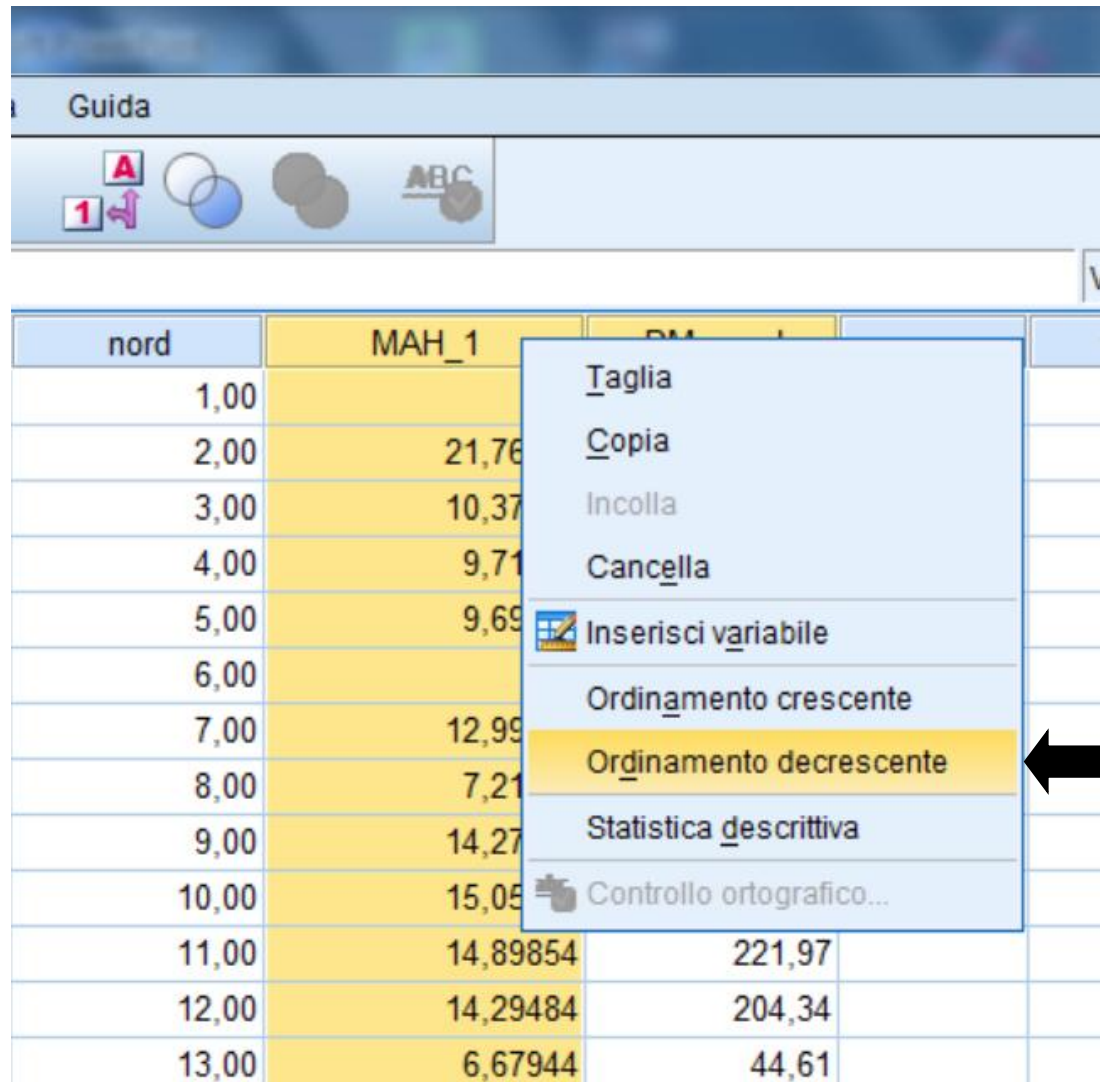
Calcolando la media della variabile DM_quad si ottiene il coefficiente di Mardia

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



Guida

nord	MAH_1	PM
1,00		
2,00	21,76	
3,00	10,37	
4,00	9,71	
5,00	9,69	
6,00		
7,00	12,99	
8,00	7,21	
9,00	14,27	
10,00	15,05	
11,00	14,89854	221,97
12,00	14,29484	204,34
13,00	6,67944	44,61



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

Seleziona casi: Se

sex
age
att
ns
contco
compas
int
contco_2
Zatt
Zns
Zcompas
Zint
Zcontco_2
filter_5
nord
MAH_1
DM_quad

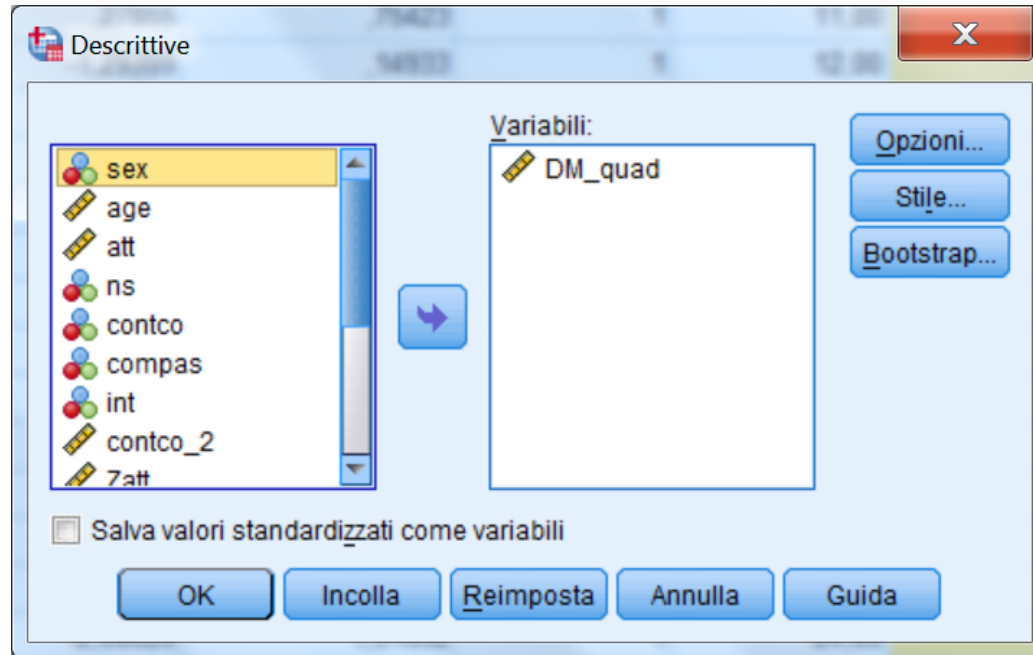
Zatt > -3 & Zns > -3 & MAH_1 < 20

Gruppo di funzioni:
Tutto
Aritmetico
CDF e CDF noncentrale
Conversione
Data/Ora corrente
Aritmetica data
Creazione data

Funzioni e variabili speciali:

Continua Annulla Guida

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,
COMPORAMENTO 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**



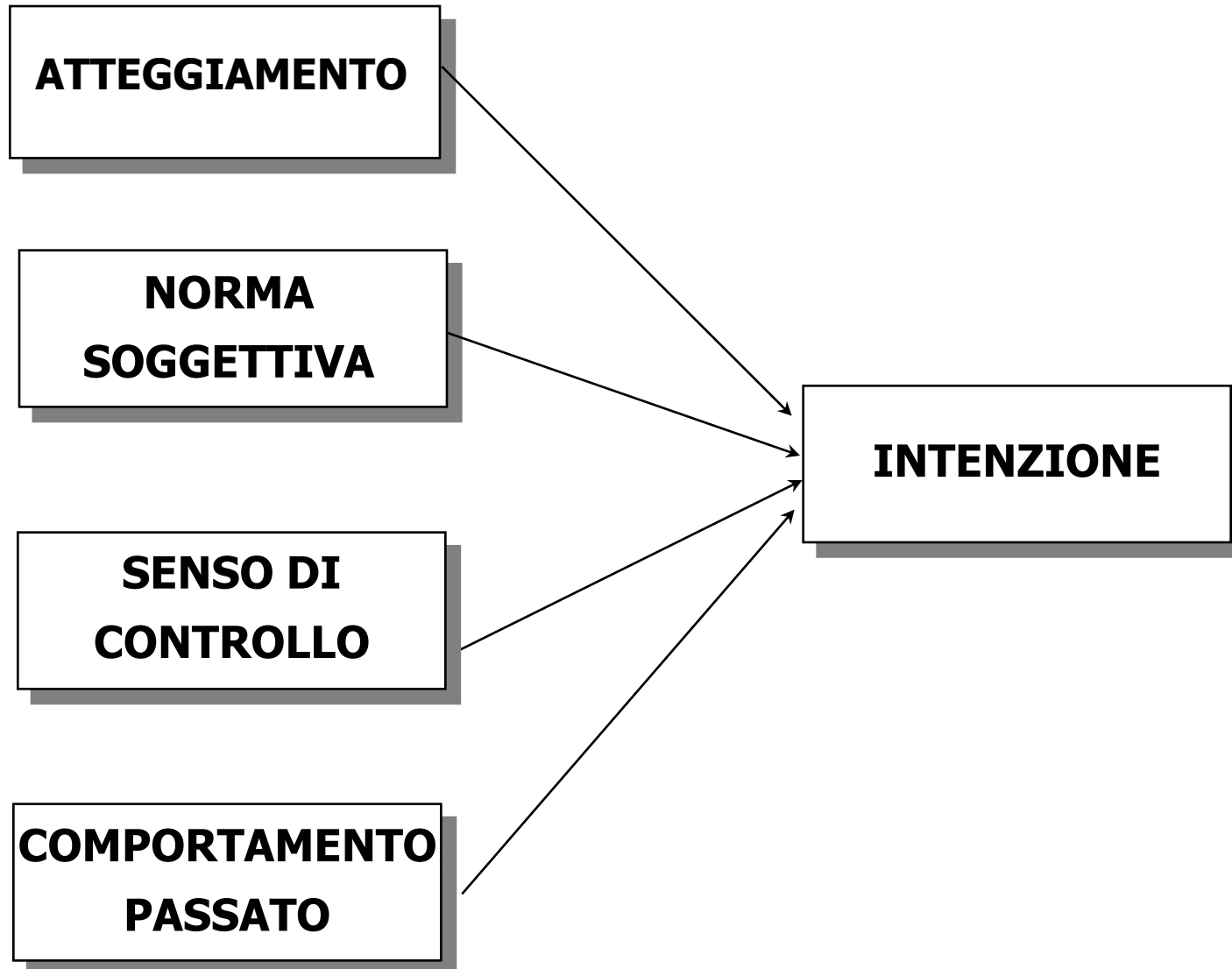
La Regressione esamina la relazione lineare tra una o più variabili esplicative (o indipendenti, VI, o “predittori”) e una variabile criterio (o dipendente, VD).

Dupliche 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

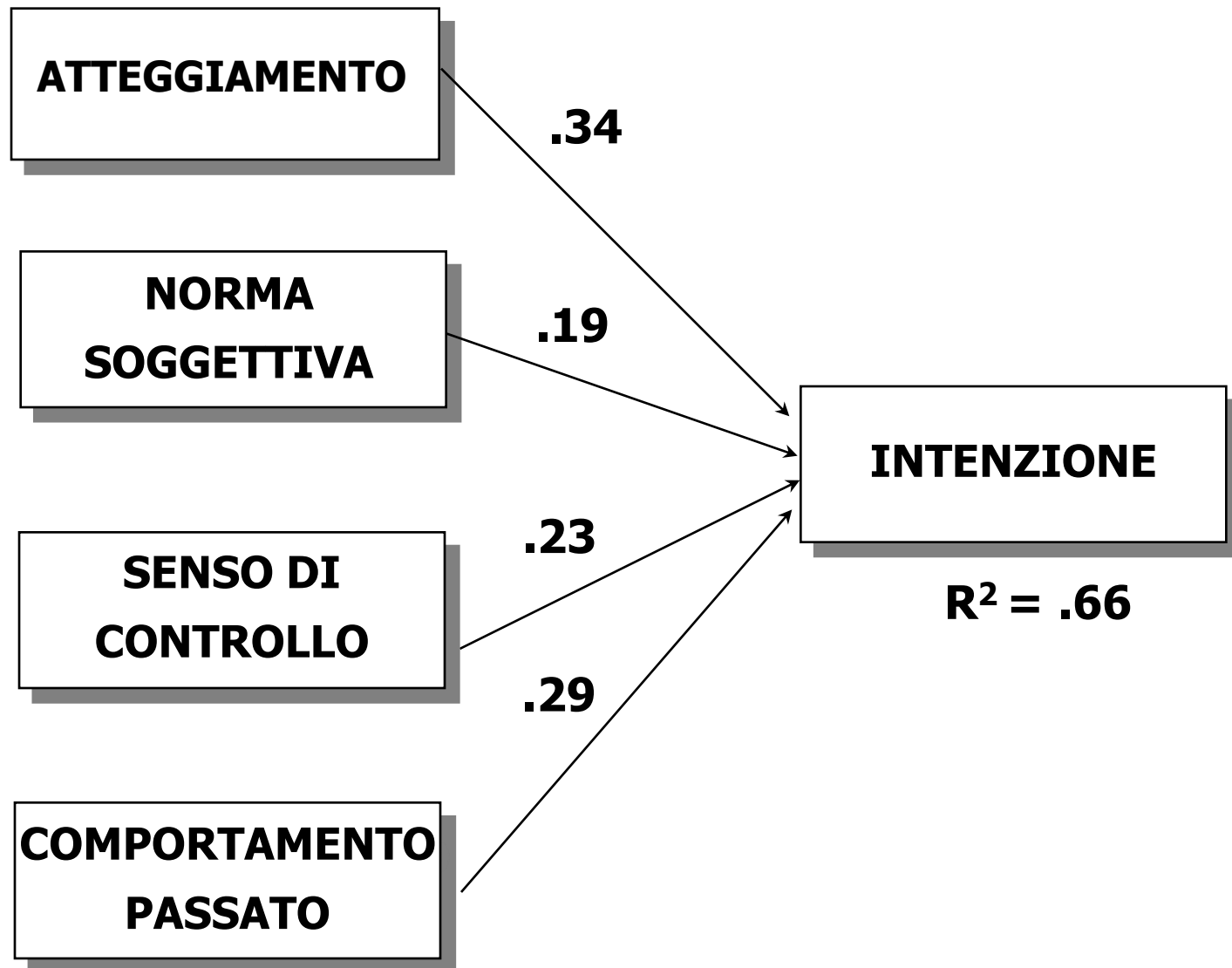




Da dove si parte: Matrice delle covarianze

	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

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



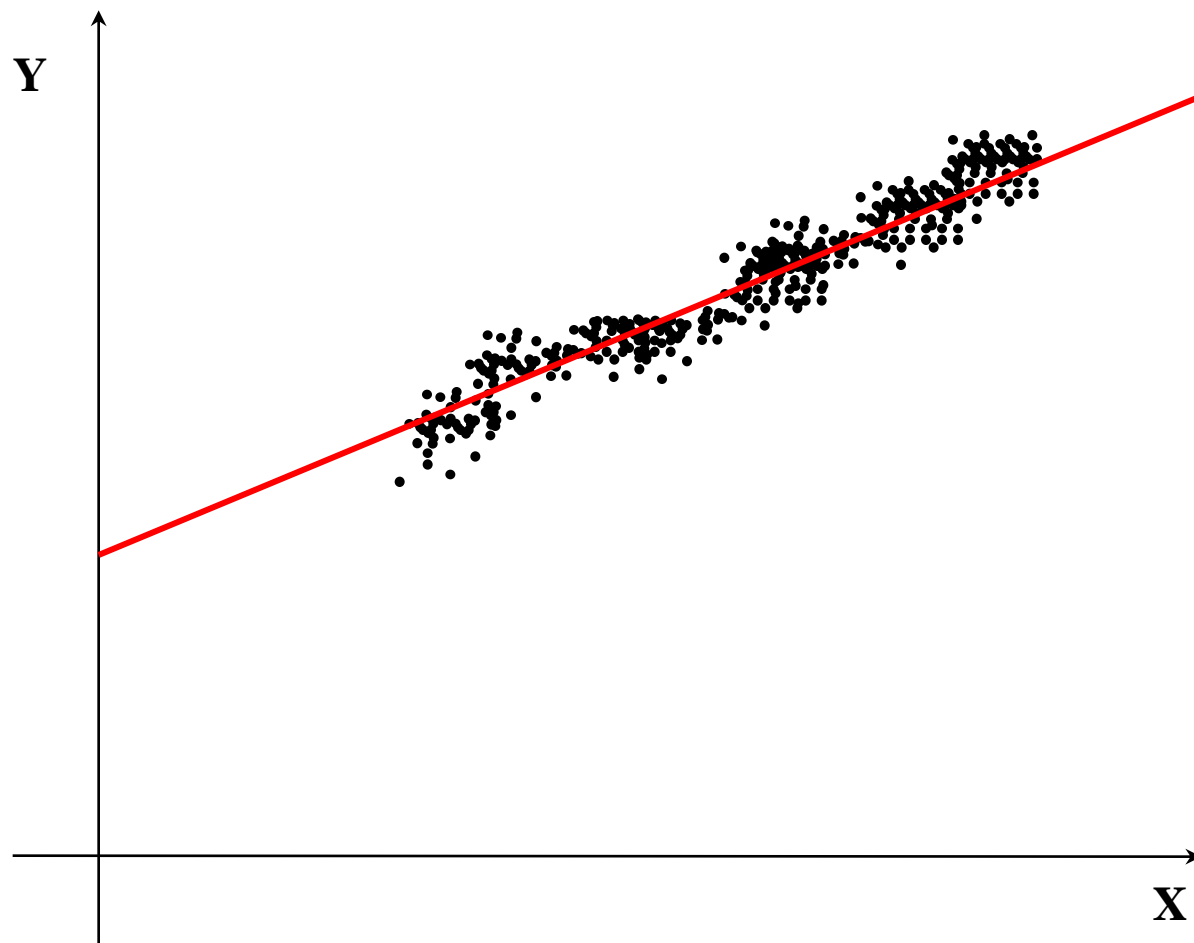
Regressione bivariata (o semplice)

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:**

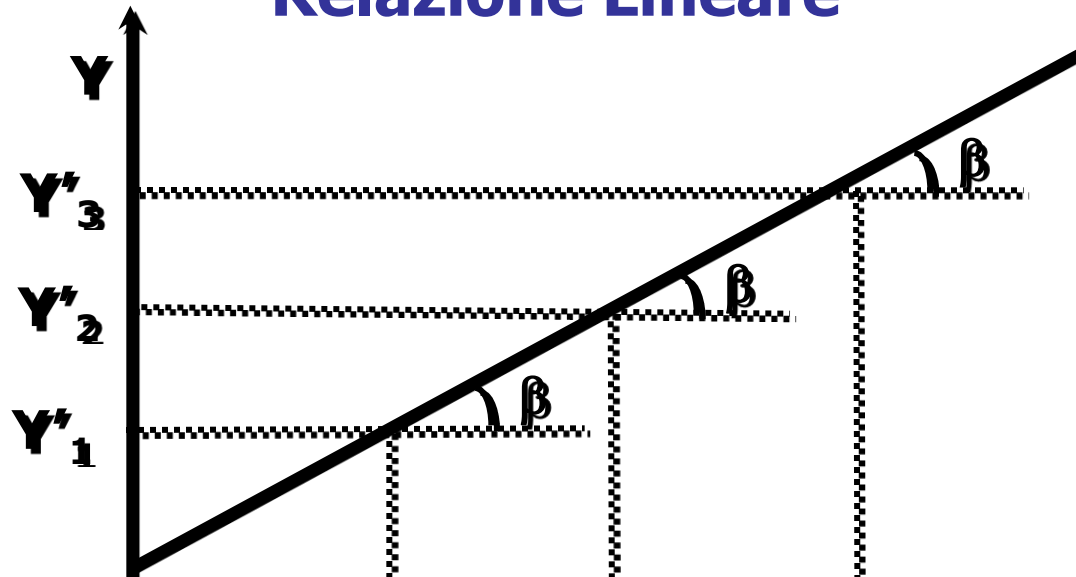
$$Y = \alpha + \beta X$$

Parametri dell'equazione:

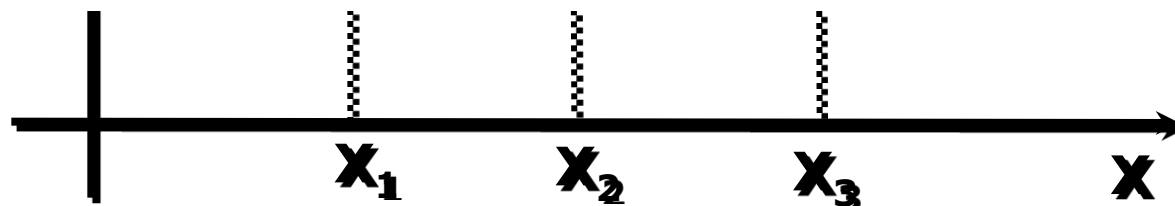
Intercetta: α , punto in cui la retta incrocia l'asse delle ordinate (altezza della linea).

