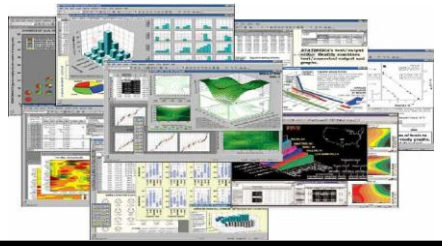
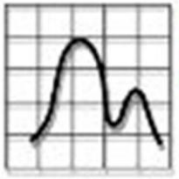
UNIVERSITÀ
DEGLI STUDI
DI PADOVA





Facoltà di Psicologia - Corso FSE – Ottobre 2009

Marco Vicentini
info@marcovicentini.it

Introduzione a Statistica 6

Sommario

- Statistica vs. SPSS
- Importare, costruire e manipolare un file
- Statistiche descrittive e grafici
- Analisi correlazionale
- Confronto tra medie
  - *t*-test
  - ANOVA
- Cenni di statistiche non parametrica

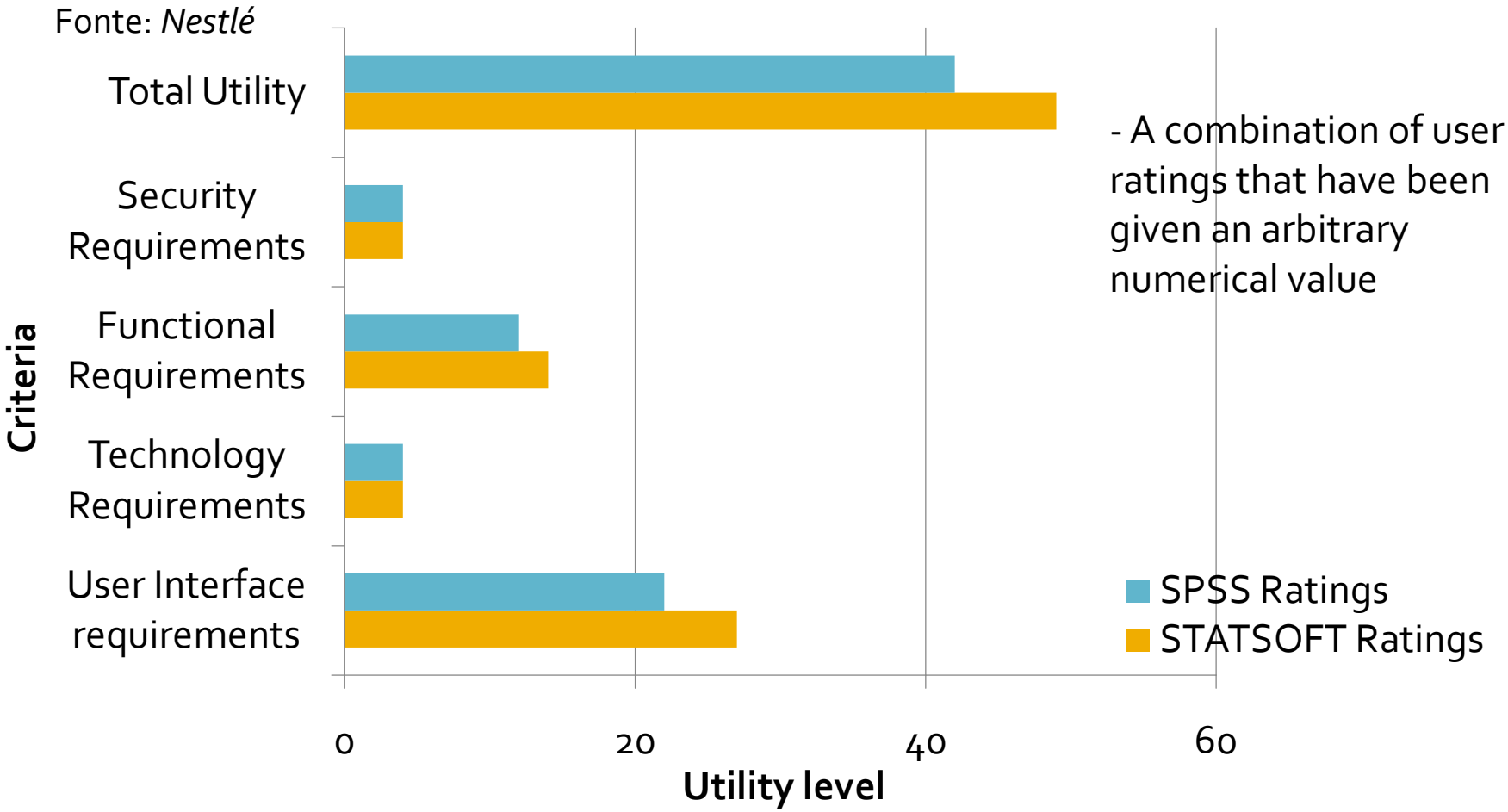
e tante esercitazioni …

Perché Statistica

- 1984 StatSoft nasce da un gruppo di professori universitari che necessitano di uno strumento per le analisi dei dati.
- 1985 Primo prodotto statistico per Lotus 123 e versione standalone STATS+
- 2009 rilasciata la versione 9 di Statistica, nelle versioni Enterprise, Web e Desktop. Sviluppi nelle capacità di data mining.

Perché Statistica

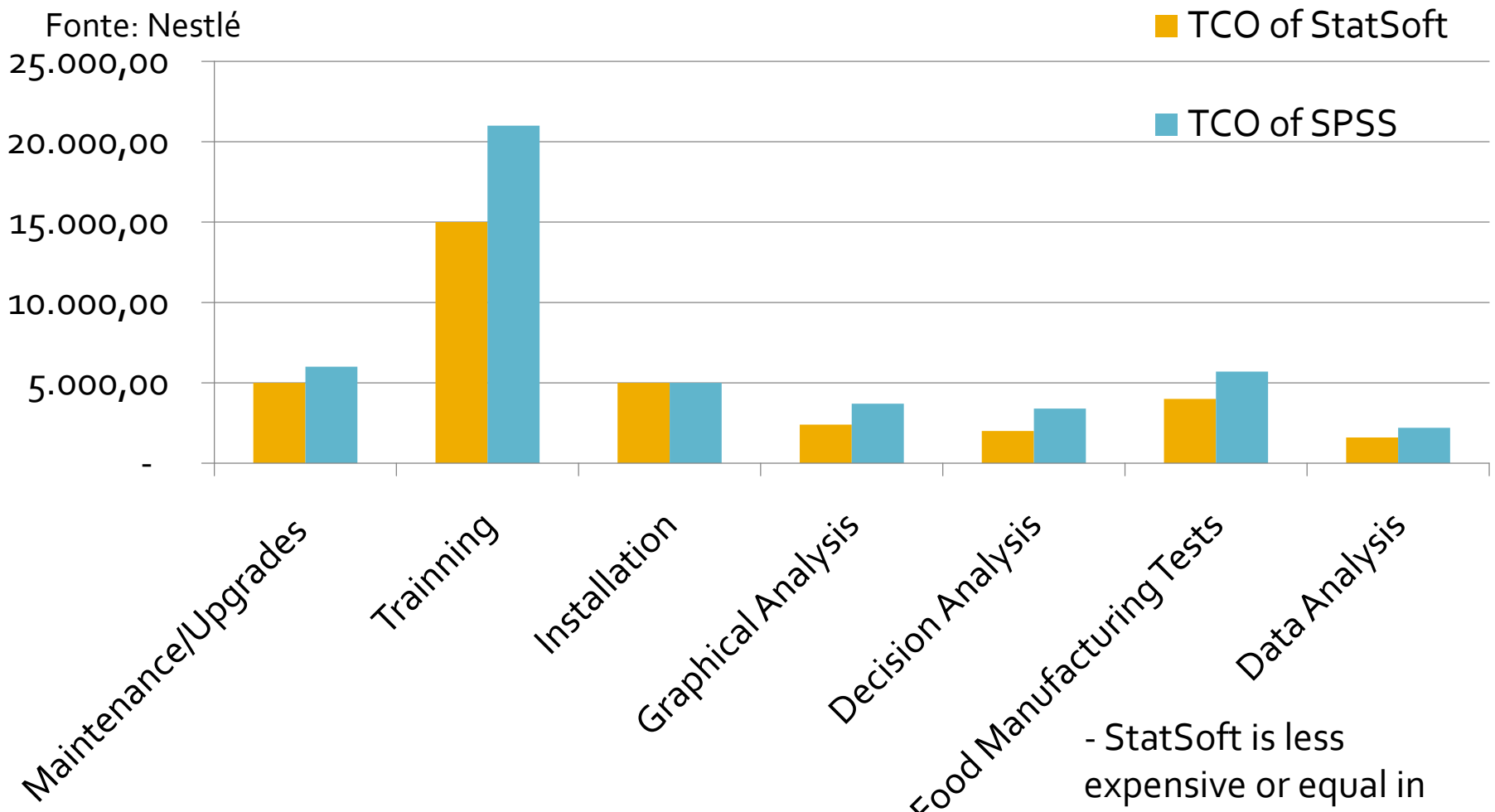
Fonte: *Nestlé*



- A combination of user ratings that have been given an arbitrary numerical value

Perché Statistica

Fonte: Nestlé



- StatSoft is less expensive or equal in price to SPSS in all areas

Statistica

- Programma generale per svolgere differenti analisi statistiche
- Organizzato a moduli
  - Statistiche descrittive …
  - Tecniche esplorative
  - Modelli lineari
  - …
  - Data Mining
  - Reti neurali
- Versioni desktop, enterprise, e web

Nozioni fondamentali

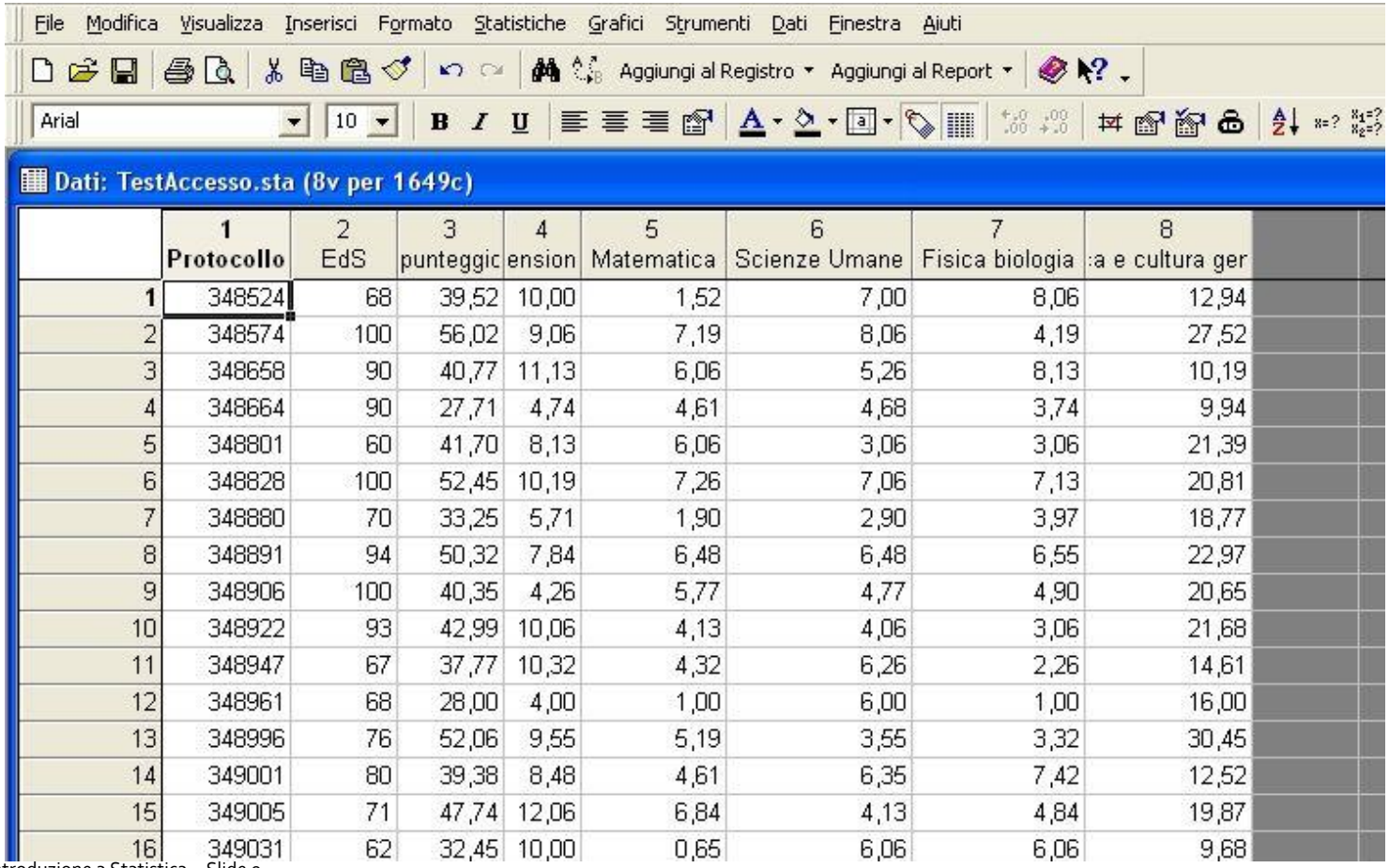
Nozioni fondamentali

SPREADSHEET:

- gli spreadsheet sono i fogli di lavoro di STATISTICA. Essi si basano sulla tecnologia delle tabelle multimediali e sono usati per gestire sia i dati di input (dati da elaborare) che l'output numerico/di testo (risultati di un'analisi, che possono a loro volta essere usati come dati di input per un'ulteriore analisi).
- La forma di base dello spreadsheet è una semplice tabella bidimensionale che può gestire un numero (virtualmente) illimitato di *casi* (righe) e *variabili* (colonne).

