La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

Elementi di Analisi Multivariata

V SCUOLA ESTIVA AISV
5 -- 9 ottobre 2009 - Soriano nel Cimino (VT)

Sabrina Giordano
Dipartimento di Economia e Statistica
Università della Calabria
sabrina.giordano@unical.it

Quale metodo?

	Confronto tra 2 gruppi	Confronto tra più di 2 gruppi	Associazione tra variabili
Dati quantitativi	t-test ▪ per campioni indipendenti ▪ per dati appaiati ▪ Test di Mann-Whitney ▪ Test di Wilcoxon	ANOVA ▪ per campioni indipendenti ▪ per misure ripetute	Regressione lineare
Dati qualitativi	▪ z-test ▪ Test chi-quadro ▪ McNemar	▪ Test chi-quadro	Regressione logistica

Confronto tra due gruppi

- Dati
  - quantitativi: confronto tra due medie
  - qualitativi: confronto tra due proporzioni

- Campioni
  - indipendenti
  - accoppiati

Confronto tra due medie

(campioni indipendenti)

- Esempio motivante

gli psicologi hanno dimostrato che associare delle immagini ad alcune parole favorisce la memorizzazione di queste. Un esperimento è condotto per verificare tale fenomeno. A 40 partecipanti, divisi in due gruppi da 20, è richiesto di ricordare il massimo di parole possibili riportate su una lista leggendole in 5 minuti. Solo ai partecipanti del gruppo 1 viene esplicitamente richiesto di creare delle immagini per legare le parole mentalmente. Alla fine dell'esperimento, i singoli partecipanti elencano le parole ricordate e sulla base di questi dati gli psicologi concludono che l'ausilio delle immagini può influire sulla memoria.

memory.sav [InsiemeDati1] - SPSS Data Editor

File Modifica Visualizza Dati Trasforma Analizza Grafici S

31 :

	parole	gruppo	var	var
1	19	1		
2	20	1		
3	24	1		
4	30	1		
5	31	1		
6	32	1		
7	30	1		
8	27	1		
9	22	1		
10	25	1		
11	30	1		
12	19	1		
13	18	1		
14	20	1		
15	20	1		
16	22	1		
17	24	1		
18	26	1		
19	25	1		
20	25	1		
21	23	2		
22	22	2		
23	15	2		
24	16	2		
25	18	2		
26	12	2		
27	16	2		
28	19	2		

dati

t-test per il confronto tra due medie
(campioni indipendenti)

Problema: mediamente il valore assunto dalla variabile numero di parole è diverso nei due gruppi di lettori? Se si, la differenza è dovuta al caso? Oppure è da attribuire all'utilità delle immagini?

TEST t

Cosa occorre per costruire il test t

- $n_i, \overline{X}_i, S_i^2$ l'ampiezza, la media e la varianza campionaria, i=1,2
- m_1 e m_2 sono le medie delle due popolazioni (Normali)
- Oggetto del test: $m_1 - m_2$

In teoria

- Assunzioni del t-test:

1. La variabile deve essere quantitativa
2. Distribuzione Normale
3. Campioni con numerosità >= 20
4. Omogeneità delle varianze (se c'è eterogeneità si riducono i gdl)

t-test per il confronto tra due medie
(campioni indipendenti)

Ipotesi: H$_0$: $m_1 - m_2 = 0$

t-statistic:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

con stimatore *pooled della varianza*

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 - (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Regione di rifiuto:

$$\begin{cases} t : |t| > t_{\alpha/2} & H_1 : m_1 - m_2 \neq 0 \\ t : t > t_{\alpha} & H_1 : m_1 - m_2 > 0 \\ t : t < -t_{\alpha} & H_1 : m_1 - m_2 < 0 \end{cases}$$

IC:

$$\overline{X}_1 - \overline{X}_2 \mp t_{\alpha/2} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

3 modi per una decisione

Dopo aver calcolato il valore della statistica test con i dati campionari si procede in 3 strade equivalenti, in particolare, si riterrà che i dati non siano coerenti con l'ipotesi nulla se:

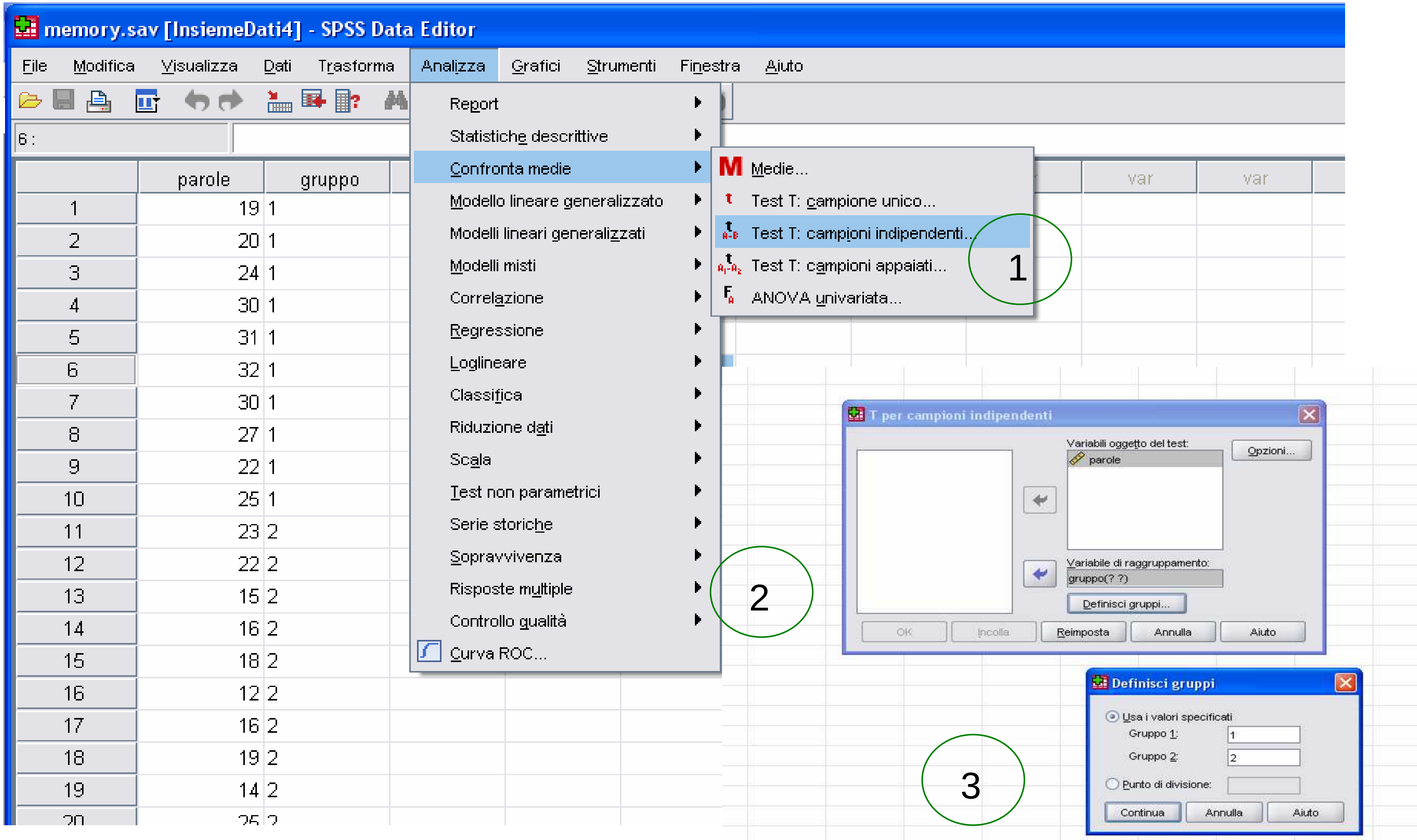
1. se il t calcolato ricade nella regione di rifiuto
2. se il p-value associato al t calcolato è inferiore ad un livello di significatività (solitamente 0.05)
3. l'intervallo di confidenza per la differenza tra le medie non contiene lo zero

Nell'esempio: la differenza è significativa:

$t=5.342$, $gdl=38$, $p<0.05$ si rif. H_0

In particolare, osservando l'intervallo di confidenza si desume che il gruppo che usa la tecnica di associare le immagini alle parole mediamente ne ricorda di più.

In spss


memory.sav [InsiemeDati4] - SPSS Data Editor
File Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
6 :
parole gruppo
1 19 1
2 20 1
3 24 1
4 30 1
5 31 1
6 32 1
7 30 1
8 27 1
9 22 1
10 25 1
11 23 2
12 22 2
13 15 2
14 16 2
15 18 2
16 12 2
17 16 2
18 19 2
19 14 2
var var
Report
Statistiche descrittive
Confronta medie
Modello lineare generalizzato
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Classifica
Riduzione dati
Scala
Test non parametrici
Serie storiche
Sopravvivenza
Risposte multiple
Controllo qualità
Curva ROC...
Medie...
Test T: campione unico...
Test T: campioni indipendenti...
Test T: campioni appaiati...
ANOVA univariata...
1
2
3
T per campioni indipendenti
Variabili oggetto del test:
parole
Opzioni...
Variabile di raggruppamento:
gruppo(? ?)
Definisci gruppi...
OK Incolla Reimposta Annulla Aiuto
Definisci gruppi
Usa i valori specificati
Gruppo 1: 1
Gruppo 2: 2
Punto di divisione:
Continua Annulla Aiuto

Output

p1] - SPSS Viewer

Dati Trasforma Inserisci Formato Analizza Grafici Strumenti Finestra Aiuto

```
T-TEST GROUPS=gruppo(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=parole
  /CRITERIA=CI(.9500).
```

pendenti

Test t

[InsiemeDati3] C:\Documents and Settings\Hp\Desktop\fonetica\scuolafonetica\memory.sav

Statistiche di gruppo

	gruppo	N	Media	Deviazione std.	Errore std. Media
parole	1	20	24,45	4,442	,993
	2	20	17,55	3,692	,825

Test per campioni indipendenti

		Test di Levene di uguaglianza delle varianze		Test t di uguaglianza delle medie						
									Intervallo di confidenza per la differenza al 95%	
		F	Sig.	t	df	Sig. (2-code)	Differenza fra medie	Differenza errore standard	Inferiore	Superiore
parole	Assumi varianze uguali	,846	,363	5,342	38	,000	6,900	1,292	4,285	9,515
	Non assumere varianze uguali			5,342	36,769	,000	6,900	1,292	4,282	9,518

Vale l'omogeneità

La differenza è significativa

t-test per il confronto tra due medie

(dati appaiati)

Esempio: Si comparano le medie della frequenza fondamentale F0 estrapolata dalle registrazioni effettuate su 2 diversi stili: voce alta e normale sullo stesso soggetto LR

dati aisv F0lezione.sav [InsiemeDati1] - SPSS Data Editor

File Modifica Visualizza Dati Trasforma Analizza Grafici

1 : LRvoceA 119,29691

	LRvoceA	LRvoceN	var
1	119,30	137,97	
2	119,43	139,21	
3	119,40	142,27	
4	120,12	143,34	
5	121,06	144,64	
6	122,86	145,20	
7	126,64	146,05	
8	127,75	147,68	
9	128,95	149,54	
10	129,63	151,49	
11	128,01	153,55	
12	126,73	155,10	
13	124,02	157,80	
14	125,14	159,84	
15	125,97	161,94	
16	125,96	163,40	
17	126,47	166,23	
18	126,75	166,89	
19	136,81	167,40	
20	138,79	167,67	
21	148,20	164,47	
22	142,01	164,47	
23	142,90	168,08	
24	143,09	170,70	
25	143,90	172,08	
26	144,39	174,35	
27	144,40	174,94	
28	145,04	178,13	
29	145,24	180,08	
30	146,05	181,30	
31	147,29	182,70	

dati

t-test per il confronto tra due medie

(dati appaiati)

- Si usa per il confronto di medie misurate sul medesimo campione in due diversi istanti di tempo o situazioni
- NB. non va utilizzato quando si voglion comparare medie riferite a variabili diverse sebbene rilevate sullo stesso campione!!
- d è la variabile differenza
- È come un t-test per un campione

Ipotesi: H_0: $m_d = 0$

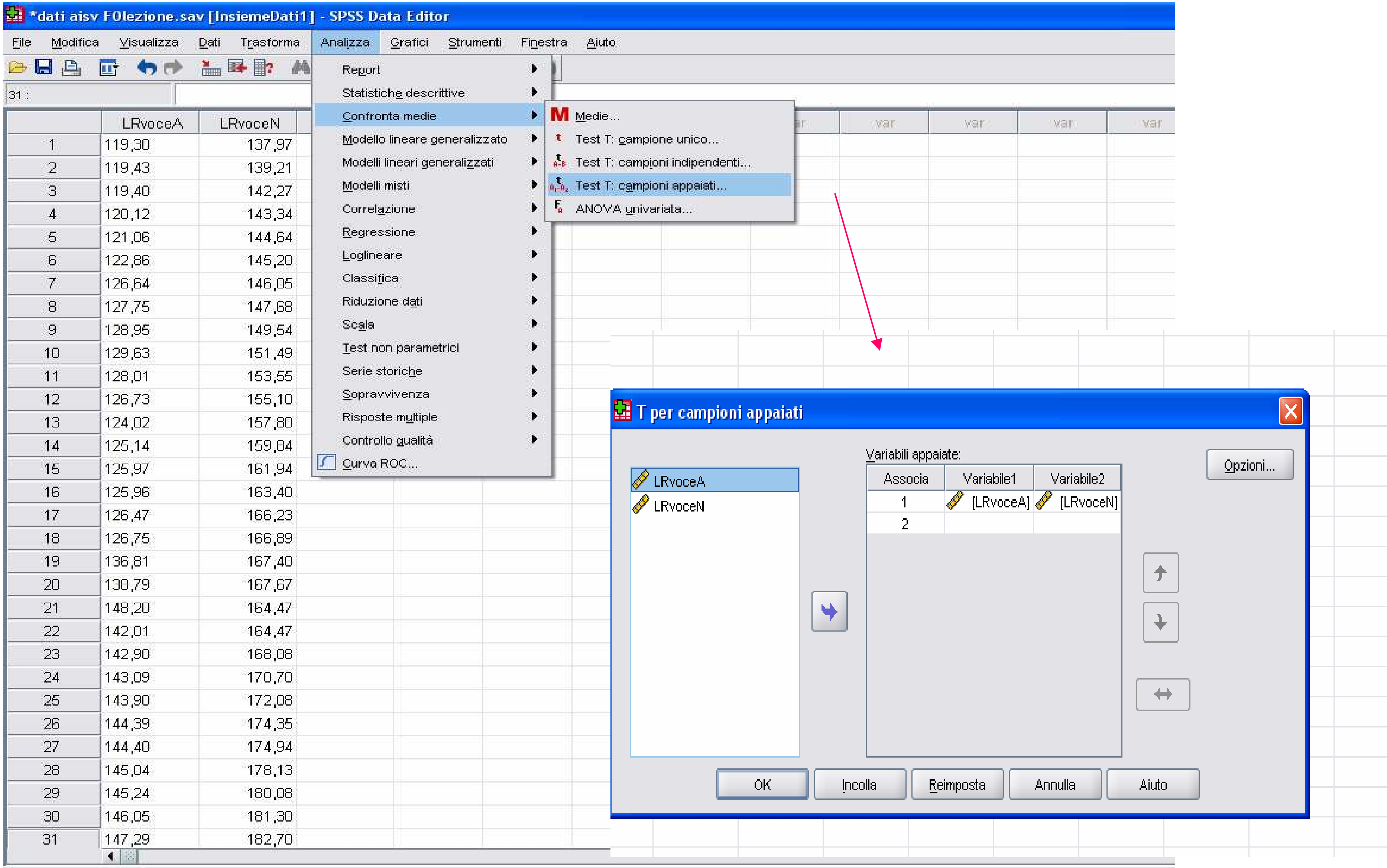
Statistica:

$$t = \frac{\overline{X}_d}{\sqrt{\frac{S_d^2}{n}}} \sim t(n-1)$$

Regione di rifiuto:

$$\begin{cases} t : |t| > t_{\alpha/2} & H_1 : m_d \neq 0 \\ t : t > t_\alpha & H_1 : m_d > 0 \\ t : t < -t_\alpha & H_1 : m_d < 0 \end{cases}$$

Su spss



*dati aisv F0lezione.sav [InsiemeDati1] - SPSS Data Editor
File
Modifica
Visualizza
Dati
Trasforma
Analizza
Grafici
Strumenti
Finestra
Aiuto
31 :
LRvoceA
LRvoceN
1 119,30 137,97
2 119,43 139,21
3 119,40 142,27
4 120,12 143,34
5 121,06 144,64
6 122,86 145,20
7 126,64 146,05
8 127,75 147,68
9 128,95 149,54
10 129,63 151,49
11 128,01 153,55
12 126,73 155,10
13 124,02 157,80
14 125,14 159,84
15 125,97 161,94
16 125,96 163,40
17 126,47 166,23
18 126,75 166,89
19 136,81 167,40
20 138,79 167,67
21 148,20 164,47
22 142,01 164,47
23 142,90 168,08
24 143,09 170,70
25 143,90 172,08
26 144,39 174,35
27 144,40 174,94
28 145,04 178,13
29 145,24 180,08
30 146,05 181,30
31 147,29 182,70
Report
Statistiche descrittive
Confronta medie
Modello lineare generalizzato
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Classifica
Riduzione dati
Scala
Test non parametrici
Serie storiche
Sopravvivenza
Risposte multiple
Controllo qualità
Curva ROC...
Medie...
Test T: campione unico...
Test T: campioni indipendenti...
Test T: campioni appaiati...
ANOVA univariata...
var
var
var
var
T per campioni appaiati
Variabili appaiate:
LRvoceA
LRvoceN
Associa
Variabile1
Variabile2
1 [LRvoceA] [LRvoceN]
2
Opzioni...
OK
Incolla
Reimposta
Annulla
Aiuto

*Output1 [Documento1] - SPSS Viewer

File Modifica Visualizza Dati Trasforma Inserisci Formato Analizza Grafici Strumenti Finestra Aiuto

Output
Registro
Test t
Titolo
Nota
File di dati attivo
Statistiche per can
Correlazioni per ca
Test per campioni

```
SAVE OUTFILE='C:\Documents and Settings\Hp\Desktop\fonetica\dati aisv F0lezione.sav' /COMPRESSED.
T-TEST PAIRS=LRvoceA WITH LRvoceN (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.
```

Test t

[InsiemeDati1] C:\Documents and Settings\Hp\Desktop\fonetica\dati aisv F0lezione.sav

Statistiche per campioni appaiati

		Media	N	Deviazione std.	Errore std. Media
Coppia 1	LRvoceA	155,6433	98	21,72253	2,19431
	LRvoceN	188,7603	98	21,63093	2,18505

Correlazioni per campioni appaiati

		N	Correlazione	Sig.
Coppia 1	LRvoceA e LRvoceN	98	,840	,000

Test per campioni appaiati

		Differenze a coppie					t	df	Sig. (2-code)
					Intervallo di confidenza per la differenza al 95%				
		Media	Deviazione std.	Errore std. Media	Inferiore	Superiore			
Coppia 1	LRvoceA - LRvoceN	-33,11694	12,24976	1,23741	-35,57286	-30,66101	-26,763	97	,000

Fo è mediamente diversa!

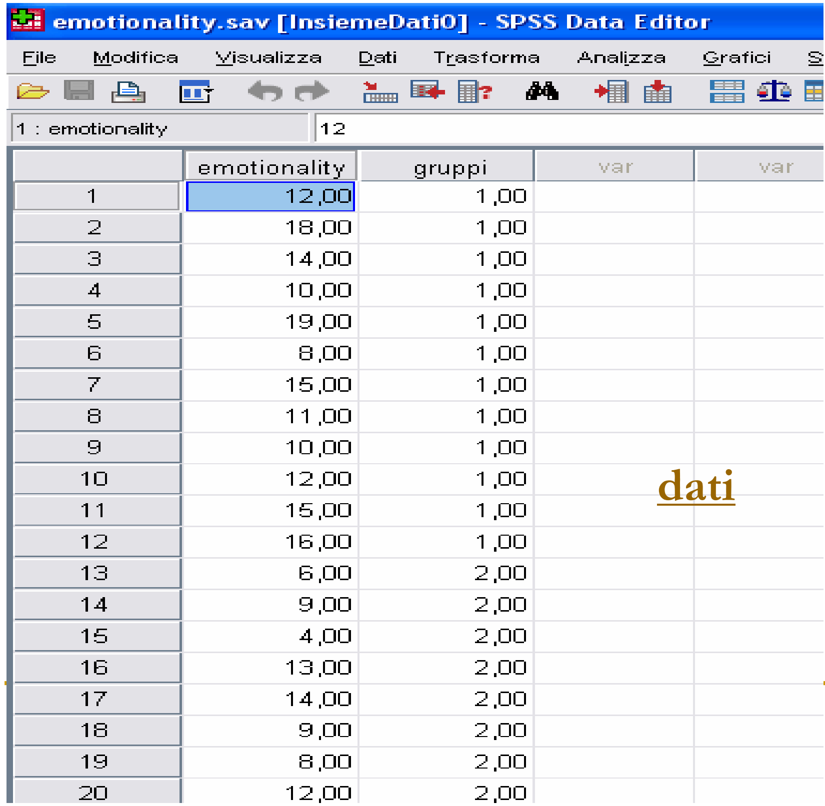
Test non parametrici

- Se le ipotesi sottostanti l'impiego del test t sono violate si può ricorrere ad approcci non parametrici
- Se, ad esempio, risulta poco realistica l'ipotesi di normalità oppure il campione è molto piccolo (n<25) si può ricorrere ai test di
  - **Mann-Withney** (per campioni indipendenti)
  - **Wilcoxon** (per dati appaiati)
- L'ipotesi nulla è che i due campioni provengano dalla stessa popolazione senza specificare la distribuzione
- Questi test sono utili anche con variabili ordinali
- Nel test di Mann vengono ordinate tutte le osservazioni insieme ed a ognuna si assegna un rango (da 1 a $n_1 + n_2$), poi si calcolano le medie dei ranghi nei due gruppi e si comparano. In Wilcoxon test si assegnano dei ranghi alle differenze tra le coppie di valori riferite alle stesse unità, si sommano i ranghi + e -, e si comparano

Test non parametrici (campioni indipendenti)

- Mann-Whitney U-test

Esempio: vengono rilevati dei punteggi che riguardano reazioni di emotività in bambini che hanno entrambi i genitori ed in altri che, invece, vivono con uno solo di loro. Si vuol valutare se mediamente i punteggi variano nelle due diverse situazioni familiari.