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

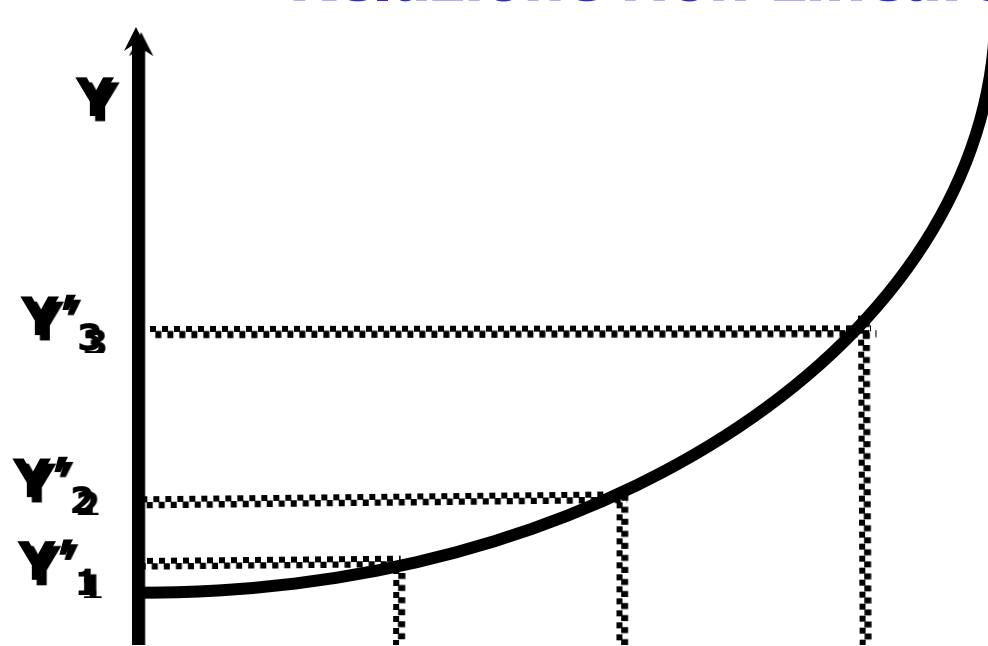


Per ogni variazione in X si determina sempre la stessa variazione in Y qualunque sia il valore di X sull'asse delle ascisse.

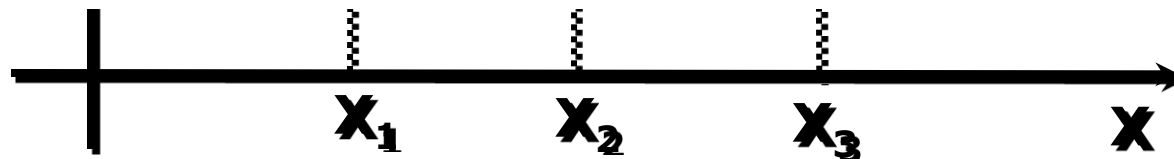


$$(X_3 - X_2) = (X_2 - X_1) \Rightarrow (Y'_3 - Y'_2) = (Y'_2 - Y'_1)$$

Relazione Non Lineare



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



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

Errore o residuo

I punti sono dispersi intorno alla retta di regressione perché:

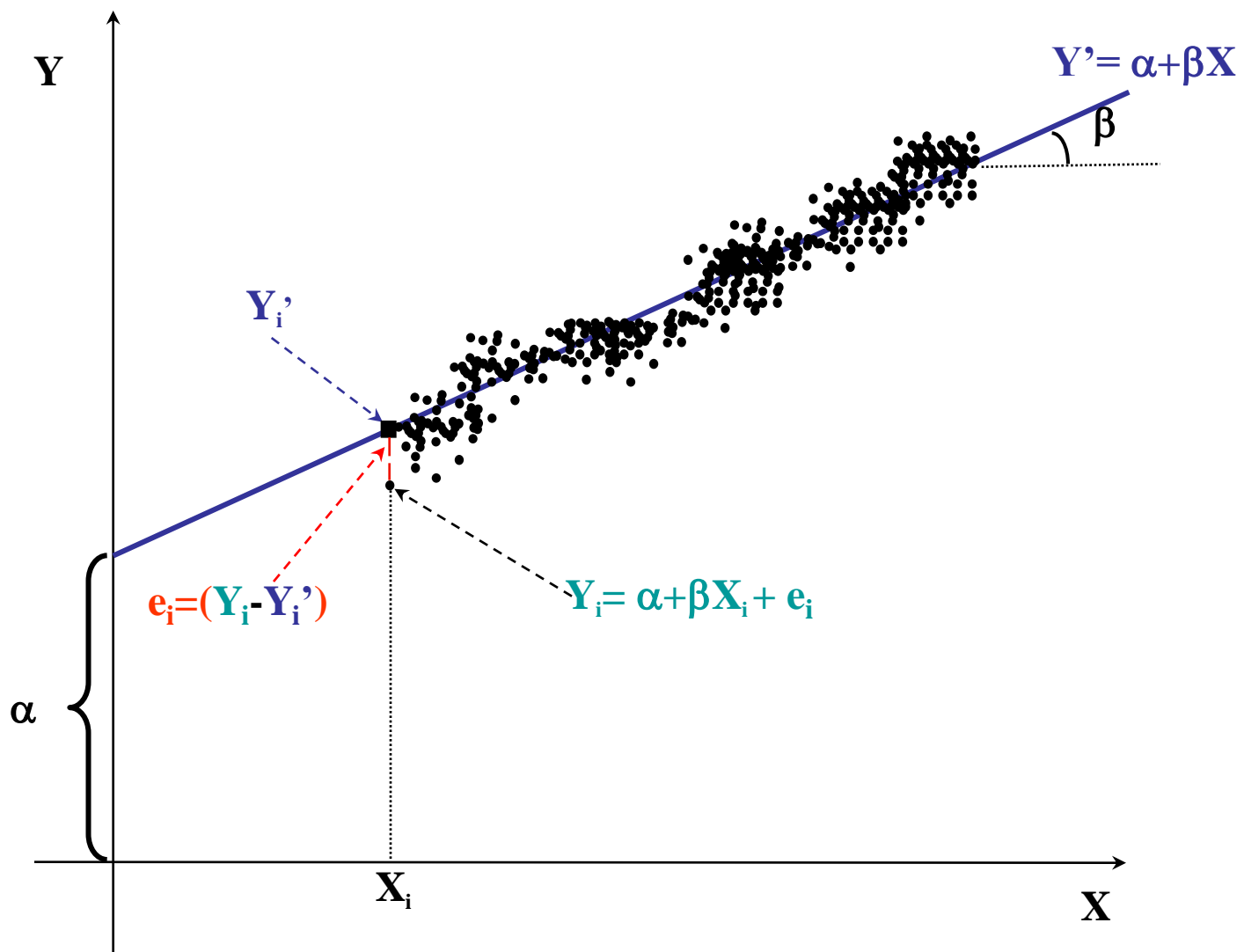
- le variabili sono misurate con errore;**
- la relazione può non essere perfettamente lineare;**
- predittori importanti possono essere omessi.**

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' .





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:

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

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:

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

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



La regressione multipla

Una variabile dipendente che regredisce su almeno due variabili indipendenti. Equazione di regressione:

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

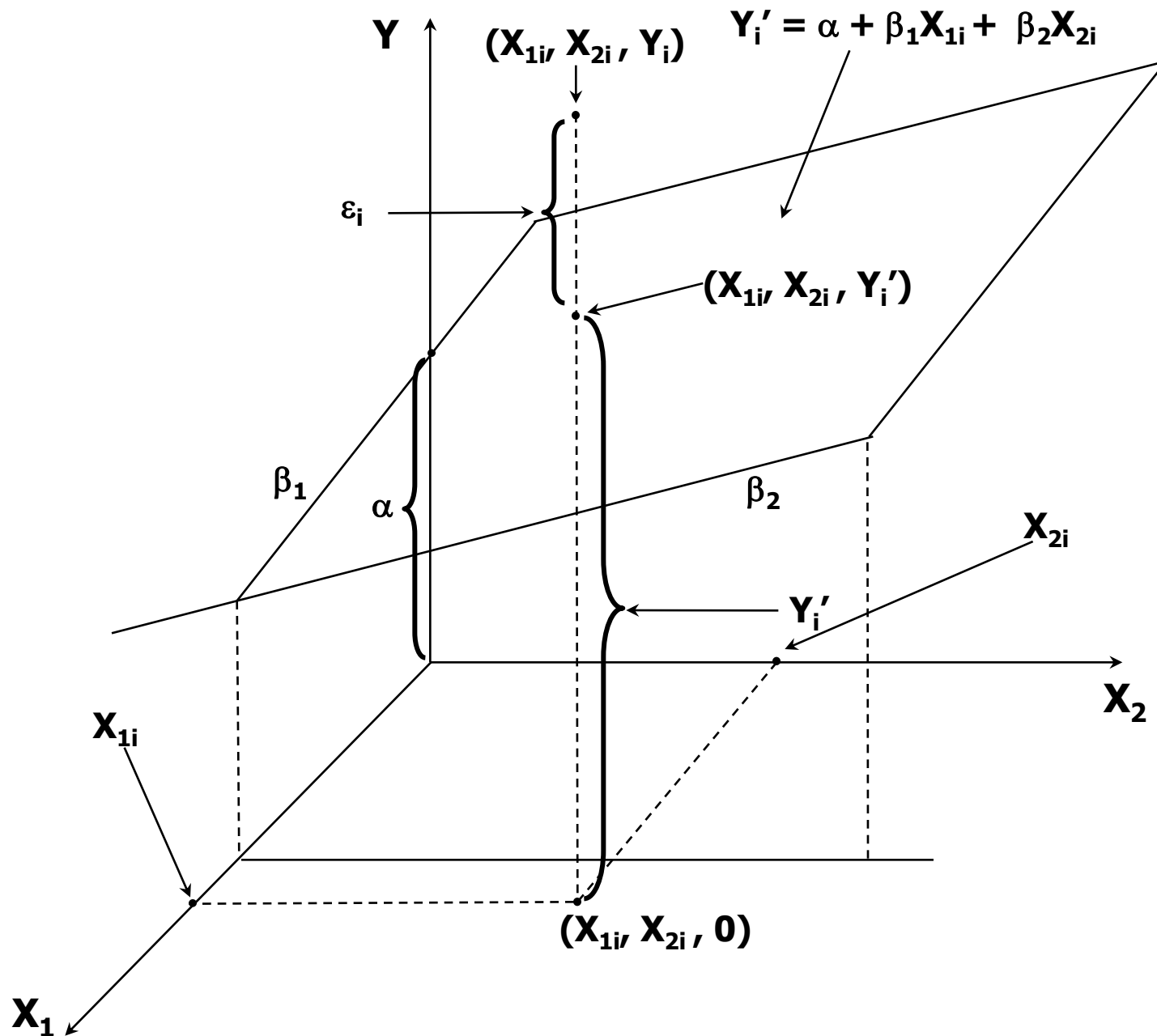
Piano di regressione (due VI);

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).

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

Espressioni matriciali delle equazioni:

$$y = bX + e \quad \text{equazione di regressione} \quad (1)$$

$$b = (X'X)^{-1} X'Y \quad \text{coefficienti di regressione} \quad (2)$$

$$e = Y - (Xb + a) \quad \text{residui} \quad (3)$$

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

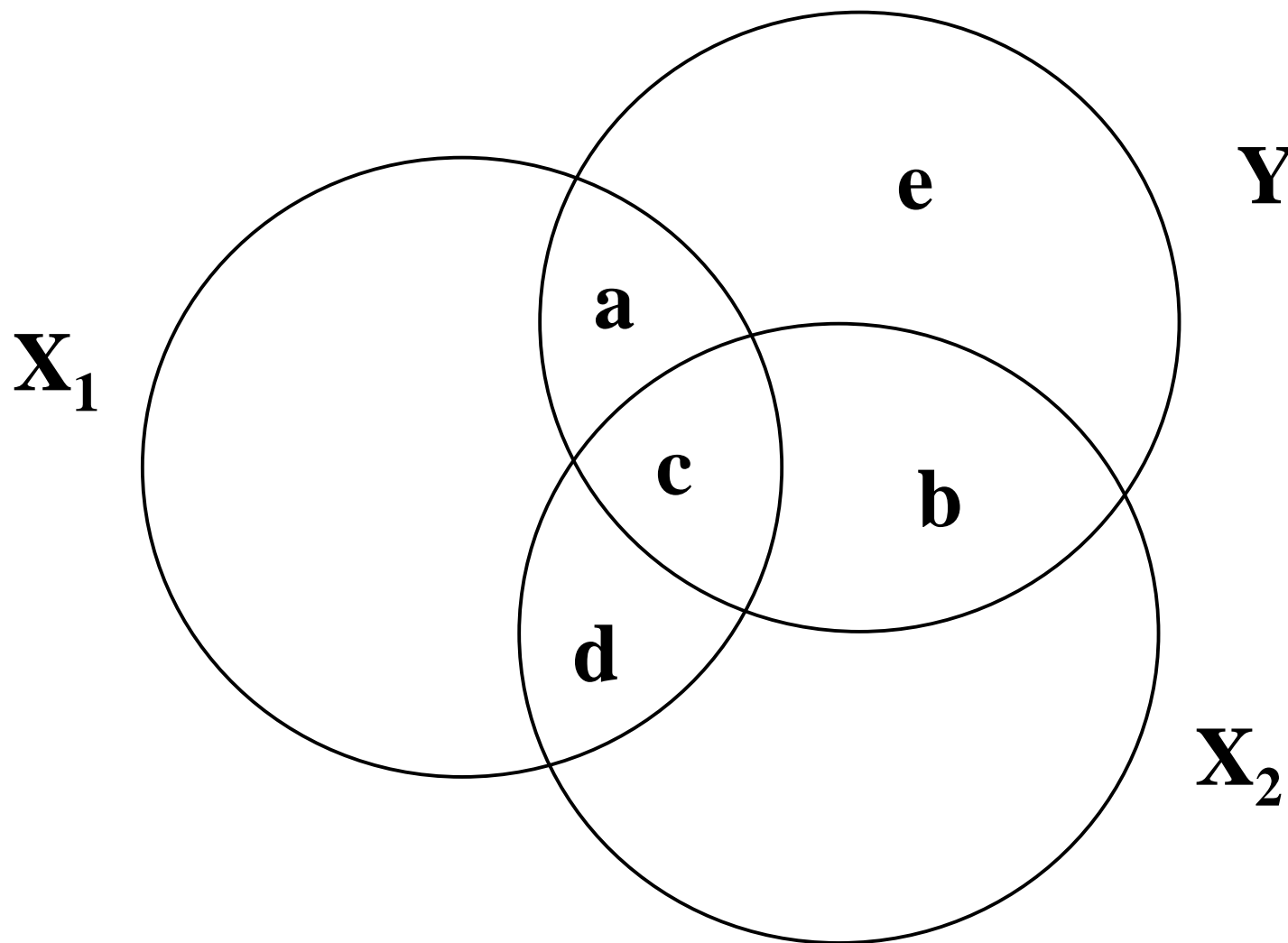


Relazioni tra una VD Y e due VI X_1 e X_2 , espresse in termini della varianza che condividono:

- " $a+c$ ": varianza in comune tra X_1 e Y , e " a ": varianza che Y condivide solo con X_1 ;
- " $c+b$ ": varianza in comune tra X_2 e Y , e " b ": che Y condivide solo con X_2 ;
- " $c+d$ ": varianza in comune tra X_1 e X_2 ;
- " e " var. che Y non condivide né con X_1 né con X_2 ;



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.

$$sr_{y1.2} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{1 - r_{12}^2}} \quad sr_{y1.2\dots k}^2 = R_{y.12..i\dots k}^2 - R_{y.12..(i)..k}^2$$

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

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

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



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:

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



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^\wedge) in X_1 al netto di X_2 .

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

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 = coefficiente di determinazione = 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 non è spiegata dalla regressione di Y su X.

E' possibile dimostrare infatti che:

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

**$\sqrt{(1-r^2)}$ = 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.

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

Nel caso di due variabili indipendenti la formula è:

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

Somma dei prodotti delle correlazioni semplici (o “di ordine zero”) e dei coefficienti $\hat{\beta}$ 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:

$$R_{y.12\dots k} = \sqrt{R_{y.12\dots k}^2}$$

**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 $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$)

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

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

Verifica delle ipotesi (test di significatività)

Significatività statistica dei singoli b:

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

$t = (b - 0) / S_b$, con $N-k-1$ 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, $VAR(\varepsilon_i)=s^2$
 - 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

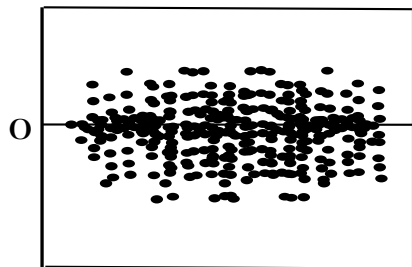


Violazione delle assunzioni:

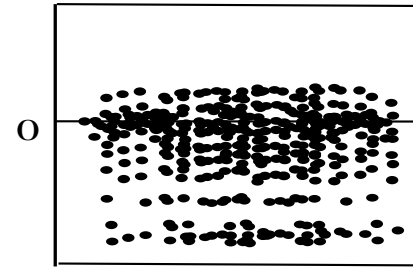
**Esame della distribuzione dei residui
 $e=(Y-Y')$ rispetto ai punteggi teorici Y' .**

Utile per rilevare:

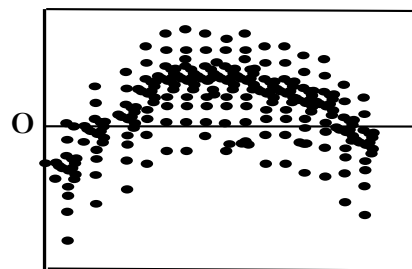
- La non linearità della relazione tra VI e VD, e tra VI,**
- La non omogeneità della varianza**
- La non normalità dei residui**



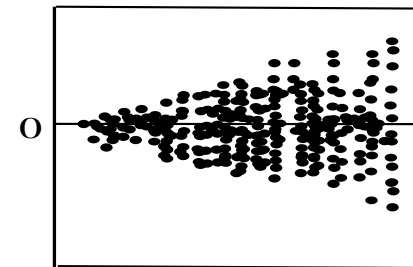
1. Assunzioni rispettate



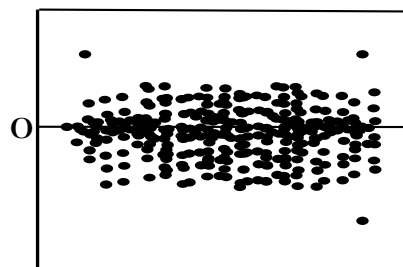
2. Non normalità



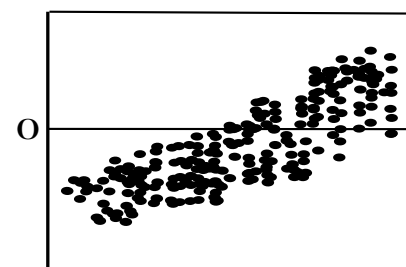
3. Non linearità



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 Inflation 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 $n \geq 100$ 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.**

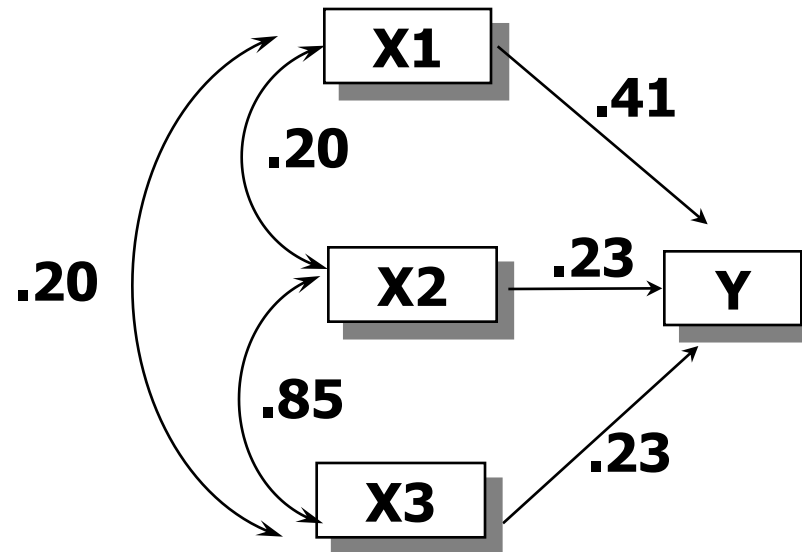
Rimedi per risolvere le violazioni: trasformazione delle variabili originali (logaritmo, reciproco, radice quadr.).

Scomposizione degli effetti

La **ridondanza** riguarda il caso in cui i coefficienti di correlazione semiparziale (s_r), parziale (p_r) 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



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

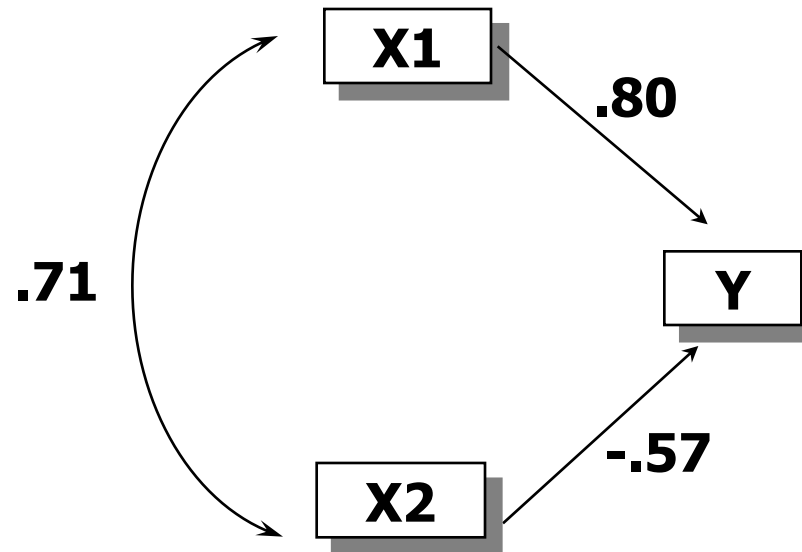
Scomposizione degli effetti

La **soppressione** riguarda il caso in cui i coefficienti s_r , p_r 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).