Nozioni fondamentali

- Un esempio di SPREADSHEET (file *.sta)



	1 Protocollo	2 EdS	3 punteggic	4 ension	5 Matematica	6 Scienze Umane	7 Fisica biologia	8 :a e cultura ger
1	348524	68	39,52	10,00	1,52	7,00	8,06	12,94
2	348574	100	56,02	9,06	7,19	8,06	4,19	27,52
3	348658	90	40,77	11,13	6,06	5,26	8,13	10,19
4	348664	90	27,71	4,74	4,61	4,68	3,74	9,94
5	348801	60	41,70	8,13	6,06	3,06	3,06	21,39
6	348828	100	52,45	10,19	7,26	7,06	7,13	20,81
7	348880	70	33,25	5,71	1,90	2,90	3,97	18,77
8	348891	94	50,32	7,84	6,48	6,48	6,55	22,97
9	348906	100	40,35	4,26	5,77	4,77	4,90	20,65
10	348922	93	42,99	10,06	4,13	4,06	3,06	21,68
11	348947	67	37,77	10,32	4,32	6,26	2,26	14,61
12	348961	68	28,00	4,00	1,00	6,00	1,00	16,00
13	348996	76	52,06	9,55	5,19	3,55	3,32	30,45
14	349001	80	39,38	8,48	4,61	6,35	7,42	12,52
15	349005	71	47,74	12,06	6,84	4,13	4,84	19,87
16	349031	62	32,45	10,00	0,65	6,06	6,06	9,68

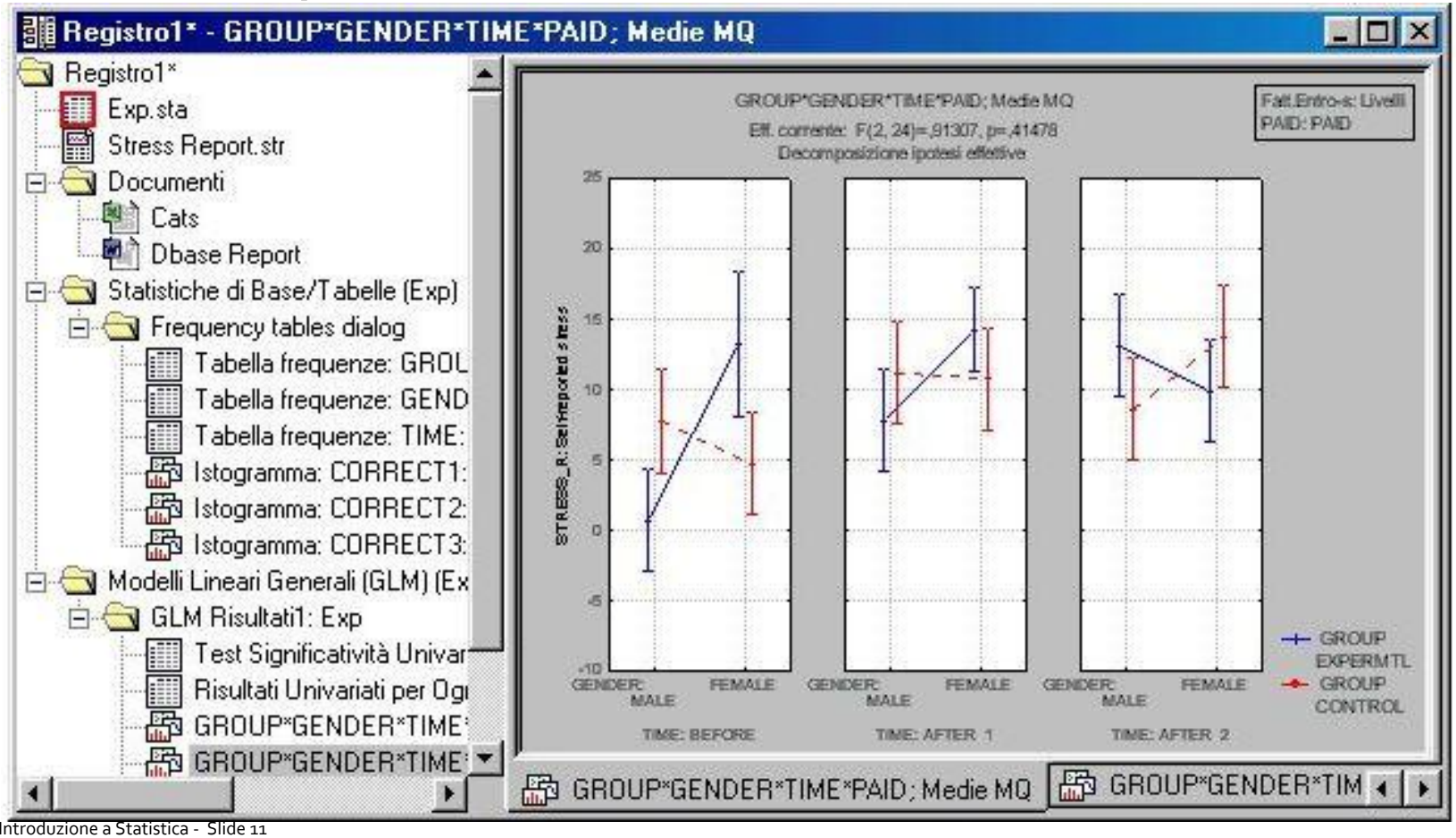
Nozioni fondamentali

REGISTRI

- i registri permettono di gestire l'output archiviandolo in forma di schede.

Nozioni fondamentali

- Un esempio di REGISTRO (file *.stw)



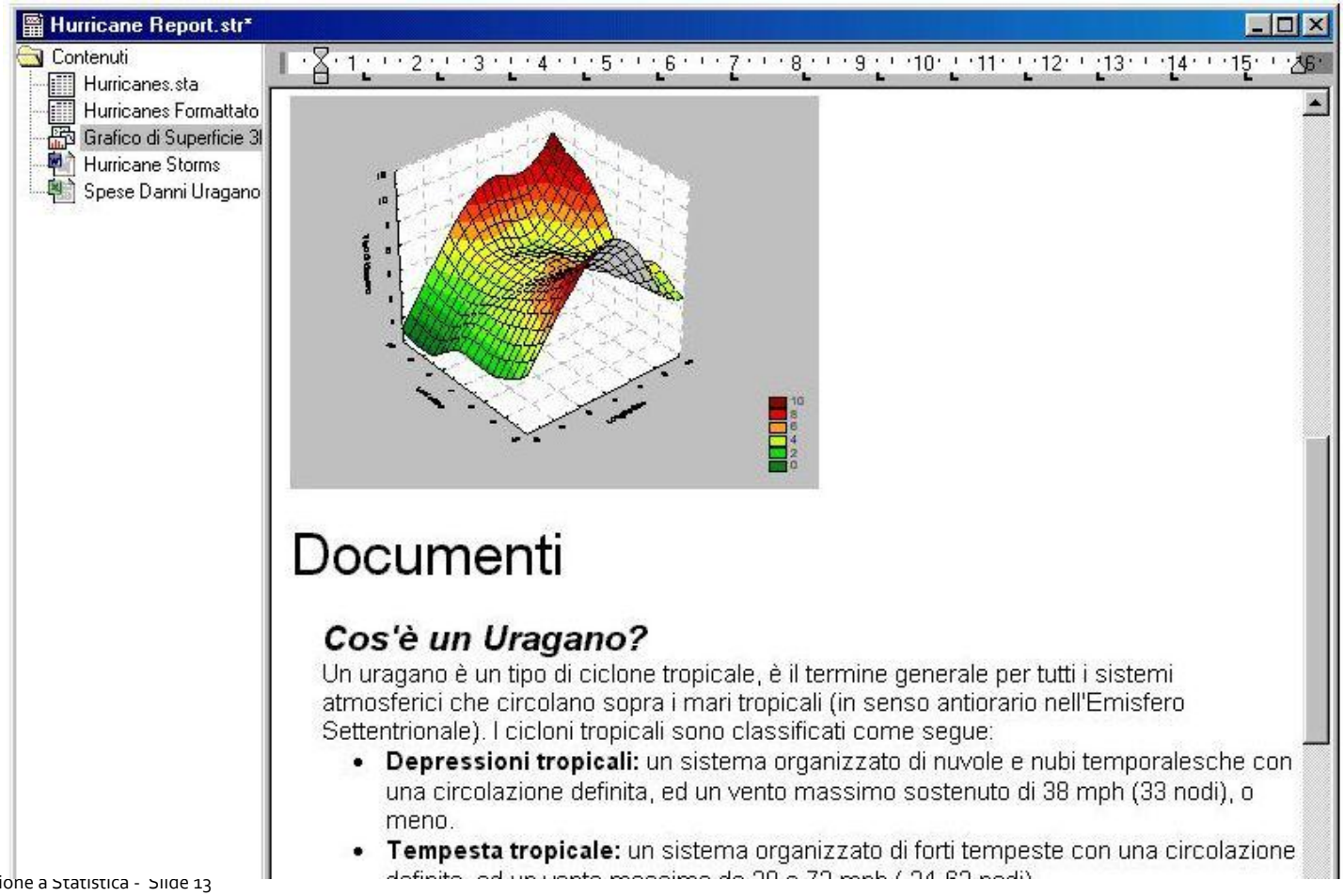
Nozioni fondamentali

REPORT

- i report permettono di gestire l'output di statistica visualizzando sequenzialmente gli oggetti (spreadsheet, grafici, etc).
- La gestione dei report consiste sostanzialmente in un documento di testo nel quale vengono inseriti i risultati di analisi ed altri oggetti.

Nozioni fondamentali

- Un esempio di REPORT (file *.str)



Importare, costruire e manipolare un file di dati

- STATISTICA può importare un file di dati in formato testo (.txt, .csv) o file di dati creati dalle più comuni applicazioni (Excel, SPSS …)

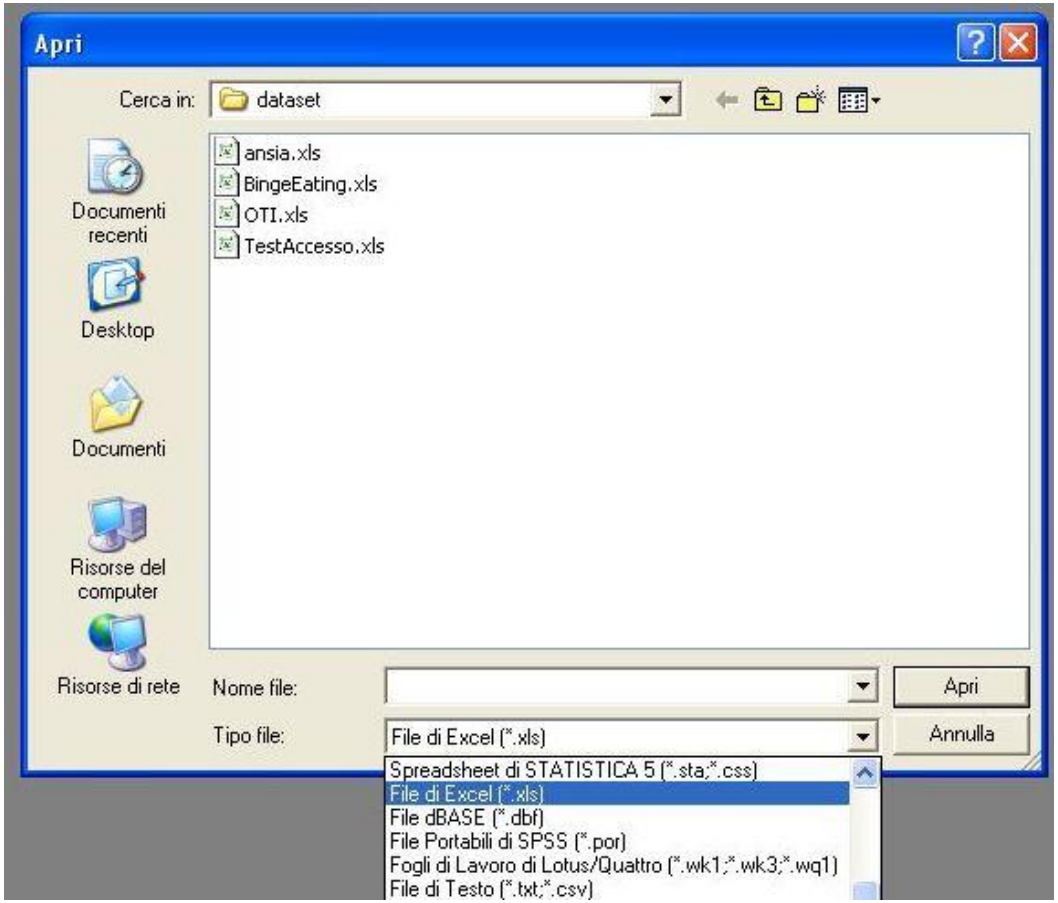
- File ➔ Apri
- Selezionare il file di interesse
  - Se file **xls**: Importare file di Excel
  - Se file **csv**: Importare come file di testo

- È possibile inoltre specificare una serie di utili opzioni per personalizzare il più possibile l'operazione di importazione
  - specificare che il nome delle variabili si trova nella prima riga del file da importare
  - specificare il separatore in uso nel file da importare

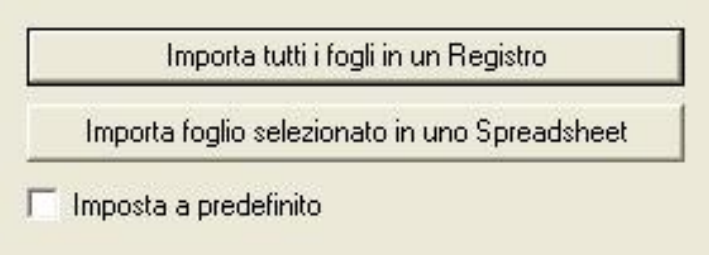
Nota: dopo aver importato il file è sempre buona norma controllare, anche rapidamente, la corretta riuscita dell'operazione.