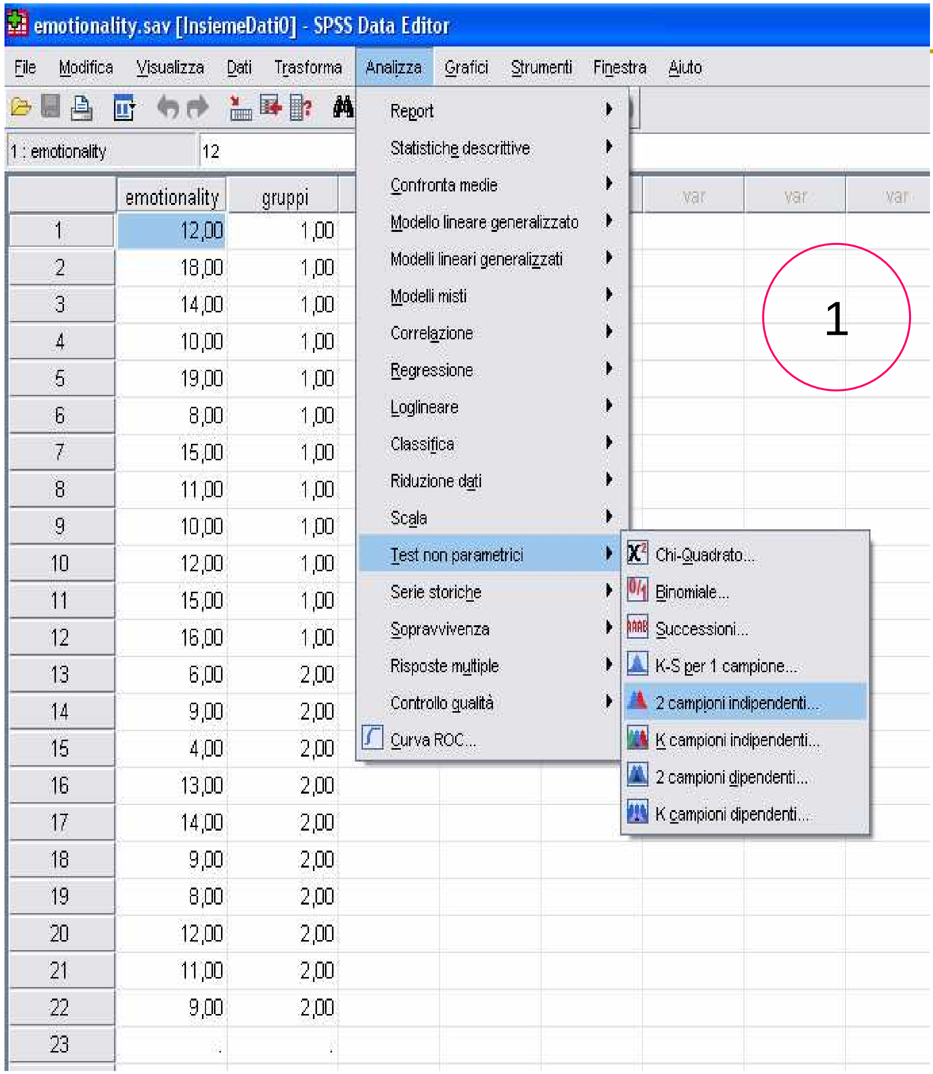
	emotionality	gruppi	var	var
1	12,00	1,00		
2	18,00	1,00		
3	14,00	1,00		
4	10,00	1,00		
5	19,00	1,00		
6	8,00	1,00		
7	15,00	1,00		
8	11,00	1,00		
9	10,00	1,00		
10	12,00	1,00		
11	15,00	1,00		
12	16,00	1,00		
13	6,00	2,00		
14	9,00	2,00		
15	4,00	2,00		
16	13,00	2,00		
17	14,00	2,00		
18	9,00	2,00		
19	8,00	2,00		
20	12,00	2,00		

dati

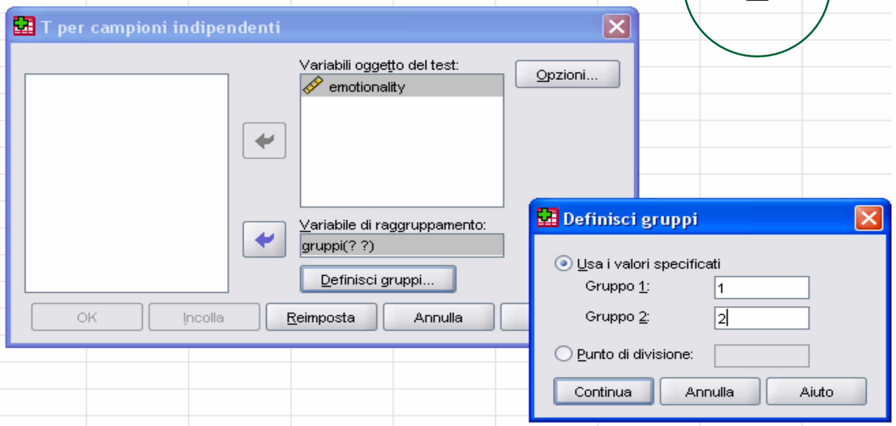
$$U = n_1 * n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

R_1 è la somma dei ranghi per il gruppo 1

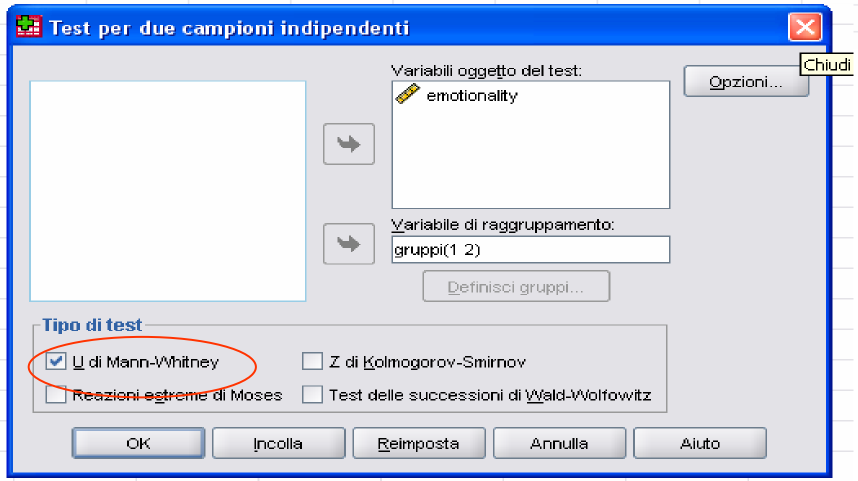
1


emotionality.sav [InsiemeDati0] - SPSS Data Editor
File Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
1 : emotionality 12
emotionality gruppi var var var
1 12,00 1,00
2 18,00 1,00
3 14,00 1,00
4 10,00 1,00
5 19,00 1,00
6 8,00 1,00
7 15,00 1,00
8 11,00 1,00
9 10,00 1,00
10 12,00 1,00
11 15,00 1,00
12 16,00 1,00
13 6,00 2,00
14 9,00 2,00
15 4,00 2,00
16 13,00 2,00
17 14,00 2,00
18 9,00 2,00
19 8,00 2,00
20 12,00 2,00
21 11,00 2,00
22 9,00 2,00
23
Report
Statistiche descrittive
Confronta medie
Modello lineare generalizzato
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Classifica
Riduzione dati
Scala
Test non parametrici
Serie storiche
Sopravvivenza
Risposte multiple
Controllo qualità
Curva ROC...
Chi-Quadrato...
Binomiale...
Successioni...
K-S per 1 campione...
2 campioni indipendenti...
K campioni indipendenti...
2 campioni dipendenti...
K campioni dipendenti...

2


T per campioni indipendenti
Variabili oggetto del test:
emotionality
Opzioni...
Variabile di raggruppamento:
gruppi(? ?)
Definisci gruppi...
OK Incolla Reimposta Annulla
Definisci gruppi
Usa i valori specificati
Gruppo 1: 1
Gruppo 2: 2
Punto di divisione:
Continua Annulla Aiuto

3


Test per due campioni indipendenti
Chiudi
Variabili oggetto del test:
emotionality
Opzioni...
Variabile di raggruppamento:
gruppi(1 2)
Definisci gruppi...
Tipo di test
U di Mann-Whitney
Z di Kolmogorov-Smirnov
Reazioni estreme di Moses
Test delle successioni di Wald-Wolfowitz
OK Incolla Reimposta Annulla Aiuto

*Output1 [Documento1] - SPSS Viewer

File Modifica Visualizza Dati Trasforma Inserisci Formato Analizza Grafici Strumenti Finestra Aiuto

- Output
  - Test non parametrici
    - Titolo
    - Nota
    - File di dati attivo
    - Test di Mann-Whit
      - Titolo
      - Ranghi
      - Test

Test non parametrici

[InsiemeDati0]

Test di Mann-Whitney

Ranghi del gruppo 1 >del gruppo 2

Ranghi

	gruppi	Numerosità	Rango medio	Somma dei ranghi
emotionality	1,00	12	14,46	173,50
	2,00	10	7,95	79,50
	Totale	22		

Test[b]

	emotionality
U di Mann-Whitney	24,500
W di Wilcoxon	79,500
Z	-2,349
Sig. Asint. a 2 code	,019
Significatività esatta [2*(Significatività a 1 coda)]	,017[a]

a. Non corretto per valori pari merito.

b. Variabile di raggruppamento: gruppi

ciò implica che risultano mediamente più emotivi i bambini in famiglie con un solo genitore

Si ragiona come nel test z

Test non parametrici (campioni dipendenti)

- Test di Wilcoxon

Sui dati F_0 →

dati

Test non parametrici

[InsiemeDati1] C:\Documents and Settings\Hp\Desktop\fonetica\dati aisv F0lezione.sav

Test di Wilcoxon

Ranghi

		Numerosità	Rango medio	Somma dei ranghi
LRvoceN - LRvoceA	Ranghi negativi	0[a]	,00	,00
	Ranghi positivi	98[b]	49,50	4851,00
	Valori pari merito	0[c]		
	Totale	98		

a. LRvoceN < LRvoceA

b. LRvoceN > LRvoceA

c. LRvoceN = LRvoceA

Test[b]

	LRvoceN - LRvoceA
Z	-8,595[a]
Sig. Asint. a 2 code	,000

a. Basato su ranghi negativi.

b. Test di Wilcoxon

Si rifiuta l'ipotesi che non ci sia diversità sui valori medi di F_0 a voce alta e normale

Dati categoriali: confronto tra proporzioni
(campioni indipendenti)

Esempio:

Il compito di ripetizione di non-parole è tradizionalmente usato come misura della memoria fonologica a breve termine dei bambini. Ad alcuni bambini della stessa età viene richiesto di ripetere 10 non-parole e si registra il punteggio ottenuto contando gli errori dei bambini, suddivisi tra maschi e femmine, nel pronunciare le non-parole. In proporzione sbagliano meno i maschi o le femmine?

Numero di errori

	<5	>=5	
femmine	**37**	**10**	**47**
maschi	**12**	**32**	**44**

<5 è "successo"
>=5 è "insuccesso"

Test per il confronto tra due proporzioni

(campioni indipendenti)

Ipotesi: $H_0 : p_1 - p_2 = 0$

Cosa occorre: $\hat{p}_1, \hat{p}_2$ e n_1, n_2 proporzioni e numerosità campionarie

Z-statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \overset{as}{\sim} N(0,1)$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{as}{\sim} N(0,1)$$

con stimatore *pooled* $\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$

Regione di rifiuto:

$$\begin{cases} z : |z| > z_{\alpha/2} & H_1 : p_1 - p_2 \neq 0 \\ z : z > z_{\alpha} & H_1 : p_1 - p_2 > 0 \\ z : z < -z_{\alpha} & H_1 : p_1 - p_2 < 0 \end{cases}$$

NB. a volte al numeratore si utilizza la correzione di continuità di Yates

$$\hat{p}_1 - \hat{p}_2 - 0.5\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Intervallo di confidenza per la differenza tra proporzioni

- L'IC al 100(1-α)% per la differenza p_1-p_2 è:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} * \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- Quando l'intervallo di confidenza per p_1-p_2 contiene lo 0 è plausibile ritenere che p_1-p_2=0; se l'intervallo ha solo valori positivi (negativi) allora si può dedurre che la p_1 è maggiore di p_2 (p_1 è minore di p_2)

Un po' di calcoli

- Per il gruppo femmine: $n_1=47$ e $\hat{p}_1 = {}^{37}\!/_{47} = 0.79$
- Per il gruppo maschi: $n_2=44$ e $\hat{p}_2 = {}^{12}\!/_{44} = 0.27$

- Valore calcolato della statistica z:

$$z = \frac{0.79 - 0.27}{\sqrt{\frac{0.27(1-0.27)}{44} + \frac{0.79(1-0.79)}{47}}} = 5.81$$

- NB $z = 5.81 > z_{\alpha/2} = 1.96$ quindi si rifiuta l'ipotesi che le due proporzioni siano uguali, (il valore di z può variare se si usa la correzione, ma l'esito è uguale)
- IC al 100(1-α)% è (0.34; 0.69), la proporzione riscontrata nelle femmine è più alta di almeno un terzo e al più due terzi rispetto a quella dei maschi

Dati categoriali: confronto tra proporzioni
(campioni dipendenti)

Esempio motivante: riconoscimento vocale automatico

Prima di parlare con un operatore in un call-center spesso bisogna pronunciare delle parole per indirizzare la procedura. Un ricercatore è interessato a confrontare due diversi sistemi di riconoscimento e valuta quanto ciascun sistema sbagli nel riconoscere la parola. Bisogna testare se le proporzioni di corretti (errati) riconoscimenti siano significativamente differenti (2000 *parole*)

		CDHMM *corretto*	CDHMM *sbagliato*	
GMDS	*corretto*	1921	58	1979
GMDS	*sbagliato*	16	5	21
		1937	63	2000

Test di McNemar (non parametrico)

(campioni dipendenti) tabelle 2x2

Ipotesi: $H_0 : p_{riga} - p_{col} = 0$

Statistica:

$$z = \frac{b-c}{\sqrt{b+c}}$$

	Si	No	
Si	a	b	a+b
No	c	d	c+d
	a+c	b+d	n

- test di omogeneità marginale a+b=a+c; b+d=c+d ovvero b=c
- situazione tipo: *favore/sfavore* e *prima/dopo*
- quando b+ c >20, z ha distribuzione Normale std
- *Nell'esempio:* $z = {(58-16)}\big/{\sqrt{58+16}} = 4.88$ *con p-value=0.000, si rif* H_0
- Per grandi campioni l'IC è: $(\hat{p}_{riga} - \hat{p}_{col}) \pm z_{\alpha/2} * \frac{1}{n}\sqrt{(b+c) - {(b-c)^2}\big/{n}}$
- *Nell'esempio IC al 95% è (0.013,0.029) (significatività statistica ma poco pratica)*

Confronto tra proporzioni con il chi-quadro (tabelle 2 x 2)

- Risposta binaria (in colonna) ed esplicativa (2 gruppi) in riga
- p_1 è la proporzione di successi nella popolazione 1, $1 - p_1$ è la proporzione di insuccessi
- Ipotesi: H_0: $p_1 - p_2 = 0$ è di omogeneità
- Statistica chi-quadro ($\chi^2 = z^2$)

Nell'es: $\chi^2 = 24, p = 0.000$, si rif. H_0

Proporzioni di risposte

	successi	insuccessi
gruppo 1	p_1	$1-p_1$
gruppo 2	p_2	$1- p_2$

Numero di errori

	<5	>=5	
femmine	**37**	**10**	**47**
maschi	**12**	**32**	**44**

Test chi-quadro di indipendenza

- Per tabelle r x s
- Ipotesi: H_0: le variabili sono indipendenti
  H_1: le variabili sono dipendenti

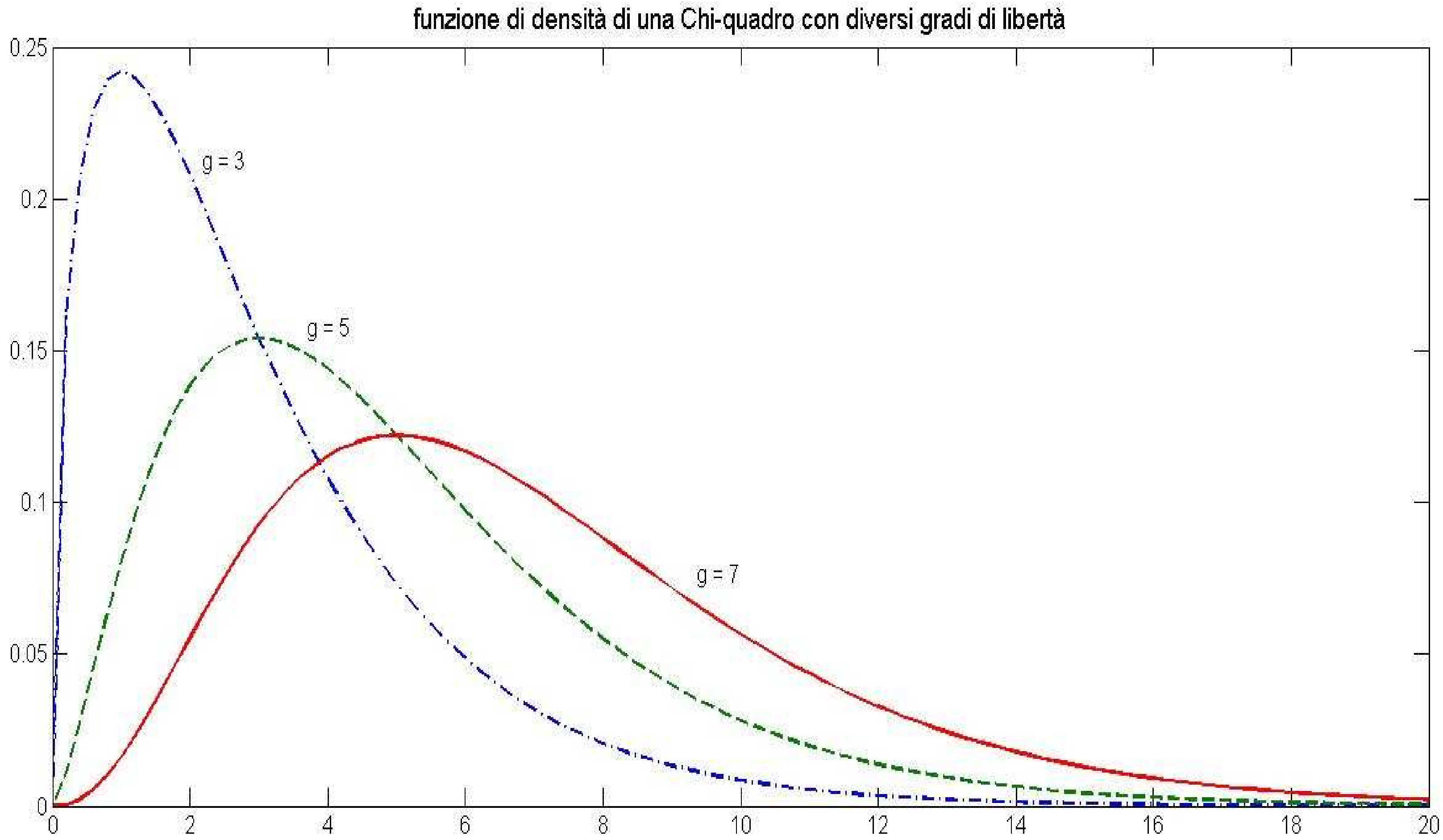
Esempio:

		Comportamento nel gioco		
		collaborativo	**competitivo**	
Stile educativo	**permissivo**	9	15	24
	equilibrato	24	9	33
	autoritario	8	19	27
		41	43	84