The image displays two overlapping windows from the IBM SPSS Statistics Editor. The left window, titled 'reg_dati.sav [Dataset4] - IBM SPSS Statistics Editor dei dati', shows a data grid with columns 'sex' and 'age'. The right window, titled 'reg_dati.sav [Dataset4] - IBM SPSS Statistics Editor dei dati', shows a variable list table with columns 'Nome', 'Tipo', 'Larghezza', 'Decimali', and 'Etichetta'.

	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

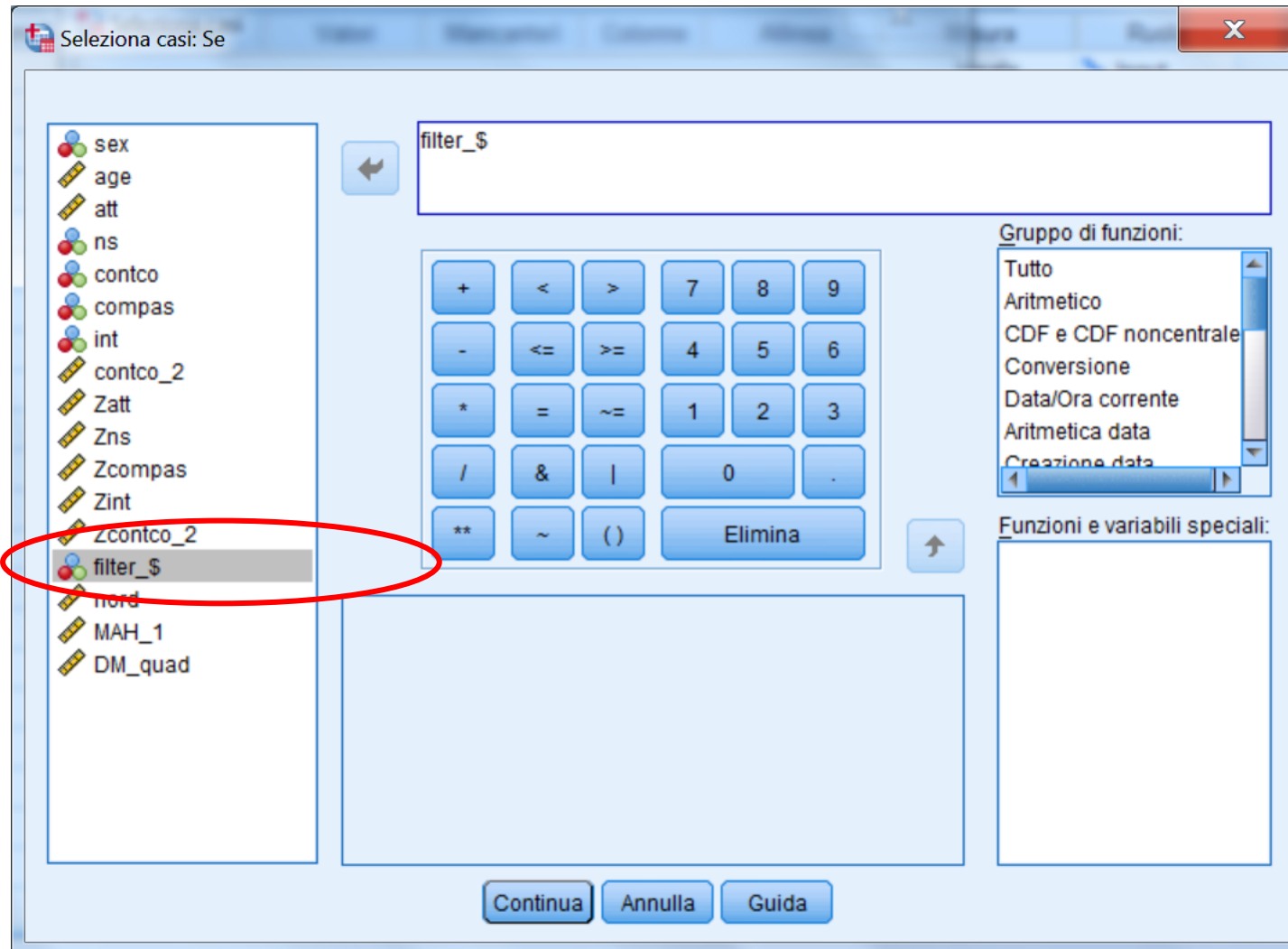
Senza titolo5 [Dataset5] - IBM SPSS Statistics Editor dei dati

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



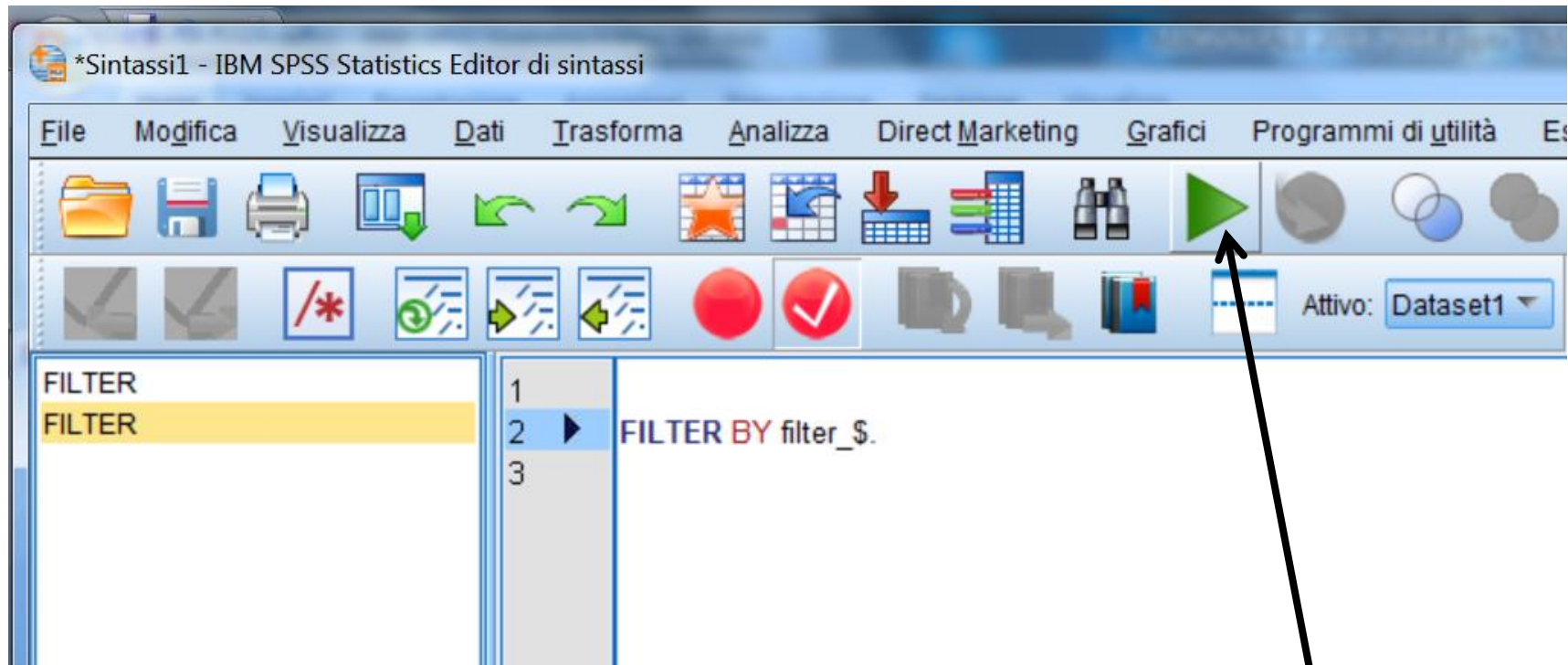
	Nome	Tipo	Larghezza	Decimali	Etichetta	Valori
1	sex	Numerico	12	0		{1, MA
2	age	Numerico	12	0		Nessu
3	att	Numerico	12	0		Nessu
4	ns	Numerico	12	0		Nessu
5	contco	Numerico	12	0		Nessu
6	compas	Numerico	12	0		Nessu
7	int	Numerico	12	0		Nessu
8	contco_2	Numerico	8	2		Nessu
9	Zatt	Numerico	11	5	Punteggio 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.

iv [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...

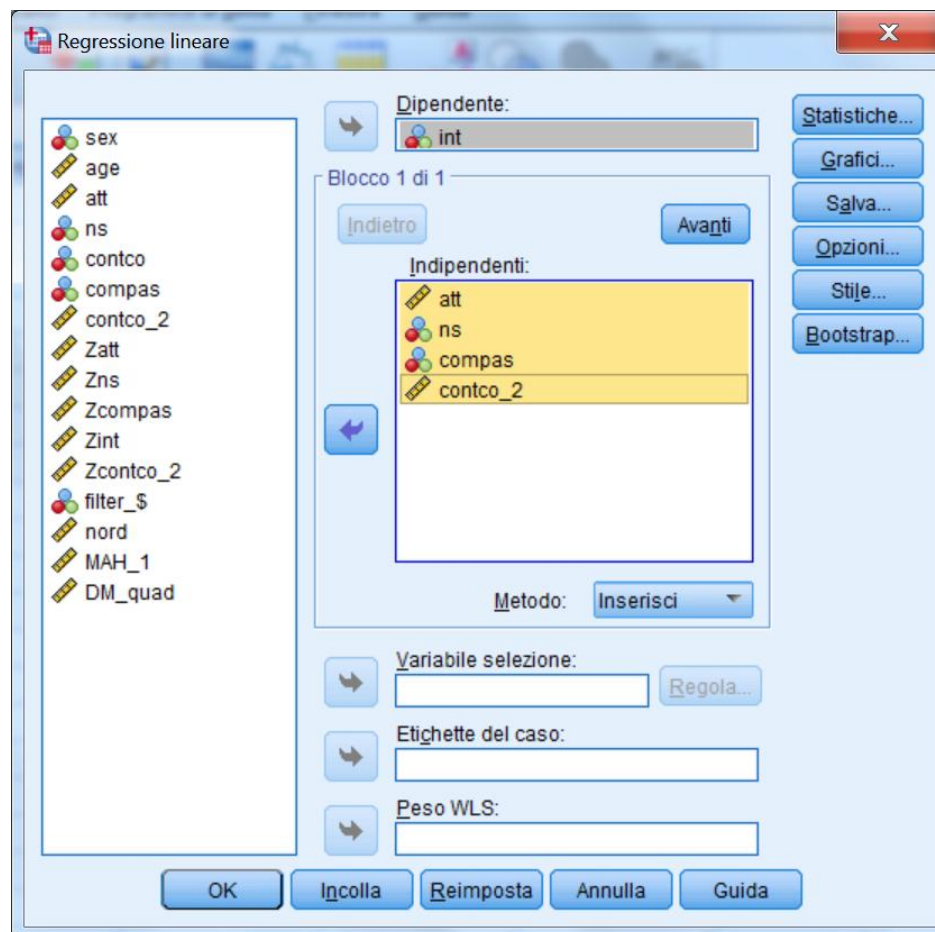
contco compas int cont

contco	compas	int	cont
3	0		3
10	0		2
8	8		8
9	0		4

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)...

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 X_1 è inserita dopo la VI X_2 , 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 lineare: Statistiche

Coefficienti di regressi...

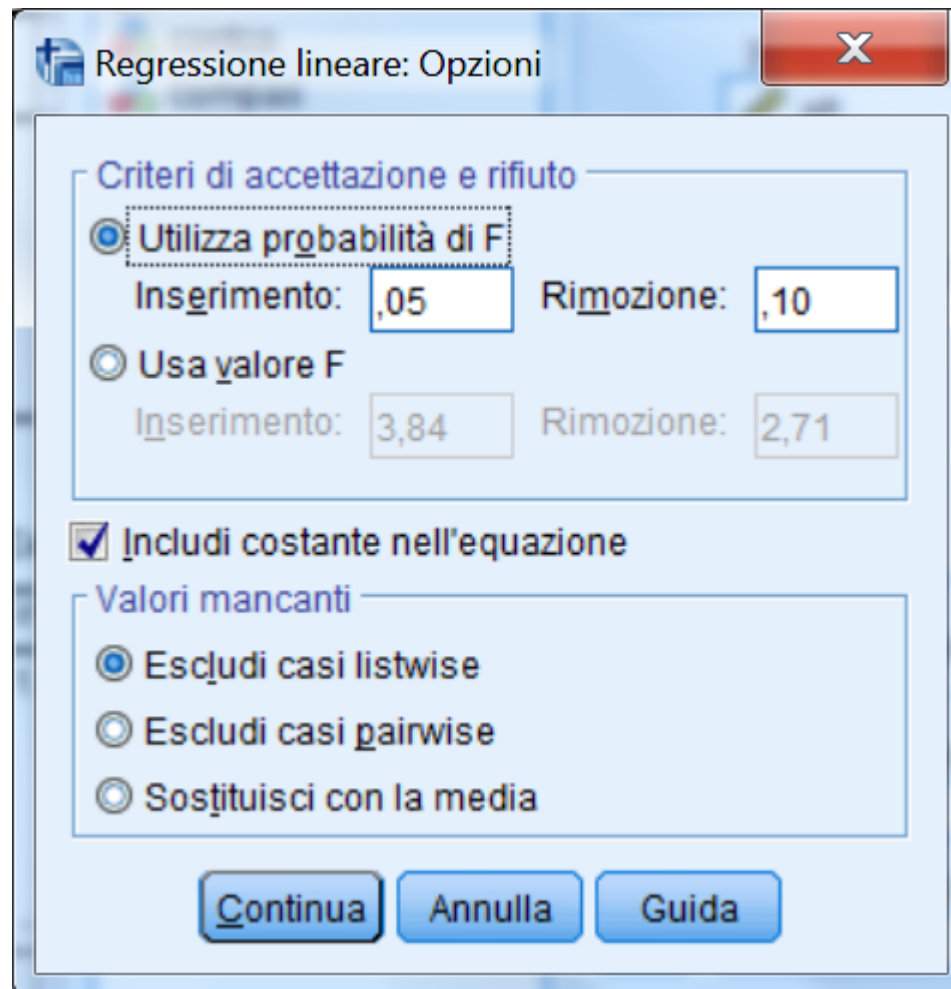
- Stime
- Intervalli di confidenza
- Livello (%):
- Matrice di covarianza
- Adattamento del modello
- Cambiamento di R quadrato
- Descrittive
- Correlazioni di ordine zero e parziali
- Diagnostiche di collinearità

Residui

- Durbin-Watson
- Diagnostiche casewise
- Valori anomali oltre: deviazioni standard
- Tutti i casi

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.

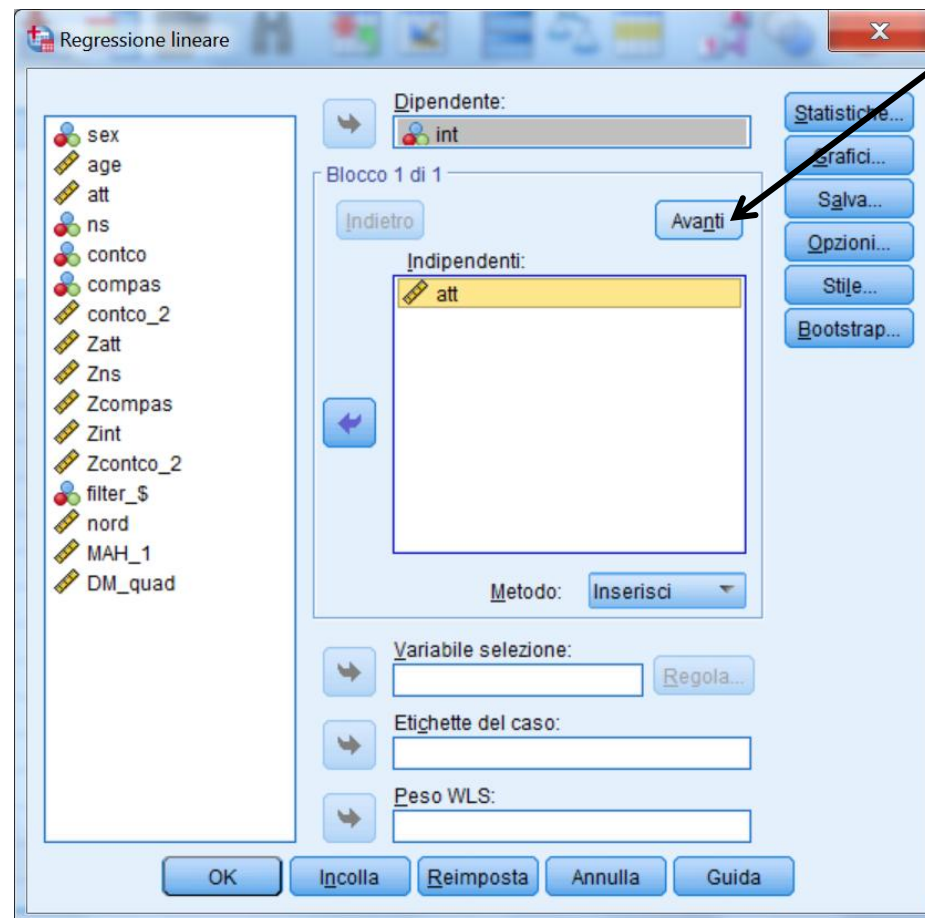
L'analisi procede attraverso "passi" sequenziali. Ogni VI è valutata per quanto aggiunge, nello spiegare la VD, rispetto a quanto è stato spiegato dalle VI inserite precedentemente. **Partizione ordinata della varianza di VD spiegata dalle VI.**

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 lineare

Dipendente: int

Blocco 2 di 2

Indietro Avanti

Indipendenti: ns

Metodo: Inserisci

Variabile selezione: Regola...

Etichette del caso:

Peso WLS:

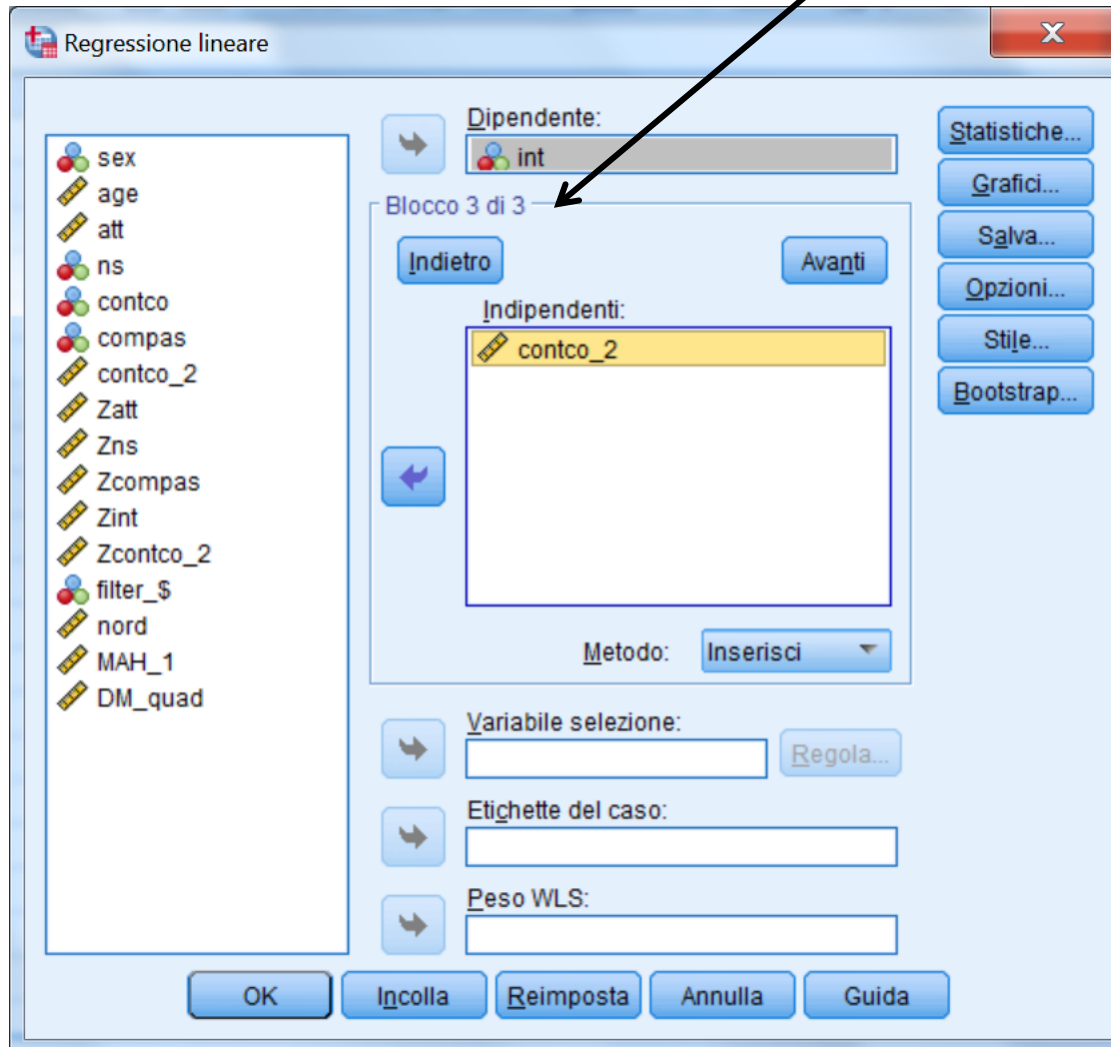
OK Incolla Reimposta Annulla Guida

Statistiche...
Grafici...
Salva...
Opzioni...
Stile...
Bootstrap...