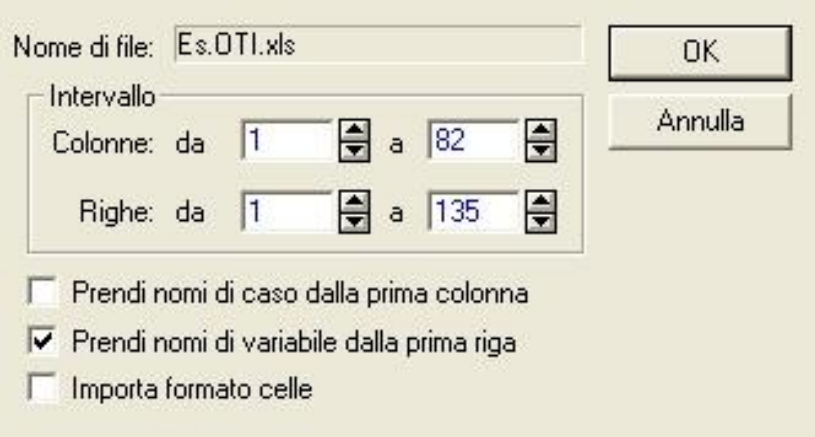
- Nel file "ansia.xls" sono contenuti i dati rilevati su un gruppo di 85 soggetti.
  - Le variabili misurate sono: il sesso, l'età, il reddito, il punteggio ottenuto in un test riguardante l'ansia e il punteggio ottenuto in un test riguardante la depressione.
- Importare il dataset "ansia.xls" e controllare la correttezza dell'operazione eseguita.
- Salvare il dataset nel formato associato a STATISTICA (.sta).

- Esempio di importazione di un file di dati (I)



Esempio di importazione di un file di dati (II)





Costruzione di un file di dati

Oltre ad importare un file di dati è possibile costruire direttamente in STATISTICA un file di dati.

- File ➔ Nuovo
- Si apre una finestra di dialogo che chiede se si vuole creare
  - uno Spreadsheet (foglio di calcolo)
  - un Report (modulo che permette di gestire l'output visualizzando sequenzialmente gli oggetti – spreadsheet, grafici ecc. –)
  - un Programma Macro (per registrare una macro in Visual Basic)
  - un Registro (strumento che permette di gestire l'output archiviandolo in forma di schede).

- Selezionare la scheda "Spreadsheet" per creare un nuovo file di dati con estensione ".sta", utile se si vogliono inserire o copiare i dati direttamente in STATISTICA

Manipolare un file di dati

- Menu **Modifica**

- Sotto questo menù si trovano diverse opzioni utili, alcune delle quali comuni alla maggior parte dei programmi:
  - Le prime tre opzioni permettono di annullare un comando precedentemente digitato o di ripristinarlo.
  - Le successive cinque permettono di tagliare e incollare il contenuto delle celle, oltre che di copiarlo con o senza le intestazioni di casi e variabili.
  - Le opzioni più sotto permettono di eliminare o spostare casi o variabili oppure di cancellare valori o formati.

- All'interno del menù Modifica, esiste una insieme di opzioni utili per modificare direttamente dei "blocchi di dati" precedentemente selezionati.

- Riempi/Standardizza blocco
  - Riempi con valori casuali: riempie il blocco selezionato con valori casuali compresi tra 0 e 1
  - Riempi/copia in basso: copia i valori della riga più in alto in tutte le righe sottostanti selezionate
  - Riempi/copia a destra: copia i valori della colonna più a sinistra in tutte le colonne selezionate alla sua destra
  - Standardizza Colonne (Righe): Standardizza i valori della colonna o riga selezionata trasformandoli in punti z (valori con media 0 e varianza 1).

Impostazioni di visualizzazione

- Menu **Visualizza**

- Questo menù permette di visualizzare alcuni attributi di casi e variabili, oppure alcune barre degli strumenti, le intestazioni a piè pagina ecc.
- Permette inoltre di modificare alcune caratteristiche delle linee della griglia.

Inserire oggetti, casi o variabili

- Menu **Inserisci**

- Questo menù permette di inserire nel foglio di lavoro nuove variabili o casi oppure di spostarli all'interno del foglio di lavoro.
- Permette inoltre di inserire oggetti quali pagine Word, grafici Excel, immagini, diapositive PowerPoint

Modificare la formattazione

- Menu **Formato**

- Questo menù permette di modificare alcune caratteristiche del formato delle celle
  - (formato del numero, allineamento, font, bordi) e di casi e variabili (larghezza di righe e colonne).

Lavorare sui dati

- Menu **Dati**
- Questo menù è uno dei più importanti in quanto permette di lavorare in vario modo sui dati.
- Dati → Spreadsheet di input
  - Permette di effettuare analisi su uno spreadsheet di output rendendolo spreadsheet di input.

Manipolare un file di dati

- Dati → Trasponi
  - Permette di trasporre i valori che sono in riga in colonna e viceversa (il blocco selezionato deve avere ugual numero di casi e variabili).
  - Se si seleziona "Trasponi file" il comando rovescerà tutta la struttura del file mettendo le variabili al posto dei casi e viceversa.

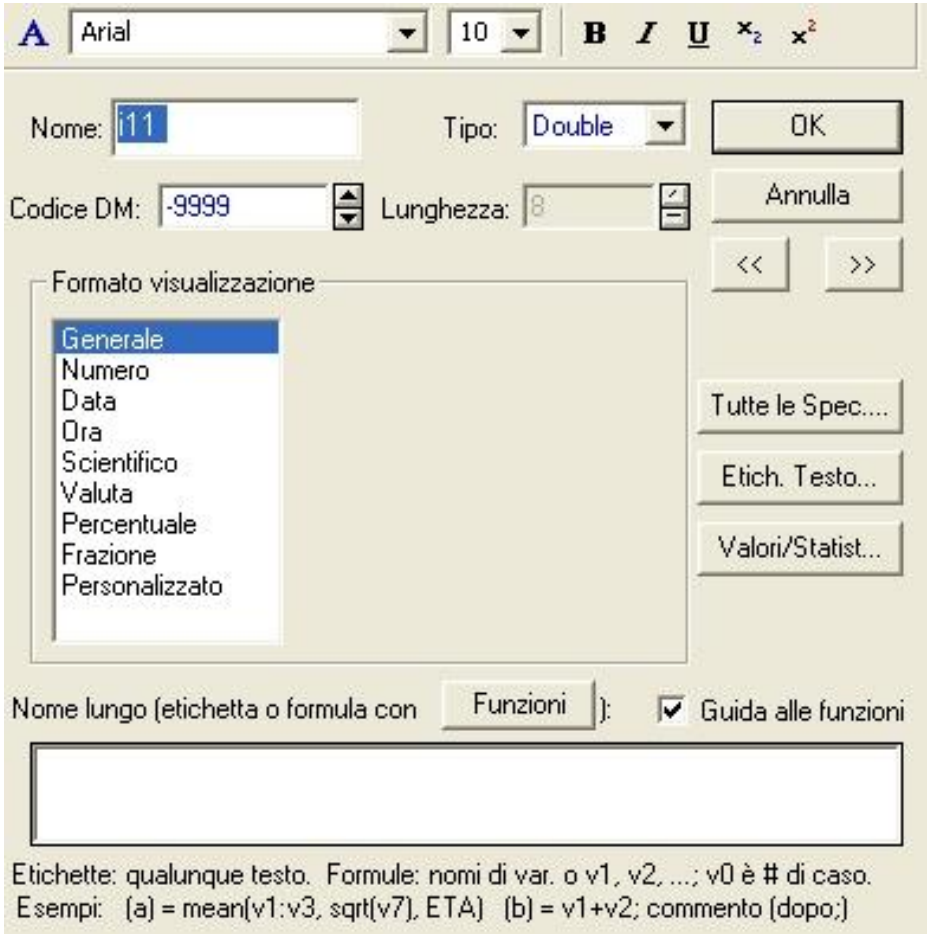
- Dati ➔ Unisci
  - Permette di mettere insieme due file per colonna o per riga, a partire da quello già aperto:
  - Se si seleziona “Variabili”, si aprirà una finestra in cui viene richiesto il nome del file da cui prendere le variabili da unire.
  - Se si seleziona “Casi”, si aprirà una finestra in cui chiede il nome del file da cui prendere i casi da unire. È possibile effettuare questa operazione solo se il numero di variabili nei due file è uguale.

- Dati → Ordina
  - Ordina le righe in ordine crescente o decrescente in base alla/e variabile/i selezionata/e.

- Dati → Campionamento sottoinsieme / Casuale
  - Crea un nuovo file, che può essere considerato un sottoinsieme del file già aperto:
  - Cliccare su "Variabili" e selezionare le colonne da inserire nel nuovo file.
  - Cliccare su "Casi" se si vogliono inserire solo una parte delle righe (soggetti) in base a specifiche condizioni di selezione.
  - È possibile anche creare un sottoinsieme con dati campionati casualmente dal file originale.

- Dati → Verifica Dati
  - Permette di considerare una serie di condizioni che devono essere rispettate dai dati e di marcare i dati non validi.
- Dati → Specifiche variabile
  - Permette di modificare gli attributi di una variabile come: il nome, il tipo, il codice associato ai dati mancanti, il formato, eventuali etichette di testo da associare ai valori delle variabile …

- Dati → Tutte le specifiche variabile
  - Permette di visualizzare e modificare alcune specifiche (“Nome”, “Tipo”, “Codice DM”, “Lunghezza”, “Nome lungo”) di tutte le variabili presenti nel dataset.


Arial
10
Nome: i11
Tipo: Double
OK
Codice DM: -9999
Lunghezza: 8
Annulla
<<
>>
Formato visualizzazione
Generale
Numero
Data
Ora
Scientifico
Valuta
Percentuale
Frazione
Personalizzato
Tutte le Spec....
Etich. Testo...
Valori/Statist...
Nome lungo (etichetta o formula con Funzioni):
Guida alle funzioni
Etichette: qualunque testo. Formule: nomi di var. o v1, v2, ...; v0 è # di caso.
Esempi: (a) = mean(v1:v3, sqrt(v7), ETA) (b) = v1+v2; commento (dopo;)

- Dati → Editor etichette di testo
  - Serve per creare etichette di testo che accompagnano i valori numerici di una variabile selezionata (ad esempio maschio = 1, femmina = 2 ecc). Le etichette di testo si possono visualizzare o meno selezionando:
  - Visualizza → Mostra etichette di testo

- Dati → Variabili (Casi)
  - Permette di aggiungere, spostare, copiare, eliminare Variabili (o Casi)
- Dati → Formule di trasformazione in lotti
  - Permette di effettuare operazioni tra variabili. Le formule devono essere scritte con sintassi del tipo: v3=v1+v2 (dove ad esempio v3 indica la terza variabile del dataset. Alternativamente può essere usato direttamente il nome della variabile.)

- Dati ➔ Ricalcola formule di Spreadsheet …
  - Permette di ricalcolare una variabile (o un gruppo di variabili) sulla base delle formule immesse nella casella "Nome lungo" della finestra di dialogo "Variabile" di ogni variabile.
  - Se si effettuano delle operazioni su una variabile che dipendono dai valori di altre variabili (come ad es., v3=v1+v2) che si prevede possano cambiare, è conveniente scrivere la formula nella casella "Nome lungo" piuttosto che nello spazio "Formule di trasformazione in lotti" e barrare "Ricalcola automaticamente quando i dati cambiano". In questo modo infatti il ricalcolo successivo ai cambiamenti di v1 o v2 sarà effettuato. Si deve infatti ricordare che altrimenti STATISTICA non aggiorna automaticamente i valori di v3, come avviene per altri programmi (ad es. Excel).
  - N.B. Nella casella "Nome lungo" le formule devono essere scritte con sintassi del tipo: =v1+v2 (se si sta scrivendo nella v3).

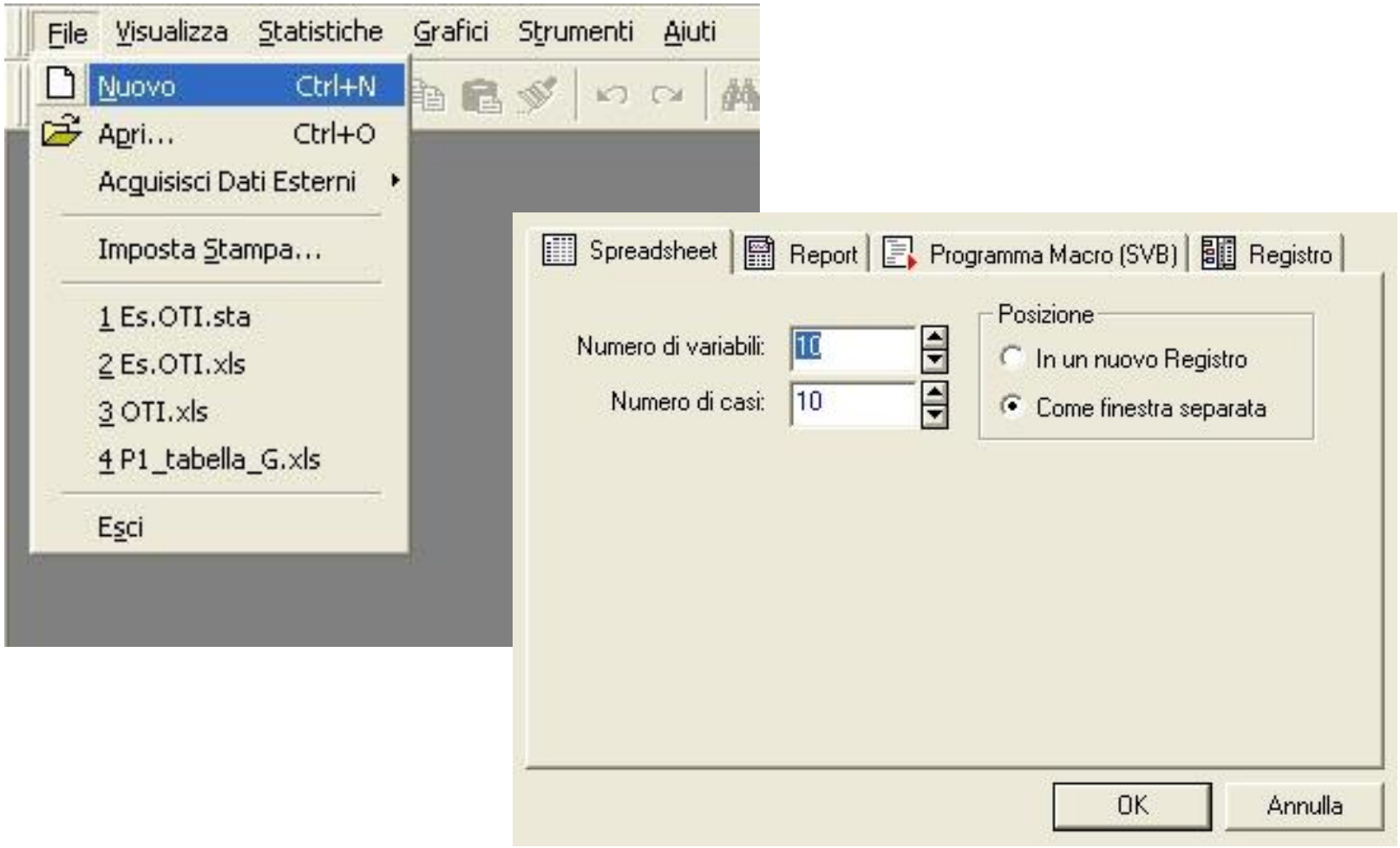
- Dati → Ricodifica
  - Permette di assegnare un nuovo valore ai dati di una variabile che rispettano le condizioni che qui si indicano.
  - Di particolare utilità se si vogliono effettuare dei sottogruppi (ex. Maschi con età inferiore a 24 anni = 1, maschi con età superiore o uguale a 24 anni = 2).
- Dati → Sostituisci dati mancanti
  - Permette di sostituire i Dati Mancanti con il valore medio della colona selezionata.