Test chi-quadro di indipendenza

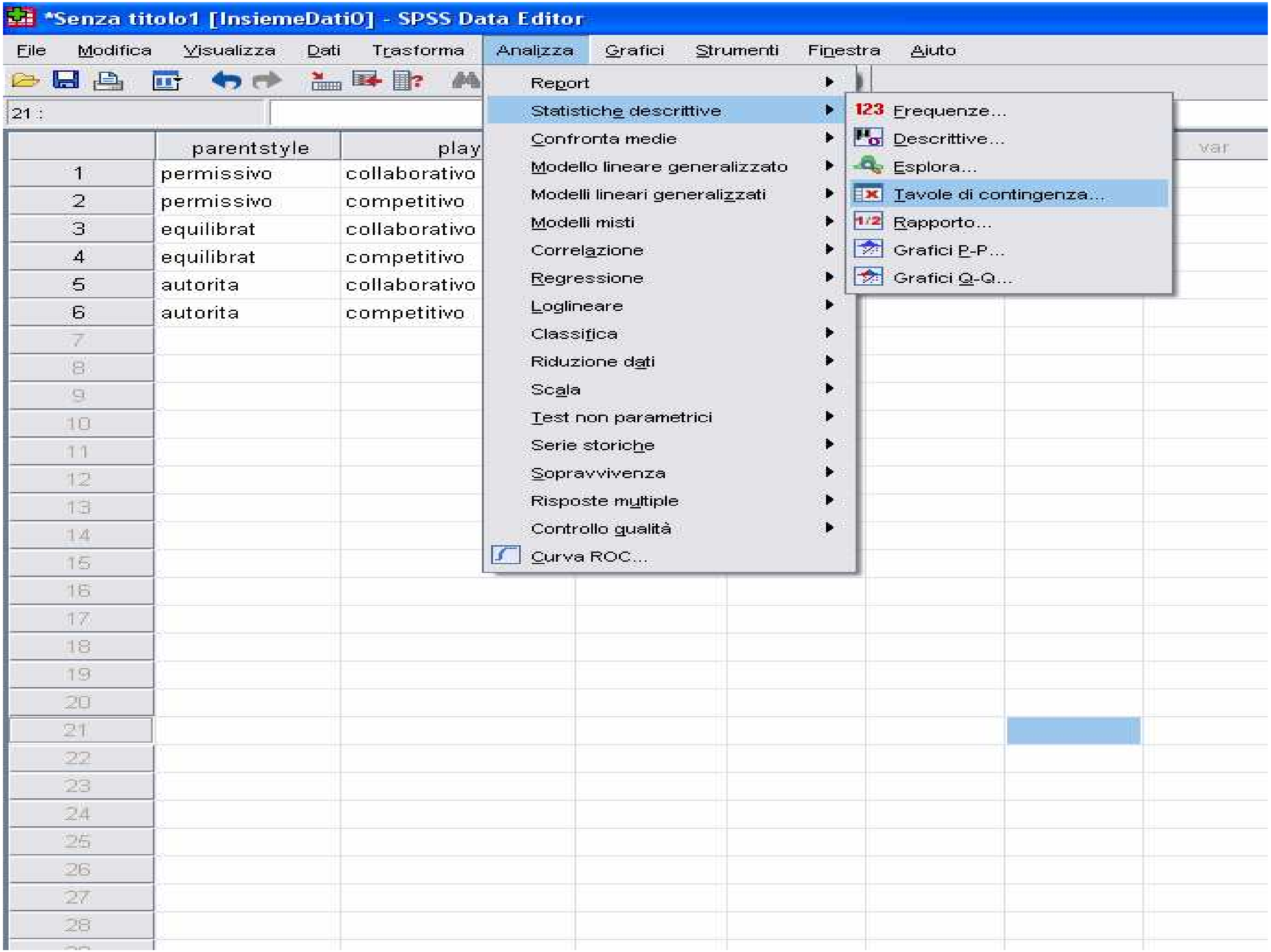
- Statistica $$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$
- f_0 frequenze osservate, f_e frequenze attese nel caso di indipendenza (quindi sotto H_0)
- le f_e devono essere almeno pari a 5 (non oltre il 20% delle frequenze attese deve essere <5), altrimenti si ricorre al test esatto di Fisher
- La statistica χ^2 ha distribuzione chi-quadro con (r-1)(s-1) gdl
- Si rifiuta H_0 di indipendenza quando il chi-quadro calcolato supera il valore critico

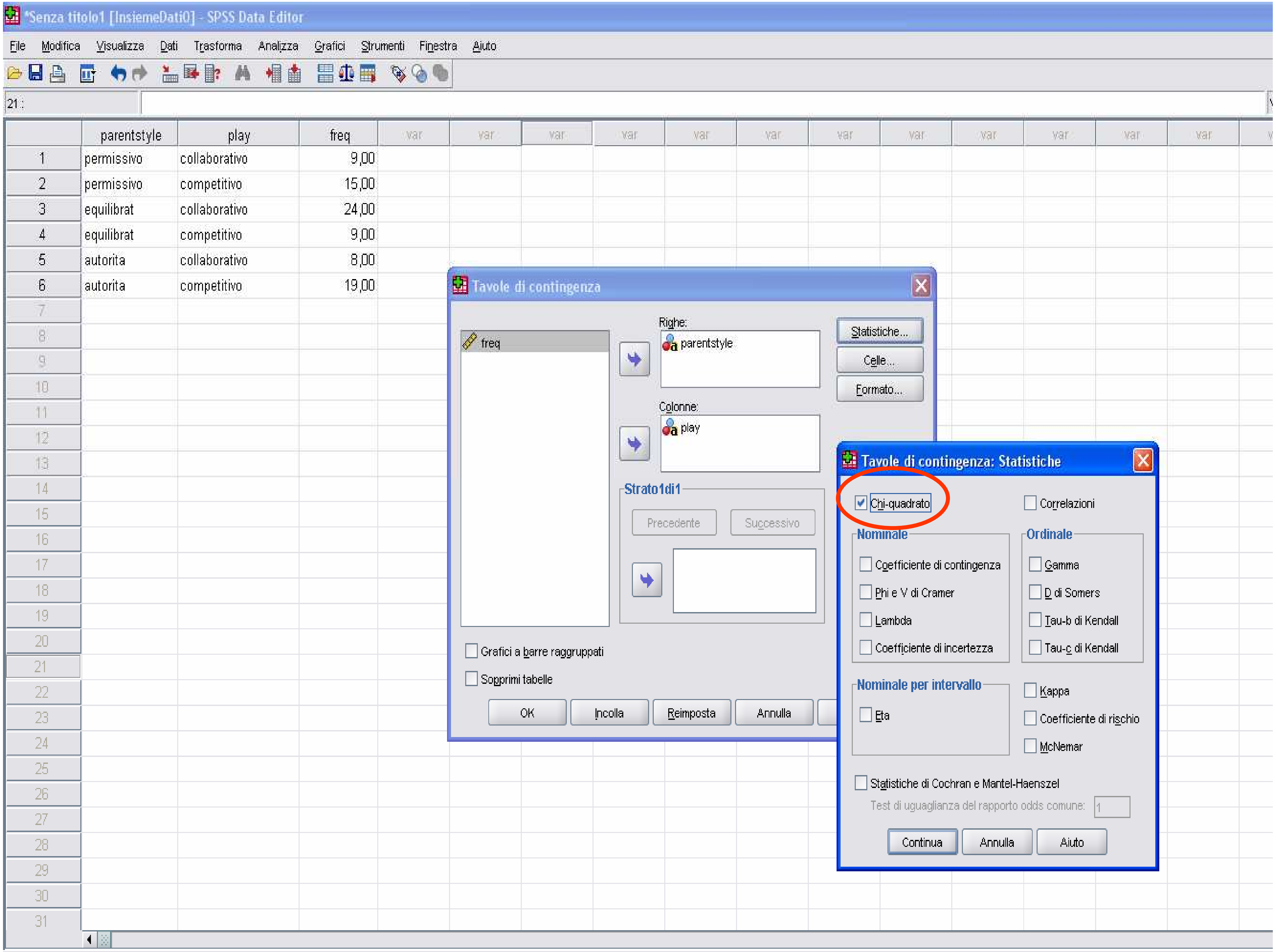
Chi-quadro


funzione di densità di una Chi-quadro con diversi gradi di libertà
0.25
0.2
0.15
0.1
0.05
0
0
2
4
6
8
10
12
14
16
18
20
g = 3
g = 5
g = 7

Test chi-quadro di indipendenza

- Regola: *f_e è dato dal prodotto dei totali di riga e di colonna corrispondenti alla cella diviso per il numero totale di osservazioni.* Ad esempio la freq attesa della prima cella è $f_e = \frac{41}{84} * 24 = 0.488 * 24 = 11.712$
- Significato in termini di indipendenza: *Nel campione, 41 bambini su 84 hanno un comportamento collaborativo nel gioco (48.8%). Se non ci fosse alcuna incidenza della scelta educativa dei genitori sul comportamento nel gioco dei figli ci si attenderebbe che il 48.8% di quelli che ricevono un'educazione permissiva, il 48,8% di quelli con educazione equilibrata e il 48.8% che ricevono educazione autoritaria siano collaborativi nel gioco.*
- La statistica chi-quadro sintetizza quanto siano vicine le frequenze osservate a quelle attese. Valori elevati della statistica (pvalue piccoli) indicano un allontanamento dall'ipotesi di indipendenza


*Senza titolo1 [InsiemeDati0] - SPSS Data Editor
File
Modifica
Visualizza
Dati
Trasforma
Analizza
Grafici
Strumenti
Finestra
Aiuto
21 :
parentstyle
play
var
permissivo
collaborativo
permissivo
competitivo
equilibrat
collaborativo
equilibrat
competitivo
autorita
collaborativo
autorita
competitivo
Report
Statistiche descrittive
Confronta medie
Modello lineare generalizzato
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Classifica
Riduzione dati
Scala
Test non parametrici
Serie storiche
Sopravvivenza
Risposte multiple
Controllo qualità
Curva ROC...
Frequenze...
Descrittive...
Esplora...
Tavole di contingenza...
Rapporto...
Grafici P-P...
Grafici Q-Q...


*Senza titolo1 [InsiemeDati0] - SPSS Data Editor
File Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
21 :
parentstyle play freq var
1 permissivo collaborativo 9,00
2 permissivo competitivo 15,00
3 equilibrat collaborativo 24,00
4 equilibrat competitivo 9,00
5 autorita collaborativo 8,00
6 autorita competitivo 19,00
Tavole di contingenza
freq
Righe:
parentstyle
Colonne:
play
Strato1di1
Precedente
Successivo
Statistiche...
Celle...
Formato...
Grafici a barre raggruppati
Sopprimi tabelle
OK
Incolla
Reimposta
Annulla
Tavole di contingenza: Statistiche
Chi-quadrato
Correlazioni
Nominale
Coefficiente di contingenza
Phi e V di Cramer
Lambda
Coefficiente di incertezza
Ordinale
Gamma
D di Somers
Tau-b di Kendall
Tau-c di Kendall
Nominale per intervallo
Eta
Kappa
Coefficiente di rischio
McNemar
Statistiche di Cochran e Mantel-Haenszel
Test di uguaglianza del rapporto odds comune:
1
Continua
Annulla
Aiuto

*Output2 [Documento2] - SPSS Viewer

File Modifica Visualizza Dati Trasforma Inserisci Formato Analizza Grafici Strumenti Finestra Aiuto

- Output
  - Registro
  - Tavole di contingenza
    - Titolo
    - Nota
    - File di dati attivo
    - Riepilogo dei casi
    - Tavola di continge
    - Chi-quadrato
  - Registro
  - Tavole di contingenza
    - Titolo
    - Nota
    - File di dati attivo
    - Riepilogo dei casi
    - Tavola di continge
    - Chi-quadrato

```
/CELLS=COUNT EXPECTED SRESID
/COUNT ROUND CELL.
```

Tavole di contingenza

[InsiemeDati0]

Riepilogo dei casi

	Casi					
	Validi		Mancanti		Totale	
	N	Percentuale	N	Percentuale	N	Percentuale
parentstyle * play	84	100,0%	0	,0%	84	100,0%

Tavola di contingenza parentstyle * play

			play		
			collaborativo	competitivo	Totale
parentstyle	autorita	Conteggio	8	19	27
		Conteggio atteso	13,2	13,8	27,0
		Residui stand.	-1,4	1,4	
	equilibrat	Conteggio	24	9	33
		Conteggio atteso	16,1	16,9	33,0
		Residui stand.	2,0	-1,9	
	permissivo	Conteggio	9	15	24
		Conteggio atteso	11,7	12,3	24,0
		Residui stand.	-,8	,8	
	Totale	Conteggio	41	43	84
		Conteggio atteso	41,0	43,0	84,0

Chi-quadrato

	Valore	df	Sig. asint. (2 vie)
Chi-quadrato di Pearson	12,759ª	2	,002
Rapporto di verosimiglianza	13,158	2	,001
N. di casi validi	84		

a. 0 celle (,0%) hanno un conteggio atteso inferiore a 5. Il conteggio atteso minimo è 11,71.

Con il linguaggio R

```
> tab<-matrix(c(9,15,24,9,8,19),nrow=3,byrow=T)
> colnames(tab)<-c("collaborativo","competitivo")
> rownames(tab)<-c("permissivo","equilibrato","autoritario")
> tab
            collaborativo competitivo
permissivo              9          15
equilibrato            24           9
autoritario             8          19
> chisq.test(tab)

        Pearson's Chi-squared test

data:  tab
X-squared = 12.7593, df = 2, p-value = 0.001696

> chisq.test(tab)$expected
            collaborativo competitivo
permissivo       11.71429    12.28571
equilibrato      16.10714    16.89286
autoritario      13.17857    13.82143
> 
```

Altre misure del confronto tra proporzioni

- Si possono utilizzare delle misure di associazione che comparano le proporzioni mediante rapporti
  - Rischio relativo
  - Odds ratio

- Un rapporto tra proporzioni in una tabella 2x2 è il "rischio relativo"

Esempio. *Rapportiamo le proporzioni di coloro che dichiarano di essere felici tra quelli che guadagnano molto rispetto a coloro che hanno un basso reddito*

RR =0.84/0.29 = 2.89

Pensi di essere felice?

		si	no	
Reddito	elevato	272	49	321
	basso	85	208	293

L'odds

- Un'altra misura per confrontare proporzioni è l'*odds ratio: rapporto tra odds*
- Per una variabile binaria con categorie "successo", "insuccesso", l'odds si definisce come

$$Odds = P(successo)/P(insuccesso) = P(successo)/[1 - P(successo)]$$

Per esempio,

- se $P(successo) = 0.80$, $P(insuccesso) = 0.20$, l'odds = 0.80/0.20 = 4.0
- se $P(successo) = 0.20$, $P(insuccesso) = 0.80$, l'odds = 0.20/0.80 = 0.25

La probabilità di successo che si ottiene dall'odds è:

$$Probabilità\ di\ successo = odds/(odds + 1)$$

Per es., all'odds = 4.0 corrisponde la probabilità = 4/(4+1) =4/5=0.80

L'odds ratio

- Per 2 gruppi nelle righe in una tabella 2x2

$$\textit{odds ratio} = (\text{odds nella riga 1})/(\text{odds nella riga 2})$$

Es: Indagine su alcuni studenti di scuola superiore (A. Agresti)

		alcool: Si	alcool: No
fumo	Si	1449	46
	No	500	281

- Per quelli che fumano, l'odds di aver consumato alcool è 1449/46 = 31.5
- Per quelli che non fumano, l'odds di aver consumato alcool è 500/281 = 1.78

L'odds ratio è: OR = 31.5/1.78 = 17.7

La stima dell'odds di aver consumato alcool per gli studenti che fumano è 17.7 volte l'odds che i non fumatori consumino alcool

Limiti del test chi-quadro

- Il test evidenzia se c'è o meno associazione ma non l'intensità di un'eventuale associazione: un elevato valore della statistica chi-quadro ed un basso p-value indicano una forte evidenza che ci sia associazione ma non necessariamente una forte associazione!!!!

Risposta

	1	2	1	2	1	2	1	2
Gruppo 1	15	10	30	20	60	40	600	400
Gruppo 2	10	15	20	30	40	60	400	600

χ^2:	2	4	8	80
P-value:	0.16	0.046	0.005	3.7×10^{-19}

gdl =1, nota che $\hat{p}_1 - \hat{p}_2 = 0.60 - 0.40 = 0.20$ in ogni tabella

Possiamo aver un valore elevato del chi quadro test (e piccoli p-value) per una debole associazione, quando n è grande

Analisi della varianza ad una via

- Studia l'effetto di variabili qualitative su un variabile quantitativa

- È usata per testare la differenza tra più di due medie di una variabile quantitativa (risposta) al variare dei k livelli (o trattamenti) di una variabile qualitativa (fattore)
  - Con campioni differenti per ogni livello (dati indipendenti)
  - Lo stesso campione nei diversi livelli (misure ripetute)

- Assunzioni del t-test:

1. La variabile risposta deve essere quantitativa
2. Distribuzione Normale
3. Omogeneità delle varianze

Analisi della varianza ad una via

Esempio:Alcuni soggetti vengono sottoposti ad un test per verificare se la mancanza di sonno possa alterare la vista. Il campione è diviso in tre gruppi a seconda che si limiti il sonno di 3, 12 e 24 ore. In seguito a tale privazione i soggetti rispondono ad un test visivo e si calcola il numero di errori commesso da ciascuno.

Mediamente il numero di errori cambia al variare delle ore di sonno perse?

Cosa occorre:

- 1 sola variabile risposta per k livelli del fattore (k popolazioni con medie m_1, m_2, …., m_k)
- k campioni indipendenti di numerosità n_1, n_2, …., n_k (tutte uguali se il disegno è bilanciato)
- dati: x_{ji} i-esima unità, j-esimo gruppo $j=1,\ldots,k$ e $i=1,\ldots,n_j$

- Ipotesi: $H_0 : m_1 = m_2 = \ldots = m_k$ vs H_1 :almeno una è differente

In teoria

- SQ_{nei} è la **devianza NEI gruppi** $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ji}-\overline{X}_j)^2$

(misura la variabilità insita nei dati campionari, var. errore)

- SQ_{tra} è la **devianza TRA gruppi** $\sum_{j=1}^{k}(\overline{X}_j-\overline{X})^2 n_j$

(misura la variabilità attribuibile alla differenza tra dati non dovuta al caso)

- Scomposizione della varianza totale $\frac{SQ_{tot}}{n-1}=\frac{SQ_{tra}}{k-1}+\frac{SQ_{nei}}{n-k}$
- Statistica $F=\frac{SQ_{tra}/k-1}{SQ_{nei}/n-k}\sim F(k-1,n-k)$
- Se $F>F_\alpha$ si rifiuta l'ipotesi nulla (maggiore variabilità *tra* i gruppi rispetto a quella *nei* dà evidenza contro l'ipotesi nulla che le medie siano tutte uguali)

Se non valgono le assunzioni

- Violazione delle assunzioni
  - per testare l'assunzione di omogeneità: test di Levene
  - se non c'è omogeneità delle varianze si ricorre all'uso della statistica *F* di Brown-Forsythe (con aggiustamento dei gdl)

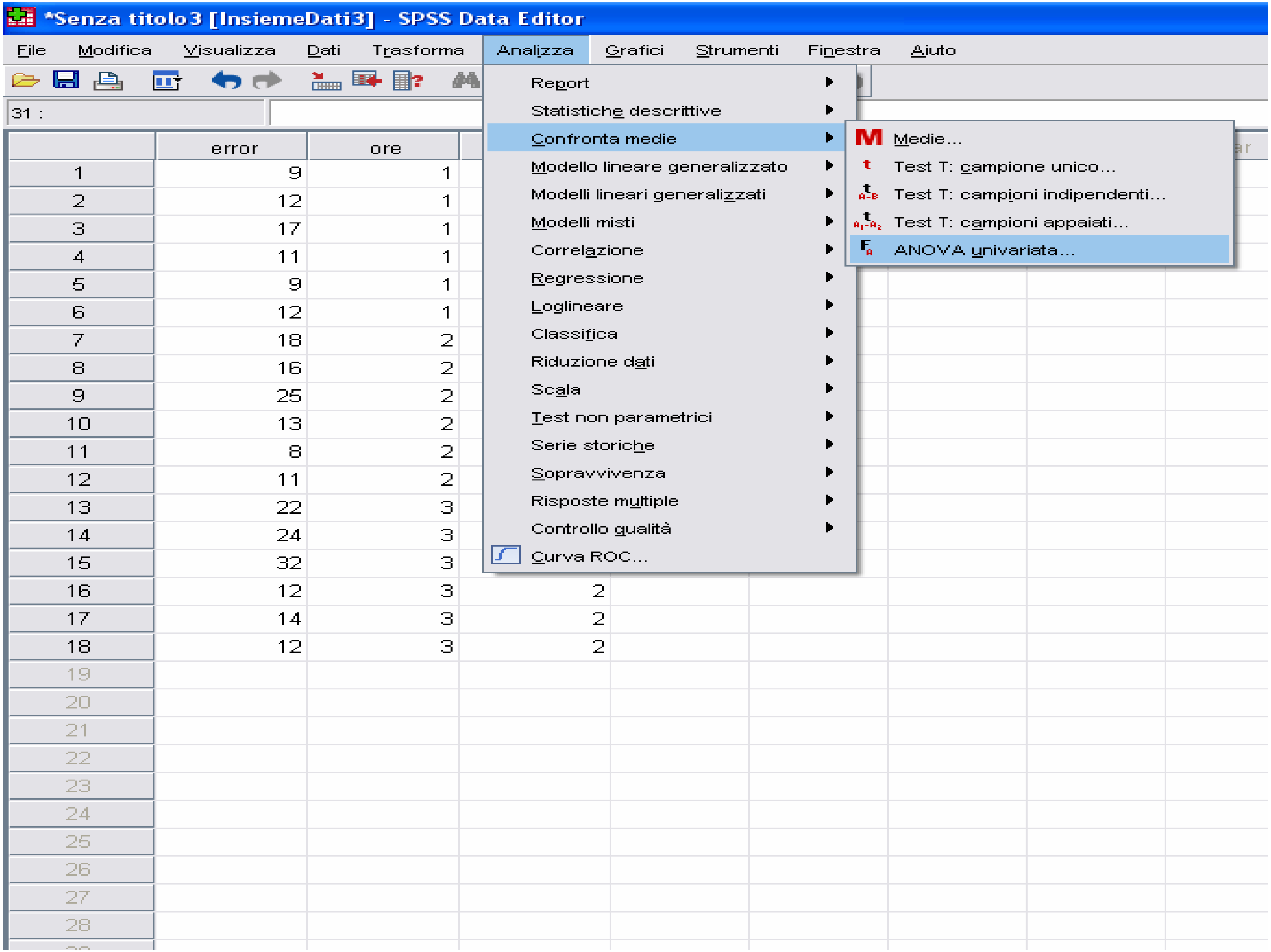
- Se le ipotesi dell'ANOVA sono violate si può usare l'approccio non parametrico di Kruscal-Wallis