sex
age
att
ns
contco
compas
contco_2
Zatt
Zns
Zcompas
Zint
Zcontco_2
filter_\$
nord
MAH_1
DM_quad

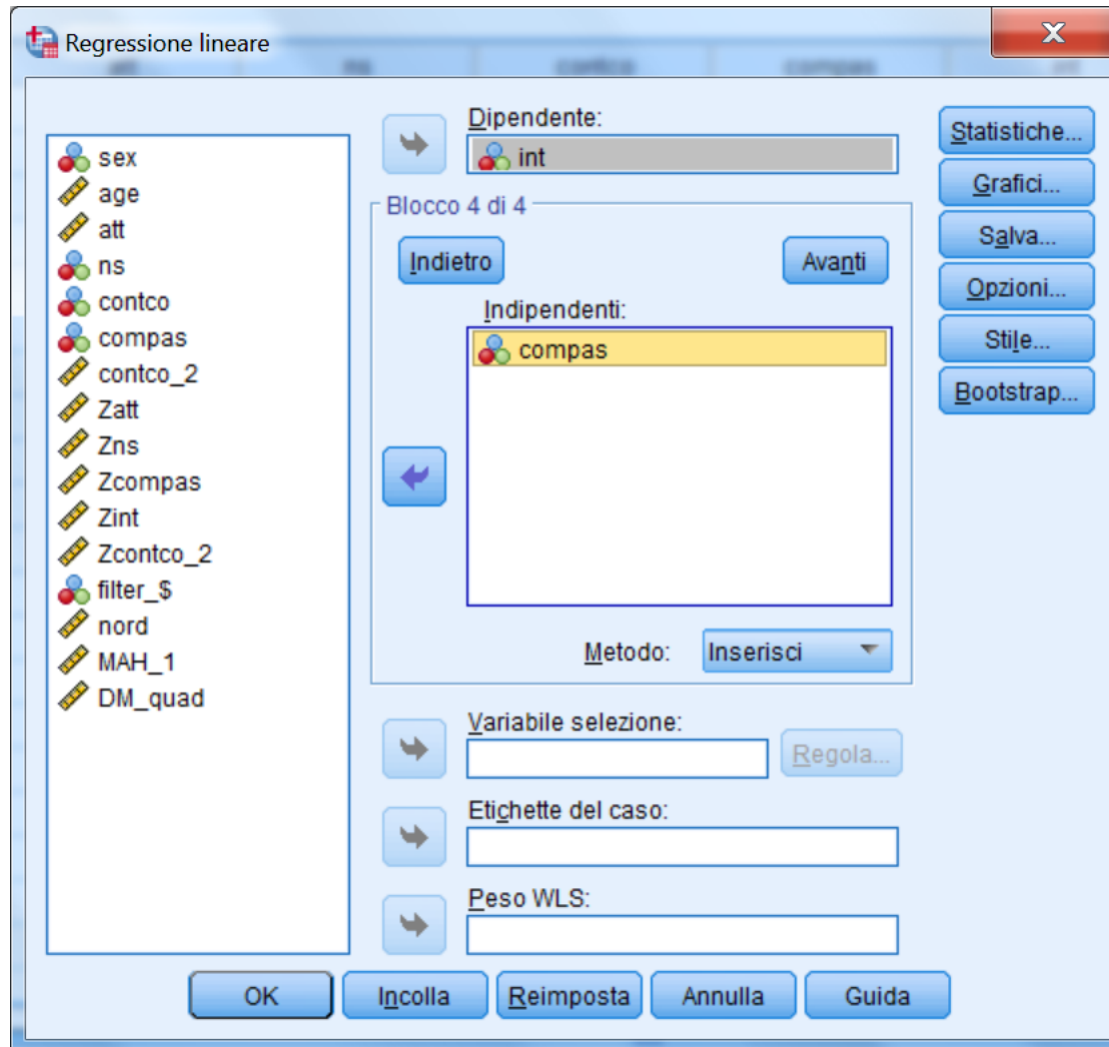
Regressione gerarchica

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



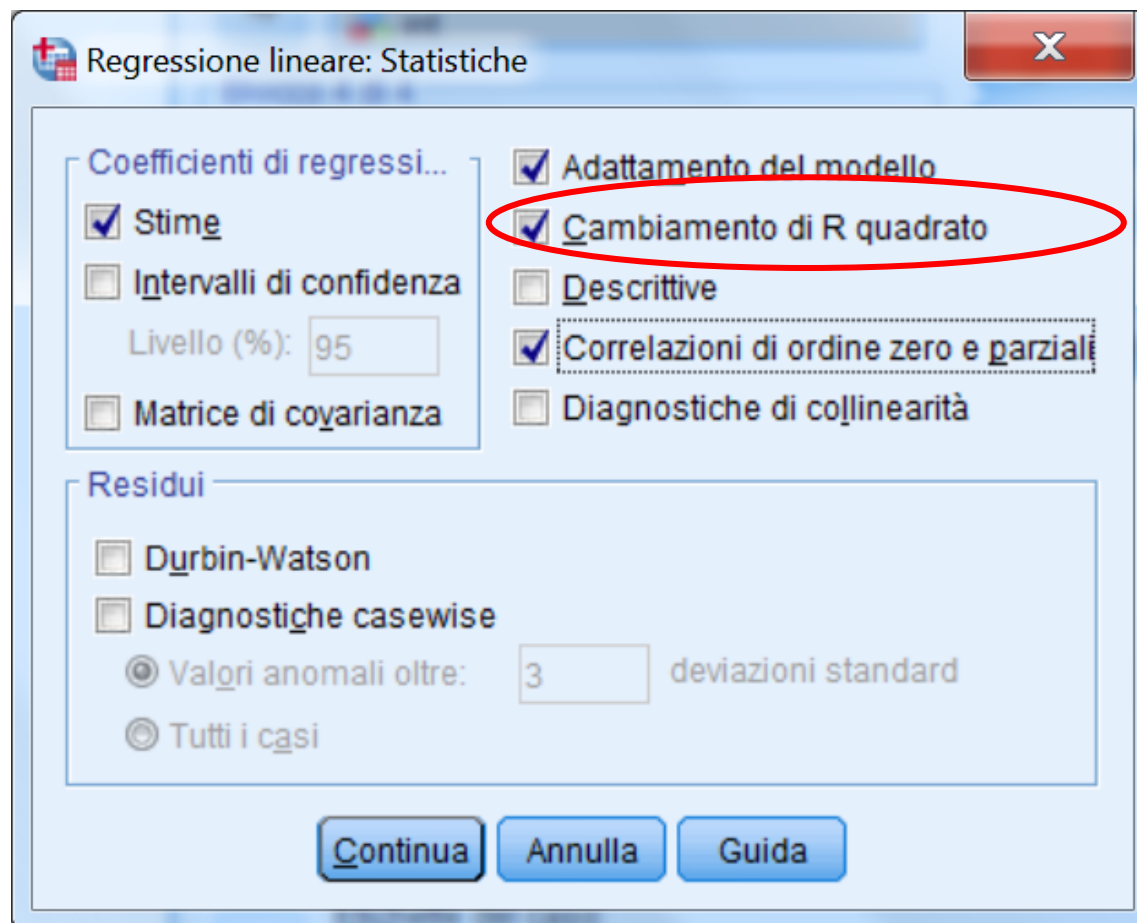
Regressione gerarchica

**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 R² attraverso i riversi passi

Step	Variabile	R	R ²	R ² C	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

sr²: quantità di varianza aggiunta all' R² da ciascuna VI nel punto in cui la VI entra nell'equazione ("incremental sr²" o cambiamento in R²).

La somma degli sr² è uguale al valore di R².



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

(Tabachnik & Fidell, 2007, p. 149)

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

$R_{wi}^2 = R^2$ ottenuta dall'inserimento della nuova variabile

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

$m =$ numero di variabili nel nuovo blocco

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



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

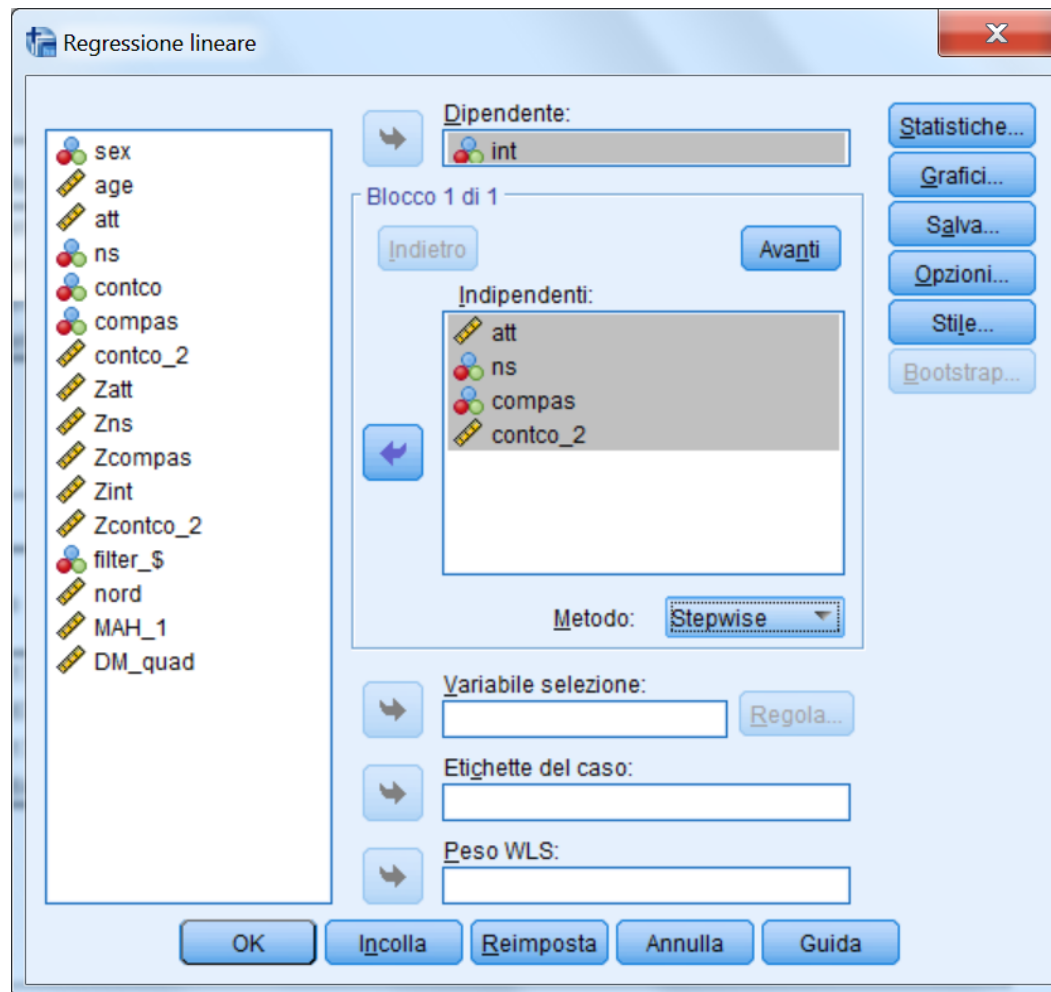
Regressione forward: equazione inizialmente “vuota”; ad ogni step viene aggiunta la VI che presenta la correlazione più elevata con la VD. Se una VI entra in equazione, vi rimane

Regressione backward: 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).

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 \Rightarrow Differenti risultati

Standard \Rightarrow 48% di varianza non attribuibile a nessuna variabile.

Gerarchica \Rightarrow Norma Soggettiva spiega più varianza del comportamento passato

Stepwise \Rightarrow 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 *multicollinearità*: 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,
COMPORAMENTO 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 $SS_T = SS_B + SS_W$:

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

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

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

GRADI DI LIBERTA' E "QUADRATI MEDI"

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

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

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

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)=

$$\text{Devianza totale}/(n-1) = SS_T/(n-1)$$

Varianza tra i gruppi (MS_B)=

$$\text{Devianza tra i gruppi}/(k-1) = SS_B/(k-1)$$

Varianza entro i gruppi (MS_W)=

$$\text{Devianza entro i gruppi}/(n-k) = SS_W/(n-k)$$

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:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

(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.**

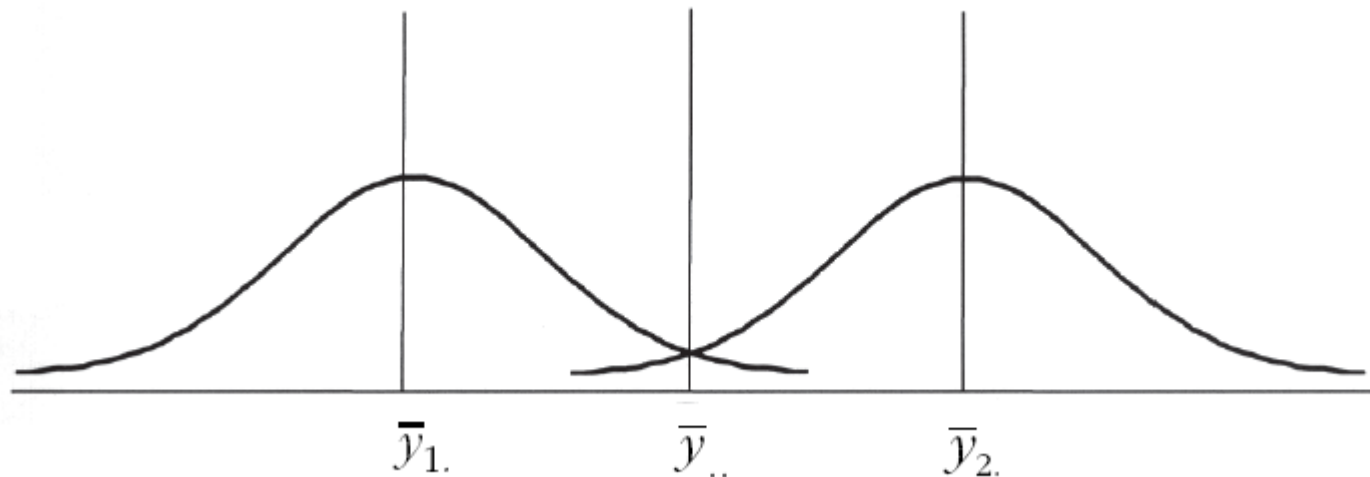
H_0 falsa:

**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 $H_0: \mu_1 = \mu_2 = \dots = \mu_k$)

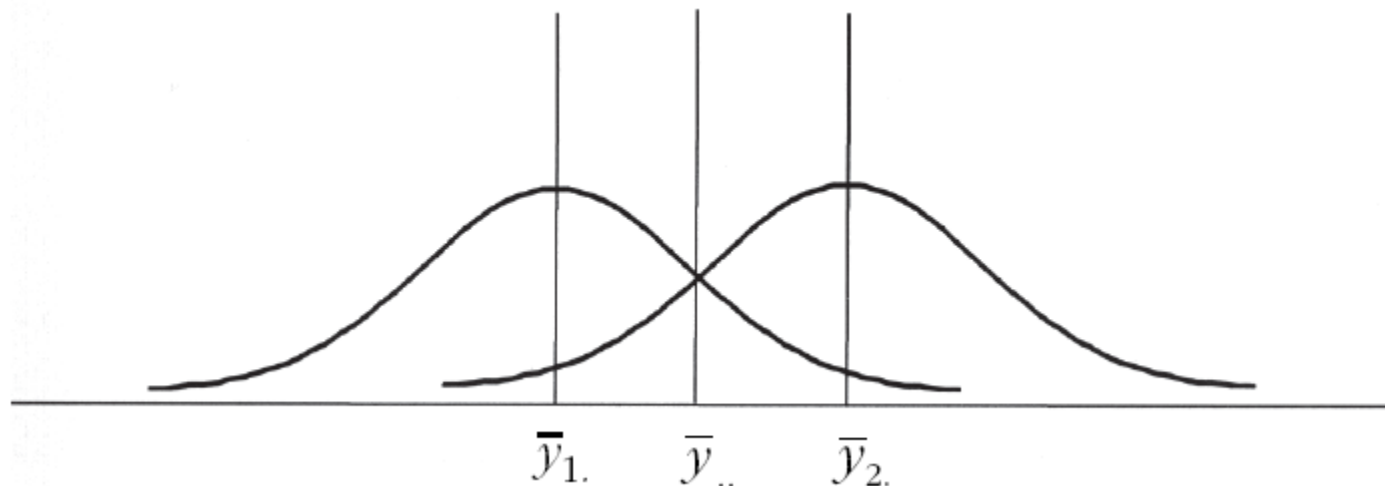
Se la varianza tra i gruppi è maggiore della varianza entro i gruppi, le medie dei gruppi saranno piuttosto distanziate.



RAPPORTO "F"

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

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 intraclassa (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

$$\eta^2 = SS_B / SS_T \text{ eta quadro}$$

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

Effect size nell'ANOVA:

$$\omega^2, \eta^2 = .01 - .05 \rightarrow \text{Basso}$$

$$\omega^2, \eta^2 = .06 - .13 \rightarrow \text{Moderato}$$

$$\omega^2, \eta^2 = .14 \rightarrow \text{Elevato}$$



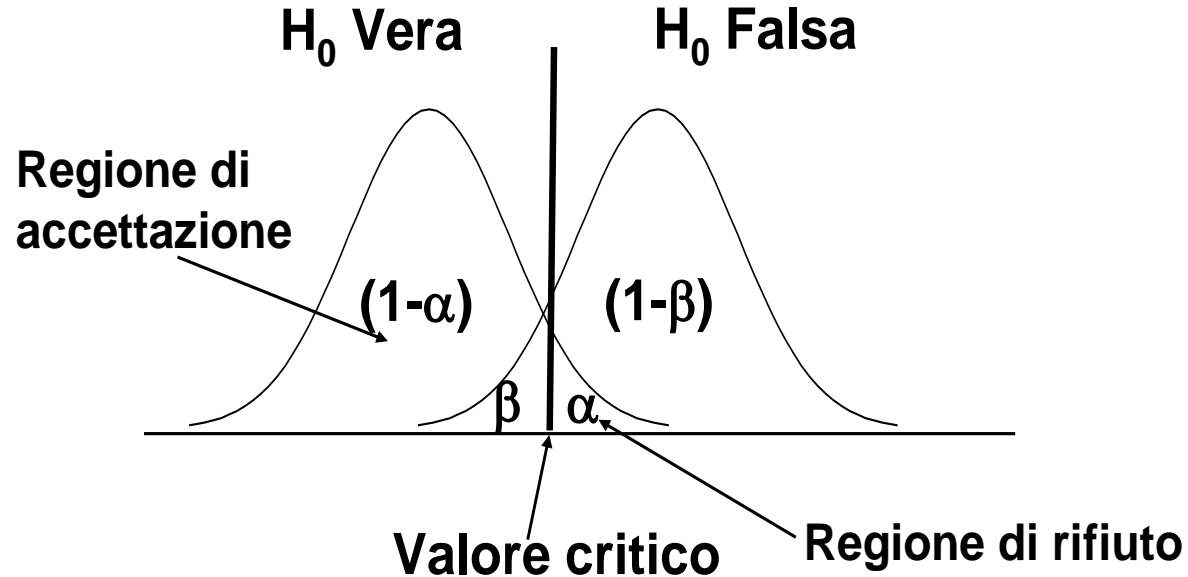
POTENZA DELLA VERIFICA

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	$1-\beta$
.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 ($1-\beta$) 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.

<u>n per gruppo</u>	<u>1-β</u>
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 ($1-\beta$) 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 ($1-\beta$) 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)

$$H_1: \mu_1 \neq \mu_2, \text{ o } \mu_1 \neq \mu_3, \text{ o } \mu_2 \neq \mu_3$$

ovvero, almeno due μ diverse

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

ANOVA IN SPSS

ONEWAY.B.sav [InsiemeDati1] - IBM SPSS Statistics Data Editor

File Modifica Visualizza Dati Trasforma **Analizza** Direct marketing Grafici Utilità Finestra Aiuto

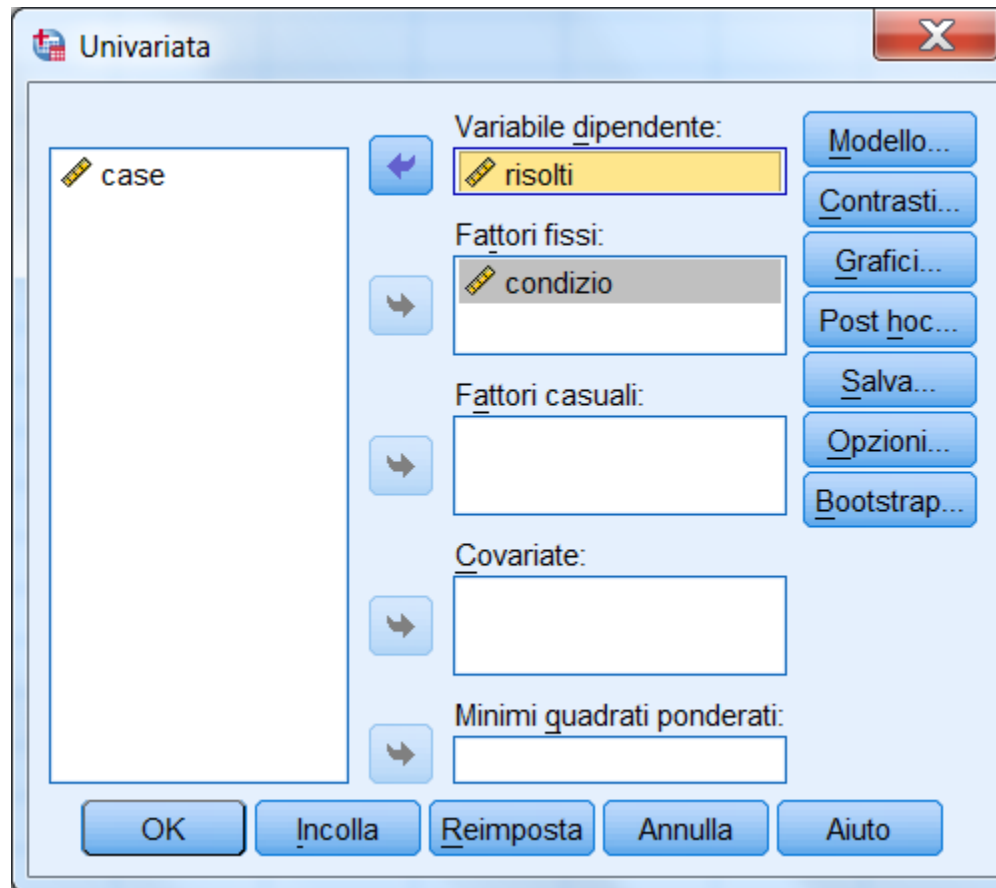
Report
 Statistiche descrittive
 Tabelle
 Confronta medie
Modello lineare generalizzato
 Modelli lineari generalizzati
 Modelli misti
 Correlazione
 Regressione
 Loglineare
 Reti neurali
 Classifica
 Riduzioni dimensione
 Scala
 Test non parametrici
 Previsioni
 Sopravvivenza
 Risposte multiple
 Analisi dei valori mancanti...
 Assegnazione multipla
 Campioni complessi
 Controllo qualità
 Curva ROC...