- Supponiamo di aver rilevato le seguenti variabili su un campione di 10 turisti a Verona:

id	Età	Nazionalità	Macchina fotografica
1	18	Italiana	Canon
2	25	Giapponese	Nikon
3	36	Italiana	Sony
4	29	Tedesca	Nikon
5	24	Giapponese	Nikon
6	45	Tedesca	Canon
7	67	Giapponese	Nikon
8	23	Italiana	Sony
9	51	Tedesca	Canon
10	38	Giapponese	Sony

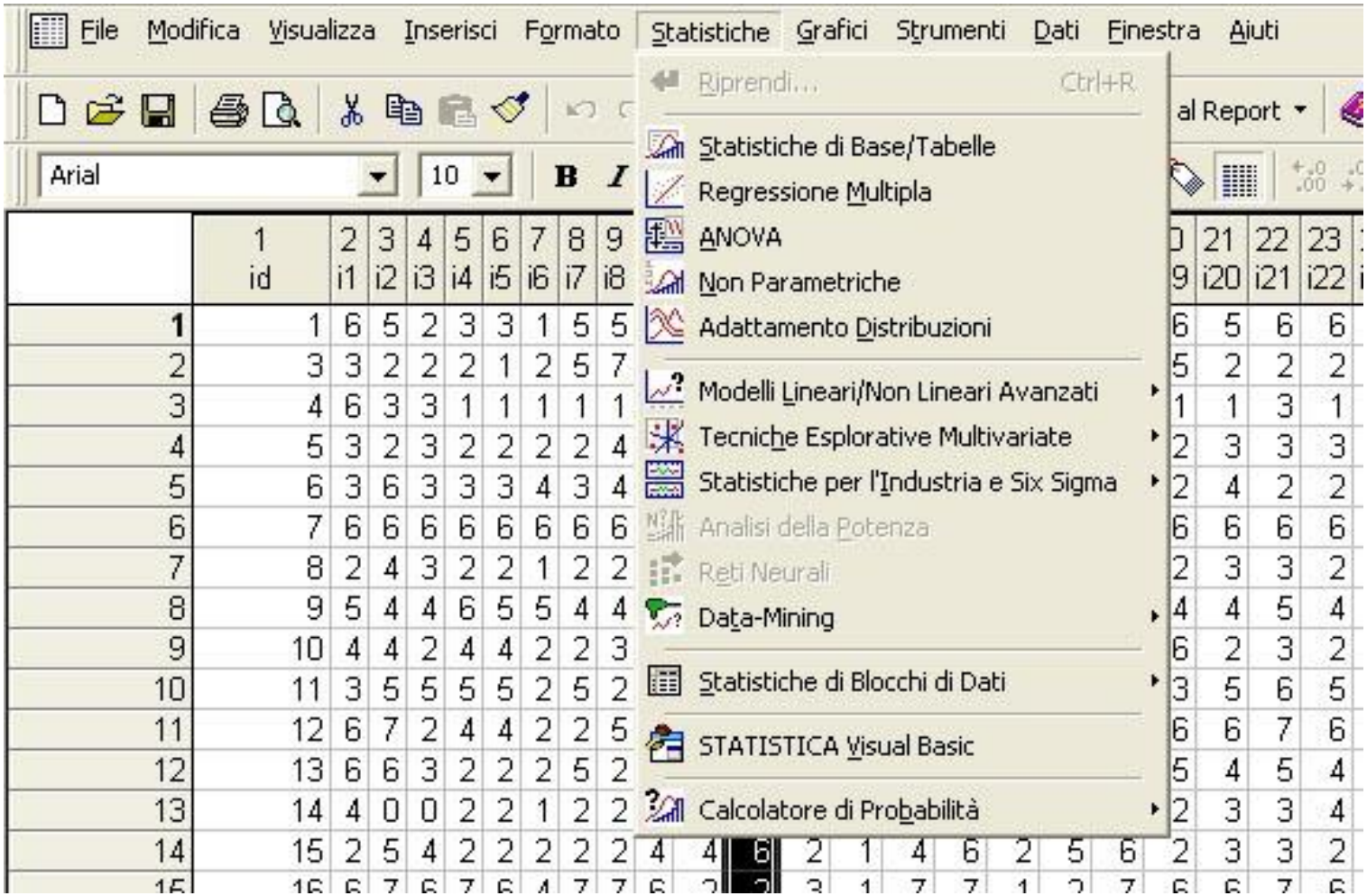
Esercitazione

- Costruire il relativo dataset in STATISTICA e salvarne il contenuto.
- Attraverso le funzioni di modifica dei dati:
  - Creare una nuova variabile in cui venga suddivisa l'età nelle seguenti categorie: "al di sotto dei 25 anni" e "da 25 anni in su"
  - Selezionare e salvare un nuovo dataset contenente solo i turisti giapponesi.
  - Sbizzarrirsi, a piacere, nell'utilizzare le funzioni di manipolazione di un dataset appena viste


File
Visualizza
Statistiche
Grafici
Strumenti
Aiuti
Nuovo Ctrl+N
Apri... Ctrl+O
Acquisisci Dati Esterni
Imposta Stampa...
1 Es.OTI.sta
2 Es.OTI.xls
3 OTI.xls
4 P1_tabella_G.xls
Esci
Spreadsheet
Report
Programma Macro (SVB)
Registro
Numero di variabili:
10
Numero di casi:
10
Posizione
In un nuovo Registro
Come finestra separata
OK
Annulla

Grafici e statistiche descrittive

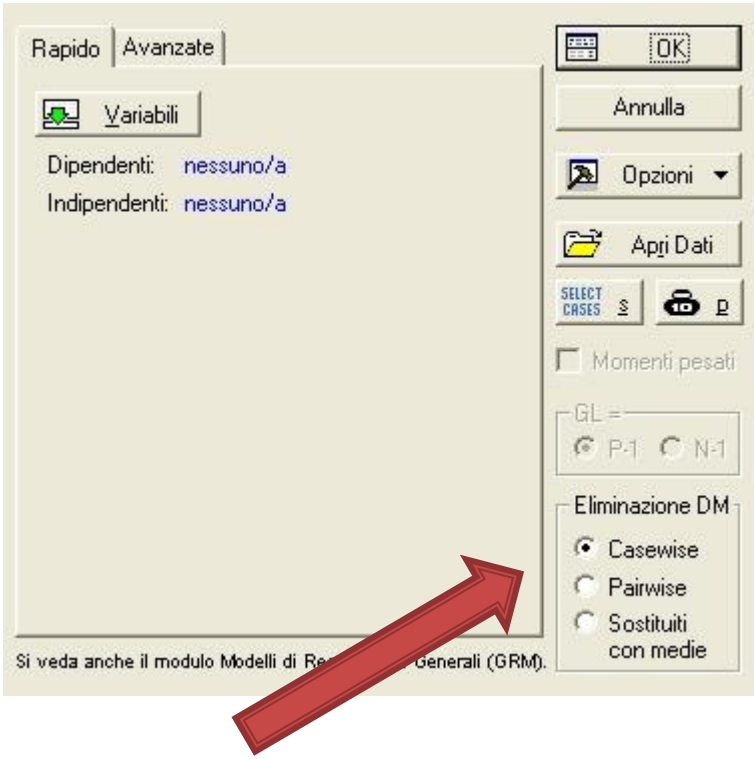
- Menu **Statistiche**
- Questo menù permette di effettuare un vasta gamma di tipologie di analisi statistiche.
- Da notare è il modo in cui viene gestito l'output: di default i risultati delle analisi in corso vengono presentati nel Registro, che può contenere anche analisi provenienti da diversi Spreadsheet.


File Modifica Visualizza Inserisci Formato Statistiche Grafici Strumenti Dati Finestra Aiuti
Riprendi... Ctrl+R
Statistiche di Base/Tabelle
Regressione Multipla
ANOVA
Non Parametriche
Adattamento Distribuzioni
Modelli Lineari/Non Lineari Avanzati
Tecniche Esplorative Multivariate
Statistiche per l'Industria e Six Sigma
Analisi della Potenza
Reti Neurali
Data-Mining
Statistiche di Blocchi di Dati
STATISTICA Visual Basic
Calcolatore di Probabilità
Arial
10
al Report
1 id
2 i1
3 i2
4 i3
5 i4
6 i5
7 i6
8 i7
9 i8
21 i20
22 i21
23 i22

- Prima di procedere con i Grafici e le Statistiche descrittive, è bene sottolineare una serie di opzioni utile per la maggior parte delle tecniche di analisi statistica presenti nel menù Statistiche.

- Sostituzione dati mancanti
  - STATISTICA non considera nelle analisi i soggetti in cui ci sono dati mancanti. È perciò possibile sostituire il valore assente con il valore medio della variabile selezionata. Oppure è possibile gestire la presenza di dati mancanti in uno dei seguenti modi:
  - **Pairwise**: si escludono dai calcoli i casi in cui, per le variabili selezionate,sono presenti dati mancanti (ad ex., se si effettua un'analisi su 3 variabili e manca un dato nella prima, il caso viene escluso solo per la prima variabile)
  - **Casewise**: si escludono dai calcoli i casi in cui sono presenti dati mancanti in almeno una delle variabili selezionate (ad ex., se si effettua un'analisi su 3 variabili e manca un dato nella prima, il caso vieneescluso da tutta l'analisi).

- Selezione dei casi
  - Per ciascuna tipologia di analisi, è possibile includere/escludere soltanto i casi che soddisfano determinate condizioni.
  - Cliccare il tasto "*Select cases*". Si apre una finestra: abilitare le condizioni di selezione spuntando "Abilita condizioni di selezione".
  - Per includere casi, cliccare su "Specifici, selezionati" e scrivere all'interno del campo "Includi casi" - "Tramite espressione" oppure "o numeri di caso".
  - Per escludere casi, scrivere all'interno del campi "Escludi casi" -"Tramite espressione" oppure "o numeri di caso".



- Alcuni esempi di espressioni per includere/escludere casi da un'analisi
  - se di 20 soggetti si vogliono escludere dall'analisi tutti i soggetti aventi meno di 18 anni, digitare su "escludi se": anni < 18, oppure su "includi se": anni > 17 dove anni è la variabile che contiene l'età dei soggetti.
  - se di 3 gruppi si vogliono escludere tutti i soggetti appartenenti ad un gruppo, digitare su "escludi se": v1 = 2, dove v1 è la variabile gruppo e 2 è il codice assegnato al gruppo da escludere (ovviamente può essere qualunque altro codice).
  - Varie combinazioni, a seconda delle esigenze, si possono ottenere con gli operatori logici "AND", "OR", NOT, <, >, =, <>.