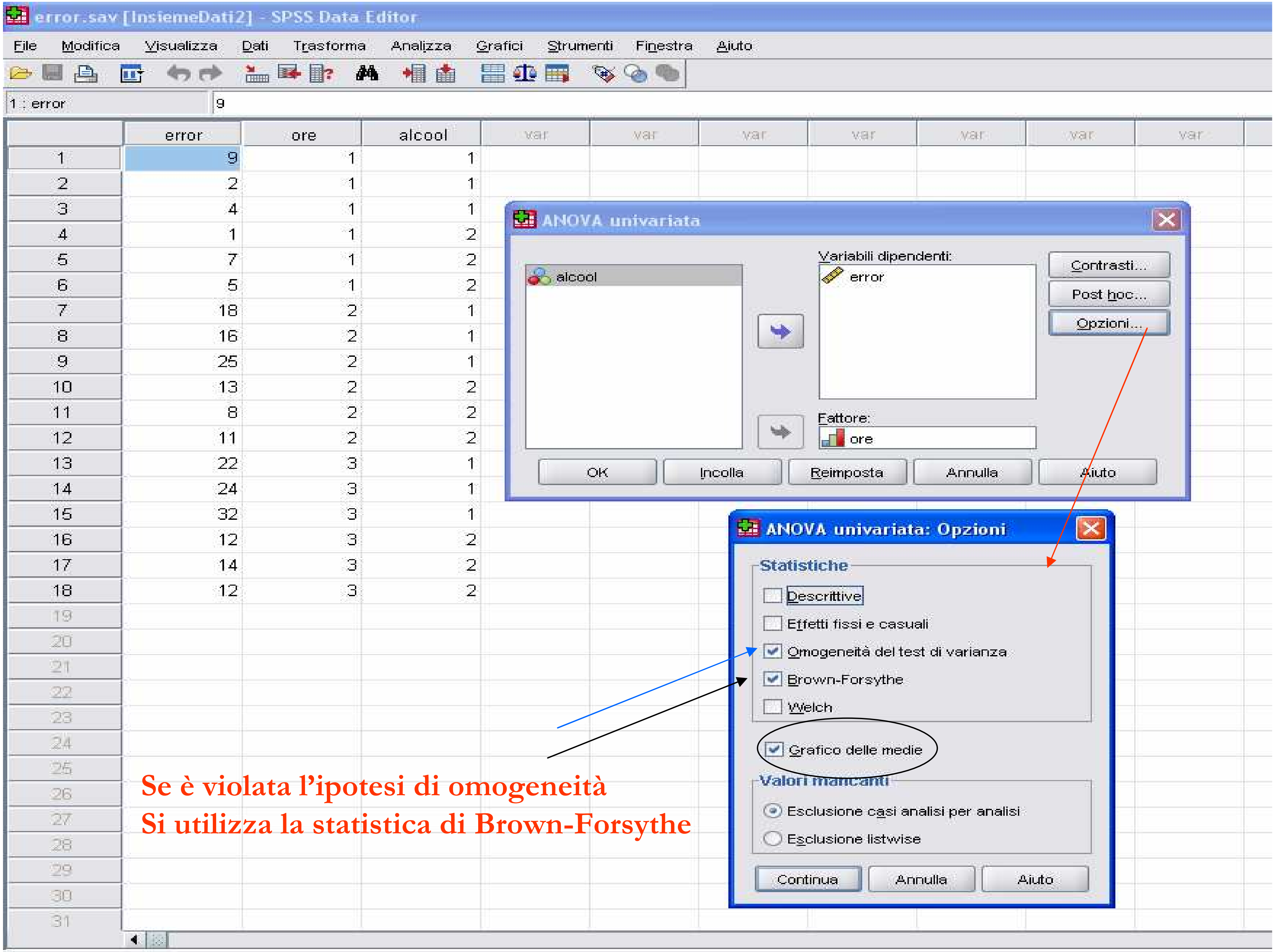
Cosa fare dopo aver rifiutato H_0

- Se si rifiuta l'ipotesi nulla si coglie evidenza nei dati che almeno due medie siano diverse
- Per scoprire a quali medie sia da attribuire il rifiuto si conducono
  - **Contrasti (pianificati)** contrasti lineari tra medie con pesi che sommano a 0

Ad es. si può voler confrontare la media del primo gruppo con quella dei gruppi 2 e 3 congiuntamente. I pesi saranno -2, 1, 1. Esempio Placebo, basso e alto dosaggio

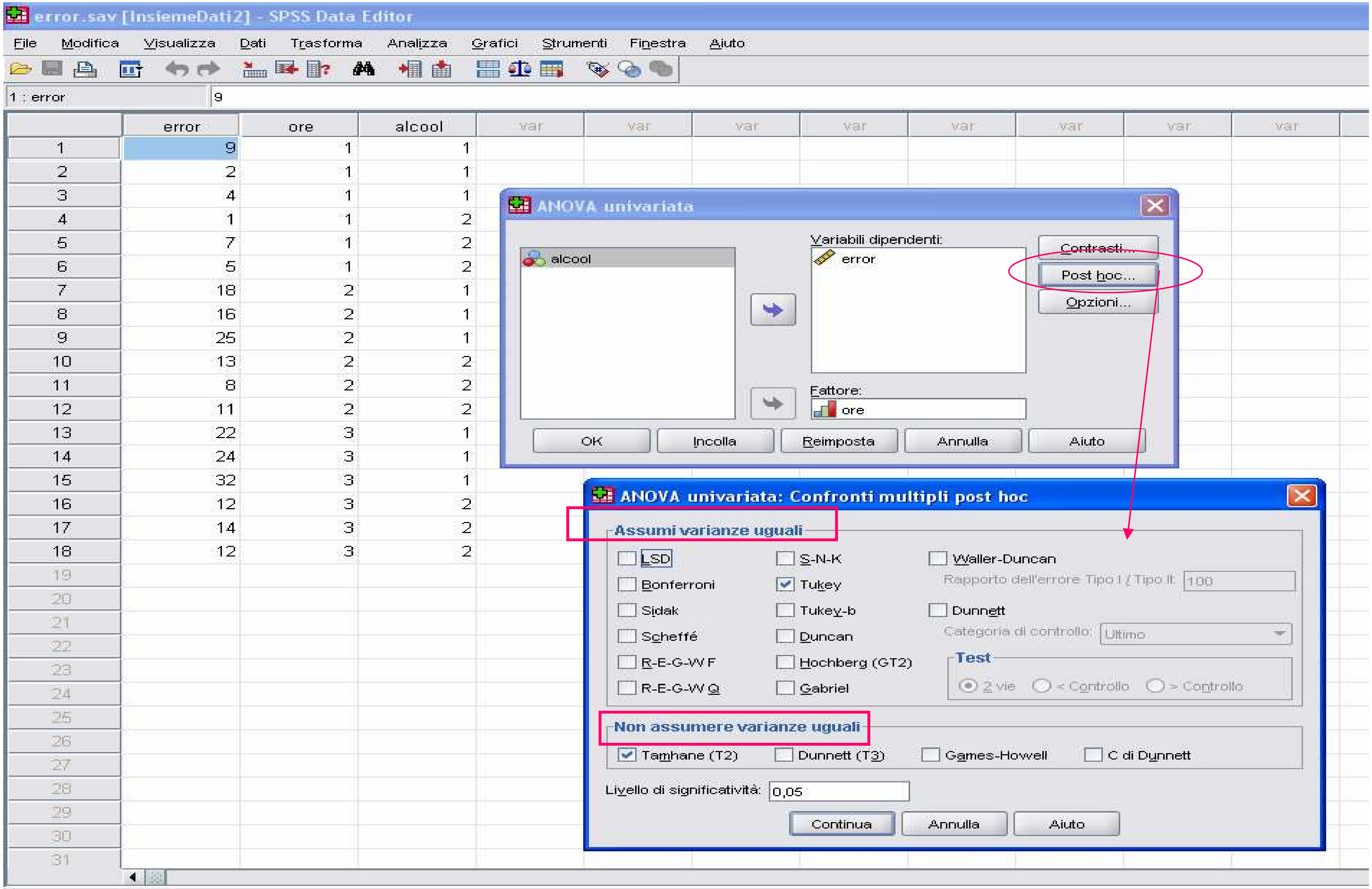
  - **Test post hoc** confronti multipli (ogni media contro le altre)
    - per rimediare al problema sull'errore di primo tipo che scaturisce da più test simultanei sugli stessi dati (Bonferroni, Scheffè, etc…)
    - la statistica di Dunnett sceglie una media di riferimento-controllo
    - se non c'è omogeneità delle varianze si ricorre all'uso di statistiche di Tamhane, Dunnett, Games-Howell etc..)
    - con campioni con diverse numerosità: statistiche Gabriel e Hochberg GT2
    - se le assunzioni del test valgono è consigliabile la statistica di Tukey


*Senza titolo3 [InsiemeDati3] - SPSS Data Editor
File
Modifica
Visualizza
Dati
Trasforma
Analizza
Grafici
Strumenti
Finestra
Aiuto
31 :
error
ore
Report
Statistiche descrittive
Confronta medie
Modello lineare generalizzato
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Classifica
Riduzione dati
Scala
Test non parametrici
Serie storiche
Sopravvivenza
Risposte multiple
Controllo qualità
Curva ROC...
Medie...
Test T: campione unico...
Test T: campioni indipendenti...
Test T: campioni appaiati...
ANOVA univariata...



Se è violata l'ipotesi di omogeneità
Si utilizza la statistica di Brown-Forsythe

Confronti multipli: post hoc



*Output1 [Documento1] - SPSS Viewer

File Modifica Visualizza Dati Trasforma Inserisci Formato Analizza Grafici Strumenti Finestra Aiuto

ANOVA univariata

[InsiemeDati2] C:\Documents and Settings\Hp\Desktop\fonetica\error.sav

Test di omogeneità delle varianze

error

Statistica di Levene	df1	df2	Sig.
3,284	2	15	,066

c'è omogeneità

ANOVA univariata

error

	Somma dei quadrati	df	Media dei quadrati	F	Sig.
Fra gruppi	685,444	2	342,722	9,355	,002
Entro gruppi	549,500	15	36,633		
Totale	1234,944	17			

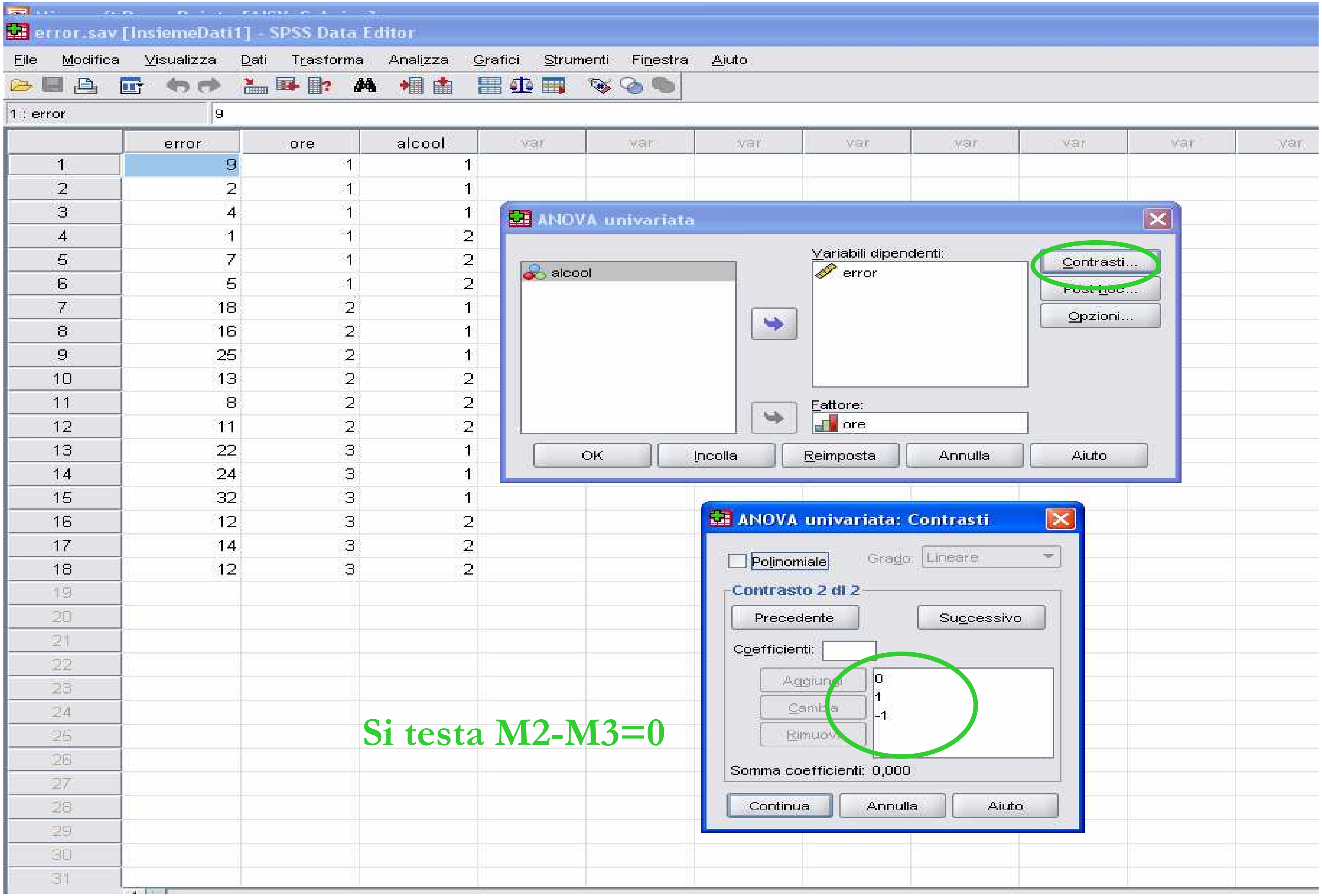
Test robusti per l'uguaglianza delle medie

error

	Statistica[a]	df1	df2	Sig.
Brown-Forsythe	9,355	2	10,793	,004

a. Distribuito a F asintoticamente

Confronti multipli: contrasti



Si testa M2-M3=0

Output sui contrasti

PSS Viewer

Trasforma Inserisci Formato Analizza Grafici Strumenti Finestra Aiuto

```
GET
  FILE='C:\Documents and Settings\Hp\Desktop\fonetica\error.sav'.
DATASET NAME InsiemeDati1 WINDOW=FRONT.
ONEWAY error BY ore
  /CONTRAST=2 -1 -1
  /CONTRAST=0 1 -1
  /MISSING ANALYSIS.
```

ANOVA univariata

[InsiemeDati1] C:\Documents and Settings\Hp\Desktop\fonetica\error.sav

ANOVA univariata

error

	Somma dei quadrati	df	Media dei quadrati	F	Sig.
Fra gruppi	685,444	2	342,722	9,355	,002
Entro gruppi	549,500	15	36,633		
Totale	1234,944	17			

Coefficienti di contrasto

Contrasto	ore 1	ore 2	ore 3
1	2	-1	-1
2	0	1	-1

Test di contrasto

		Contrasto	Valore di contrasto	Errore std.	t	df	Sig. (2-code)
error	Assumi varianze uguali	1	-25,17	6,053	-4,158	15	,001
		2	-4,17	3,494	-1,192	15	,252
	Non presume varianze uguali	1	-25,17	4,780	-5,265	13,764	,000
		2	-4,17	4,099	-1,016	9,221	,335

C'è differenza in media tra gruppo 1 e gruppi 2 e 3 insieme (primo contrasto), ma non c'è differenza tra M2 e M3 (secondo contrasto)

Output sui post hoc

Test post hoc

Confronti multipli

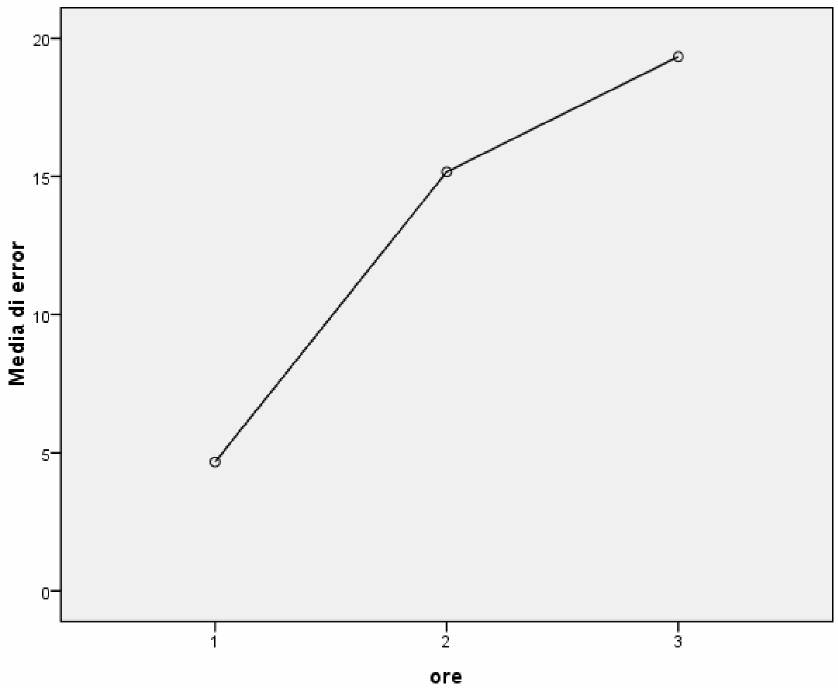
Variabile dipendente:error

	(I) ore	(J) ore	Differenza fra medie (I-J)	Errore std	Sig.	Intervallo di confidenza 95% Limite inferiore	Limite superiore
HSD di Tukey	1	2	-10,500*	3,494	,023	-19,58	-1,42
		3	-14,667*	3,494	,002	-23,74	-5,59
	2	1	10,500*	3,494	,023	1,42	19,58
		3	-4,167	3,494	,475	-13,24	4,91
	3	1	14,667*	3,494	,002	5,59	23,74
		2	4,167	3,494	,475	-4,91	13,24
Tamhane	1	2	-10,500*	2,734	,017	-18,89	-2,11
		3	-14,667*	3,515	,015	-25,94	-3,39
	2	1	10,500*	2,734	,017	2,11	18,89
		3	-4,167	4,099	,706	-16,09	7,75
	3	1	14,667*	3,515	,015	3,39	25,94
		2	4,167	4,099	,706	-7,75	16,09

*. La differenza media è significativa al livello 0.05

Sottoinsiemi omogenei

Grafici delle medie



error

	ore	N	Sottoinsieme per alfa = 0.05	
			1	2
HSD di Tukey[a]	1	6	4,67	
	2	6		15,17
	3	6		19,33
	Sig.		1,000	,475

Sono visualizzate le medie per i gruppi di sottoinsiemi omogenei.

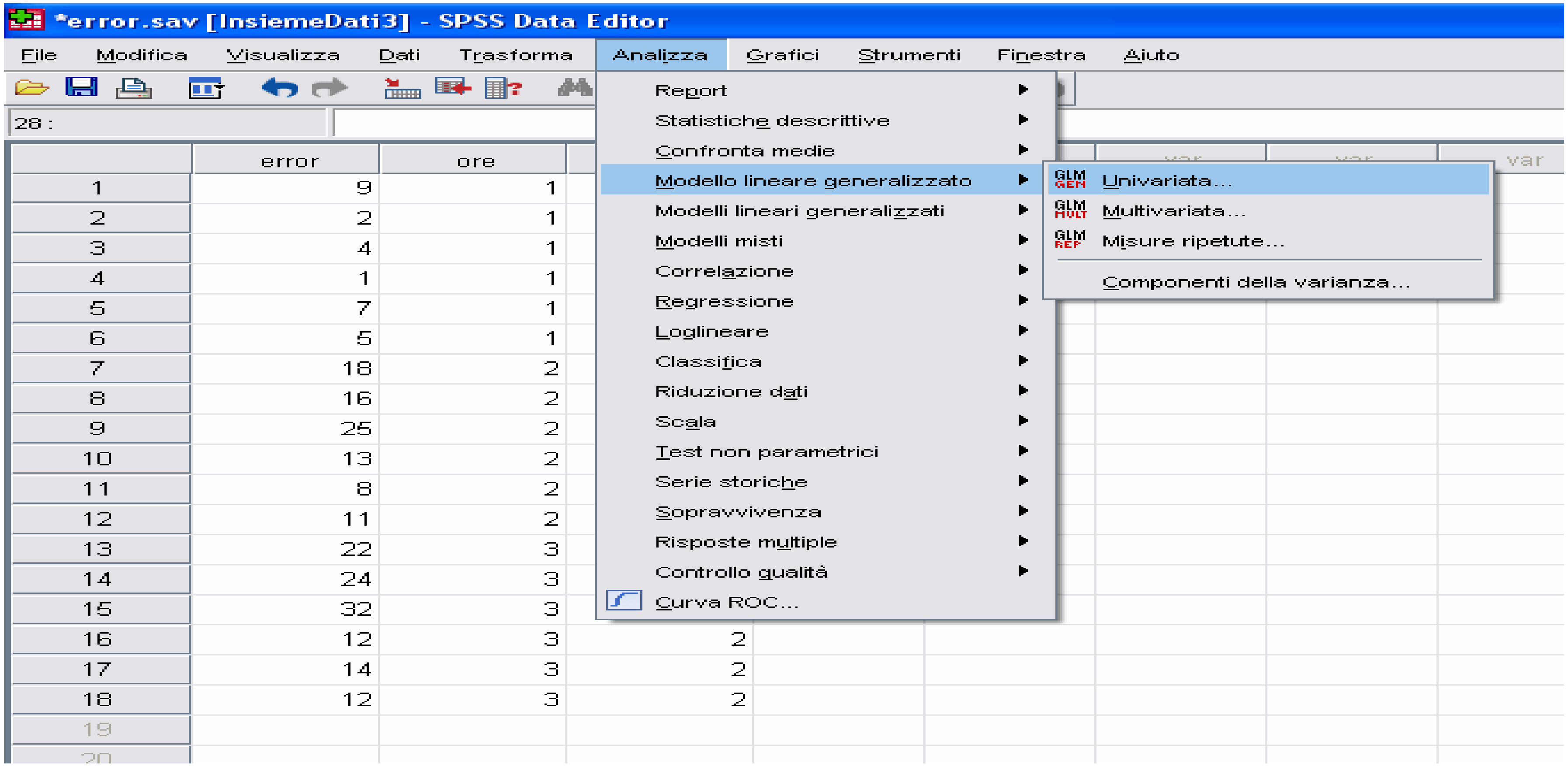
a. Utilizza dimensione campionaria media armonica = 6,000.