Univariata...
 Multivariata...
 Misure ripetute...
 Componenti della varianza...

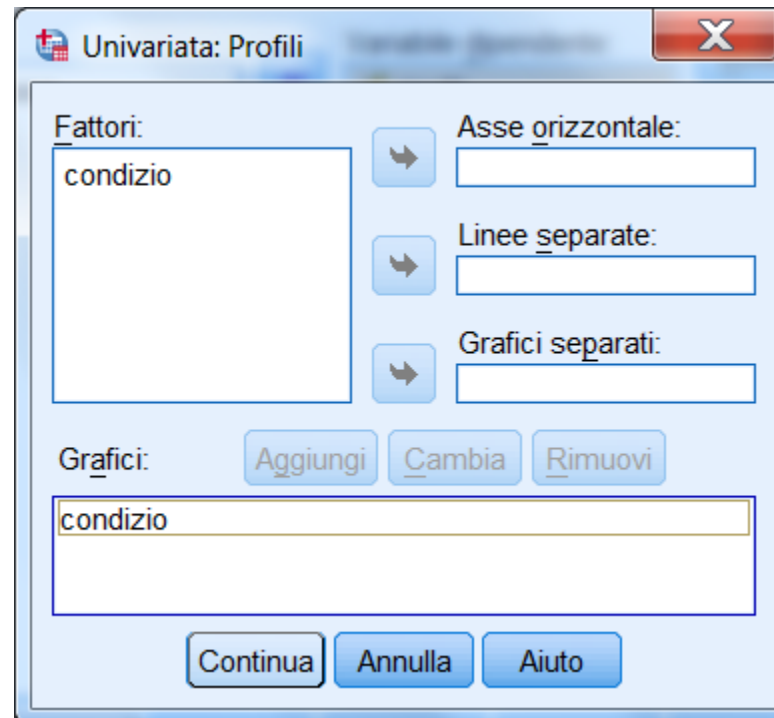
	case	condizio	ri	var	var	var	var	var	var	var	var	var	var
1	1,00	1,00											
2	2,00	1,00											
3	3,00	1,00											
4	4,00	1,00											
5	5,00	1,00											
6	6,00	2,00											
7	7,00	2,00											
8	8,00	2,00											
9	9,00	2,00											
10	10,00	2,00											
11	11,00	3,00											
12	12,00	3,00											
13	13,00	3,00											
14	14,00	3,00											
15	15,00	3,00											
16													
17													
18													
19													
20													
21													
22													
23													

Visualizzazione dati Visualizzazione variabili

ANOVA A UNA VIA "BETWEEN" IN SPSS



ANOVA A UNA VIA "BETWEEN" IN SPSS



ANOVA A UNA VIA "BETWEEN" IN SPSS

Univariata: Opzioni

Medie marginali stimate

Fattori e interazioni fra fattori :

(OVERALL)
condizio

Visualizza medie per:
condizio

Confronta effetti principali

Correzione intervallo di confidenza:
Sidak

Visualizza

Statistiche descrittive
 Stima della dimensione degli effetti
 Potenza osservata
 Stime dei parametri
 Matrice dei coefficienti di contrasto

Test di omogeneità
 Grafico di variabilità contro densità
 Grafici dei residui
 Mancanza di adattamento
 Forma funzionale generalizzata

Livello di significatività: .05 Gli intervalli di confidenza sono 95,0 %

Continua Annulla Aiuto



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

Test di Levene di uguaglianza delle varianze dell'errore^a

Variabile dipendente:risolti

F	df1	df2	Sig.
2,626	2	12	,113

Verifica l'ipotesi nulla per la quale la varianza dell'errore della variabile dipendente è uguale tra i gruppi.

a. Disegno: Intercetta + condizio

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 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 ^b 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.



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 $\alpha = 0.05$ 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 $k\alpha$.

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

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

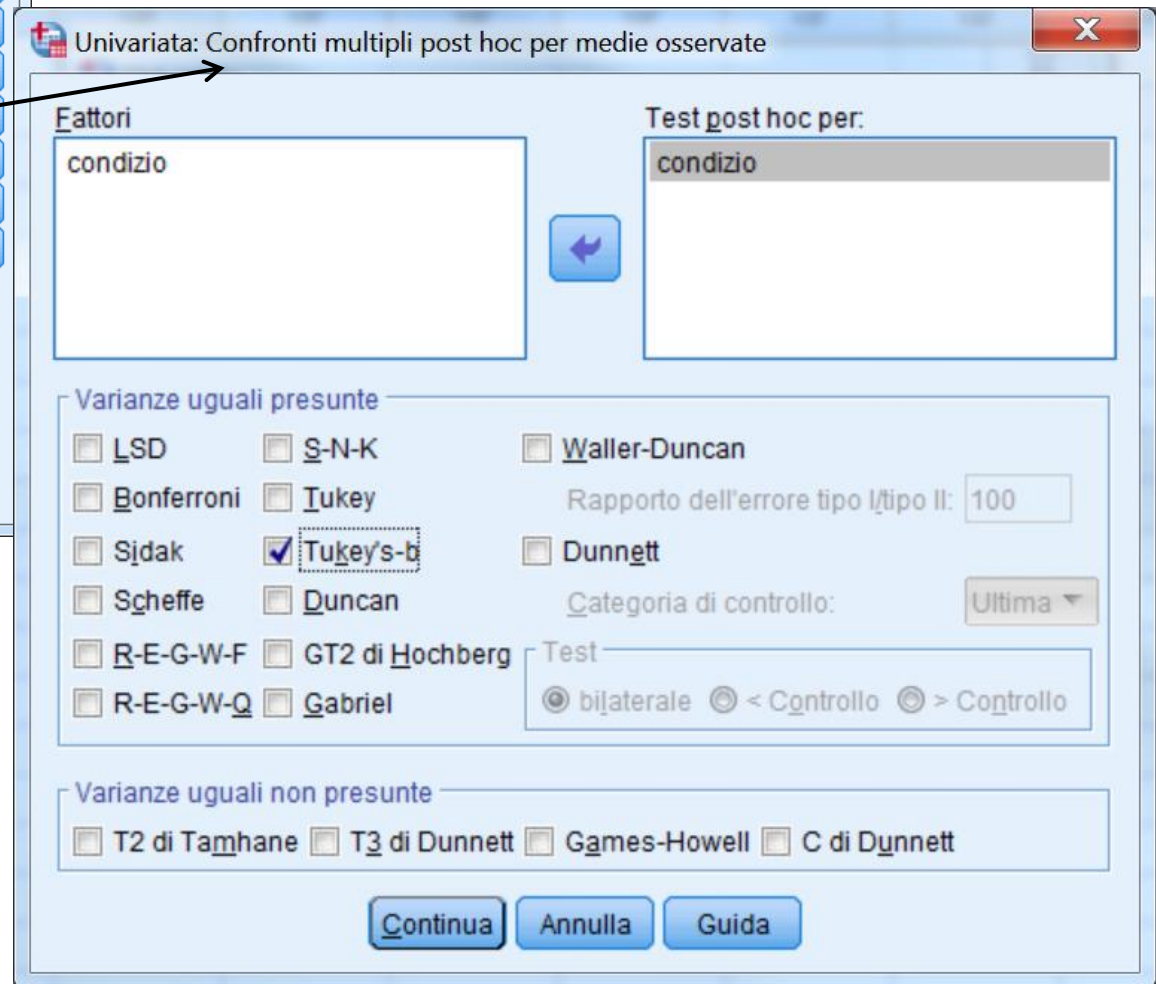
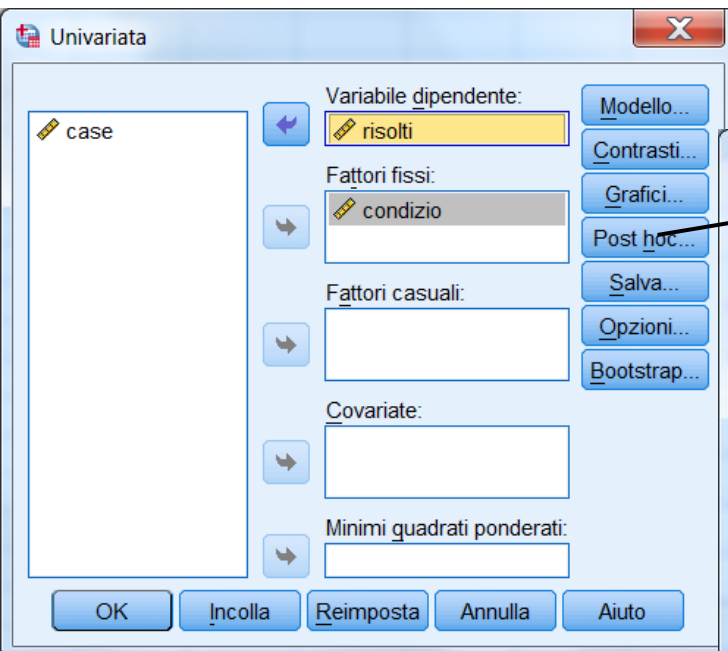
(I) condizio	(J) condizio	Intervallo di confidenza per la differenza al 95% ^a	
		Limite inferiore	Limite superiore
1,00 obiettivi	2,00 obiettivi + feedback	-3,833	3,433
	3,00 controllo	,967	8,233
2,00 obiettivi + feedback	1,00 obiettivi	-3,433	3,833
	3,00 controllo	1,167	8,433
3,00 controllo	1,00 obiettivi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





ANOVA IN SPSS

Sottoinsiemi omogenei

risolti

B di Tukey^{a,b}

condizio	N	Sottoinsieme	
		1	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.

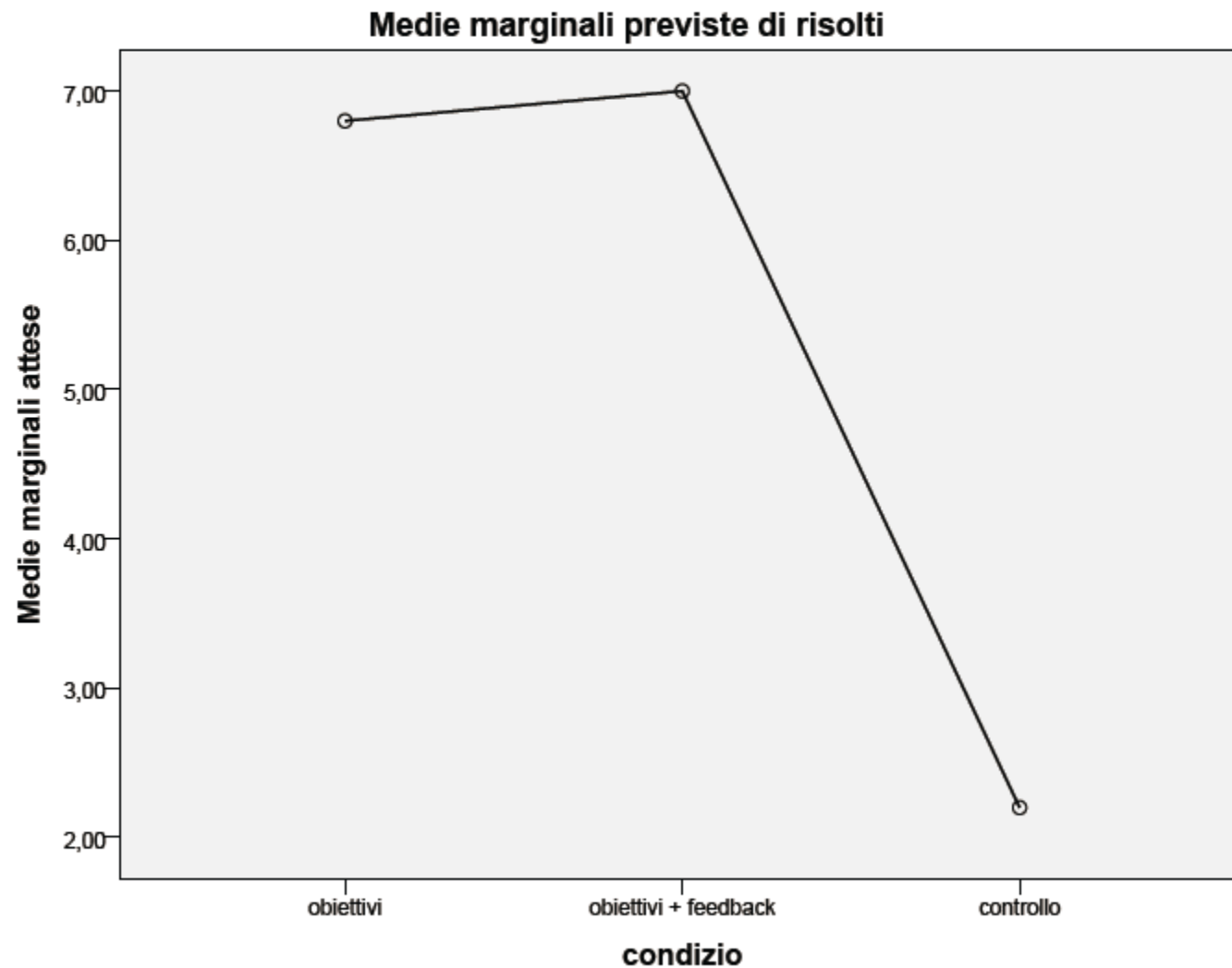
Il termine di errore è media quadratica(errore) = 4,300.

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

b. Alfa = 0,05.

ANOVA A UNA VIA "BETWEEN" IN SPSS

Grafici di profilo





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).



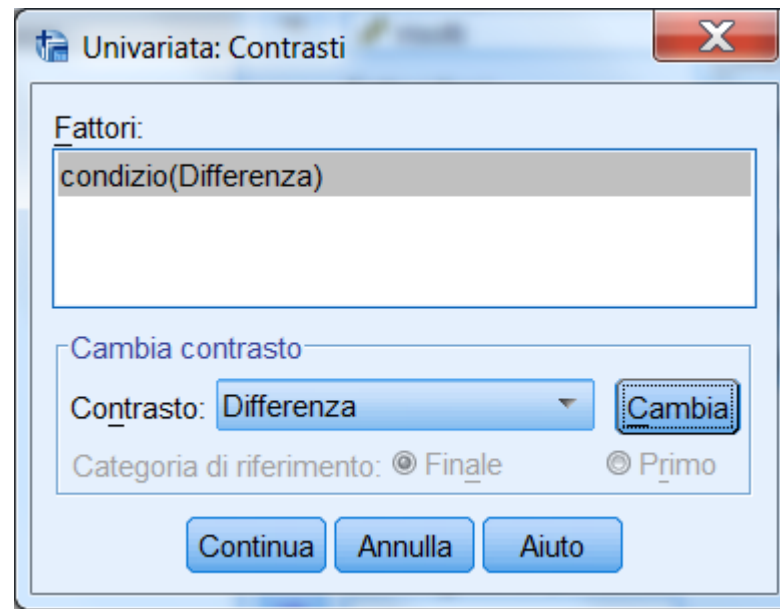
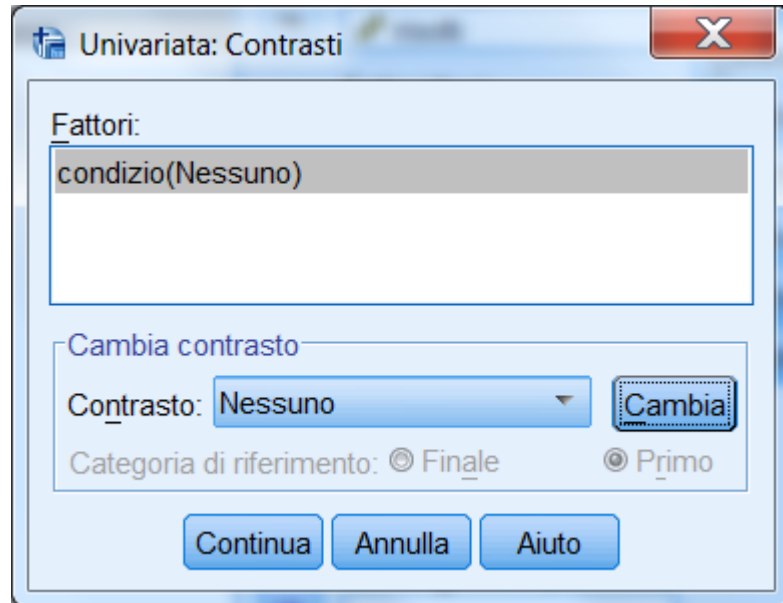
Set di confronti non ortogonali:

	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			$m_{i.}$
	T1	T2	T3	
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 Marginali di Riga

Medie delle
celle

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:

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

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:

$$(85-60) = (80-63) = (76-68), \text{ cioè, } 25 = 17 = 8 ?]$$

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_6
		S_2	S_7
	
NO		S_{11}	S_{16}
		S_{12}	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:

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

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

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

$$SS_{F1 \times F2} = \sum_i \sum_j \sum_k (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \quad \text{dev. interazione}$$