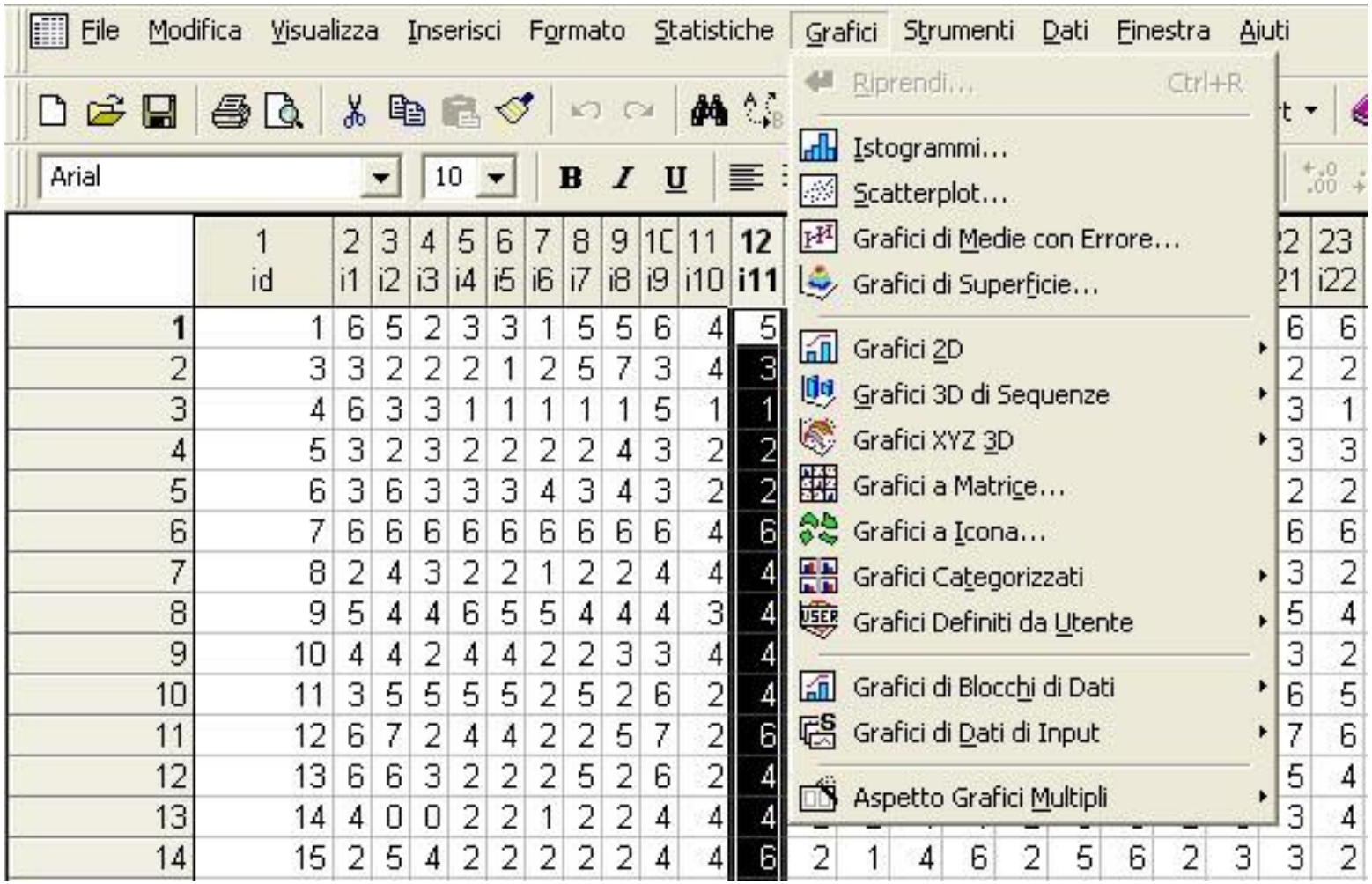
Statistiche descrittive e grafici

- Prima di procedere ad adattare dei modelli statistici sui dati a propria disposizione **è assolutamente indispensabile** **realizzare dei grafici** e delle **statistiche descrittive**. Ciò serve per:

  - avere una prima idea dei dati oggetto di studio;
  - fornire una prima descrizione dei dati ("utilizzare le statistiche descrittive è un po' come scattare delle fotografie ai dati");
  - controllare che le operazioni di costruzione del dataset siano state eseguite in modo corretto;
  - controllare la distribuzione delle variabili e valutare la presenza di possibili valori anomali.

Realizzazione di grafici

- Menu **Grafici**
- Permette di realizzare svariati tipi di grafici.
- Statistiche ➔ Statistiche di base/Tabelle
  - Permette, tra l'altro, di utilizzare una varietà di statistiche descrittive.
  - STATISTICA permette di utilizzare molte tecniche di analisi descrittive e soprattutto una grande varietà di grafici (*tra le altre cose, molto belli :D*).
  - Nel seguito vedremo solo alcune (le più utilizzate) di tecniche descrittive.



	1 id	2 i1	3 i2	4 i3	5 i4	6 i5	7 i6	8 i7	9 i8	10 i9	11 i10	12 i11
1	1	6	5	2	3	3	1	5	5	6	4	5
2	3	3	2	2	2	1	2	5	7	3	4	3
3	4	6	3	3	1	1	1	1	1	5	1	1
4	5	3	2	3	2	2	2	2	4	3	2	2
5	6	3	6	3	3	3	4	3	4	3	2	2
6	7	6	6	6	6	6	6	6	6	6	4	6
7	8	2	4	3	2	2	1	2	2	4	4	4
8	9	5	4	4	6	5	5	4	4	4	3	4
9	10	4	4	2	4	4	2	2	3	3	4	4
10	11	3	5	5	5	5	2	5	2	6	2	4
11	12	6	7	2	4	4	2	2	5	7	2	6
12	13	6	6	3	2	2	2	5	2	6	2	4
13	14	4	0	0	2	2	1	2	2	4	4	4
14	15	2	5	4	2	2	2	2	2	4	4	6

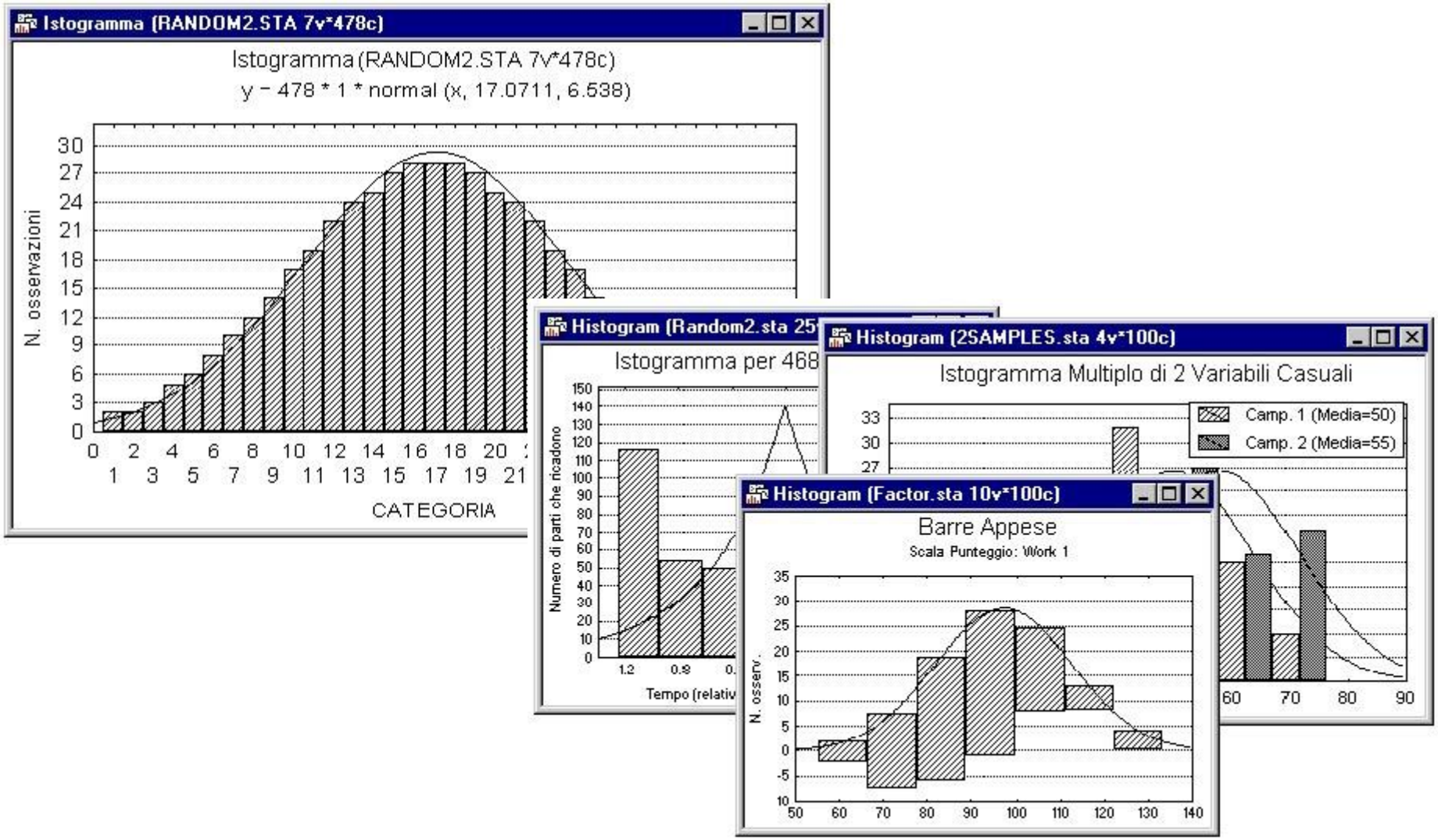
21	22	23
6	6	
2	2	
3	1	
3	3	
2	2	
6	6	
3	2	
5	4	
3	2	
6	5	
7	6	
5	4	
3	4	
2	1	4	6	2	5	6	2	3	3	2

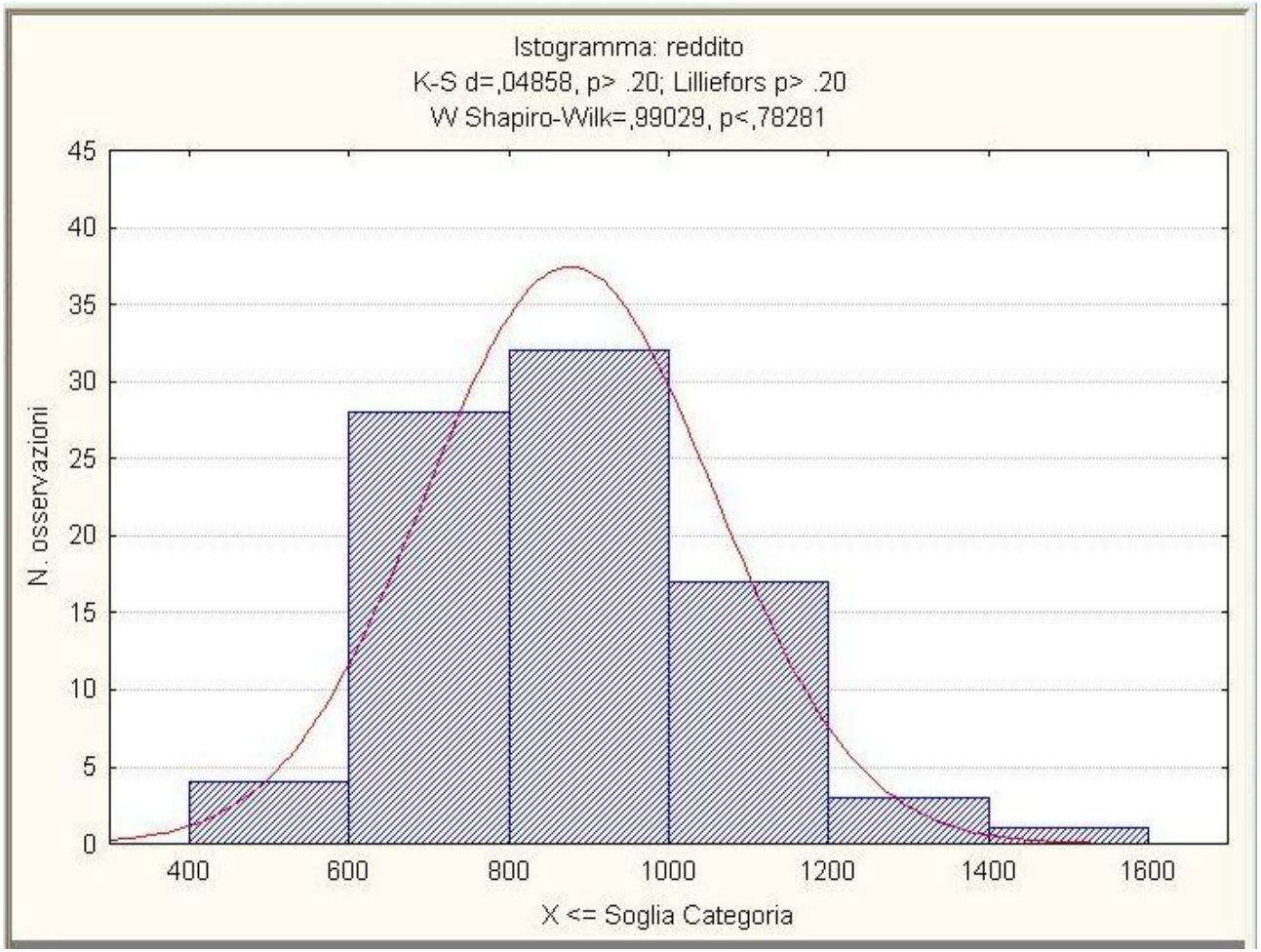
Tipo di Variabile	Tipo di Grafico	In STATISTICA
categoriale **nominale** (ad esempio il genere: M vs F)	istogramma a barre	Grafici ➔ Istogrammi Selezionare nel menù "intervalli" l'opzione "codici" e inserire i codici delle modalità da rappresentare.
categoriale **ordinale** (ad esempio il reddito: basso / medio /alto)	istogramma a barre	Grafici ➔ Istogrammi
quantitativa (ad esempio il tempo di reazione)	•istogramma • boxplot	Grafici ➔ Grafici2D ➔ Boxplot