Il numero medio di errori commessi senza 3h di sonno differisce da quello che si registra quando le privazioni siano di 12h oppure 24h, ma qs ultime non fanno riscontrare differenze significative

Analisi della varianza a due vie

- *Esempio: Nell'esperimento sui problemi alla vista creati dall'effetto della mancanza di sonno si aggiunge, per metà campione, l'effetto dell'alcool.*

Il numero medio di errori nel test visivo varia al variare delle ore di sonno e del consumo di alcool?



Analisi della varianza univariata

[InsiemeDati1] C:\Documents and Settings\Hp\Desktop\fonetica\error.sav

Fattori tra soggetti

		Etichetta di valore	N
alcool	1		9
	2		9
ore	1	3 ore	6
	2	12 ore	6
	3	24 ore	6

Test degli effetti fra soggetti

Variabile dipendente:error

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.	Eta quadrato parziale
Modello corretto	1074,278[a]	5	214,856	16,047	,000	,870
Intercetta	3068,056	1	3068,056	229,149	,000	,950
alcool	264,500	1	264,500	19,755	,001	,622
ore	685,444	2	342,722	25,598	,000	,810
alcool * ore	124,333	2	62,167	4,643	,032	,436
Errore	160,667	12	13,389			
Totale	4303,000	18				
Totale corretto	1234,944	17				

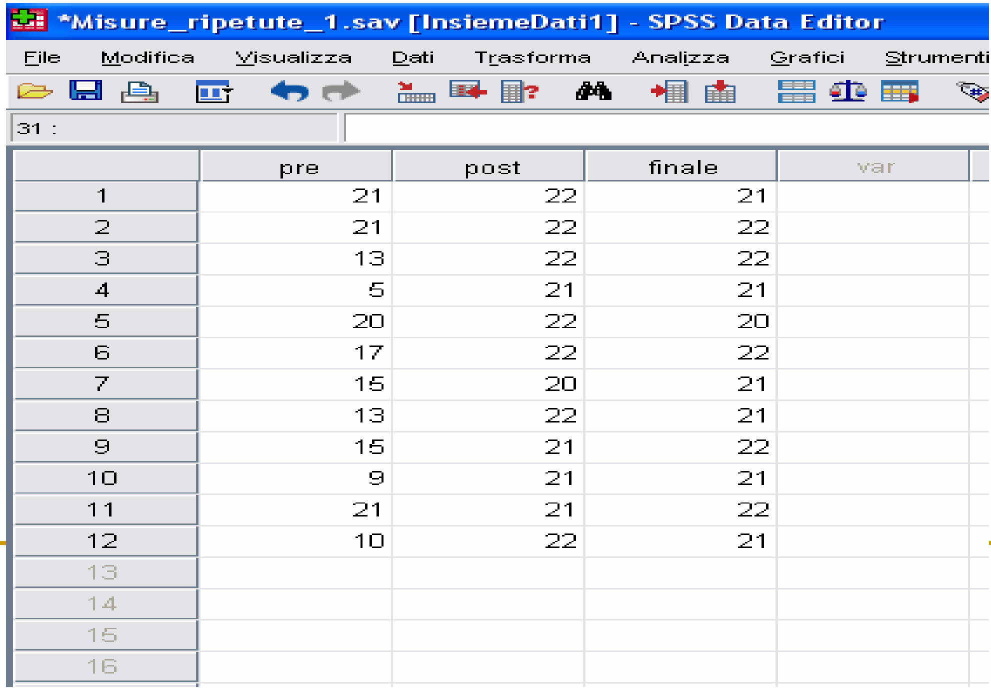
a. R quadrato = .870 (R quadrato corretto = .816)

Eta quadro misura il contributo del singolo fattore e delle interazioni, indica la varianza spiegata dall'effetto del fattore dopo aver rimosso gli altri effetti, eta^2=SQfattore/(SQfattore+SQerrore)

Analisi della varianza (misure ripetute)

- *Esempio: in uno studio condotto sull'apprendimento di bambini viene somministrato un pre-test prima di una lezione e conteggiato il punteggio raggiunto, poi viene somministrato ancora il test subito dopo una lezione e dopo un mese. Si vuol valutare se mediamente l'apprendimento sia variato, in termini di punti raggiunti nel test, nelle tre situazioni temporali.*

Stessi bambini nelle 3 prove, confronto di 3 medie e misure ripetute



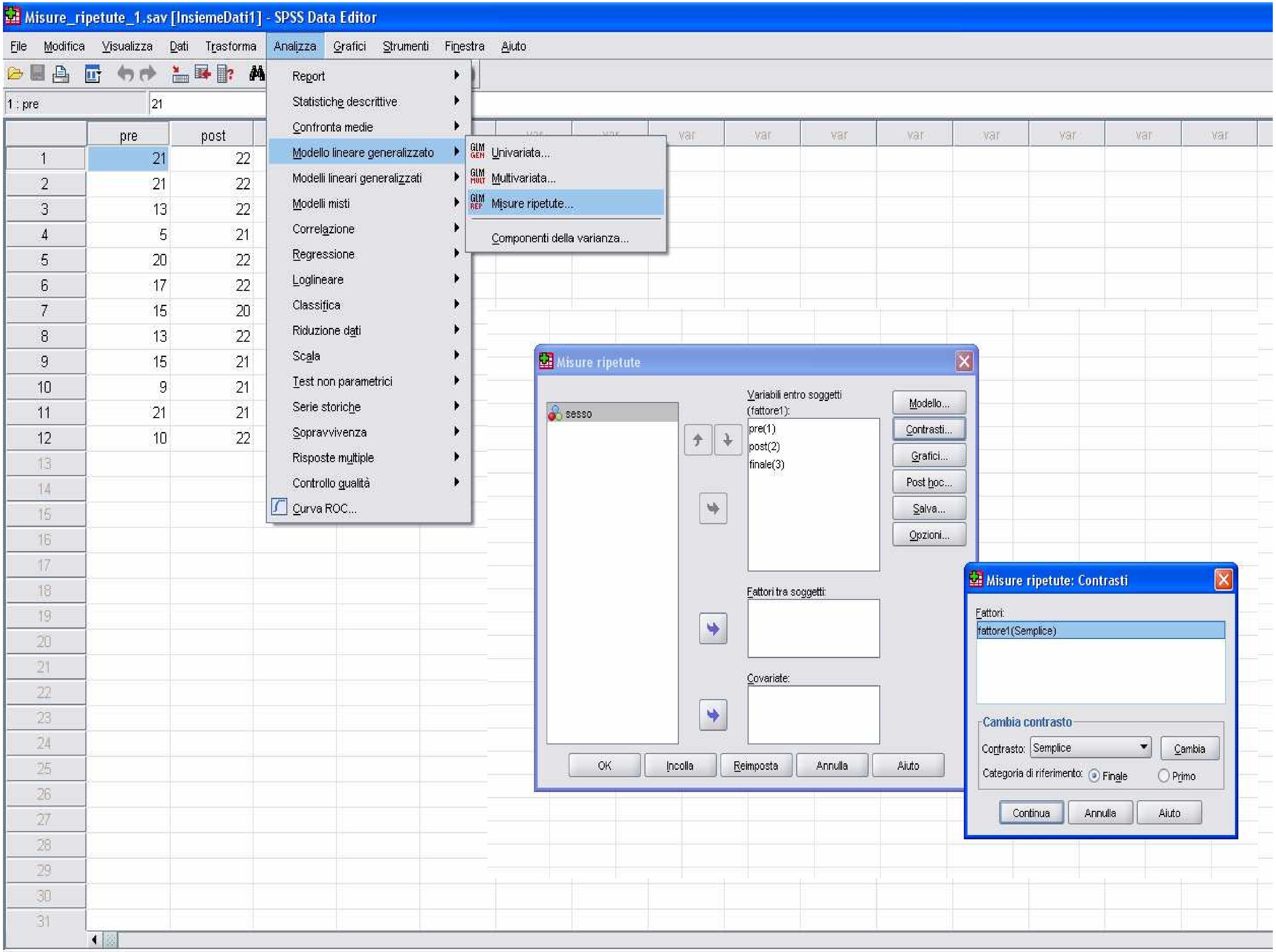
	pre	post	finale	var
1	21	22	21	
2	21	22	22	
3	13	22	22	
4	5	21	21	
5	20	22	20	
6	17	22	22	
7	15	20	21	
8	13	22	21	
9	15	21	22	
10	9	21	21	
11	21	21	22	
12	10	22	21	
13				
14				
15				
16				

Analisi della varianza (misure ripetute)

Si usa per il confronto di più medie calcolate per diversi livelli di un fattore sulle stesse unità

Cosa cambia rispetto all'ANOVA per dati indipendenti

- La variabilità totale = variabilità dovuta alle differenze da attribuire all'effetto del fattore + variabilità dovuta alle misure ripetute sulle stesse unità (differenze individuali) + varianza residua
- la componente di var NEI gruppi (componente di errore) si divide, nel caso di misure ripetute, in errore nei soggetti e errore residuo
- Vale l'ipotesi di sfericità: le varianze devono essere omogenee nei vari livelli del fattore e le covarianze per coppie di livelli devono essere uguali. Mauchly test per la sfericità


Misure_ripetute_1.sav [InsiemeDati1] - SPSS Data Editor
File Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
1 : pre 21
pre post var
Report
Statistiche descrittive
Confronta medie
Modello lineare generalizzato
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Classifica
Riduzione dati
Scala
Test non parametrici
Serie storiche
Sopravvivenza
Risposte multiple
Controllo qualità
Curva ROC...
Univariata...
Multivariata...
Misure ripetute...
Componenti della varianza...
Misure ripetute
sesso
Variabili entro soggetti (fattore1):
pre(1)
post(2)
finale(3)
Fattori tra soggetti:
Covariate:
Modello...
Contrasti...
Grafici...
Post hoc...
Salva...
Opzioni...
OK Incolla Reimposta Annulla Aiuto
Misure ripetute: Contrasti
Fattori:
fattore1(Semplice)
Cambia contrasto
Contrasto: Semplice Cambia
Categoria di riferimento: Finale Primo
Continua Annulla Aiuto

*Output1 [Documento1] - SPSS Viewer

Test di sfericità di Mauchly[b]

Misura:MEASURE_1

Effetto entro soggetti	W di Mauchly	Approssimazione chi-quadrato	df	Sig.	Epsilon[a] Greenhouse-Geisser	Huynh-Feldt	Limite inferiore
fattore1	,094	23,595	2	,000	,525	,532	,500

Verifica l'ipotesi nulla per la quale la matrice di covarianza dell'errore della variabile dipendente trasformata ortonormalizzata è proporzionale a una n

a. È possibile utilizzarlo per regolare i gradi di libertà per i test di significatività mediati. I test corretti vengono visualizzati nella tabella dei test sugli eff

b. Disegno: Intercetta
Disegno entro soggetti: fattore1

Test degli effetti entro soggetti

Misura:MEASURE_1

Sorgente		Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.
fattore1	Assumendo la sfericità	329,556	2	164,778	18,206	,000
	Greenhouse-Geisser	329,556	1,050	313,989	18,206	,001
	Huynh-Feldt	329,556	1,065	309,508	18,206	,001
	Limite inferiore	329,556	1,000	329,556	18,206	,001
Errore(fattore1)	Assumendo la sfericità	199,111	22	9,051		
	Greenhouse-Geisser	199,111	11,545	17,246		
	Huynh-Feldt	199,111	11,712	17,000		
	Limite inferiore	199,111	11,000	18,101		

Test dei contrasti entro soggetti

Misura:MEASURE_1

Sorgente	fattore1	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.
fattore1	Livello 1 vs livello 3	481,333	1	481,333	17,847	,001
	Livello 2 vs livello 3	,333	1	,333	,379	,551
Errore(fattore1)	Livello 1 vs livello 3	296,667	11	26,970		
	Livello 2 vs livello 3	9,667	11	,879		

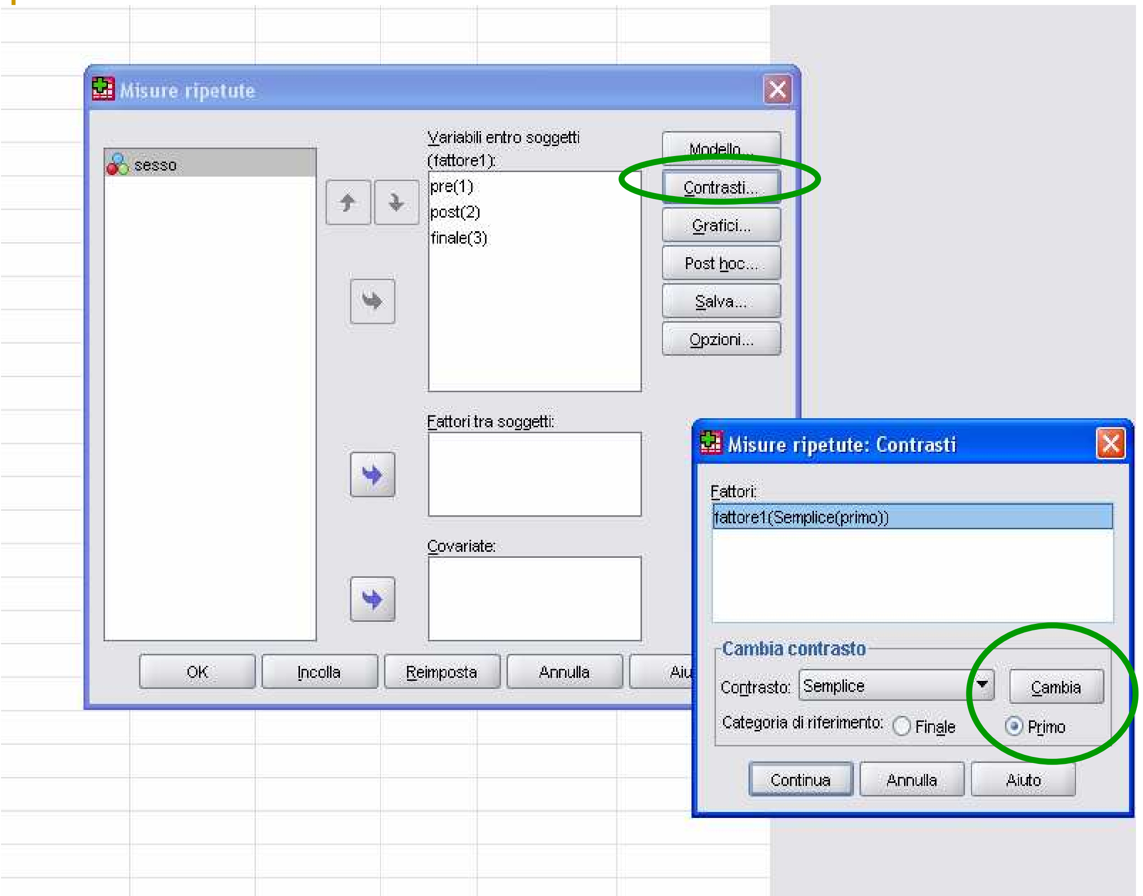
Scegliendo il contrasto semplice: M1 e M2 sono comparate con M3

Test degli effetti fra soggetti

Misura:MEASURE_1
Variabile trasformata:Media

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.
Intercetta	4459,593	1	4459,593	1262,630	,000
Errore	38,852	11	3,532		

C'è differenza tra le due situazioni estreme, ma non tra la seconda e la terza



Test dei contrasti entro soggetti

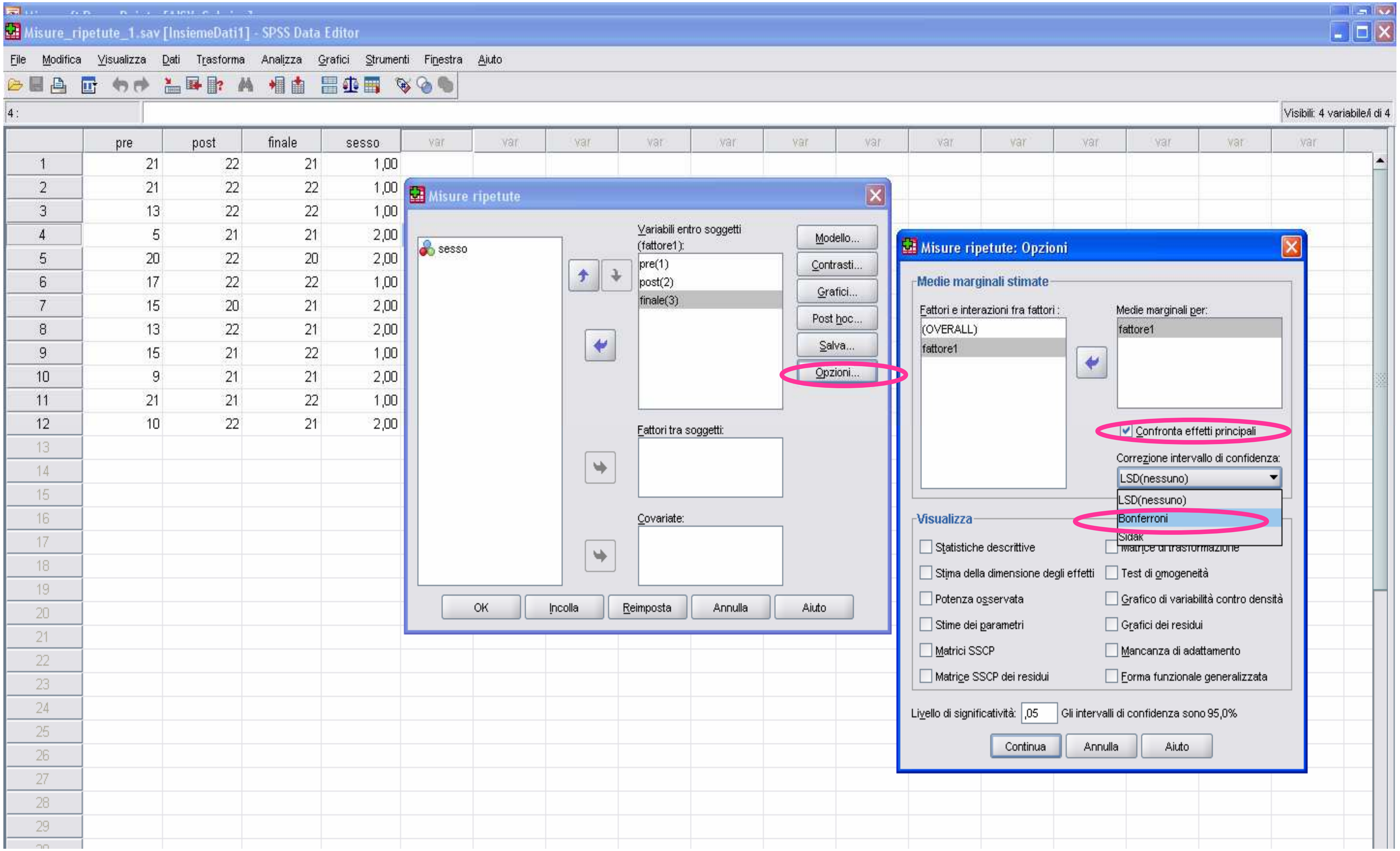
Misura:MEASURE_1

Sorgente	fattore1	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.
fattore1	Livello 2 vs livello 1	507,000	1	507,000	19,165	,001
	Livello 3 vs livello 1	481,333	1	481,333	17,847	,001
Errore(fattore1)	Livello 2 vs livello 1	291,000	11	26,455		
	Livello 3 vs livello 1	296,667	11	26,970		

Scegliendo un contrasto semplice con riferimento al primo gruppo,
M1 è comparata con M2 e M3

La lezione ha sortito il suo effetto: infatti le differenze che si riscontrano riguardano solo i confronti tra i punteggi medi raggiunti rispetto alla situazione iniziale

Post hoc (via Opzioni)



Viewer

rasforma Inserisci Formato Analizza Grafici Strumenti Finestra Aiuto

Medie marginali attese

fattore1

Stime

Misura:MEASURE_1

fattore1	Media	Errore std.	Intervallo di confidenza 95% Limite inferiore	Intervallo di confidenza 95% Limite superiore
1	15,000	1,523	11,649	18,351
2	21,500	,195	21,072	21,928
3	21,333	,188	20,919	21,747

Confronti a coppie

Misura:MEASURE_1

(I) fattore1	(J) fattore1	Differenza fra medie (I-J)	Errore std.	Sig.[a]	Intervallo di confidenza per la differenza al 95%[a] Limite inferiore	Intervallo di confidenza per la differenza al 95%[a] Limite superiore
1	2	-6,500*	1,485	,003	-10,687	-2,313
	3	-6,333*	1,499	,004	-10,561	-2,106
2	1	6,500*	1,485	,003	2,313	10,687
	3	,167	,271	1,000	-,596	,930
3	1	6,333*	1,499	,004	2,106	10,561
	2	-,167	,271	1,000	-,930	,596

Basato sulle medie marginali stimate

*. La differenza fra medie è significativa al livello .05

a. Correzione per confronti multipli: Bonferroni.