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

$$SS_T = SS_B + SS_W = SS_{F1} + SS_{F2} + SS_{F1 \times F2} + SS_W$$

Gradi di libertà e test di significatività

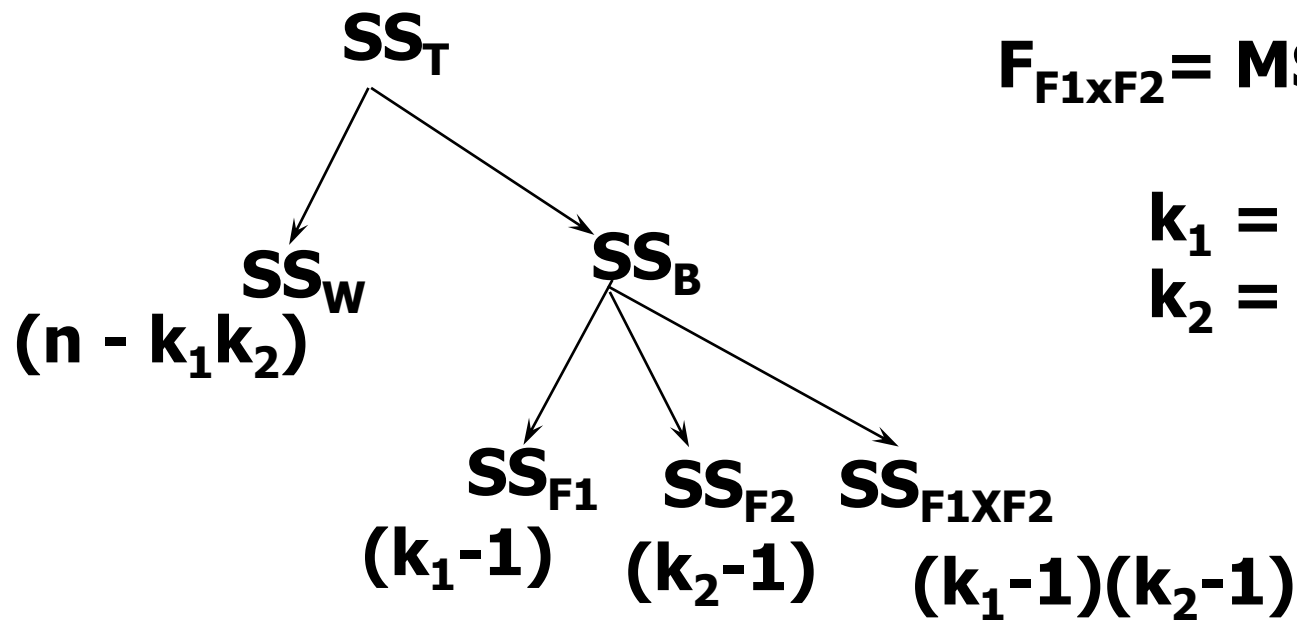
$$F_{F1} = MS_{F1} / MS_W$$

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

$$F_{F1 \times F2} = MS_{F1 \times F2} / MS_W$$

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

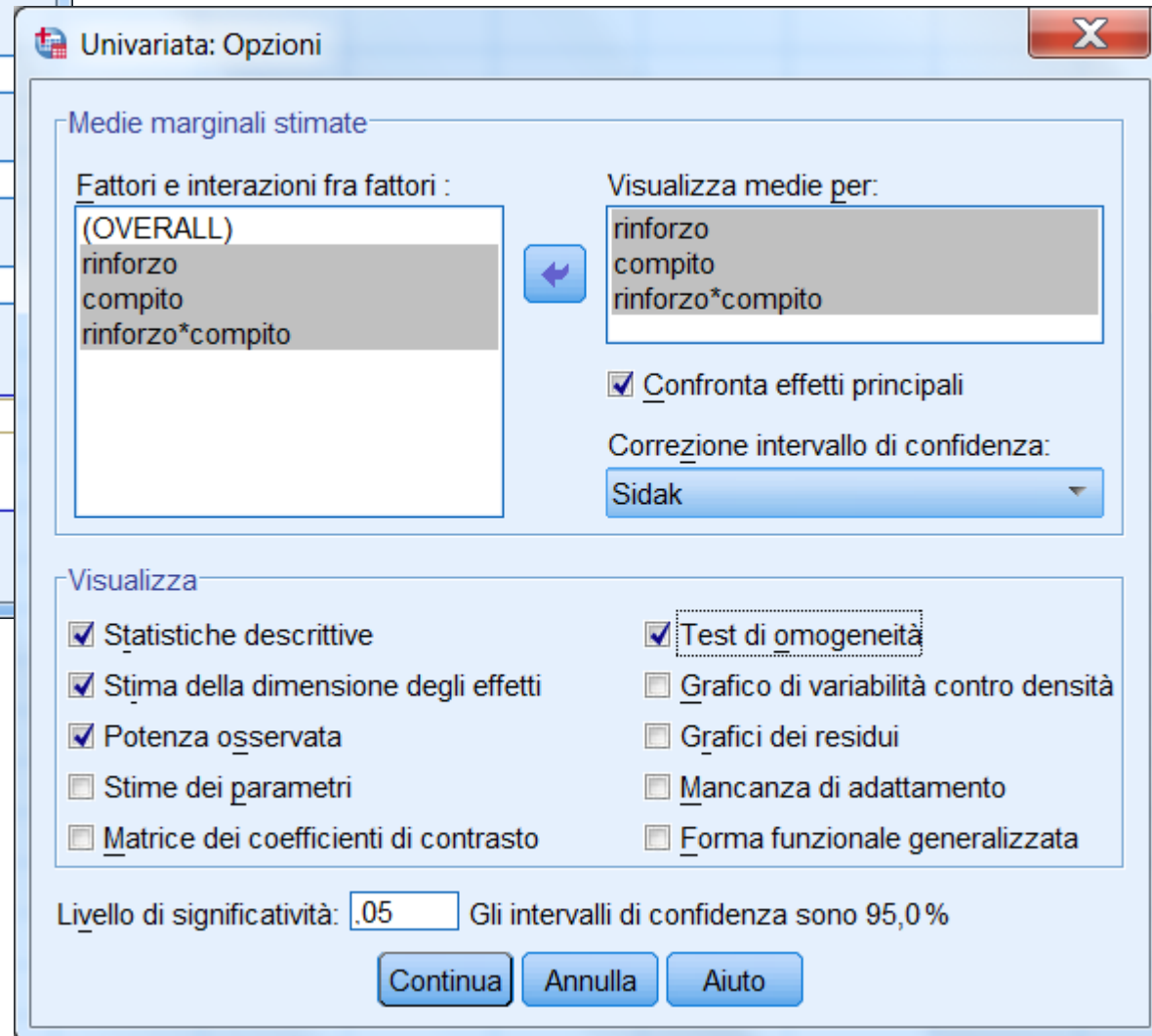
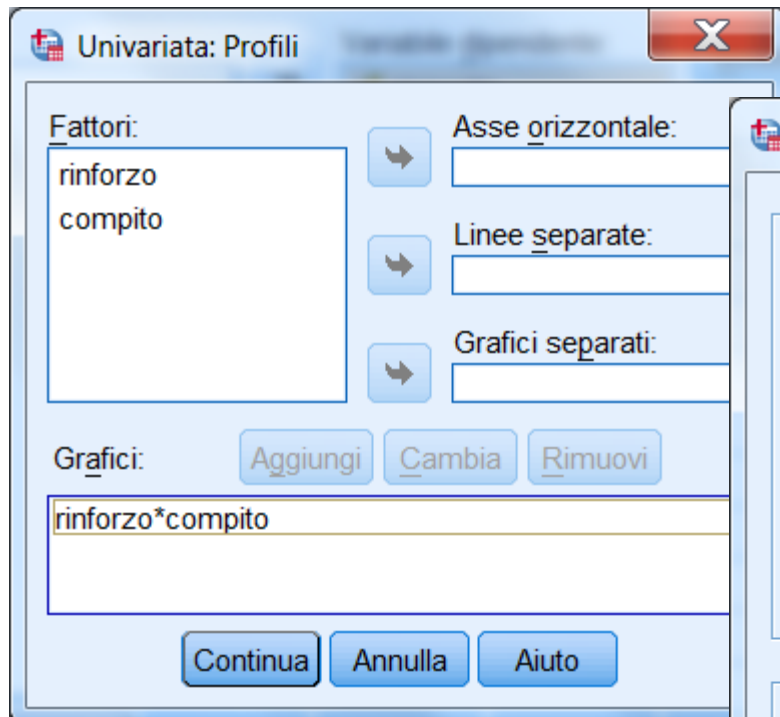
		Etichetta di valore	N
RINFORZO	1,00	LODE	10
	2,00	CRITICA	10
	3,00	SILENZIO	10
COMPITO	1,00	SEMPLICI	15
	2,00	COMPLESSI	15

ANOVA FATTORIALE BETWEEN IN SPSS



ANOVA.FAC.B.sav

ANOVA FATTORIALE BETWEEN IN SPSS



ANOVA FATTORIALE BETWEEN IN SPSS

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

Test di Levene di uguaglianza delle varianze dell'errore^a

Variabile dipendente:risposte

F	df1	df2	Sig.
,348	5	24	,879

Verifica l'ipotesi nulla per la quale la varianza dell'errore della variabile dipendente è uguale tra i gruppi.

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

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 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



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			



ANOVA FATTORIALE BETWEEN IN SPSS

Stime

Variabile dipendente:risposte

rinforzo	Media	Deviazione standard Errore	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

Stime

Variabile dipendente:risposte

compito	Media	Deviazione standard Errore	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

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

Univariata: Confronti multipli post hoc per medie osservate

Fattori:
rinforzo
compito

Test post-hoc per:
rinforzo

Assumi varianze uguali

LSD S-N-K Waller-Duncan
 Bonferroni Tukey Rapporto dell'errore Tipo I / Tipo II: 100
 Sidak Tukey-b Dunnett
 Scheffé Duncan Categoria di controllo: Ultimo
 R-E-G-W-F Hochberg (GT2) Test
 R-E-G-W-Q Gabriel 2 vie < Controllo > Controllo

Non assumere varianze uguali

Tamhane (T2) Dunnett (T3) Games-Howell C di Dunnett

Continua Annulla Aiuto



ANOVA FATTORIALE BETWEEN IN SPSS

Sottoinsiemi omogenei

risposte

B di Tukey^{a,b}

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

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.

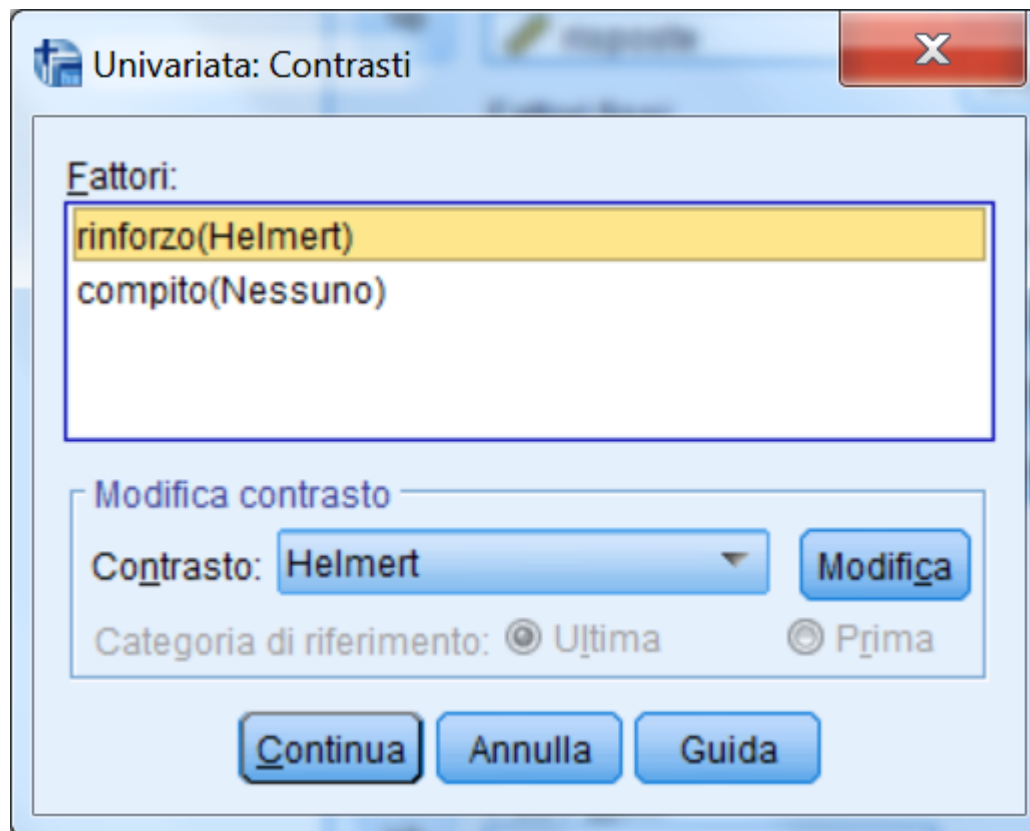
a. Utilizza dimensione campionaria media

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)

		Variabile dipendente	
		risposte	
Contrasto di Helmert rinforzo			
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 Limite superiore	1,605 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 Limite superiore	-,927 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.



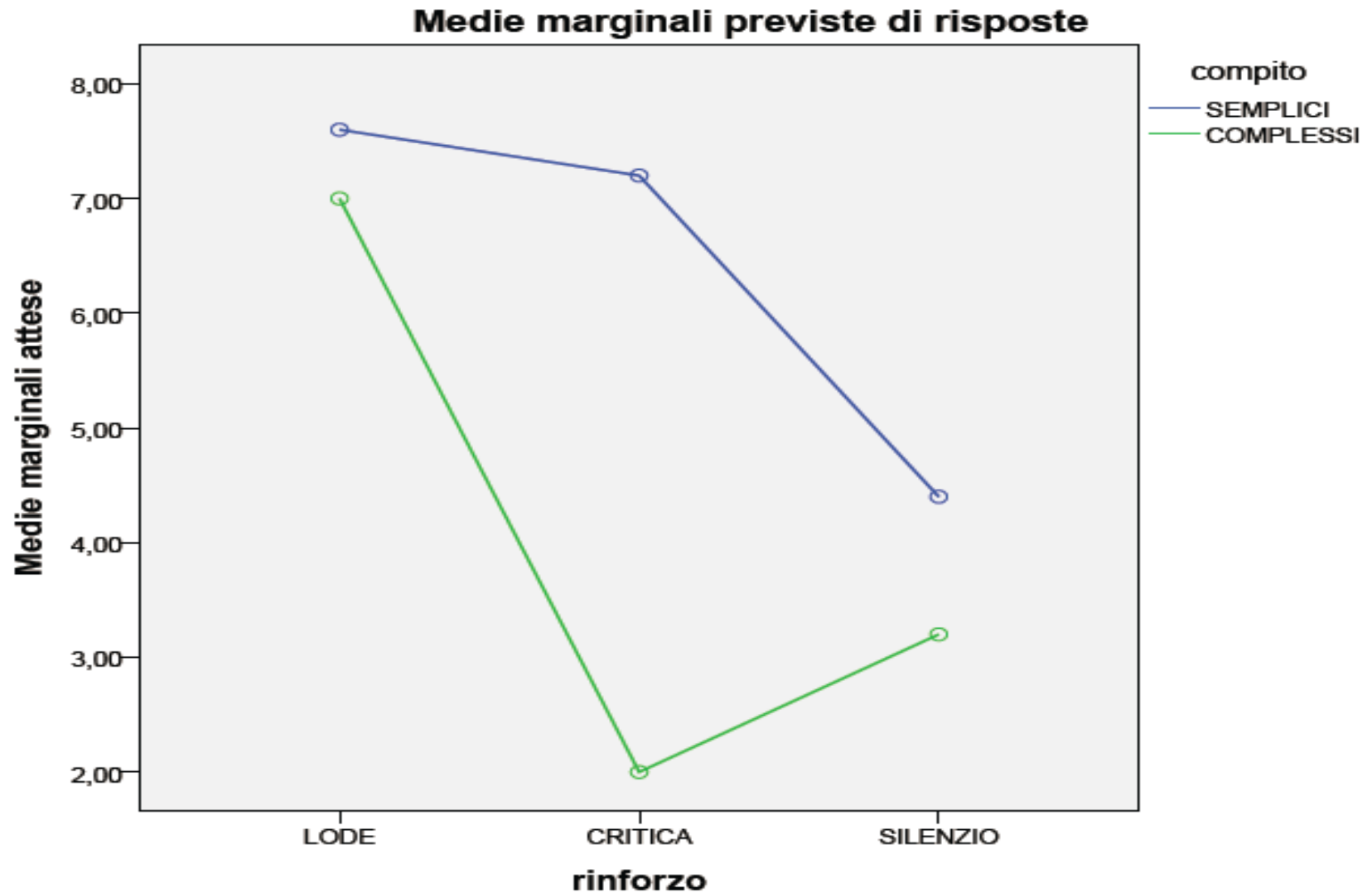
ANOVA FATTORIALE BETWEEN IN SPSS

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

ANOVA FATTORIALE BETWEEN IN SPSS





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	df	Media dei quadrati	F	Sig.
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		

F verifica l'effetto di compito. Questo test è basato sui confronti a coppie indipendenti e lineari tra le medie marginali stimate. Questi test sono basati sui confronti a coppie linearmente indipendenti tra le medie marginali stimate.

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

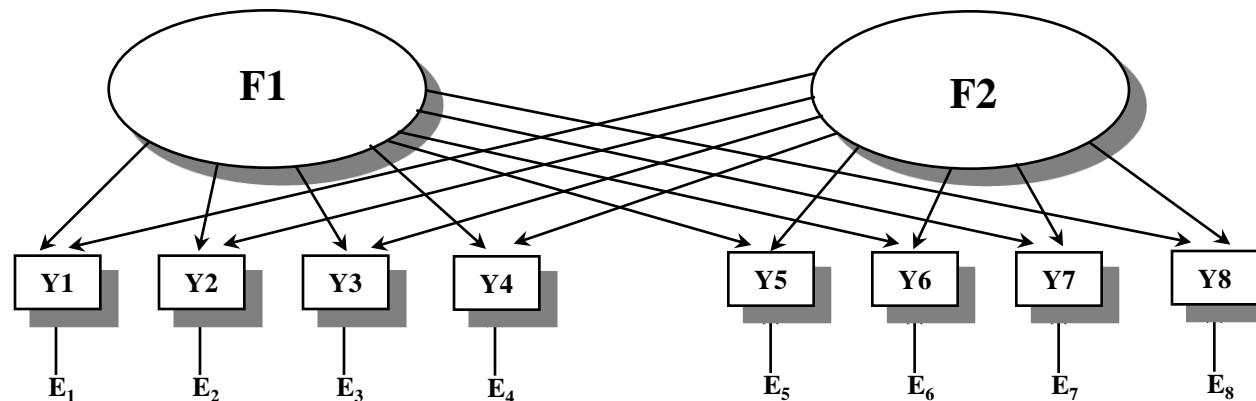


Da dove si parte.... Dati non strutturati

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

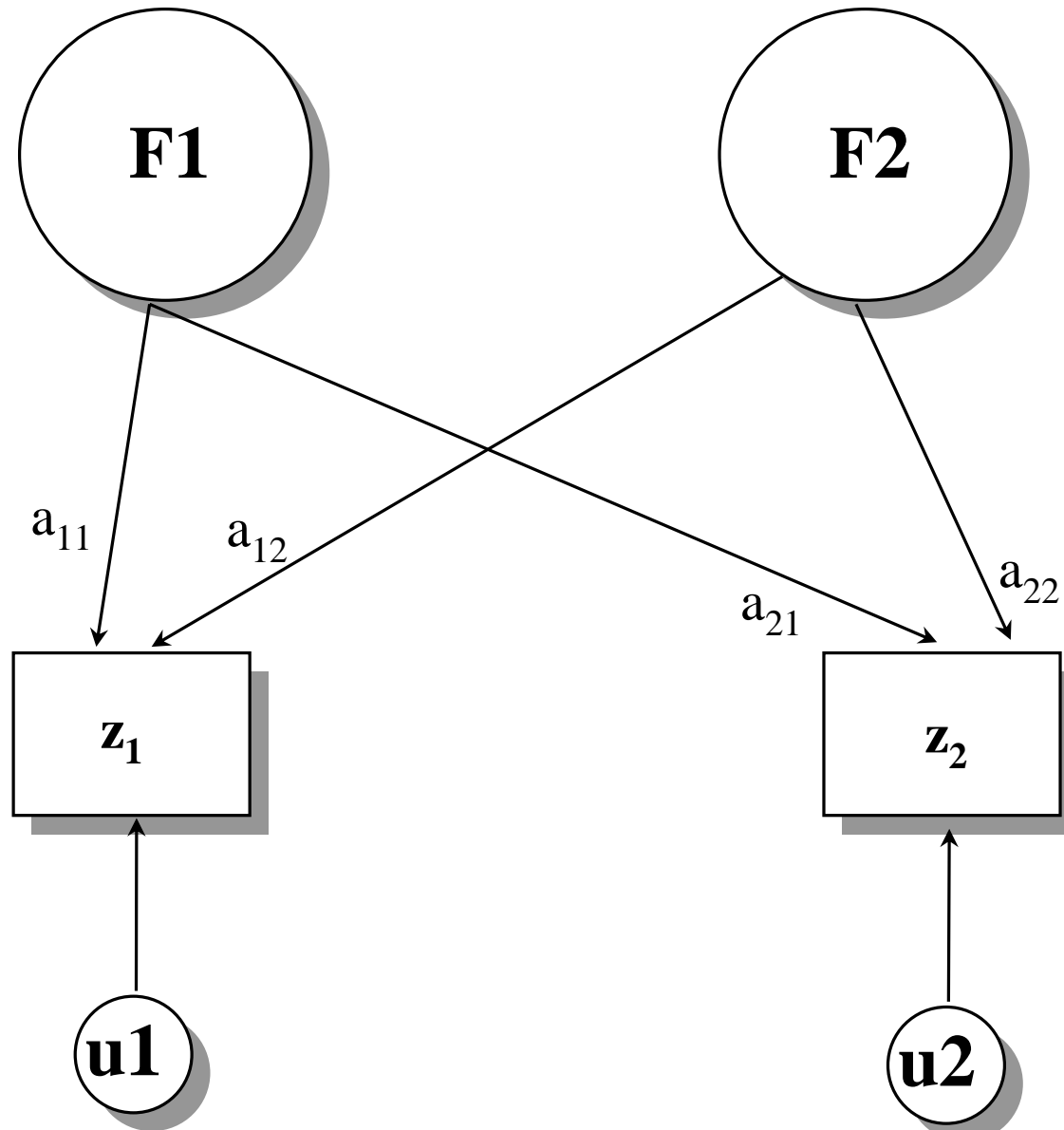
Esame della varianza che le variabili hanno in comune, ovvero della varianza comune.