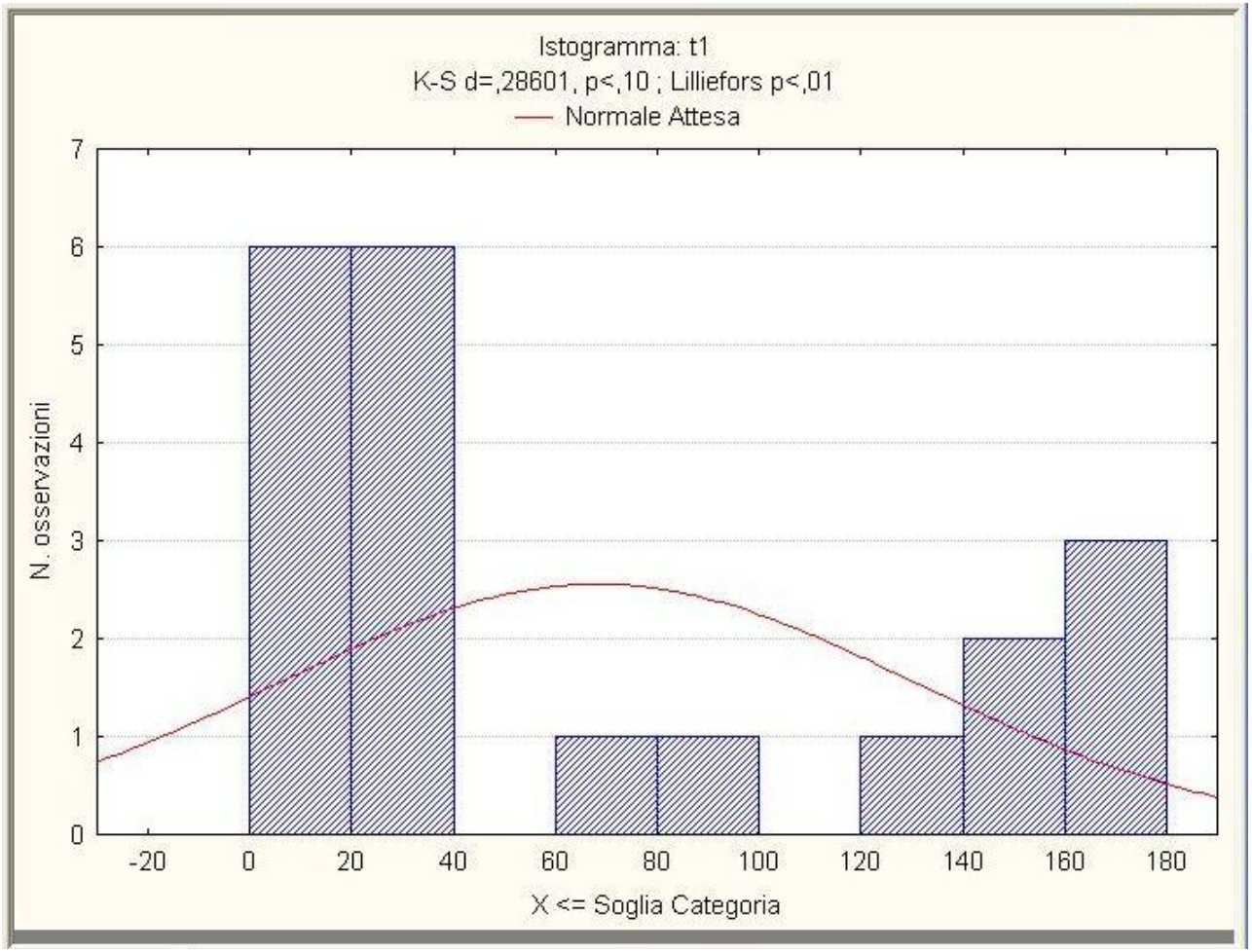
Nota sugli istogrammi per variabili quantitative

- Attraverso le "opzioni avanzate" è possibile scegliere un test per valutare la normalità della distribuzione osservata. I test possibili sono:
  - test di Kolmogorov-Smirnof
  - test di Lilliefors
  - test W di Shapiro-Wilk

  - I risultati del test selezionato vengono riportati assieme all'istogramma: se $p<.05$ la distribuzione osservata si scosta significativamente da quella normale.


Istogramma (RANDOM2.STA 7v*478c)
Istogramma (RANDOM2.STA 7v*478c)
y = 478 * 1 * normal (x, 17.0711, 6.538)
N. osservazioni
CATEGORIA
Histogram (Random2.sta 25
Istogramma per 468
Numero di parti che ricadono
Tempo (relativ
Histogram (2SAMPLES.sta 4v*100c)
Istogramma Multiplo di 2 Variabili Casuali
Camp. 1 (Media=50)
Camp. 2 (Media=55)
Histogram (Factor.sta 10v*100c)
Barre Appese
Scala Punteggio: Work 1
N. osserv.


Istogramma: reddito
K-S d=,04858, p> .20; Lilliefors p> .20
W Shapiro-Wilk=,99029, p<,78281
N. osservazioni
0
5
10
15
20
25
30
35
40
45
400
600
800
1000
1200
1400
1600
X <= Soglia Categoria


Istogramma: t1
K-S d=,28601, p<,10 ; Lilliefors p<,01
Normale Attesa
N. osservazioni
0
1
2
3
4
5
6
7
-20
0
20
40
60
80
100
120
140
160
180
X <= Soglia Categoria

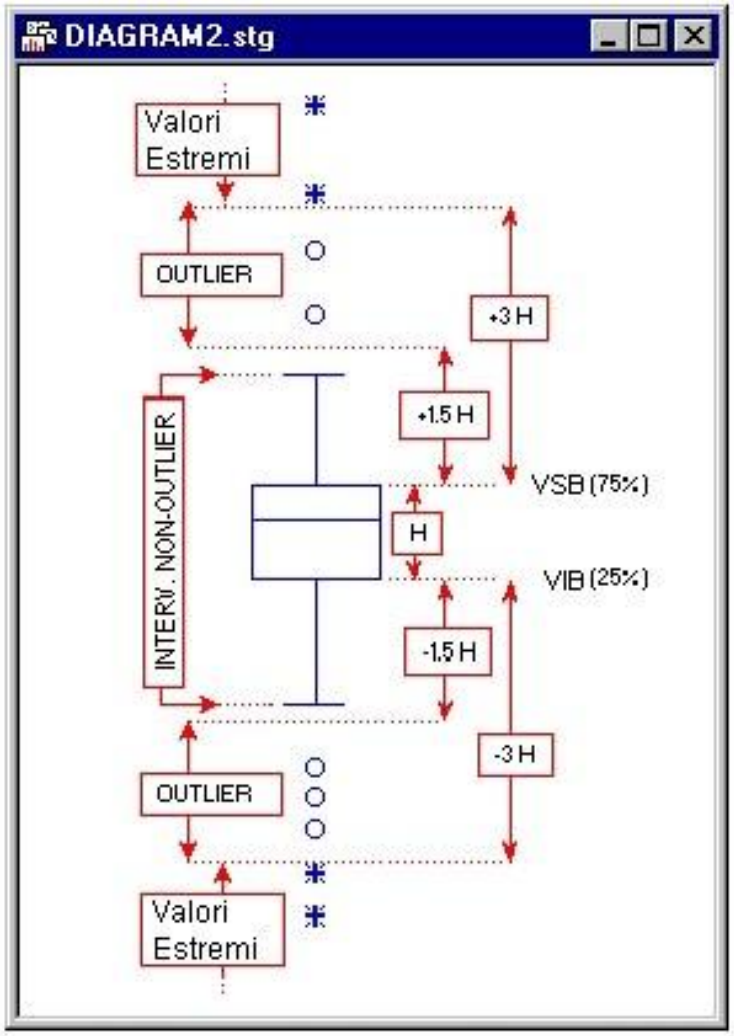
Come si legge un boxplot

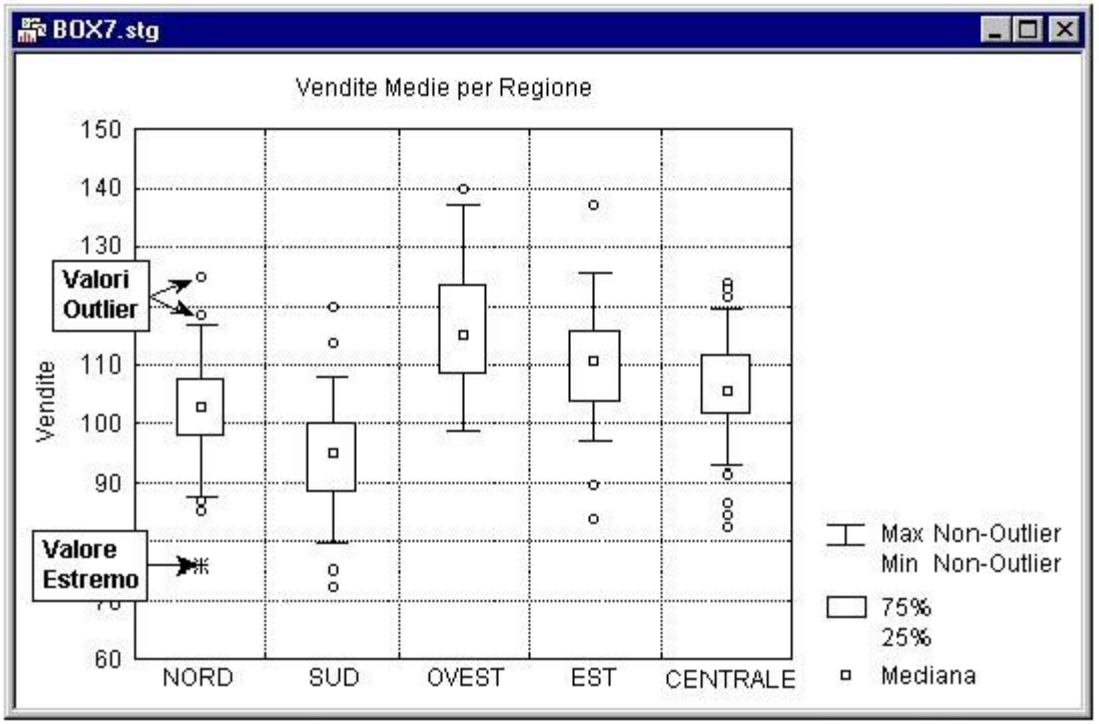
- Nei *Box Plot* (usato per la prima volta da Tukey, 1970), gli intervalli di variazione o caratteristiche distributive di valori di una o più variabili selezionate sono tracciate separatamente per gruppi di casi definiti in base ai valori di una variabile categoriale (di gruppo).
- La tendenza centrale, e le statistiche intervallo di variazione o di variabilità sono calcolate per ogni gruppo di casi, ed i valori selezionati sono presentati nello stile di box plot selezionato.

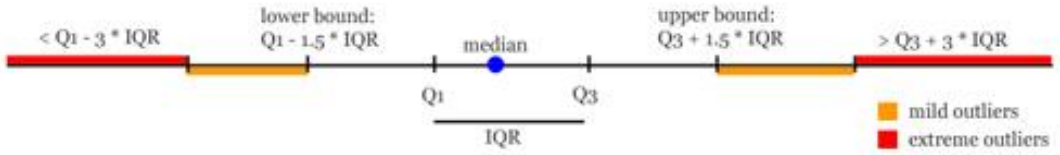
- Il boxplot contiene
  - un box (un riquadro) intorno al punto medio (cioè, la media o la mediana) che rappresenterà un intervallo selezionato (cioè, la deviazione standard, l'errore standard [1], min-max o una costante)
  - i whisker (cioè, come una linea con "baffetto" su entrambe le estremità) all'esterno del box, che a loro volta rappresenteranno un intervallo selezionato
  - Si possono anche tracciare i punti outlier.

[1] L'errore standard non è altro che la deviazione standard della media campionaria: $e.s. = s/\sqrt{n}$

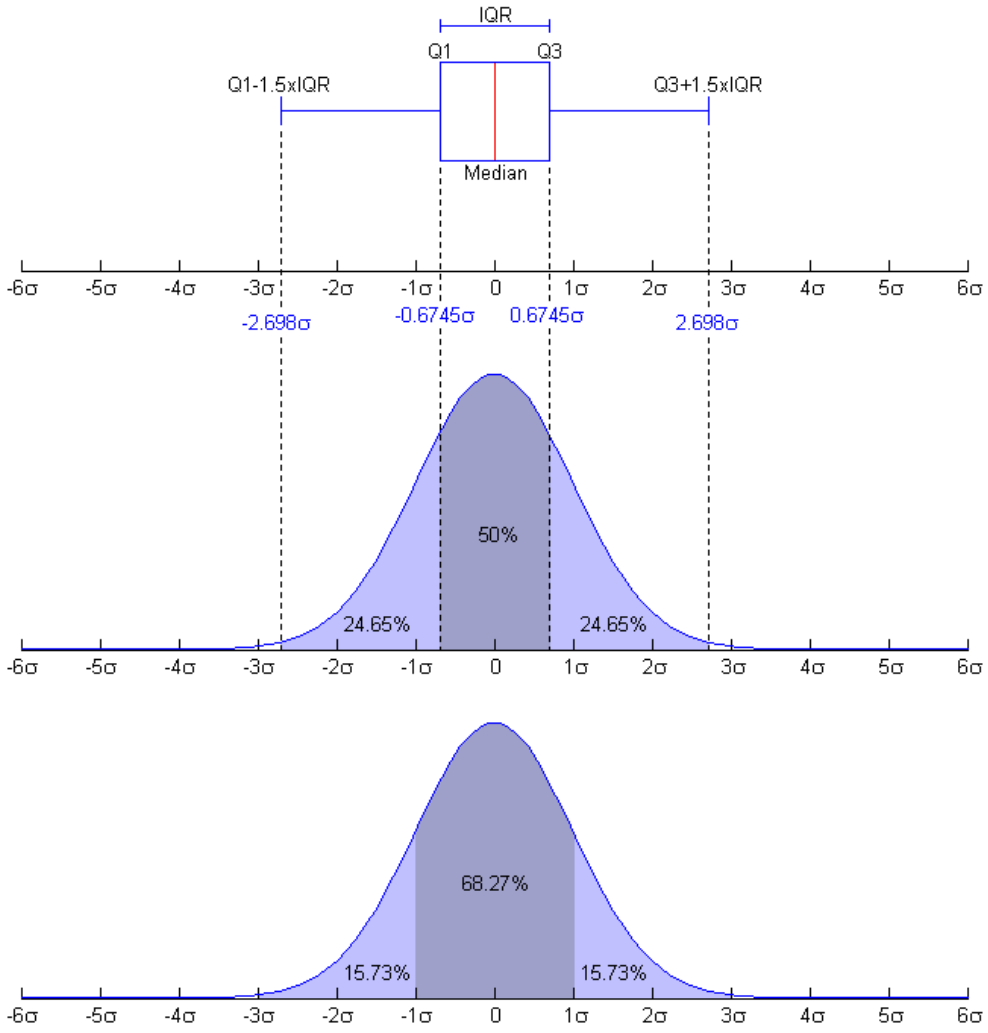
Come si legge un boxplot?