Conferma le considerazioni precedenti

Correlazione e Regressione

- L'obiettivo è l'analisi della dipendenza tra 2 variabili quantitative:
  - y (***variabile risposta***) e x (***variabile esplicativa***)
- Analizziamo come i valori di y tendano a variare in funzione dei diversi valori di x
- Una formula matematica può sintetizzare (in modo adeguato e non) il legame che esiste tra x e y per scopi di previsione e controllo
- La più semplice funzione è la retta che descrive una relazione lineare tra x e y:

$$\boxed{y = a + bx}$$

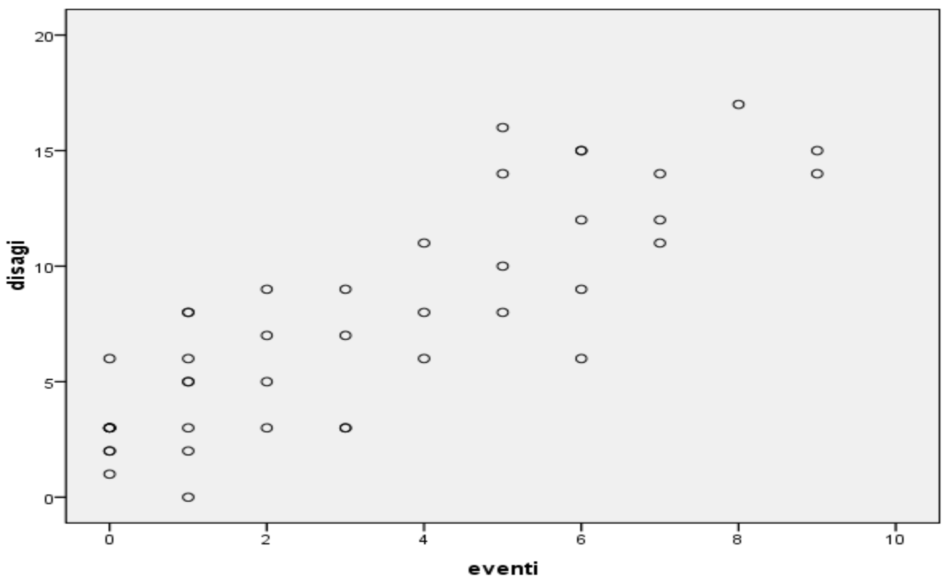
Esempio: Su un gruppo di pazienti viene rilevato il numero di visite per disagi mentali (crisi d'ansia, depressione, attacchi di panico) e il numero degli eventi di particolare rilevanza (gravi e/o felici) che hanno segnato la loro vita. Si vuole indagare se esiste un legame lineare tra disagi (risposta) ed eventi (esplicativa).

Correlazione

- Si dispone dell'elenco dei dati: n coppie di modalità relative ai caratteri quantitativi X=#eventi e Y=#disagi

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$$

- Graficamente:



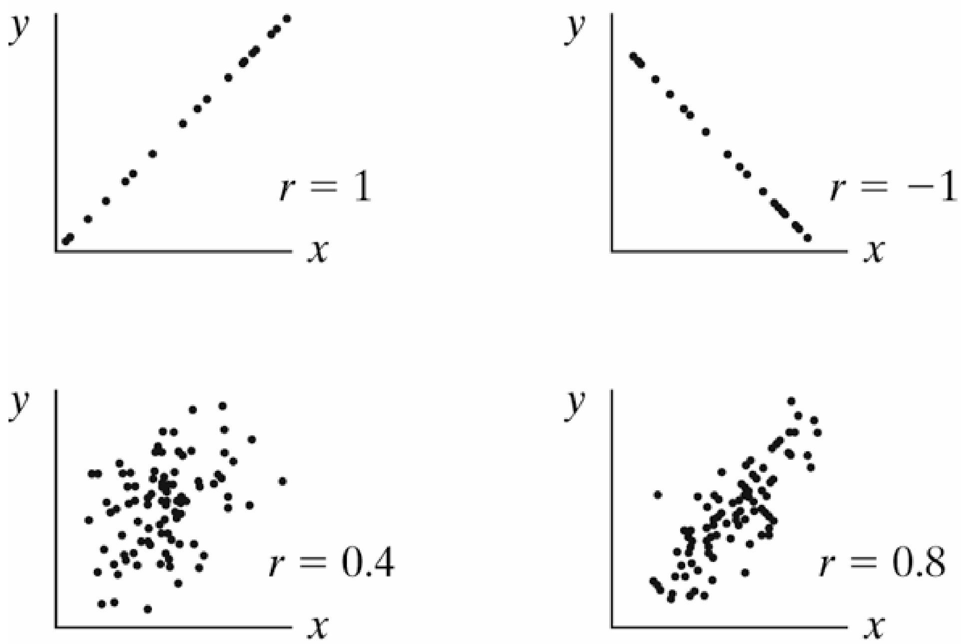
La nuvola dei punti appare caratterizzata da un trend lineare

Una misura di correlazione

- Correlazione = dipendenza lineare
- Coefficiente di correlazione di Pearson

$$r = \frac{\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\sum(x-\bar{x})^2 \; \frac{1}{n}\sum(y-\bar{y})^2}} = \frac{cov(x, y)}{\sqrt{s_x^2 s_y^2}}$$

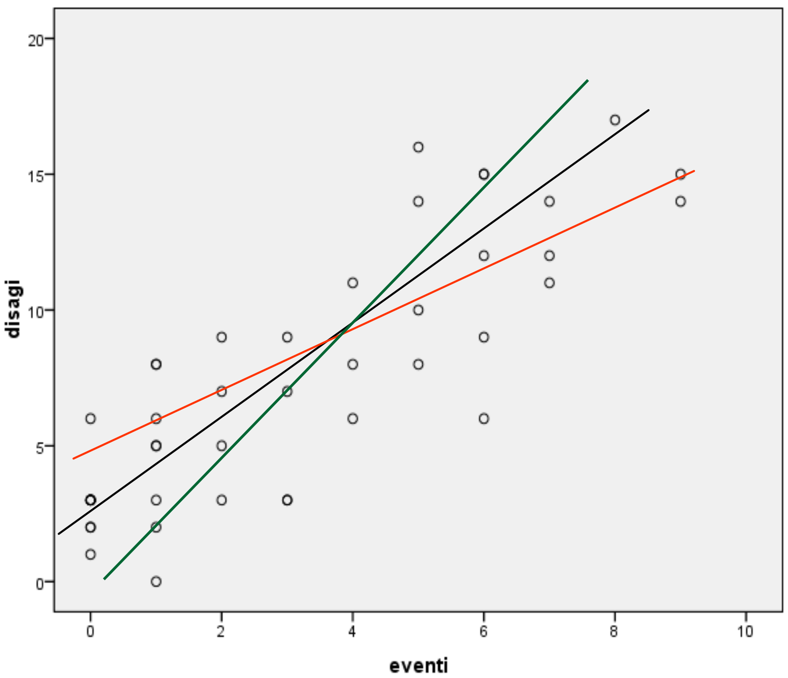
- Misura il grado di dipendenza linare
- Assume valori nell'intervallo [-1,1]
- E' pari a -1 e 1 se c'è perfetta relazione lineare



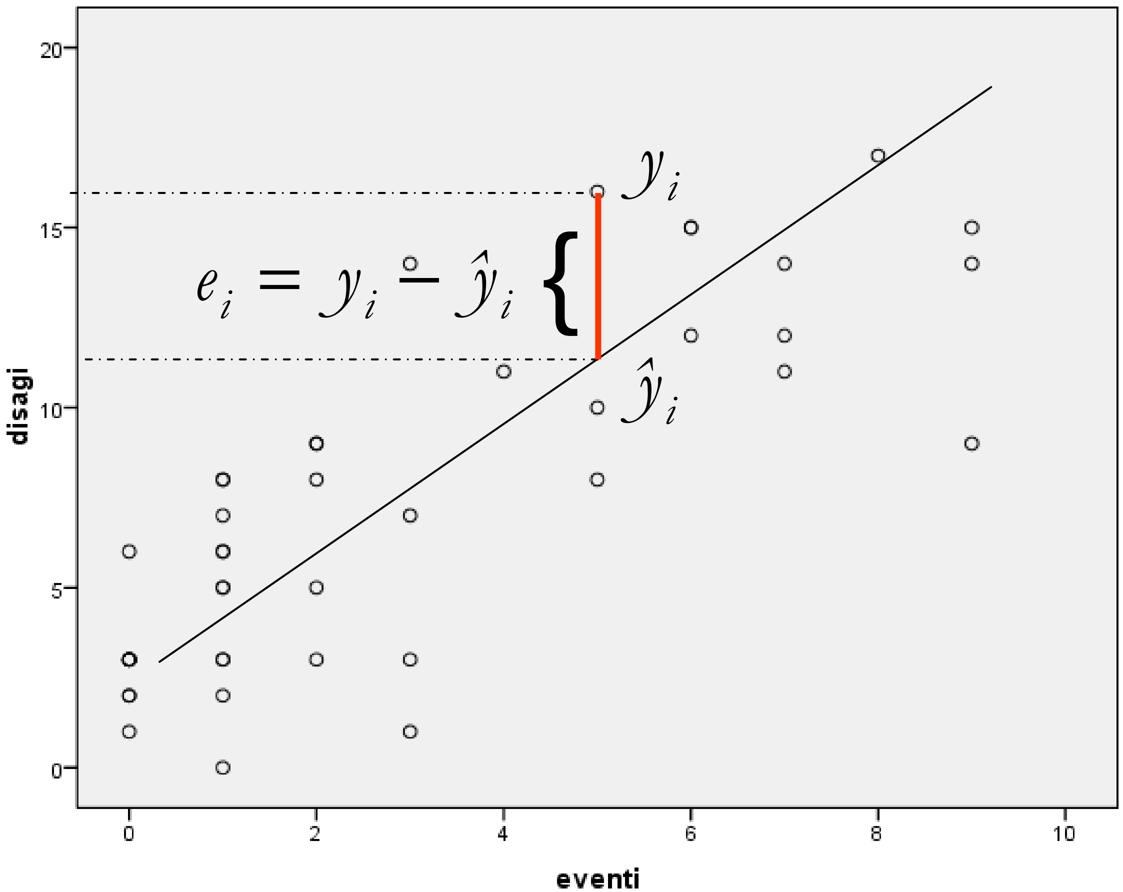
Nell'esempio: r = 0.84, p-value = 0.000

Retta di regressione

- Sembra plausibile l'idea di descrivere il trend della nuvola dei punti con una retta, e approssimare la realtà con un modello matematico, ma quale retta scegliere?



La retta dei minimi quadrati



La retta ai mini quadrati è quella che rende minima la **somma dei residui al quadrato**

$$\sum e^2 = \sum (y - \hat{y})^2$$

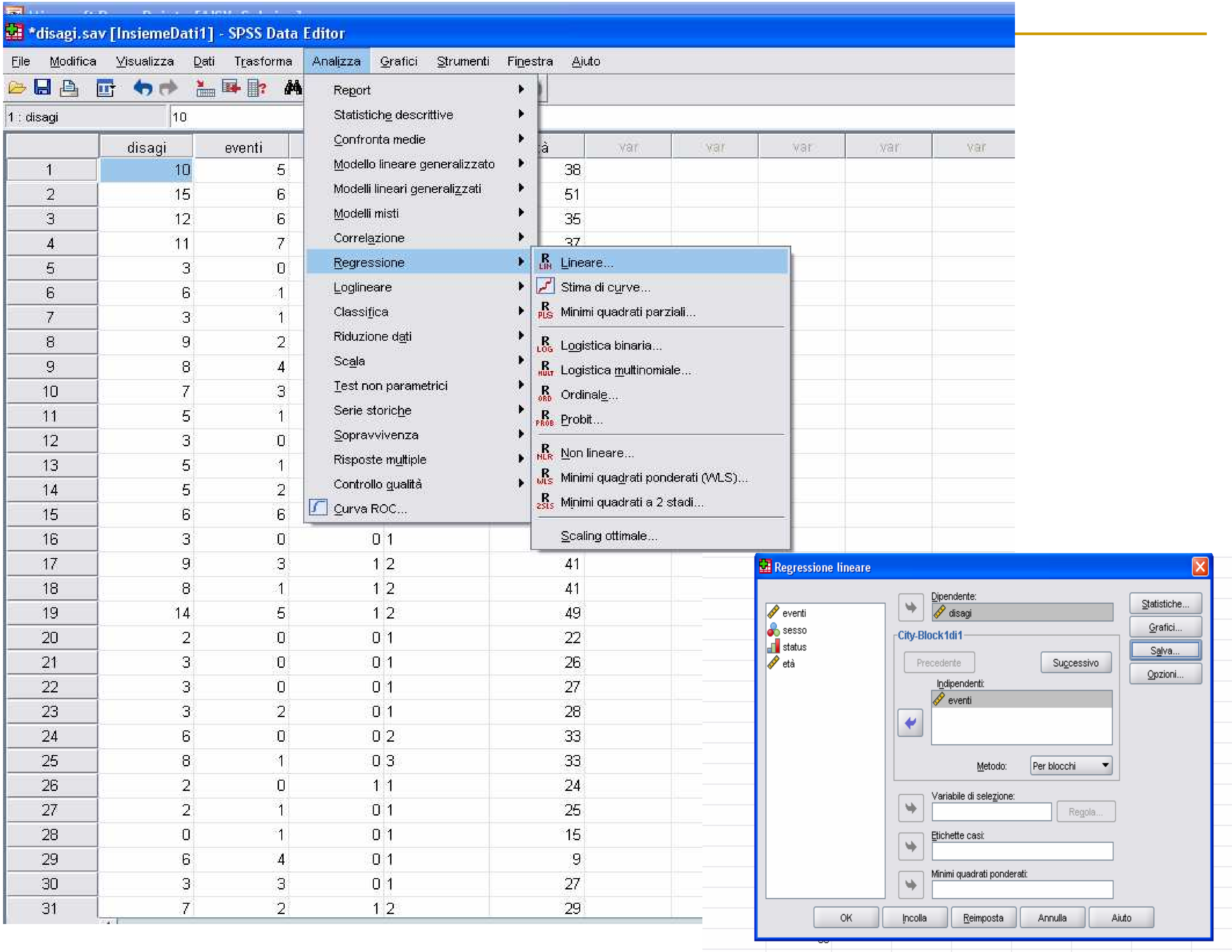
- **valori teorici**

$$\hat{y} = \hat{a} + \hat{b}x$$

- **parametri**

$$\hat{b} = \frac{cov(x, y)}{s_x^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$


*disagi.sav [InsiemeDati1] - SPSS Data Editor
File
Modifica
Visualizza
Dati
Trasforma
Analizza
Grafici
Strumenti
Finestra
Aiuto
1 : disagi
10
disagi
eventi
var
Report
Statistiche descrittive
Confronta medie
Modello lineare generalizzato
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Classifica
Riduzione dati
Scala
Test non parametrici
Serie storiche
Sopravvivenza
Risposte multiple
Controllo qualità
Curva ROC...
Lineare...
Stima di curve...
Minimi quadrati parziali...
Logistica binaria...
Logistica multinomiale...
Ordinale...
Probit...
Non lineare...
Minimi quadrati ponderati (WLS)...
Minimi quadrati a 2 stadi...
Scaling ottimale...
Regressione lineare
eventi
sesso
status
età
Dipendente:
disagi
City-Block1di1
Precedente
Successivo
Indipendenti:
eventi
Metodo:
Per blocchi
Variabile di selezione:
Regola...
Etichette casi:
Minimi quadrati ponderati:
Statistiche...
Grafici...
Salva...
Opzioni...
OK
Incolla
Reimposta
Annulla
Aiuto

Test, IC e bontà di adattamento

- Test per H_0: b=0 vs H_1: b≠0, statistica $t = \hat{b}/se$ che ha distribuzione t-Student
- Intervallo di confidenza: $\hat{b} \pm t_{\alpha/2} * se$

- Bontà di adattamento:
  - Il coefficiente di determinazione R^2 indica quanta parte della variabilità totale è spiegata dal modello

$$R^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = \frac{SQM}{SQT} = \frac{SQT - SQR}{SQT}$$

  - R^2 è la riduzione percentuale nell'errore che si ottiene sostituendo il valore teorico $\hat{y}$ anzichè la media $\bar{y}$ per prevedere y
  - Il test F ha lo stesso significato che ha nell'ANOVA

utput2 [Documento2] - SPSS Viewer

Modifica Visualizza Dati Trasforma Inserisci Formato Analizza Grafici Strumenti Finestra Aiuto

Regressione

[InsiemeDati1] C:\Documents and Settings\Hp\Desktop\fonetica\disagi.sav

Variabili inserite/rimosse[b]

Modello	Variabili inserite	Variabili rimosse	Metodo
1	eventi[a]	.	Per blocchi

a. Tutte le variabili richieste sono state inserite

b. Variabile dipendente: disagi

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1	,840[a]	,705	,698	2,594

a. Stimatori: (Costante), eventi

Significato di b: il numero di visite aumenta di 1.427per ogni evento importante in più nella vita del paziente

ANOVA[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	660,855	1	660,855	98,210	,000[a]
	Residuo	275,889	41	6,729		
	Totale	936,744	42			

a. Stimatori: (Costante), eventi

b. Variabile dipendente: disagi

La retta si adatta bene

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.	Intervallo di confidenza per B al 95%	
		B	Errore std.	Beta			Limite inferiore	Limite superiore
1	(Costante)	2,942	,606		4,856	,000	1,718	4,165
	eventi	1,427	,144	,840	9,910	,000	1,136	1,718

a. Variabile dipendente: disagi

Alcuni risultati

- Nell'esempio, l'equazione della retta è:

$$\hat{y} = 2.942 + 1.427x$$

- **Previsione**: qual è il numero di disagi che il modello stimato suggerisce per un paziente che dichiara una vita segnata da 5 eventi?

$$\hat{y} = 2.942 + 1.427 * 5 = 10$$

- **Controllo**: quanti eventi avrà subito, secondo il modello stimato, un paziente che dichiara di aver avuto 9 disagi?

$$9 = 2.942 + 1.427 * x \qquad x = {}^{(9-2.942)}\!/_{1.427} = 4.24$$

Regressione multipla

- k>1 variabili esplicative che possono spiegare la risposta

$$y = a + b_1 x_1 + \ldots + b_k x_k$$

- Il test F è analogo al test dell'ANOVA, l'ipotesi nulla è:

$$H_0 : b_1 = b_2 = \ldots = b_k = 0$$

  - Rifiutare tale H_0 significa che nessun regressore riesce a spiegare linearmente la risposta
- Il modello può comprendere tra le esplicative: variabili quantitative, nominali e ordinali. *Nell'esempio dei disagi si possono aggiungere variabili quali il sesso (nominale), lo status socio-economico (basso, medio, alto), l'età.*

NB. Le variabili ordinali e nominali vanno ricodificate usando le dummy. Ad es, la variabile status con categorie basso, medio, alto andrà sostituita da due dummy: *status1* (con valore 1 per categoria basso e 0 per categorie medio e alto) e *status 2* (con valore 1 per categoria medio e 0 per categorie basso e alto), la terza si omette perché ridondante

Nell'esempio, con soli due regressori quantitativi

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1	,902[a]	,814	,805	2,128

a. Stimatori: (Costante), età, eventi

ANOVA[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	794,930	2	397,465	87,781	,000[a]
	Residuo	181,117	40	4,528		
	Totale	976,047	42			

a. Stimatori: (Costante), età, eventi

b. Variabile dipendente: disagi

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	-1,913	1,002		-1,908	,064
	eventi	,953	,130	,571	7,306	,000
	età	,185	,031	,472	6,035	,000

a. Variabile dipendente: disagi

Sono significativi

Regressione multipla

Riepilogo del modello[b]

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1	,926[a]	,857	,837	1,944

a. Stimatori: (Costante), sesso, status2, eventi, età, status1

b. Variabile dipendente: disagi

Buon adattamento

ANOVA[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	836,243	5	167,249	44,264	,000[a]
	Residuo	139,804	37	3,778		
	Totale	976,047	42			

a. Stimatori: (Costante), sesso, status2, eventi, età, status1

b. Variabile dipendente: disagi

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	2,769	1,908		1,451	,155
	eventi	,752	,138	,451	5,461	,000
	età	,101	,038	,259	2,650	,012
	status1	-3,499	1,340	-,329	-2,611	,013
	status2	-2,284	1,232	-,177	-1,854	,072
	sesso	,357	,932	,037	,383	,704