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

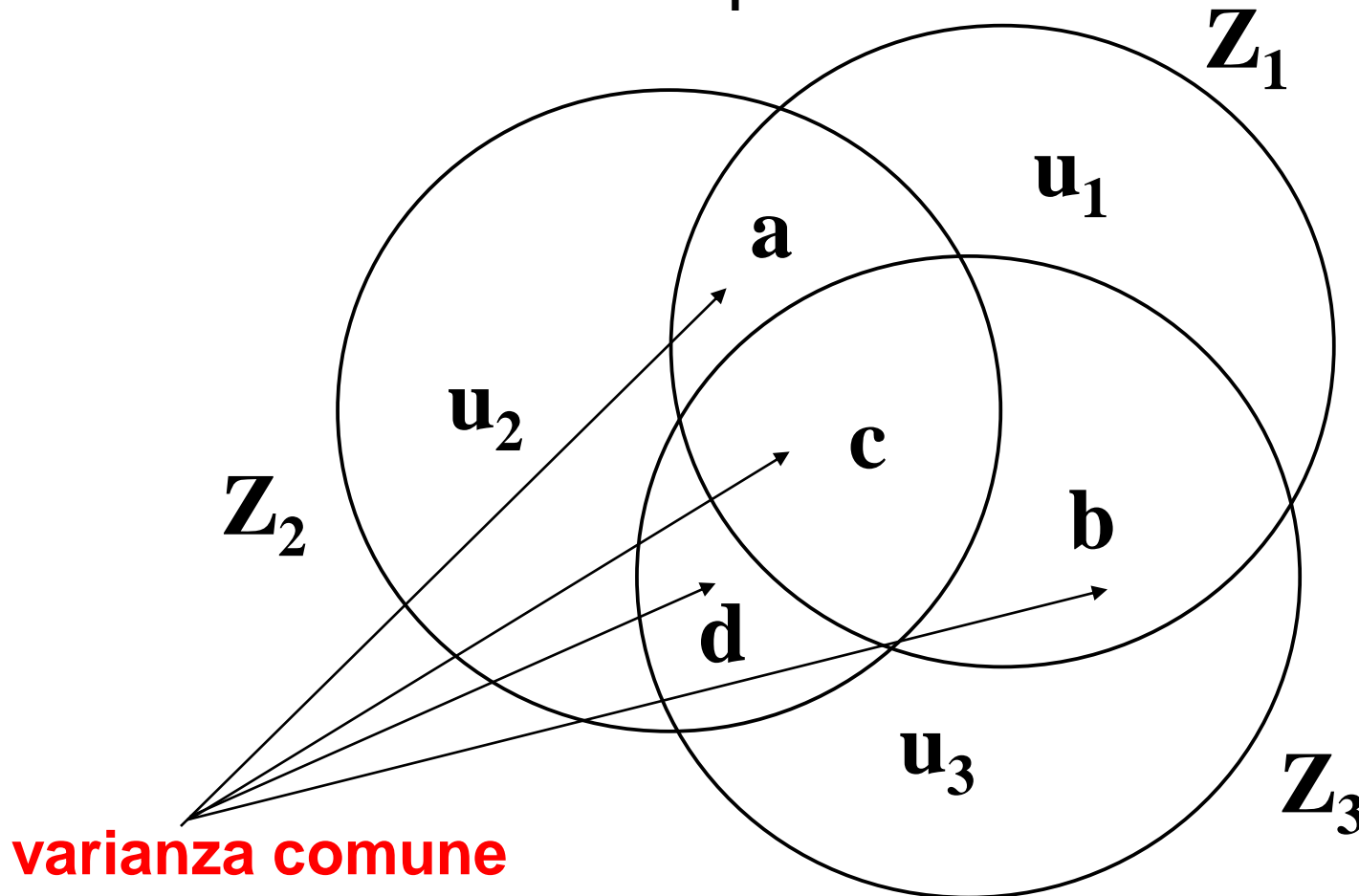
- **F = fattori comuni; rappresentano la variabilità condivisa tra le variabili in analisi. Possono influenzare più di una variabile osservata.**
- **a = saturazioni; relazioni tra variabili e fattori.**
- **u = termine unico o "unicità della variabile". Parte di varianza non condivisa. Dovuta a cause 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 .

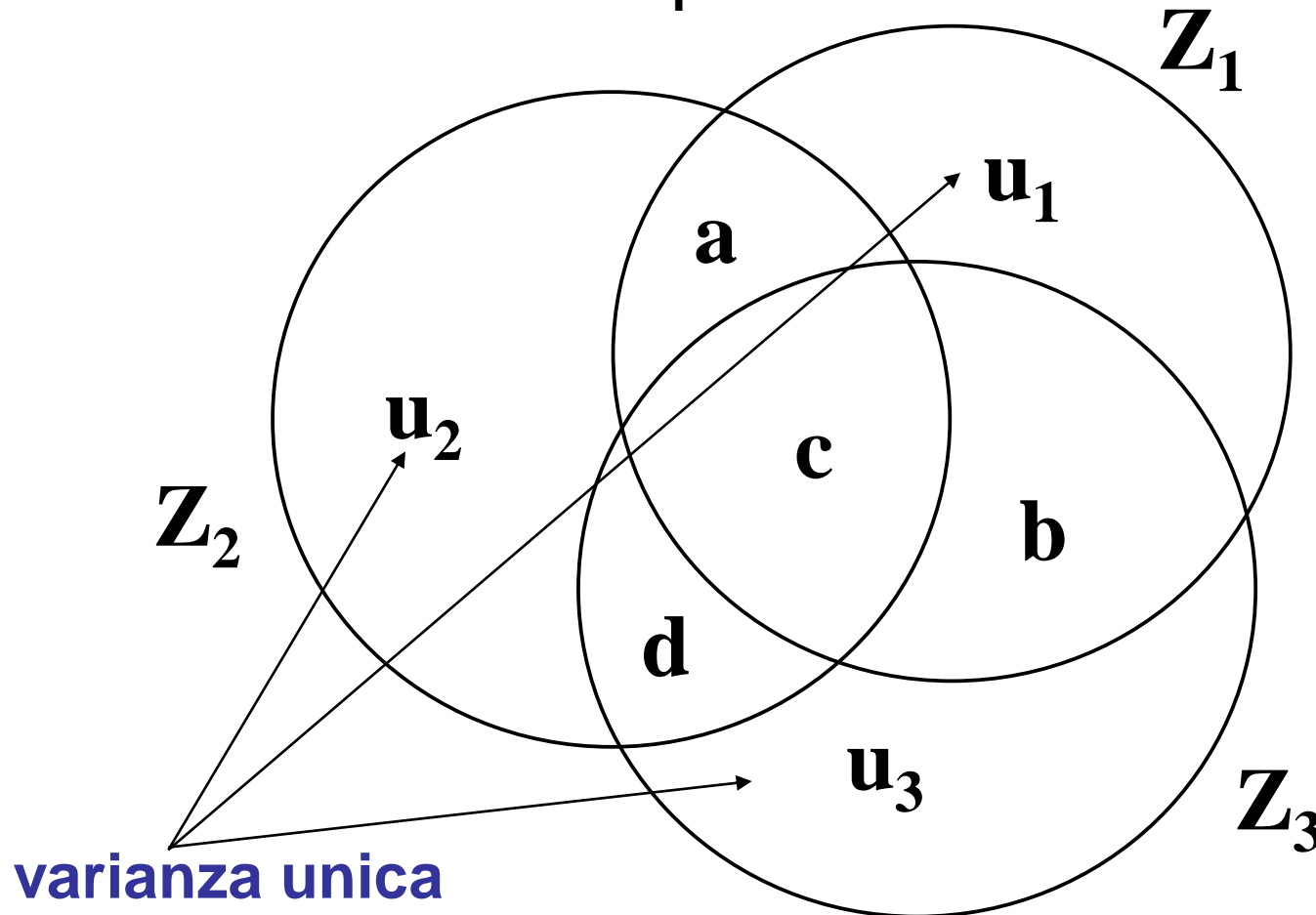




Rappresentazione grafica della varianza comune

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

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

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):

$$\mathbf{Z} = \mathbf{FA}' + \mathbf{U}$$

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:

$\text{Cov}(u_i, F_j) = 0$, per ogni i e per ogni j

$\text{Cov}(u_i, u_j) = 0$ per ogni i diversa da j

$\text{Cov}(F_i, F_j)$ diversa da 0 solo nelle soluzioni "oblique"



Espressioni matriciali

In base alle assunzioni e considerando che:

$$\mathbf{R} = \mathbf{Z}'\mathbf{Z}\mathbf{n}^{-1} \text{ e che } \mathbf{Z} = \mathbf{F}\mathbf{A}' + \mathbf{U}$$

si ha che:

$$\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{U}^2$$

A = matrice delle saturazioni nei fattori comuni

U² = matrice diagonale delle varianze uniche.

AA': rende conto degli elementi fuori della diagonale principale, e della comunaltà di ogni variabile.

U²: contribuisce a rendere conto degli elementi sulla diagonale principale di R.

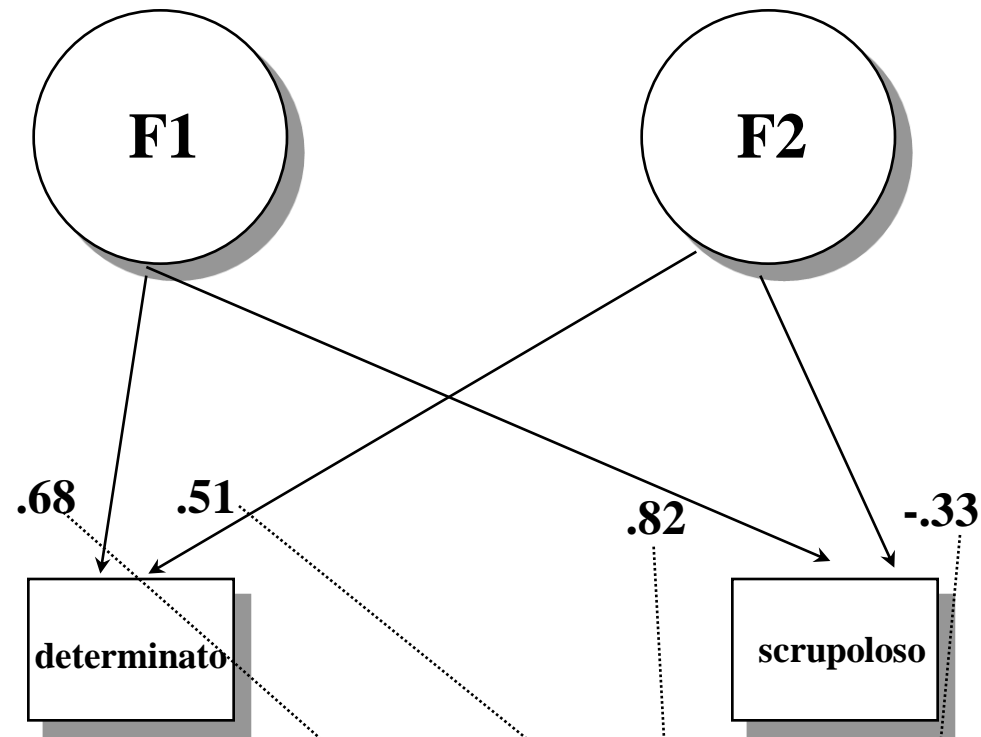


La correlazione tra due variabili i e j può essere riprodotta dalla somma dei prodotti delle loro saturazioni in ciascuno dei fattori comuni:

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

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

Rappresentiamolo con un diagramma....



$$r^{\wedge}(\text{determinato}, \text{scrupoloso}) = (.68 * .82) + (.51 * -.33) = .56 - .17 = .39$$



Correlazione residua

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

$$r(\text{determinato, scrupoloso}) = .40$$

$$r^{\wedge}(\text{determinato, scrupoloso}) = .39$$

$$r \text{ residua (e)} = .40 - .39 = .01$$



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

“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).

$$\mathbf{R}^* = \mathbf{A}\mathbf{A}' \quad (1)$$

$$\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{U}^2 \quad (2)$$

\mathbf{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:

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

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

Gli autovettori \mathbf{x}_1 relativo a λ_1 e \mathbf{x}_2 relativo a λ_2 sono:

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



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 **$R = VL V'$** .

**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

$$R = VLV'$$

$$\begin{array}{ccc|c}
 1 & .45 & .30 & \\
 .45 & 1 & .50 & \\
 .30 & .50 & 1 & \\
 \hline
 \mathbf{R} & = & & \\
 \end{array}
 =
 \begin{array}{ccc|c}
 \textcircled{-.54} & \textcircled{.76} & \textcircled{.36} & \\
 \textcircled{-.63} & \textcircled{-.08} & \textcircled{-.78} & * \\
 \textcircled{-.57} & \textcircled{-.64} & \textcircled{.52} & \\
 \hline
 & \mathbf{V} & & \\
 \end{array}
 *
 \begin{array}{ccc|c}
 \textcircled{1.84} & 0 & 0 & \\
 0 & \textcircled{.70} & 0 & \\
 0 & 0 & \textcircled{.46} & \\
 \hline
 & \mathbf{L} & & \\
 \end{array}
 *
 \begin{array}{ccc|c}
 \textcircled{-.54} & \textcircled{-.63} & \textcircled{-.57} & \\
 \textcircled{.76} & \textcircled{-.08} & \textcircled{-.64} & \\
 \textcircled{.30} & \textcircled{-.78} & \textcircled{.52} & \\
 \hline
 & \mathbf{V}' & & \\
 \end{array}$$

Primo autovettore e
primo autovalore di R

Secondo autovettore e
secondo autovalore di R

Secondo autovettore e
secondo autovalore di R



**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) = comunaltà 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 comunaltà viene analizzata la matrice di correlazione R^* , con la stima delle comunaltà (\hat{h}_j^2) sulla diagonale principale.

Nelle analisi che non utilizzano stime delle comunaltà 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 ($c_i = X_i V$, 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 comunaltà inserite nella diagonale principale.**
- **Analizza solo la varianza attribuibile ai fattori “comuni” (ovvero la comunaltà) 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 comunaltà 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 comunaltà 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:**

$$\sum_j \sum_k (r_{jk} - r_{jk}^{\wedge})^2$$



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^{\wedge} 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^{\wedge} .

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

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 $\text{tr}()$ 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.

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

$$\text{Deve valere sempre: } (n-k)^2 > (n+k)$$

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

Esempio: $n = 8$

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

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

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

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

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

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: $R = R^{\wedge}$

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 $H_0: R = R^{\wedge}$, 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)

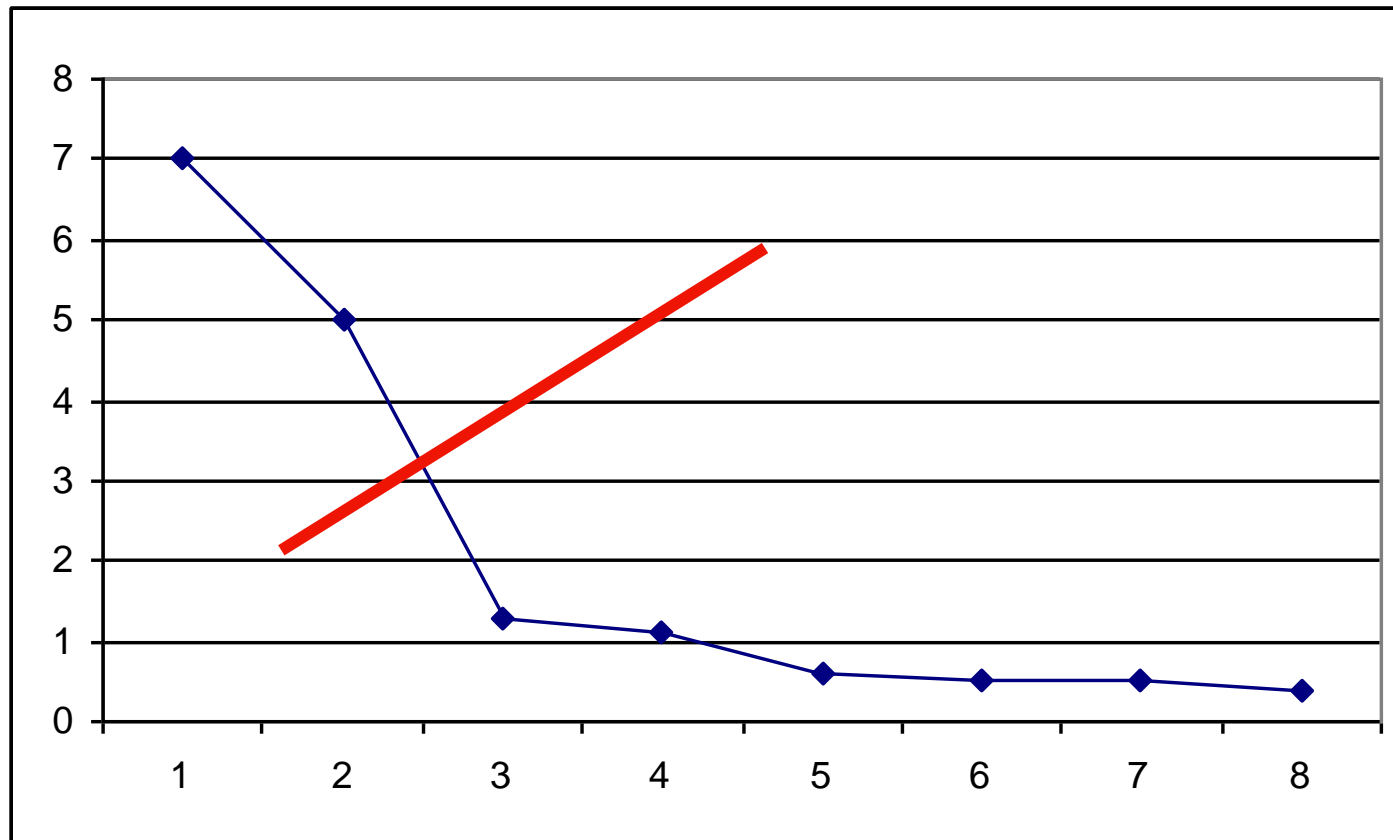
I primi fattori sono i più attendibili e i più validi, poiché spiegano una percentuale di varianza maggiore rispetto ai rimanenti, e avranno autovalori più grandi degli altri.

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

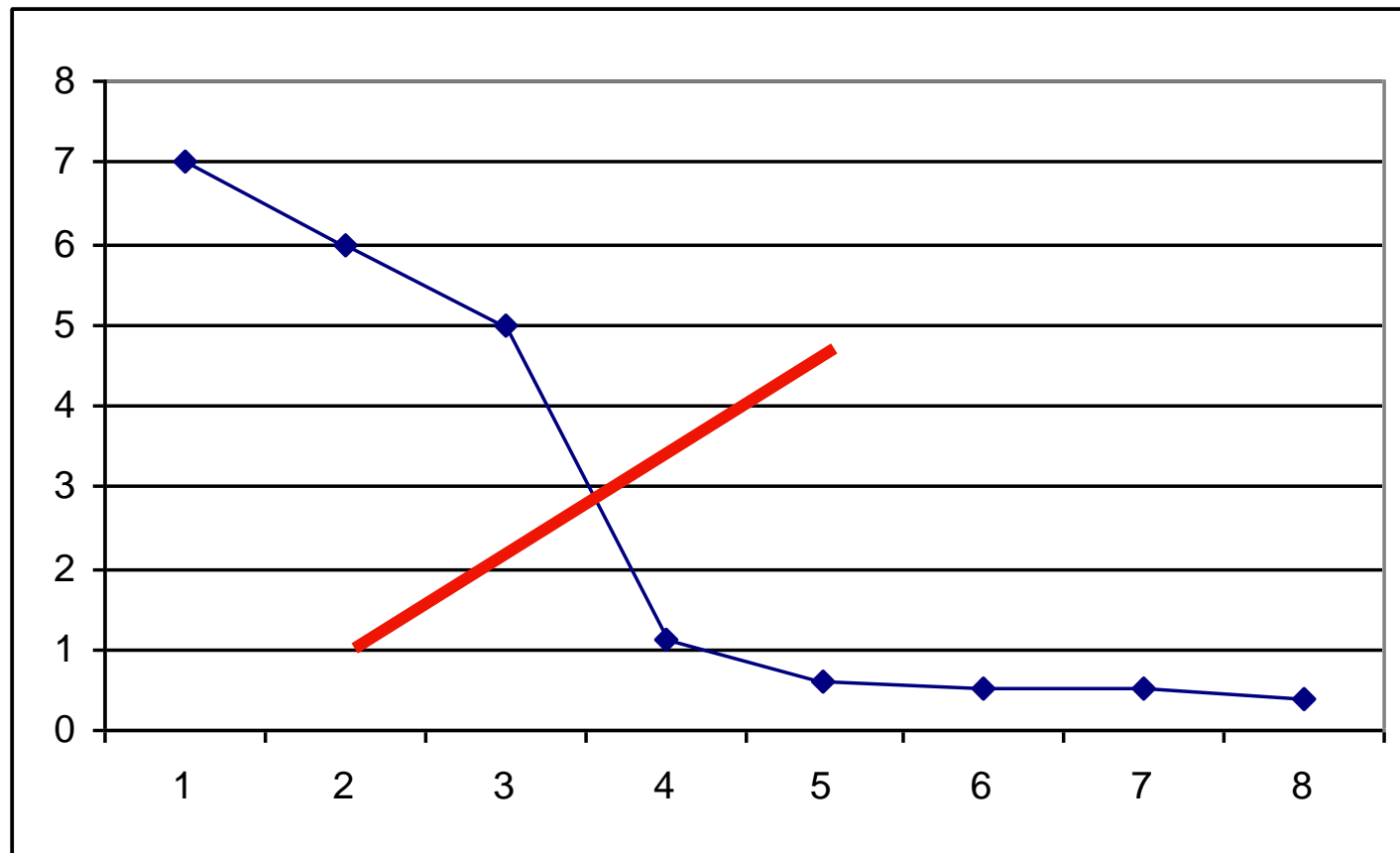
Il processo di estrazione si interrompe nel punto in cui la curva degli autovalori decrescenti cambia pendenza e diventa sostanzialmente piatta. Vanno presi quei fattori i cui autovalori sono al di sopra della linea piatta formata dagli autovalori dei fattori più piccoli.



Applicazione più attendibile quando il campione è grande, le comunaltà 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/briocconn/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 $(r - r^{\wedge})$, ovvero correlazione osservata meno correlazione riprodotta. La matrice dei residui quindi si ottiene nel modo seguente: $E = (R - R^{\wedge})$.

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:

$$\mathbf{AT} = \mathbf{B}, \text{ e } \mathbf{R} = \mathbf{BB}'$$

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 trasformazione (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, $\mathbf{AT}=\mathbf{B}$.

$\mathbf{R}=\mathbf{BB}'$, ma $\mathbf{B}=\mathbf{AT}$, quindi $\mathbf{R}=(\mathbf{AT})(\mathbf{AT})'=\mathbf{ATT}'\mathbf{A}'$

E' possibile dimostrare che $\mathbf{TT}'=\mathbf{T}'\mathbf{T}=\mathbf{I}$, quindi:

$$\mathbf{R}=\mathbf{BB}'=\mathbf{AA}'$$

Questo fenomeno viene definito indeterminatezza della soluzione fattoriale: esistono infinite matrici A tali che $\mathbf{R}=\mathbf{AA}'$, cioè che riproducono una data matrice R altrettanto bene.