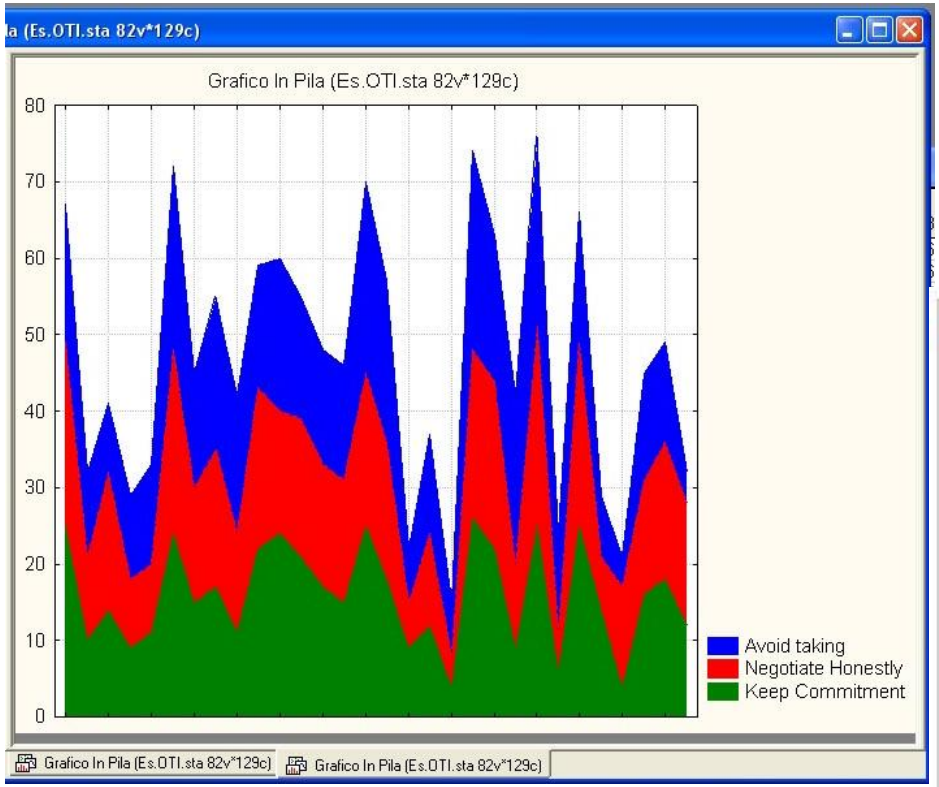


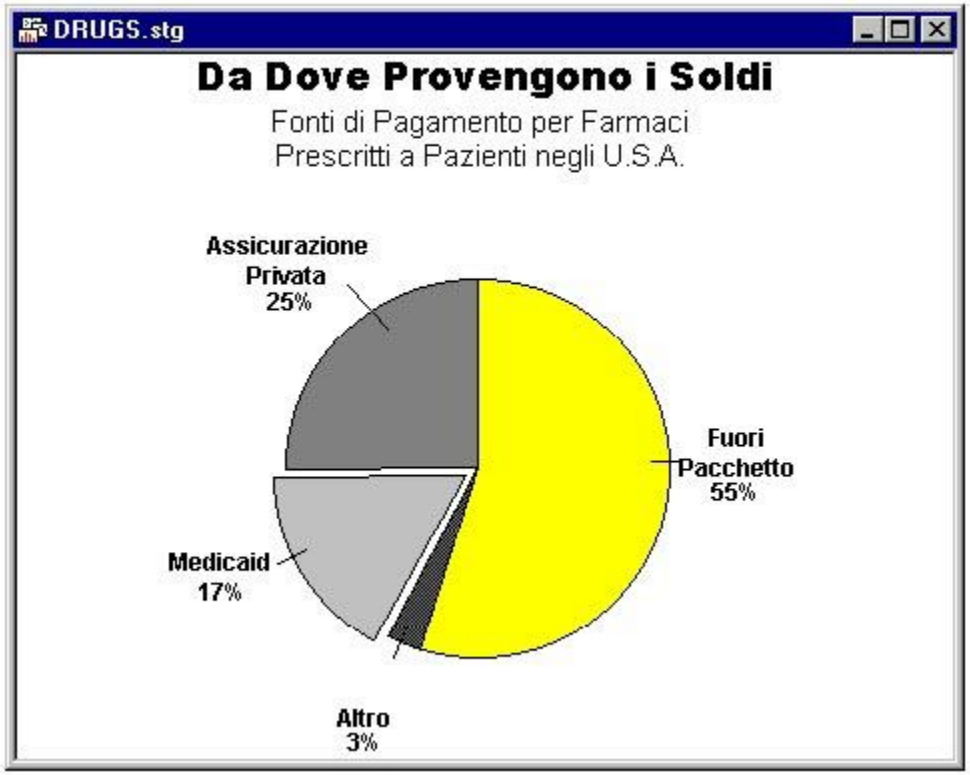


Come si legge un boxplot?



Altri grafici

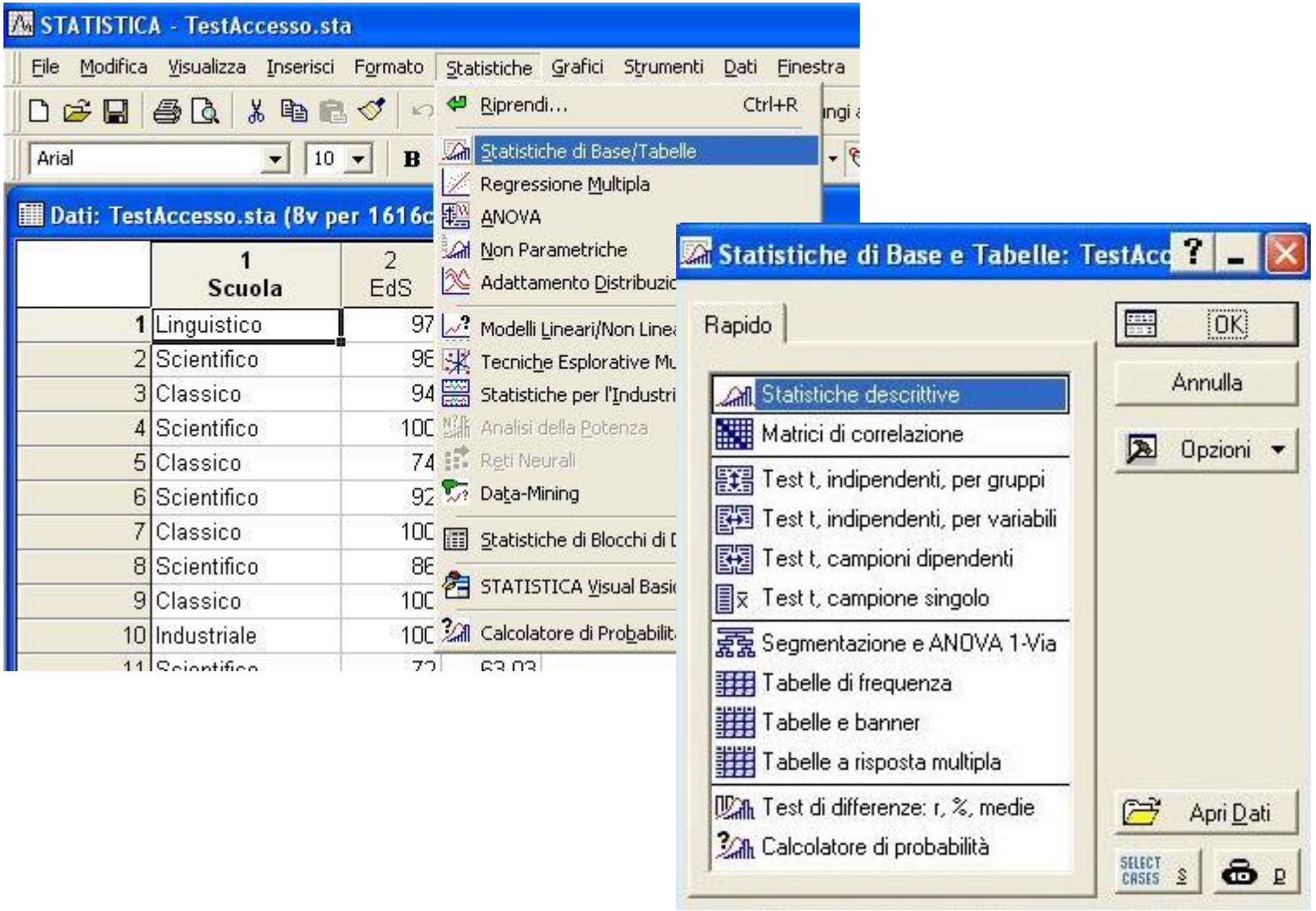




E ora proviamo ...

Statistiche descrittive (per iniziare)

Tipo di variabile	Analisi descrittiva	In Statistica
categoriale **nominale** (ad esempio il genere: M vs F)	distribuzione delle frequenze	Statistiche ➔ Statistiche di base / Tabelle ➔ Tabelle di frequenza
categoriale **ordinale** (ad esempio il reddito: basso / medio /alto)	distribuzione delle frequenze indici di posizione (percentili, quartili, mediana …)	Statistiche ➔ Statistiche di base / Tabelle ➔ Statistiche descrittive (opzione "Avanzate")
quantitativa (ad esempio il tempo di reazione)	indici di posizione (percentili, quartili, mediana …) minimo, massimo, media e deviazione standard	Statistiche ➔ Statistiche di base / Tabelle ➔ Statistiche descrittive (opzione "Avanzate")


STATISTICA - TestAccesso.sta
File Modifica Visualizza Inserisci Formato Statistiche Grafici Strumenti Dati Finestra
Riprendi... Ctrl+R
Statistiche di Base/Tabelle
Regressione Multipla
ANOVA
Non Parametriche
Adattamento Distribuzio
Modelli Lineari/Non Linea
Tecniche Esplorative Mu
Statistiche per l'Industri
Analisi della Potenza
Reti Neurali
Data-Mining
Statistiche di Blocchi di D
STATISTICA Visual Basic
Calcolatore di Probabilit
Arial
10
Dati: TestAccesso.sta (8v per 1616c
1 Scuola
2 EdS
1 Linguistico 97
2 Scientifico 98
3 Classico 94
4 Scientifico 100
5 Classico 74
6 Scientifico 92
7 Classico 100
8 Scientifico 88
9 Classico 100
10 Industriale 100
Statistiche di Base e Tabelle: TestAcc
Rapido
Statistiche descrittive
Matrici di correlazione
Test t, indipendenti, per gruppi
Test t, indipendenti, per variabili
Test t, campioni dipendenti
Test t, campione singolo
Segmentazione e ANOVA 1-Via
Tabelle di frequenza
Tabelle e banner
Tabelle a risposta multipla
Test di differenze: r, %, medie
Calcolatore di probabilità
OK
Annulla
Opzioni
Apri Dati
SELECT CASES

Esercizio

- Nel file "ansia.sta" sono contenuti i dati rilevati su un gruppo di 85 soggetti maggiorenni.
  - Le variabili misurate sono: il sesso, l'età, il reddito, il punteggio ottenuto in un test riguardante l'ansia e il punteggio ottenuto in un test riguardante la depressione.
- Svolgere un'analisi descrittiva per ciascuna variabile osservati, selezionando un opportuno grafico.

Esercizio

- Il dataset "TestAccesso.xls" contiene alcune informazioni riguardanti i risultati alla prova di ammissione ad una Facoltà.
  - Svolgere un'analisi descrittiva per ciascuna variabile inserita nel dataset.
  - Ricodificare la variabile Punteggio in 4 categorie
  - A livello descrittivo, qual è la scuola di provenienza i cui diplomati sembrano avere un maggiore punteggio alla prova di accesso?

Analisi correlazionale

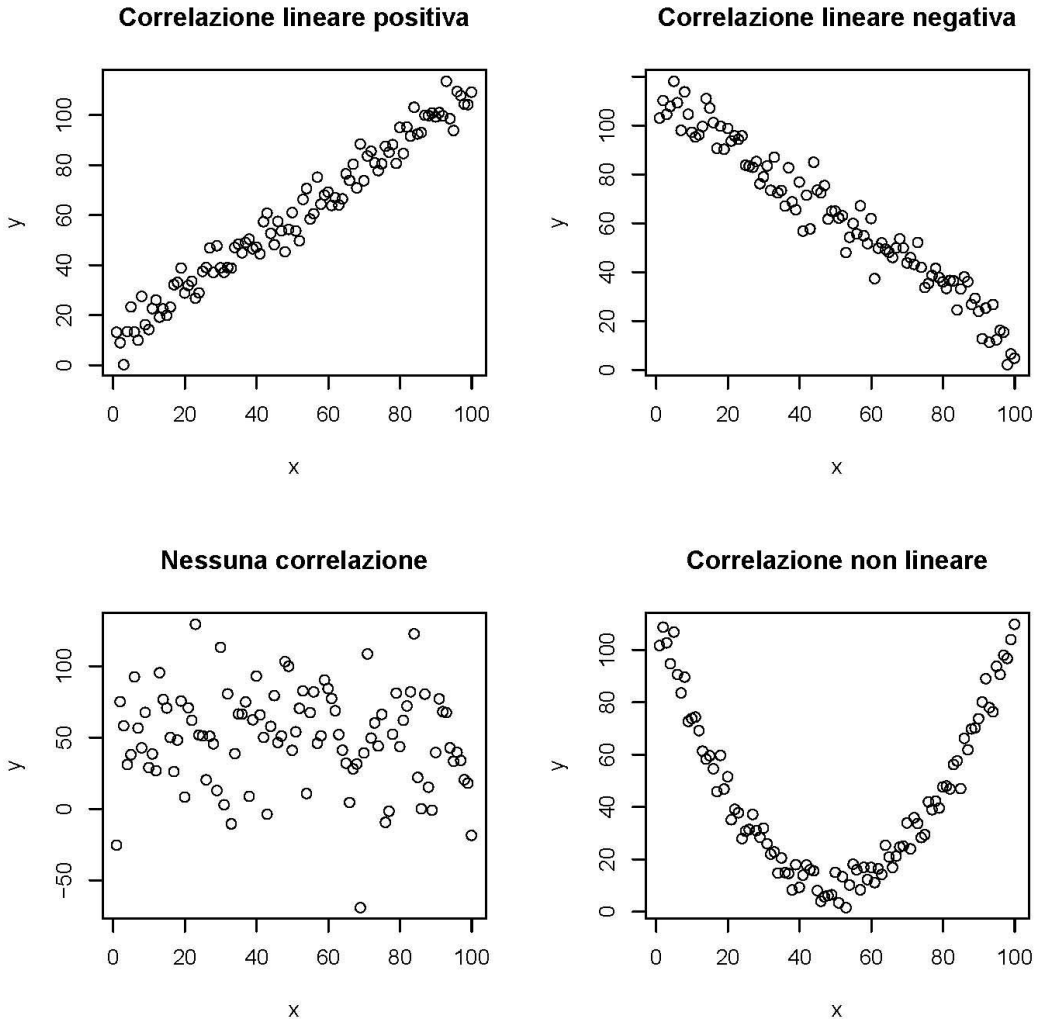
Tipologie di relazioni presenti in natura

- Scopo dell'analisi di correlazione bivariata è studiare la relazione tra due variabili quantitative X e Y.
- L'analisi di correlazione bivariata è una **metodologia simmetrica** in cui si considerano le variabili X e Y sullo **stesso piano causale**.