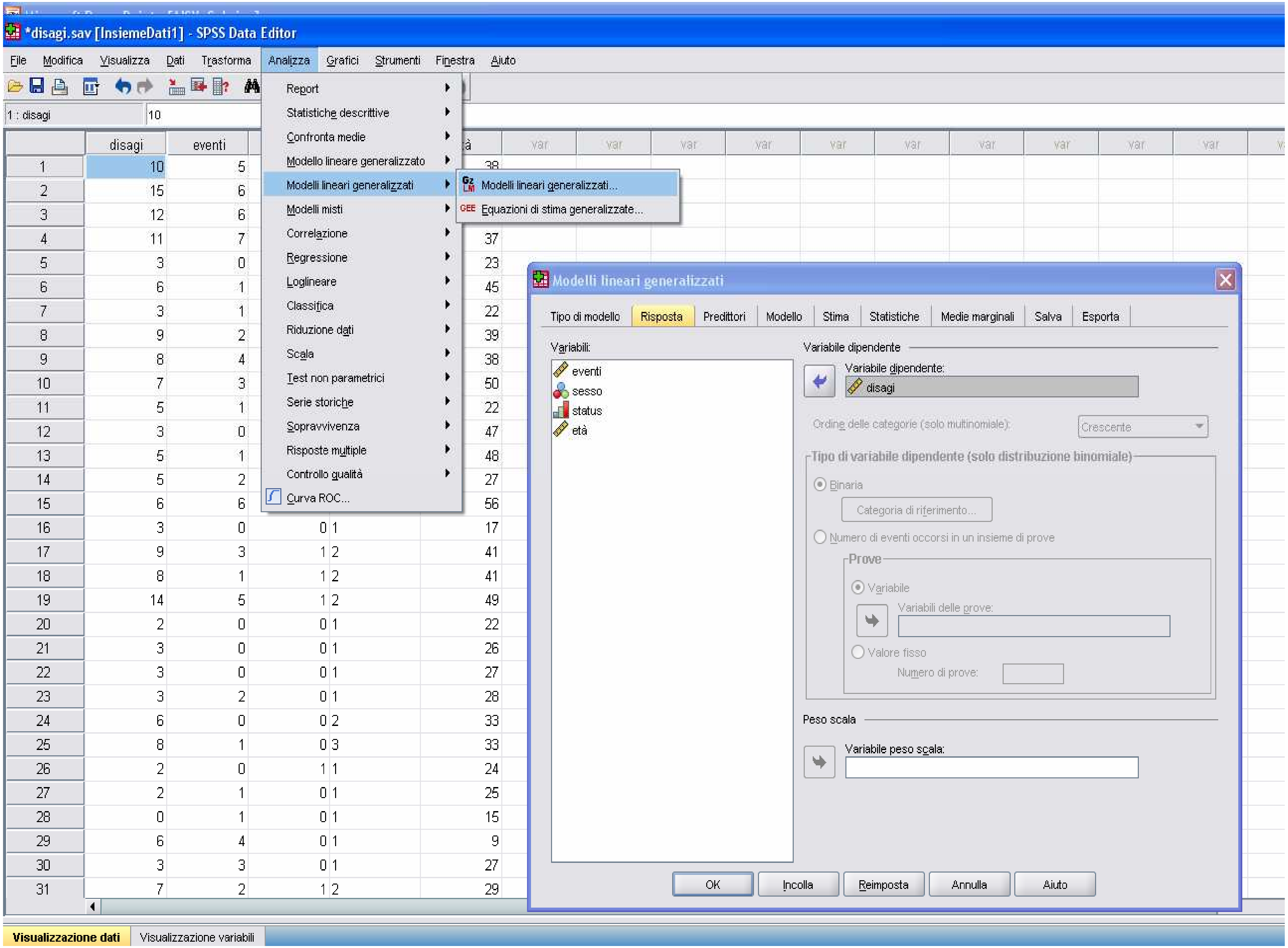
Il sesso è un regressore da eliminare

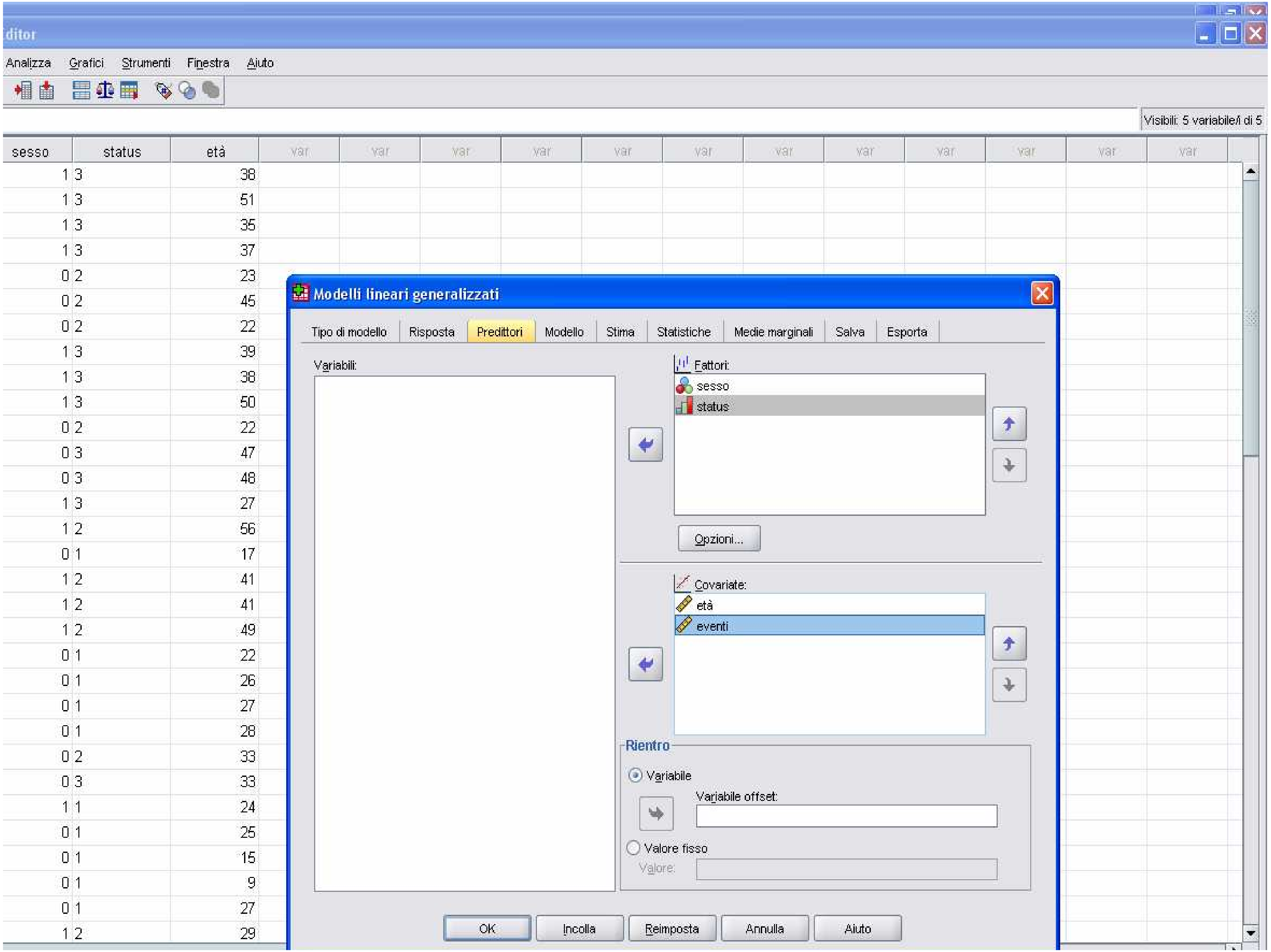
Caso particolare dei GLM

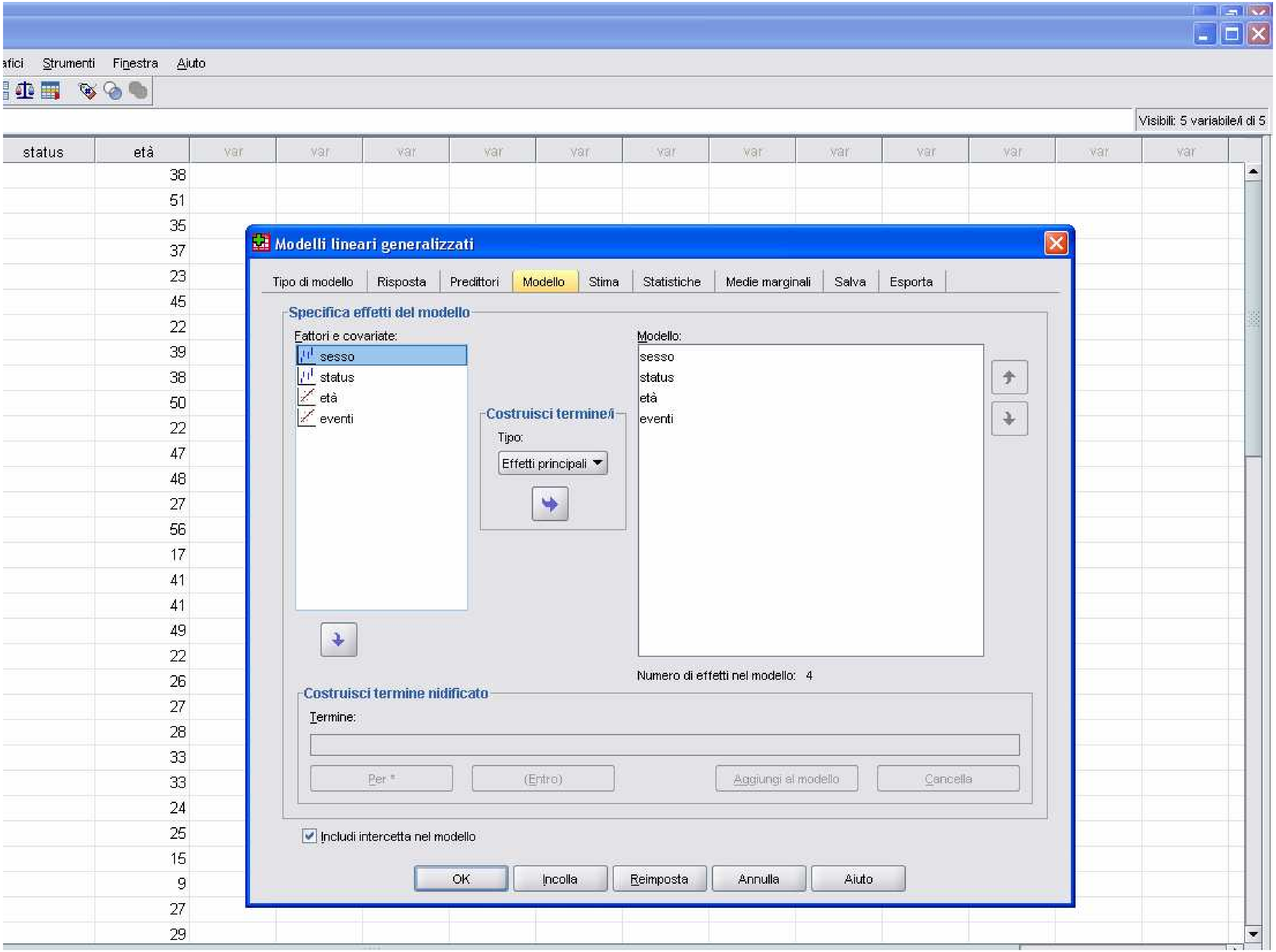
- Il modello di regressione è un caso particolare dei GLM con variabile Normale e link identità

- rispetto al modello presentato prima in cui le dummy sostituivano le variabili categoriali, le variabili qualitative vanno inserite tra i fattori e dichiarate secondo la loro natura nominale o ordinale

- La bontà di adattamento e il test su coefficienti sono effettuati usando statistiche con distribuzione chi-quadro


*disagi.sav [InsiemeDati1] - SPSS Data Editor
File
Modifica
Visualizza
Dati
Trasforma
Analizza
Grafici
Strumenti
Finestra
Aiuto
1 : disagi
10
disagi
eventi
Report
Statistiche descrittive
Confronta medie
Modello lineare generalizzato
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Classifica
Riduzione dati
Scala
Test non parametrici
Serie storiche
Sopravvivenza
Risposte multiple
Controllo qualità
Curva ROC...
Modelli lineari generalizzati...
Equazioni di stima generalizzate...
Modelli lineari generalizzati
Tipo di modello
Risposta
Predittori
Modello
Stima
Statistiche
Medie marginali
Salva
Esporta
Variabili:
eventi
sesso
status
età
Variabile dipendente
Variabile dipendente:
disagi
Ordine delle categorie (solo multinomiale):
Crescente
Tipo di variabile dipendente (solo distribuzione binomiale)
Binaria
Categoria di riferimento...
Numero di eventi occorsi in un insieme di prove
Prove
Variabile
Variabili delle prove:
Valore fisso
Numero di prove:
Peso scala
Variabile peso scala:
OK
Incolla
Reimposta
Annulla
Aiuto
Visualizzazione dati
Visualizzazione variabili


ditor
Analizza Grafici Strumenti Finestra Aiuto
Visibili: 5 variabile/i di 5
sesso status età var var var var var var var var var var var var
1 3 38
1 3 51
1 3 35
1 3 37
0 2 23
0 2 45
0 2 22
1 3 39
1 3 38
1 3 50
0 2 22
0 3 47
0 3 48
1 3 27
1 2 56
0 1 17
1 2 41
1 2 41
1 2 49
0 1 22
0 1 26
0 1 27
0 1 28
0 2 33
0 3 33
1 1 24
0 1 25
0 1 15
0 1 9
0 1 27
1 2 29
Modelli lineari generalizzati
Tipo di modello Risposta Predittori Modello Stima Statistiche Medie marginali Salva Esporta
Variabili:
Fattori:
sesso
status
Opzioni...
Covariate:
età
eventi
Rientro
Variabile
Variabile offset:
Valore fisso
Valore:
OK Incolla Reimposta Annulla Aiuto


afici Strumenti Finestra Aiuto
Visibili: 5 variabile/i di 5
status
età
var
38
51
35
37
23
45
22
39
38
50
22
47
48
27
56
17
41
41
49
22
26
27
28
33
33
24
25
15
9
27
29
Modelli lineari generalizzati
Tipo di modello
Risposta
Predittori
Modello
Stima
Statistiche
Medie marginali
Salva
Esporta
Specifica effetti del modello
Fattori e covariate:
sesso
status
età
eventi
Costruisci termine/i
Tipo:
Effetti principali
Modello:
sesso
status
età
eventi
Numero di effetti nel modello: 4
Costruisci termine nidificato
Termine:
Per *
(Entro)
Aggiungi al modello
Cancella
Includi intercetta nel modello
OK
Incolla
Reimposta
Annulla
Aiuto

Test omnibus[a]

Chi-quadrato per il rapporto di verosimiglianza	df	Correzione per confronti multipli
836,243	5	,000

Variabile dipendente: disagi
Modello: (Intercetta), sesso, status, eventi, età

a. Confronta il modello adattato con il modello con la sola intercetta

Test degli effetti del modello

Sorgente	Tipo III		
	Chi-quadrato di Wald	df	Correzione per confronti multipli
(Intercetta)	2,336	1	,126
sesso	,555	1	,456
status	25,768	2	,000
eventi	112,666	1	,000
età	26,542	1	,000

Variabile dipendente: disagi
Modello: (Intercetta), sesso, status, eventi, età

Stime dei parametri

Parametro	B	Errore standard	95% Intervallo di confidenza di Wald		Test dell'ipotesi		
			Inferiore	Superiore	Chi-quadrato di Wald	df	Correzione per confronti multipli
(Intercetta)	3,126	,9150	1,333	4,919	11,672	1	,001
[sesso=0]	-,357	,4795	-1,297	,583	,555	1	,456
[sesso=1]	0[a]	.	.	.	.	.	.
[status=1]	-3,499	,6894	-4,851	-2,148	25,760	1	,000
[status=2]	-2,284	,6339	-3,527	-1,042	12,982	1	,000
[status=3]	0[a]	.	.	.	.	.	.
eventi	,752	,0708	,613	,891	112,666	1	,000
età	,101	,0197	,063	,140	26,542	1	,000
(Scala)	1[b]						

Variabile dipendente: disagi

__Interpretazione dei coefficienti:__

- Il sesso non crea differenze significative;
- A parità di eventi, età e sesso, coloro che hanno lo status più basso fanno registrare 3.5 visite in meno rispetto agli upper class
- Ci si aspetta un disagio in più (1.01=0.101*10) per un paziente più anziano di 10 anni a parità delle altre variabili
- Per ogni evento in più che segna la vita del paziente c'è da attendersi quasi un disagio in più (0.752)

In teoria (cenni di approfondimento)

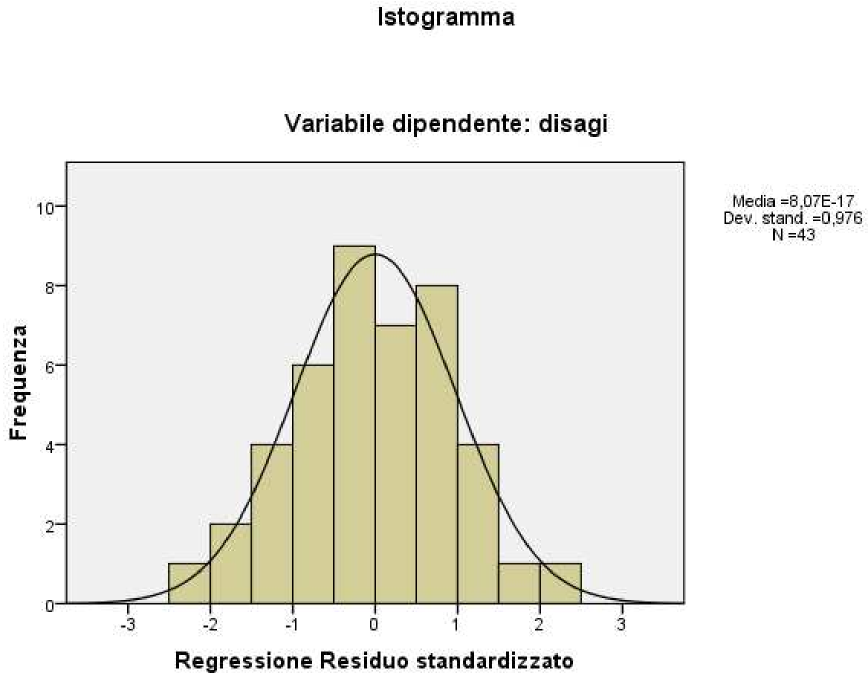
- Le assunzioni del modello di regressione multipla:

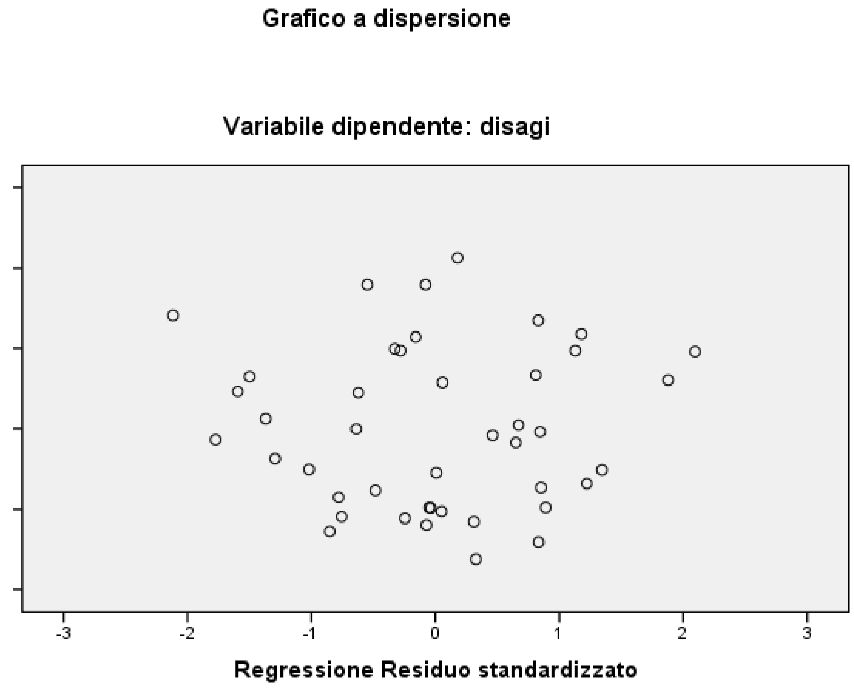
$$y = a + b_1 x_1 + \ldots + b_k x_k + \varepsilon$$

- La media $E(Y) = a + b_1 x_1 + \ldots + b_k x_k$ è la componente deterministica
- La variabile risposta è Normale, gli **errori** sono omoschedastici (con varianza costante) e incorrelati
- Le variabili esplicative non devono essere legate linearmente tra loro (multicollinearità)
- Esistono test e indici per valutare se le ipotesi del modello sono violate: Test di Bartlett e test dei residui per l'omoschedasticità, Durbin-Watson per l'incorrelazione degli errori, VIF per la multicollinearità
- Analisi grafica dei residui aiuta a comprendere se le ipotesi non sono state violate

Analisi dei residui

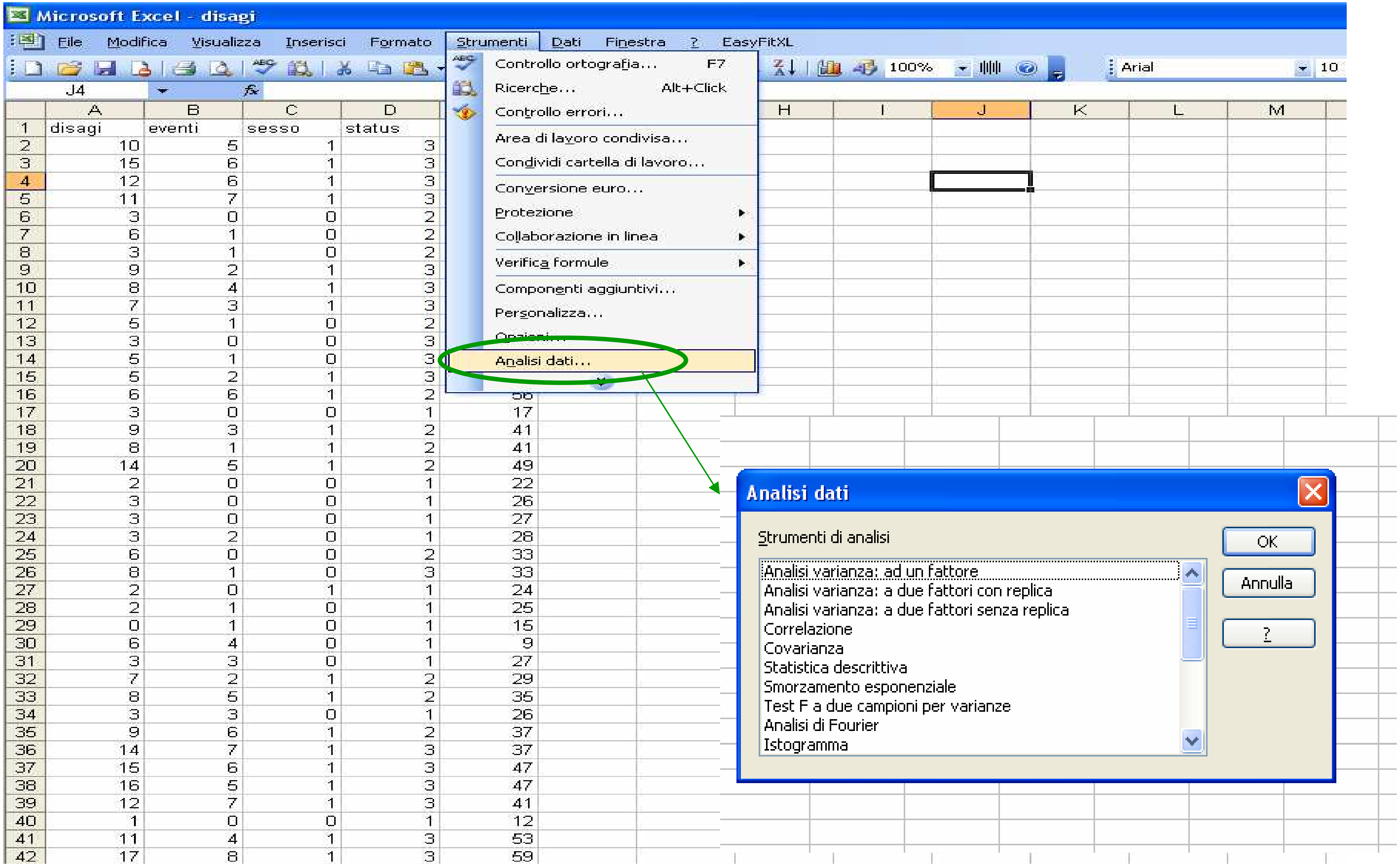
In riferimento all'esempio con risposta: disagi, esplicative: età, eventi