Soluzioni fattoriali

	Soluzione iniziale		Soluzione ortogonale		Soluzione obliqua		h ²
	F1	F2	F1	F2	F1	F2	
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).

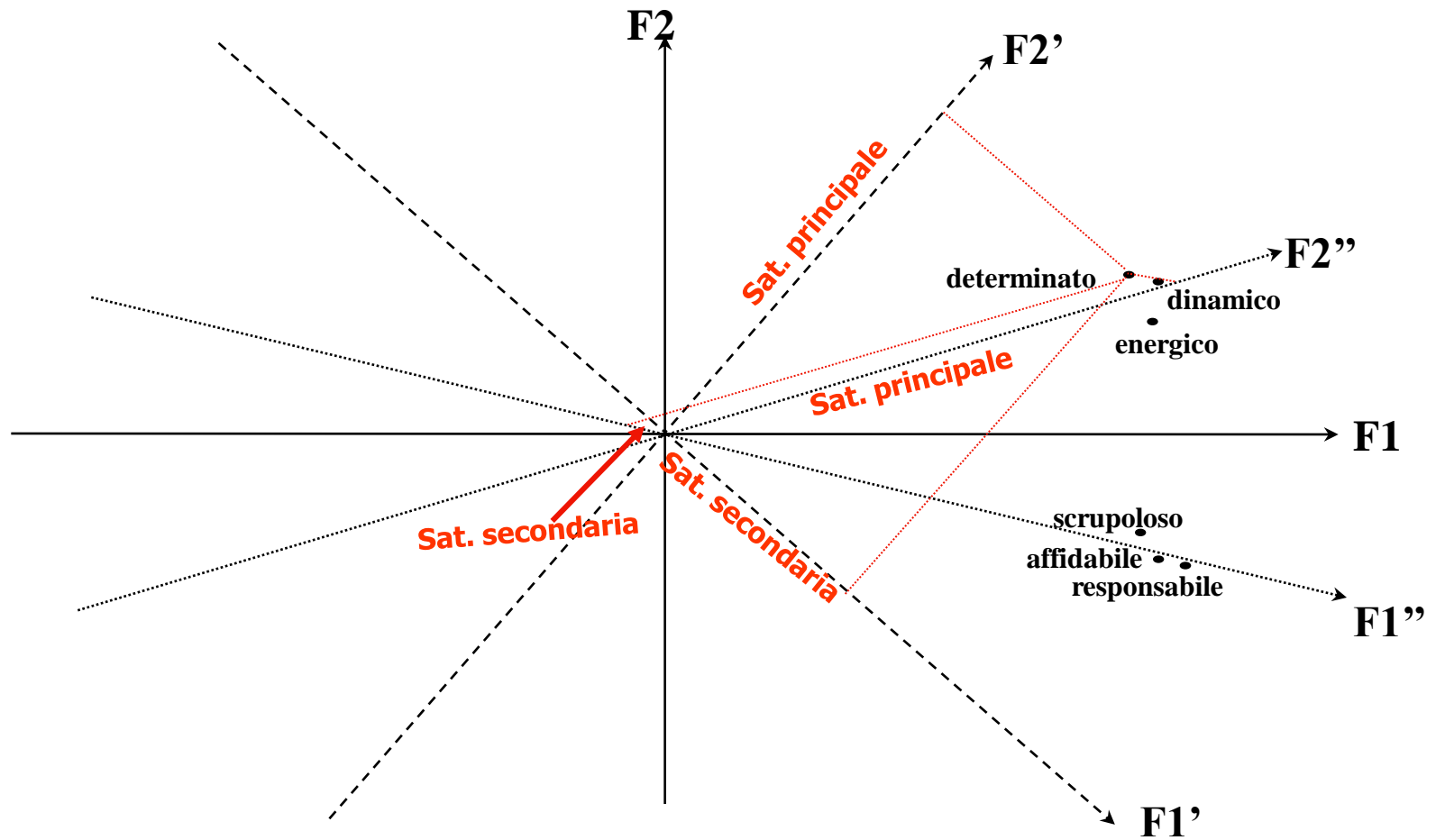


La rotazione dei fattori – La struttura semplice (Thurstone, 1947)

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.**

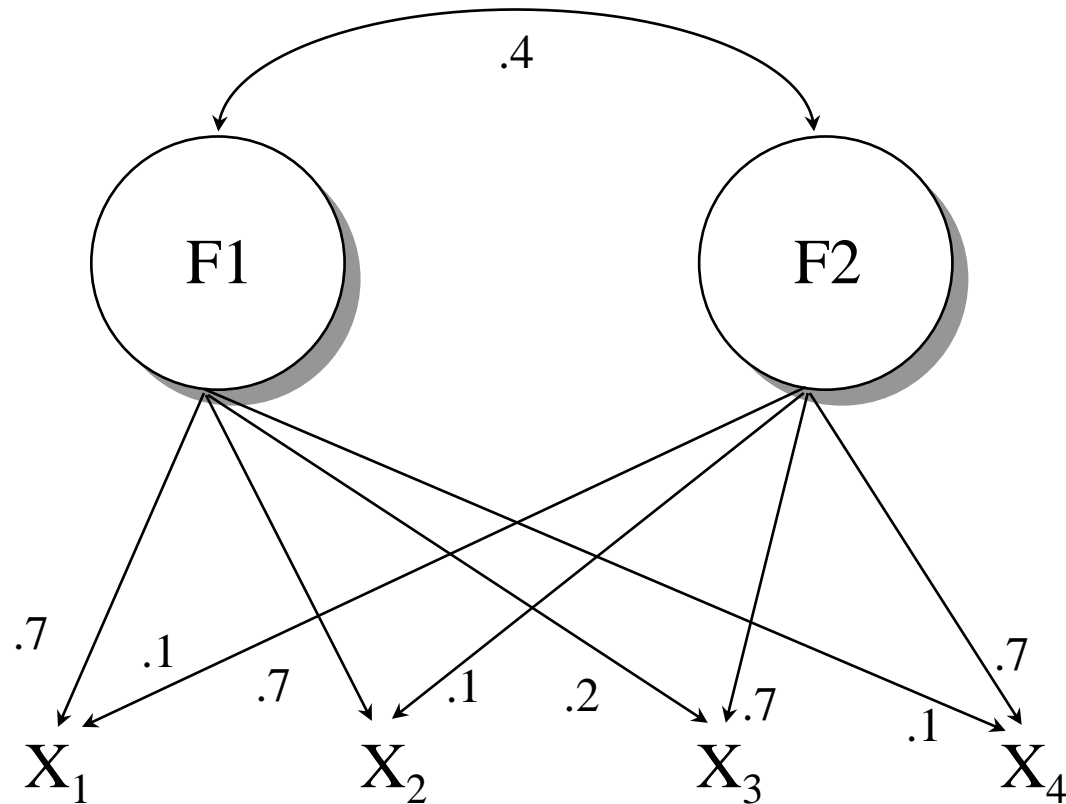
**Influenza *unica* di ciascun fattore sulle variabili
(pesi beta, β).**

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

$$R^* = SP' \quad \text{e} \quad R = SP' + U^2$$

$$R^* = P\Phi P' \quad \text{e} \quad R = P\Phi P' + U^2$$

Nelle rotazioni ortogonali invece:

$$S = P = A, \quad \Phi = I, \quad \text{quindi}$$

$$R^* = P\Phi P' = AA'$$

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)		h ²
	F1	F2	F1	F2	F1	F2	
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: comunaltà.

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:

$$KMO = \frac{\sum \sum r^2}{(\sum \sum r^2 + \sum \sum p^2)}$$

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 Preliis, 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



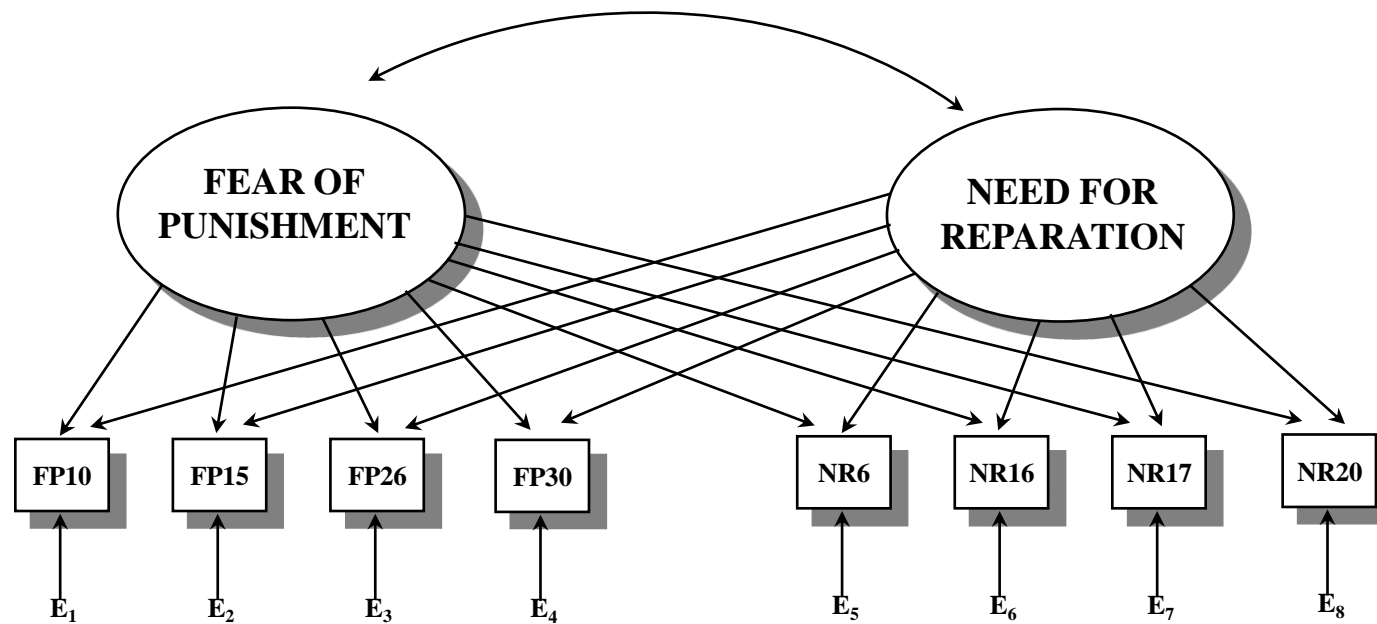
Come ottenere buone soluzioni

- 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

Calcola variabile

Variabile di destinazione:

Espressione numerica:

↳

Gruppo di funzioni:

- Tutto
- Aritmetico
- CDF e CDF noncentrale
- Conversione
- Data/ora corrente
- Aritmetica data

Funzioni e variabili speciali:

Se... (condizione di selezione dei casi facoltativa)

OK Incolla Reimposta Annulla Guida

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

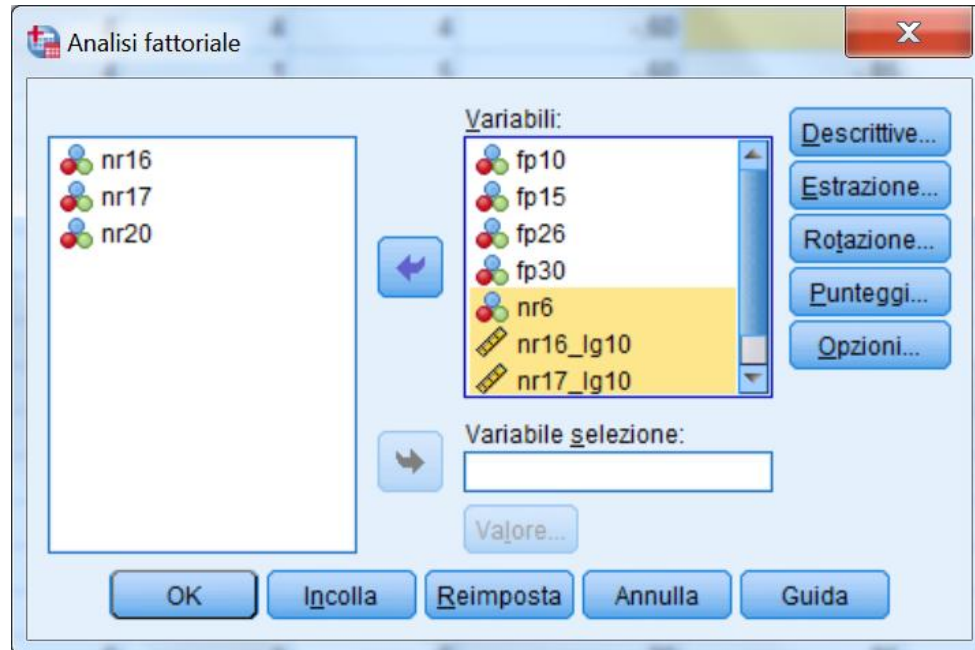
Report
 Statistiche descrittive
 Tabelle personalizzate
 Confronta medie
 Modello lineare generale
 Modelli lineari generalizzati
 Modelli misti
 Correlazione
 Regressione
 Loglineare
 Reti neurali
 Classifica
Riduzione delle dimensioni...
 Scala
 Test non parametrici
 Previsioni
 Sopravvivenza
 Risposta multipla
 Analisi valori mancanti...
 Assegnazione multipla

nr17 nr20 var va

	nr17	nr20	var	va
1	4	4		
2	1	5		
3	6	5		
4	4	6		
5	4	5		
6	6	6		
7	2	6		
8	6	5		
9	4	4		
10	5	5		
11	1	6		
12	5	5		
13	1	6		
14	5	5		
15	6	6		
16	1	6		
17	5	5		
	6	6		

Fattore...
 Analisi delle corrispondenze...
 Scaling ottimale...

EFA IN SPSS



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

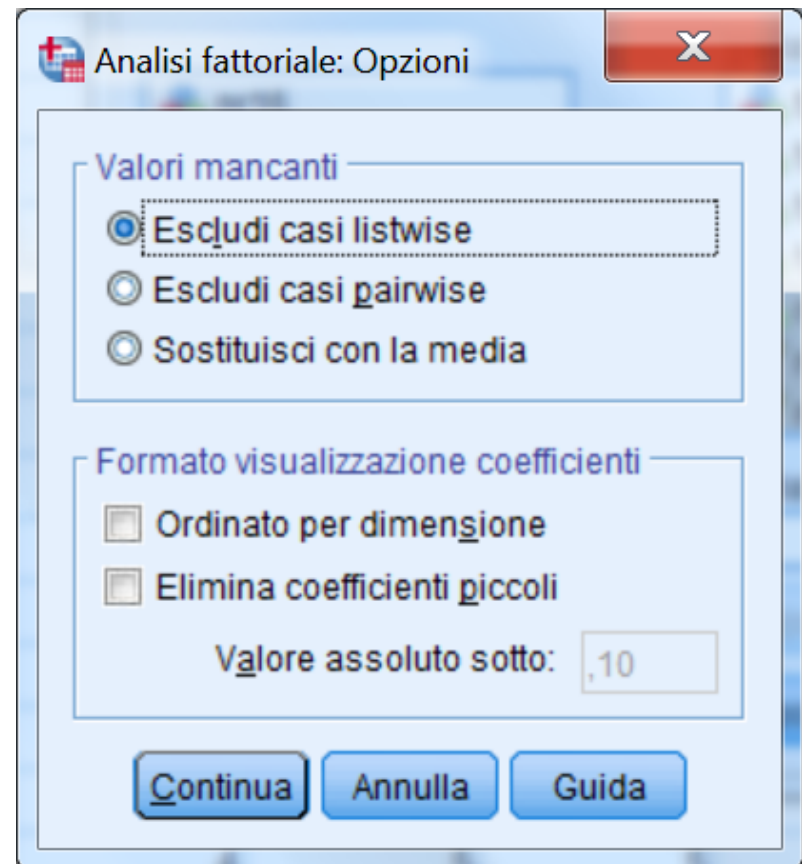
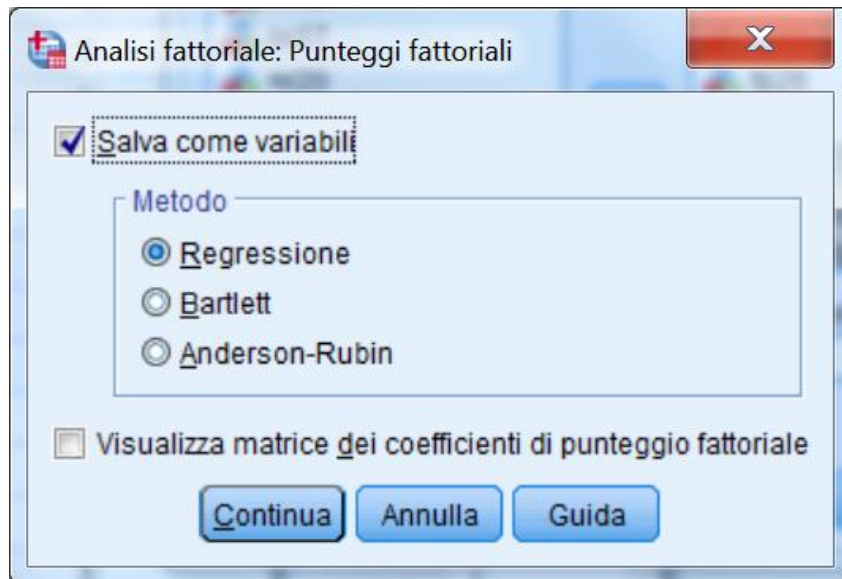
Visualizza

- Soluzione ruotata
- Grafici di caricamento

Numero massimo di iterazioni per la convergenza: 25

Continua Annulla Guida

EFA IN SPSS





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

Fattore	Autovalori iniziali			Caricamenti somme dei quadrati di estrazione			Caricamenti somme dei quadrati di rotazione ^a
	Totale	% di varianza	% cumulativa	Totale	% di varianza	% cumulativa	
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				

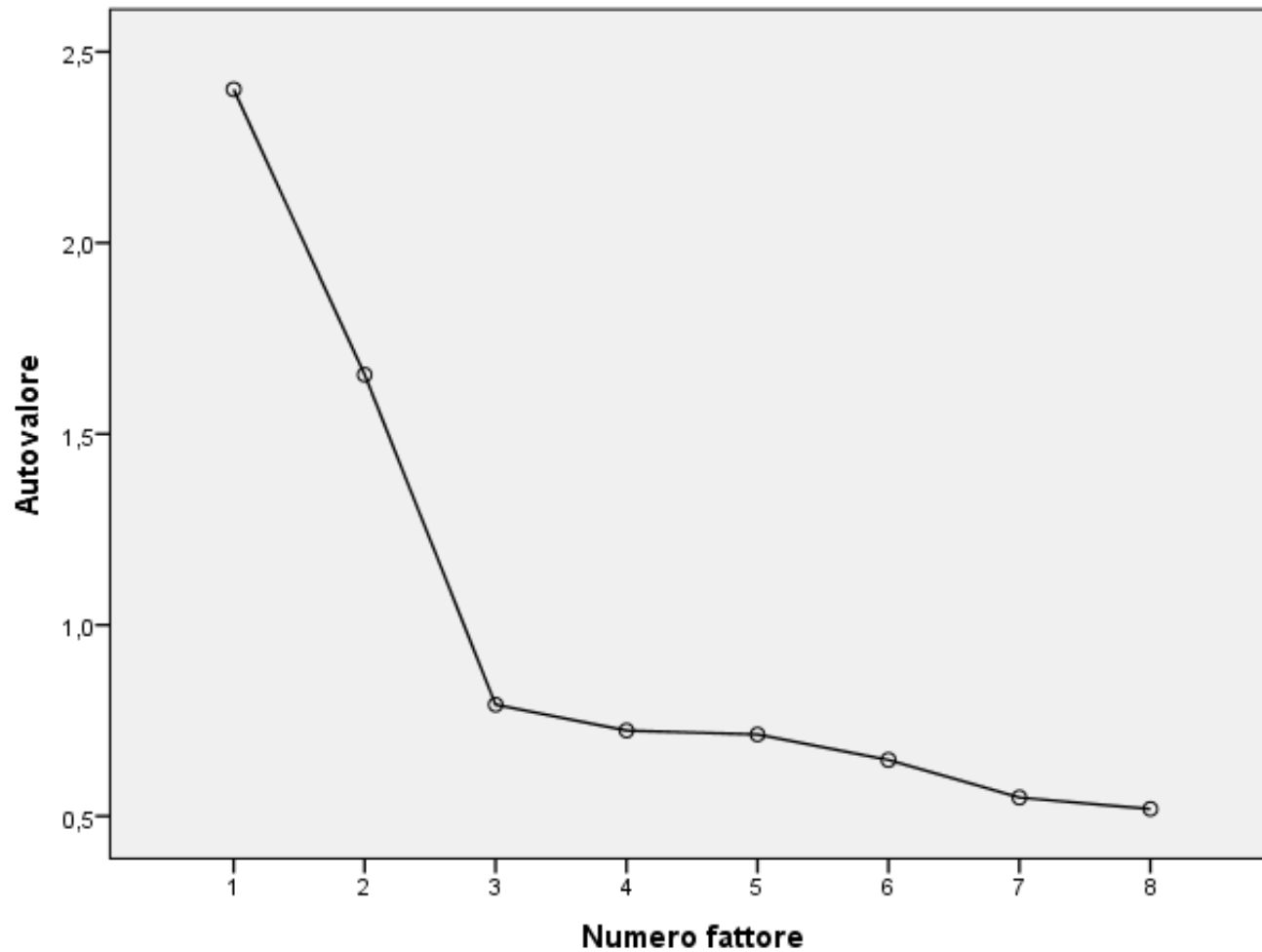
Metodo di estrazione: Fattorizzazione dell'asse principale.

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



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