L'andamento erratico dei residui indica che il modello ha un buon adattamento

Le varie analisi in excel



Regressione logistica (per risposte binarie)

- la variabile risposta è binaria (successo/insuccesso), le esplicative sono sia qualitative che quantitative
- La variabile di Bernoulli ha media $E(Y)=p$ che è la probabilità di successo quindi la relazione $E(Y)=a+b_1x_1+...+b_kx_k$ non è valida nel caso di risposta Y bernoulliana, così si utilizza una trasformazione che giustifichi la relazione:

$$g(E(Y))=a+b_1x_1+...+b_kx_k$$

- La funzione g è scelta come:

$$\boxed{g(p)=log\frac{p}{1-p}}$$

- Tale funzione è uguagliata alla combinazione lineare dei regressori
- La probabilità si ricava:

$$p=\frac{exp^{\,a+b_1x_1+...+b_kx_k}}{1+exp^{\,a+b_1x_1+...+b_kx_k}}$$

Un esempio

- *Esempio: Alcune persone intervistate per strada sono state invitate a rispondere all'interrogativo: "Cosa pensa dell'idea di dar il diritto di voto agli immigrati?" (con risposta binaria: favorevole/non favorevole). Per ogni rispondente si rilevano le informazioni in merito all'età, il sesso e il livello di scolarizzazione (basso:< =scuola media inferiore, alto: >=scuola superiore)*
- Ci si domanda se ad esser più propensi verso il riconoscimento del diritto di voto siano gli uomini o le donne, i più o i meno colti, i più giovani o gli anziani
- Si stima il modello $\log \frac{p}{1-p} = a + b_1 x_1 + b_2 x_2 + b_3 x_3$

  con x_1=età; x_2=livello di scolarizzazione; x_3=sesso

  p è la probabilità di esser *favorevole*
- la formula $p/_{1-p}$ è il rapporto tra la probabilità di *esser favorevole* rispetto a quella di essere *non favorevole*, è la propensione verso il fenomeno, è detto ***odds***

Il modello stimato

- Il modello finale include tra i regressori: *livello di scolarizzazione* e *età*

$$\log \frac{p}{1-p} = 3.449 - 0.079\,età + 1.904\,scolar$$

Stime dei parametri

opinione[a]		B	Errore std	Wald	df	Sig.	Exp(B)	Intervallo di confidenza al 95% per Exp(B) Limite inferiore	Limite superiore
favorevole	Intercetta	3,449	1,451	5,653	1	,017			
	età	-,079	,025	10,019	1	,002	,924	,880	,970
	[sesso=0]	-,756	,774	,955	1	,328	,469	,103	2,139
	[sesso=1]	0[b]	.	.	0	.	.	.	.
	[scolarizzazione=1]	1,904	,780	5,953	1	,015	6,714	1,454	30,999
	[scolarizzazione=2]	0[b]	.	.	0	.	.	.	.

a. La categoria di riferimento è: non favorevole.

non significativo

- Come calcolare le probabilità attese:

1. $$p \mid età = 20, scolar = "alto" = \frac{exp(3.449 - 0.079*20 + 1.904*1)}{1 + exp(3.449 - 0.079*20 + 1.904*1)} = 0.97$$

2. $$p \mid età = 65, scolar = "basso" = \frac{exp(3.449 - 0.079*65 + 1.904*0)}{1 + exp(3.449 - 0.079*65 + 1.904*0)} = 0.15$$

Come interpretare i coefficienti

- Il coefficiente b è il logaritmo di un odds ratio, e quindi $\frac{p|x+1}{1-p|x+1} = exp(b)\frac{p|x}{1-p|x}$

Per i regressori inclusi nel modello finale (la variabile sesso non dà contributo significativo)

- per ***x = età***, b= - 0.079

$$\frac{p|x+1}{1-p|x+1} = exp(-0.079)\frac{p|x}{1-p|x}$$

$$\frac{p|x+1}{1-p|x+1} = 0.92\frac{p|x}{1-p|x}$$

$$\frac{p|x+10}{1-p|x+10} = exp(-0.079*10)\frac{p|x}{1-p|x}$$

$$\frac{p|x+10}{1-p|x+10} = 0.45\frac{p|x}{1-p|x}$$

Per esempio: exp(10*-0.079)=0.45

Come si commenta?

la propensione verso il si (sono favorevole) per chi ha 10 anni in più è meno della metà (0.45) della propensione verso il si dichiarata da coloro che sono più giovani di 10 anni

- per ***x = livello di scolarizzazione***, b= 1.904 (livello alto)

$$\frac{p|x=alto}{1-p|x=alto} = exp(1.904)\frac{p|x=basso}{1-p|x=basso}$$

$$\frac{p|x=alto}{1-p|x=alto} = 6.96\frac{p|x=basso}{1-p|x=basso}$$

Come si commenta?

la propensione verso il si (sono favorevole) per chi ha un livello culturale più elevato è quasi 7 volte (6.96) più grande della propensione verso il si dei meno colti

NB il livello di riferimento è "basso" etichettato in spss con il valore più alto

Come giudicare il modello

- Bontà di adattamento $G^2 = -2(L_{intercetta} - L_{finale}) \sim \chi^2_{gdl}$
- Per il confronto tra modelli $G^2 = -2(L_{ridotto} - L_{finale}) \sim \chi^2_{gdl}$

- *L è la log-verosimiglianza*
- Sotto H_0 si considera il modello "più piccolo", che contiene meno parametri
- Il modello che contiene solo l'intercetta è il più parsimonioso
- Il modello detto *ridotto* contiene uno o più regressori in meno rispetto al modello detto *finale*
- *gdl = # di parametri uguagliati a zero per passare dal modello "più grande" sotto H_1 al modello "più piccolo" sotto H_0*

Come si decide?

Valori elevati della statistica G^2 (pvalue piccoli) conducono al rifiuto di H_0 e quindi sono a sostegno del modello "più grande"

G^2 ed altri modi per giudicare il modello

- Statistica G^2

Informazioni sull'adeguamento del modello

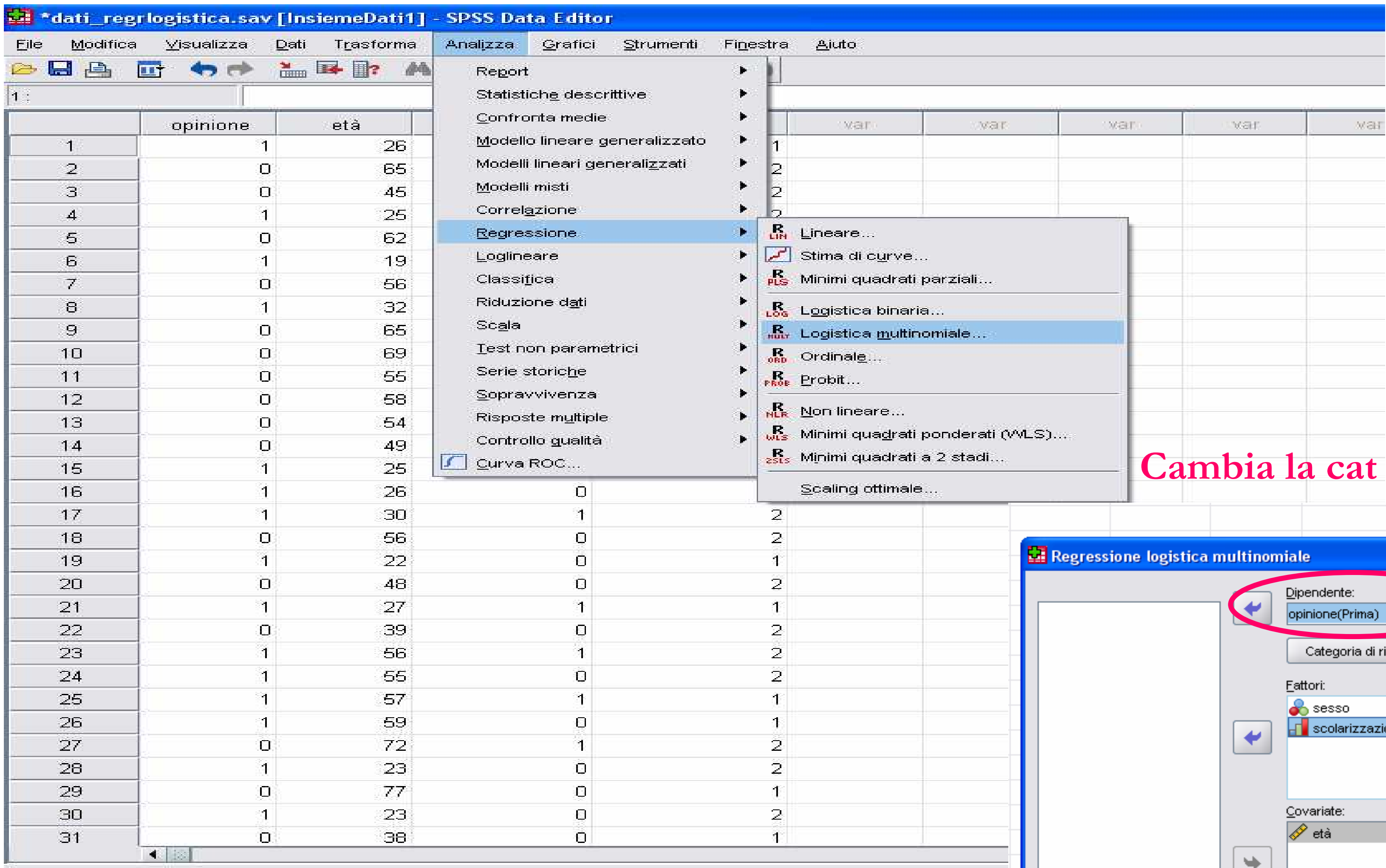
Modello	Criteri di adattamento del modello	Test del rapporto di verosimiglianza		
	-2 Log verosimiglianza	Chi-quadrato	df	Sig.
Solo intercetta	73,455			
Finale	45,114	28,341	3	,000

- L'indice R^2 (analogo a quello usato nella regressione lineare)
  - valori prossimi ad 1 indicano un buon adattamento del modello
- Probabilità di corretta attribuzione

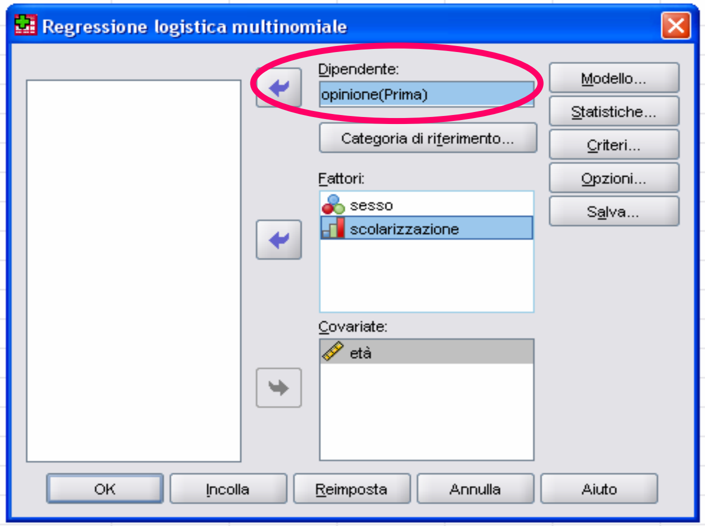
Classificazione

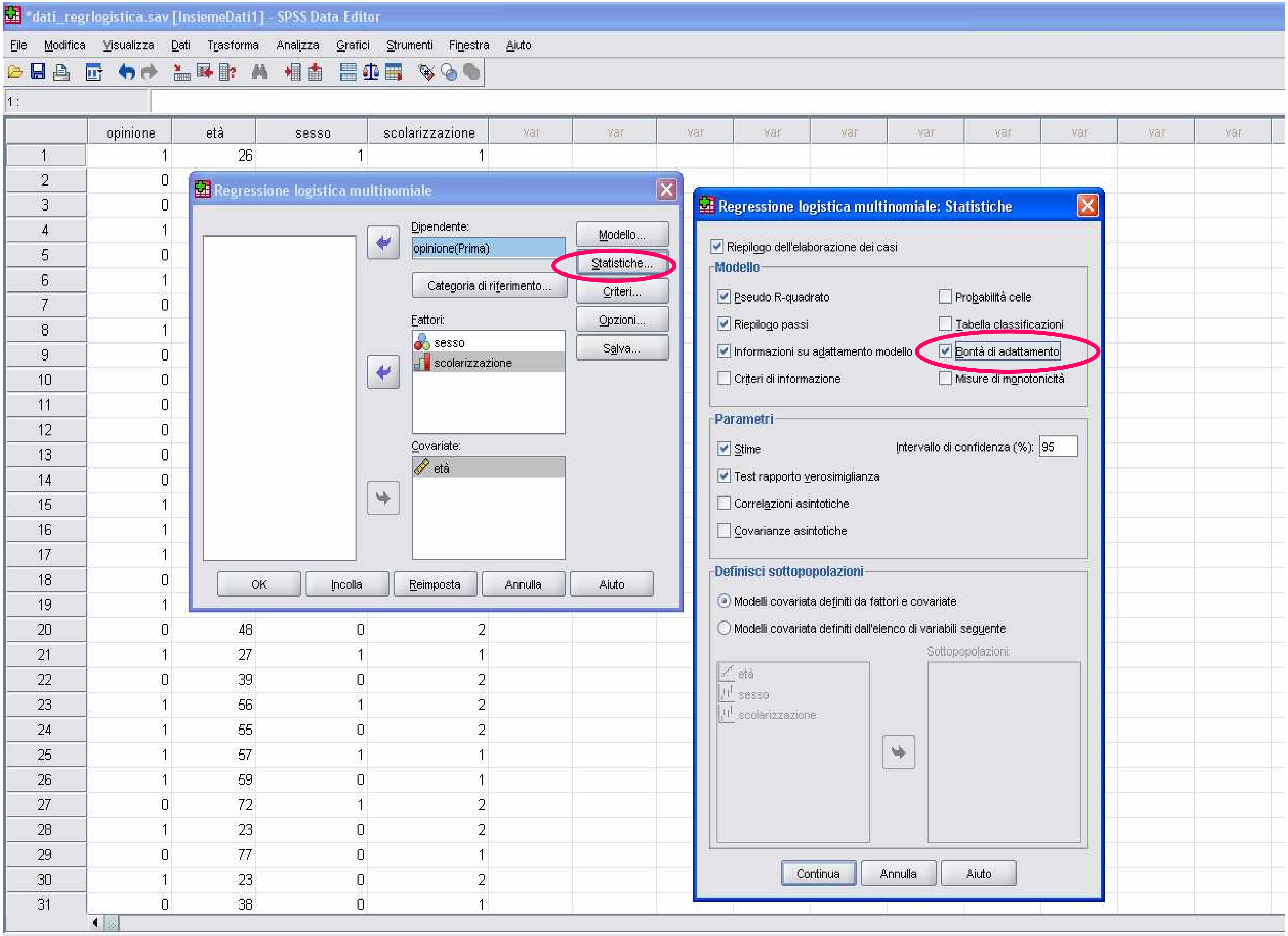
Osservato	Atteso		
	non favorevole	favorevole	Percentuale corretta
non favorevole	23	5	82,1%
favorevole	6	21	77,8%
Percentuale globale	52,7%	47,3%	80,0%

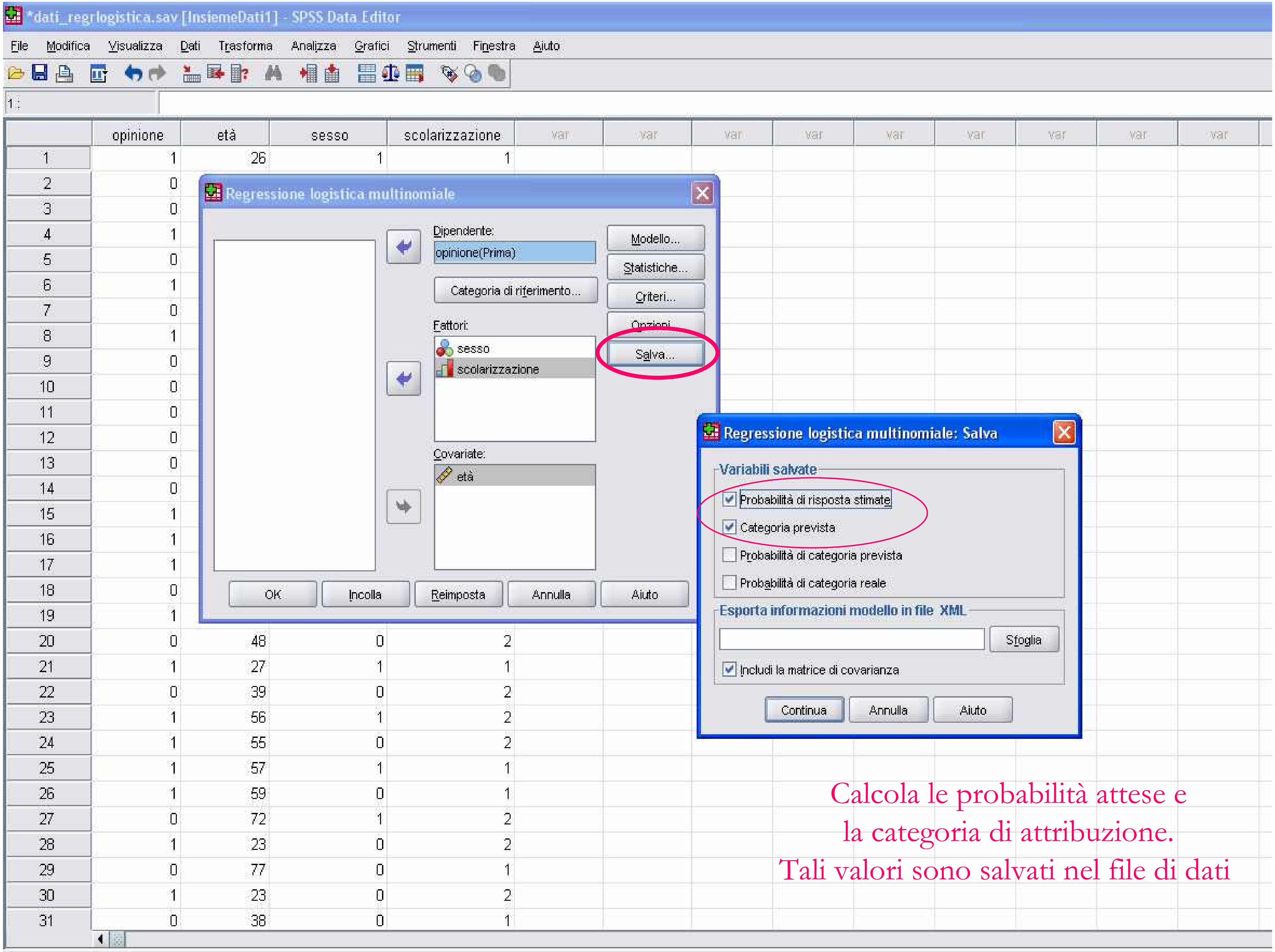
In spss (la procedura multinomiale)



Cambia la cat di riferimento




*dati_regrlogistica.sav [InsiemeDati1] - SPSS Data Editor
File Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
1 :
opinione età sesso scolarizzazione var var var var var var var var var var
1 1 26 1 1
2 0
3 0
4 1
5 0
6 1
7 0
8 1
9 0
10 0
11 0
12 0
13 0
14 0
15 1
16 1
17 1
18 0
19 1
20 0 48 0 2
21 1 27 1 1
22 0 39 0 2
23 1 56 1 2
24 1 55 0 2
25 1 57 1 1
26 1 59 0 1
27 0 72 1 2
28 1 23 0 2
29 0 77 0 1
30 1 23 0 2
31 0 38 0 1
Regressione logistica multinomiale
Dipendente:
opinione(Prima)
Modello...
Statistiche...
Categoria di riferimento...
Criteri...
Opzioni...
Salva...
Fattori:
sesso
scolarizzazione
Covariate:
età
OK
Incolla
Reimposta
Annulla
Aiuto
Regressione logistica multinomiale: Statistiche
Riepilogo dell'elaborazione dei casi
Modello
Pseudo R-quadrato
Probabilità celle
Riepilogo passi
Tabella classificazioni
Informazioni su adattamento modello
Bontà di adattamento
Criteri di informazione
Misure di monotonicità
Parametri
Stime
Intervallo di confidenza (%): 95
Test rapporto verosimiglianza
Correlazioni asintotiche
Covarianze asintotiche
Definisci sottopopolazioni
Modelli covariata definiti da fattori e covariate
Modelli covariata definiti dall'elenco di variabili seguente
Sottopopolazioni:
età
sesso
scolarizzazione
Continua
Annulla
Aiuto



Calcola le probabilità attese e
la categoria di attribuzione.
Tali valori sono salvati nel file di dati