Metodi asimmetrici vs. metodi simmetrici

- I **metodi asimmetrici** vengono utilizzati per studiare relazioni di tipo "causa ed effetto" tra le variabili.
  - Es. il ricercatore ipotizza a priori una relazione causale tra le due variabili: una viene considerata dipendente e l'altra indipendente (ad es. Analisi di Regressione).
- Nei **metodi simmetrici** non viene ipotizzata una relazione causale tra le variabili. Non esiste quindi la suddivisione tra variabile dipendente e variabile indipendente, ma le due variabili vengono considerate sullo stesso piano (ad es. Analisi di Correlazione).

Correlazione



Correlazione lineare

- Il coefficiente di correlazione lineare di Bravais - Pearson misura il tipo e l'intensità della relazione lineare tra due variabili X e Y.

Esso si indica:

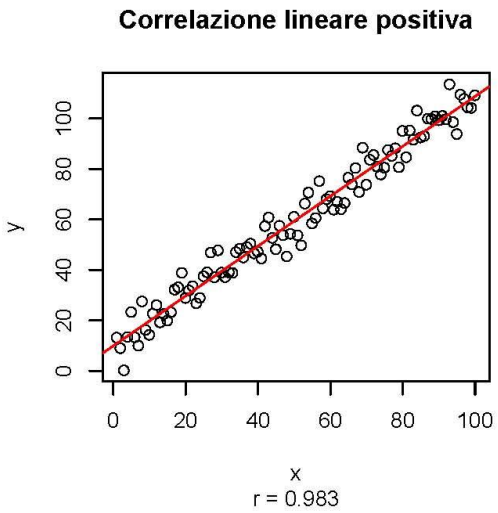
- con la lettera greca ρ se viene calcolato su tutta la popolazione oggetto dell'indagine;
- con la lettera *r* se viene calcolato su un campione rappresentativo della popolazione.

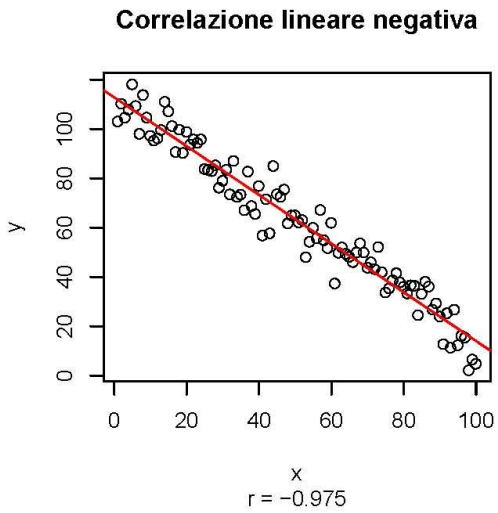
- Il coefficiente di correlazione lineare varia tra -1 e 1 (sempre !)

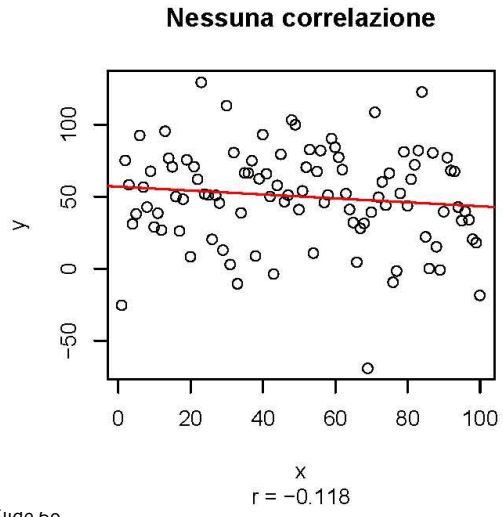
- Il segno di r (+ o -) da informazioni sul tipo di relazione:
  - il segno positivo indica che le due variabili aumentano o diminuiscono assieme (relazione lineare positiva)
  - il segno negativo indica che all'aumentare di una variabile l'altra diminuisce e viceversa (relazione lineare negativa)

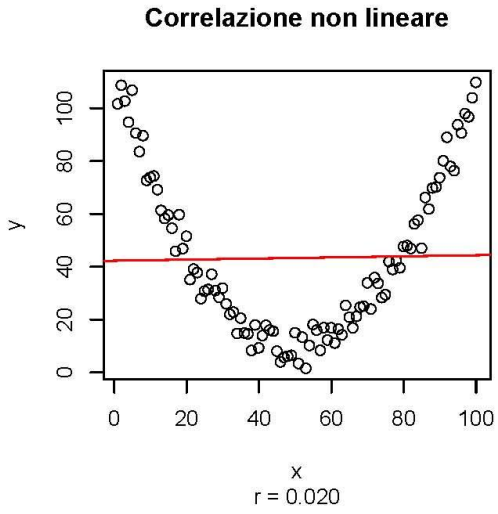
- Il valore assoluto di r, che varia tra 0 e 1, da informazioni sulla forza della relazione lineare:
  - è massimo (assume valore 1) quando esiste una perfetta relazione lineare tra le due variabili.
  - tende a ridursi al diminuire dell'intensità della relazione lineare e assume il valore 0 quando essa è nulla.

- Rissuamendo
- I valori che può assuemere r
  - r = -1 : perfetta relazione lineare negativa
  - r = 0 : assenza di relazione lineare
  - r = 1 : perfetta relazione lineare positiva


Correlazione lineare positiva
y
x
r = 0.983


Correlazione lineare negativa
y
x
r = −0.975


Nessuna correlazione
y
x
r = −0.118


Correlazione non lineare
y
x
r = 0.020

Analisi correlazionale

- Nella maggior parte dei casi il coefficiente di Correlazione di Pearson viene calcolato su un campione della popolazione.

- Obiettivo della verifica di ipotesi:
  - capire se esiste una **correlazione statisticamente significativa** tra le due variabili X e Y.

- FORMULAZIONE DEL PROBLEMA
  - H_0: non c'è una significativa correlazione lineare tra le variabili X e Y ($\rho=0$)
  - H_1: esiste una significativa correlazione lineare tra le variabili X e Y ($\rho \neq 0$)

- Per verificare la significatività statistica di un coefficiente di correlazione si ricorrere solitamente al test *t* di Student.
  - La condizione di validità per poter applicare i test t di Student è che le variabili X e Y abbiano una distribuzione approssimativamente normale bivariata.

- Nel caso in cui sia vera l'ipotesi nulla ($\rho = 0$), la statistica test

$$t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

- dove:
  - r è il coefficiente di correlazione calcolato sul campione,
  - n è la numerosità del campione

- è distribuita come una t di Student con n-2 gradi di libertà.

- Se il p-value associato alla statistica osservata t è maggiore del valore critico (che solitamente è fissato in 0.05) si accetta H_0 e quindi di conclude che non c'è una correlazione lineare statisticamente significativa tra le due variabili X e Y.
- In caso contrario si rifiuta H_0.

E se mi chiedessi … "come faccio a scriverlo nella tesi ?!"

- Devono essere specificati:
  - la numerosità del campione;
  - il valore di r;
  - la presenza (o assenza) di una relazione statisticamente significativa;
  - il valore del p osservato;
  - il tipo di test utilizzato (a una coda o a due code).

- "La ricerca ha riscontrato la presenza di una correlazione lineare positiva, statisticamente significativa, fra l'età e il grado di apprendimento dei pazienti ($r=0.82$, $n=50$, $p<0.05$, due code)."

Rappresentazione grafica

- Prima di procedere al calcolo dell'indice di correlazione è molto utile rappresentare graficamente la distribuzione congiunta delle due variabili oggetto di studio in un grafico a dispersione.
- Questo oltre che ha dare "una prima idea" sulla relazione tra le variabili è molto utile per valutare la presenza di eventuali valori anomali (outliers)

- Grafici → Grafici 2D → Scatterplot
  - Permette di visualizzare il grafico a dispersione di 2 variabili e la relativa retta di regressione.

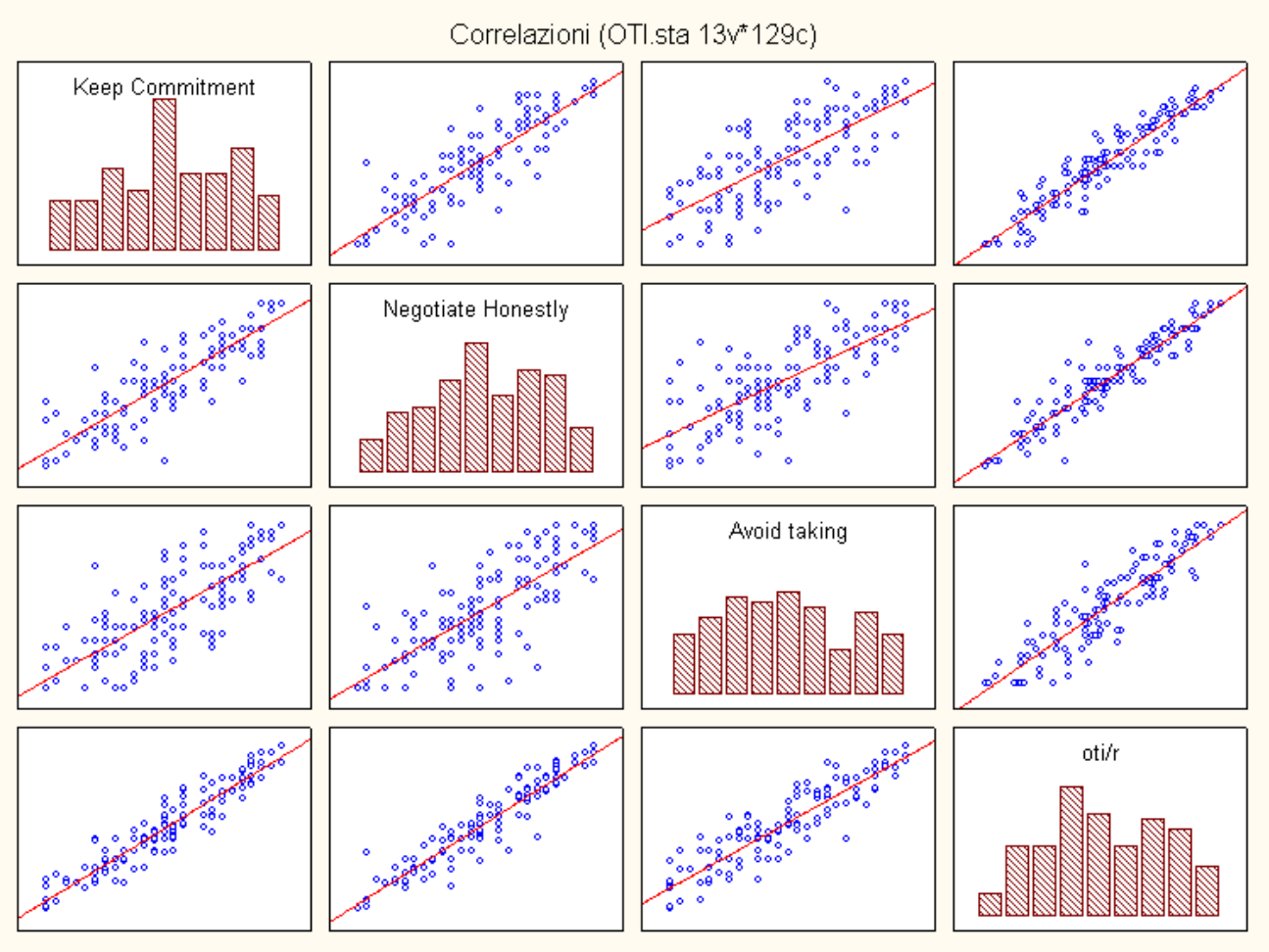
- Grafici → Grafici a Matrice
  - Permette di visualizzare i grafici di dispersione tra tutte le coppie delle variabili considerate (molto utile!).

- Statistiche → Statistiche di Base/Tabelle → Matrici di Correlazione
  - Selezionare l'opzione "Una lista di Variabili" ed inserire le variabili sulle quali calcolare i coefficienti di correlazione.

  - Nota: L' opzione "Eliminazione DM" consente di gestire i dati mancanti. Esistono due possibilità:
    - selezionando "Pairwise", un soggetto non viene considerato solo per le variabili sulle quali ha un dato mancante (metodo di default)
    - selezionando "Listwise", un soggetto che ha almeno un dato mancate sulle variabili considerate viene escluso completamente dall'analisi;

- È stato somministrato un test sulla fiducia nel mondo del lavoro ad un campione di 129 lavoratori.
- Il questionario misura le seguenti dimensioni
  - Punteggio totale (oti) e punteggio alla versione ridotta (oti/r)
  - Punteggio nelle scale: Keep committment, Negotiate honestly, avoid taking excessive advantages.
  - Commentare le relazioni tra le suddette variabili.

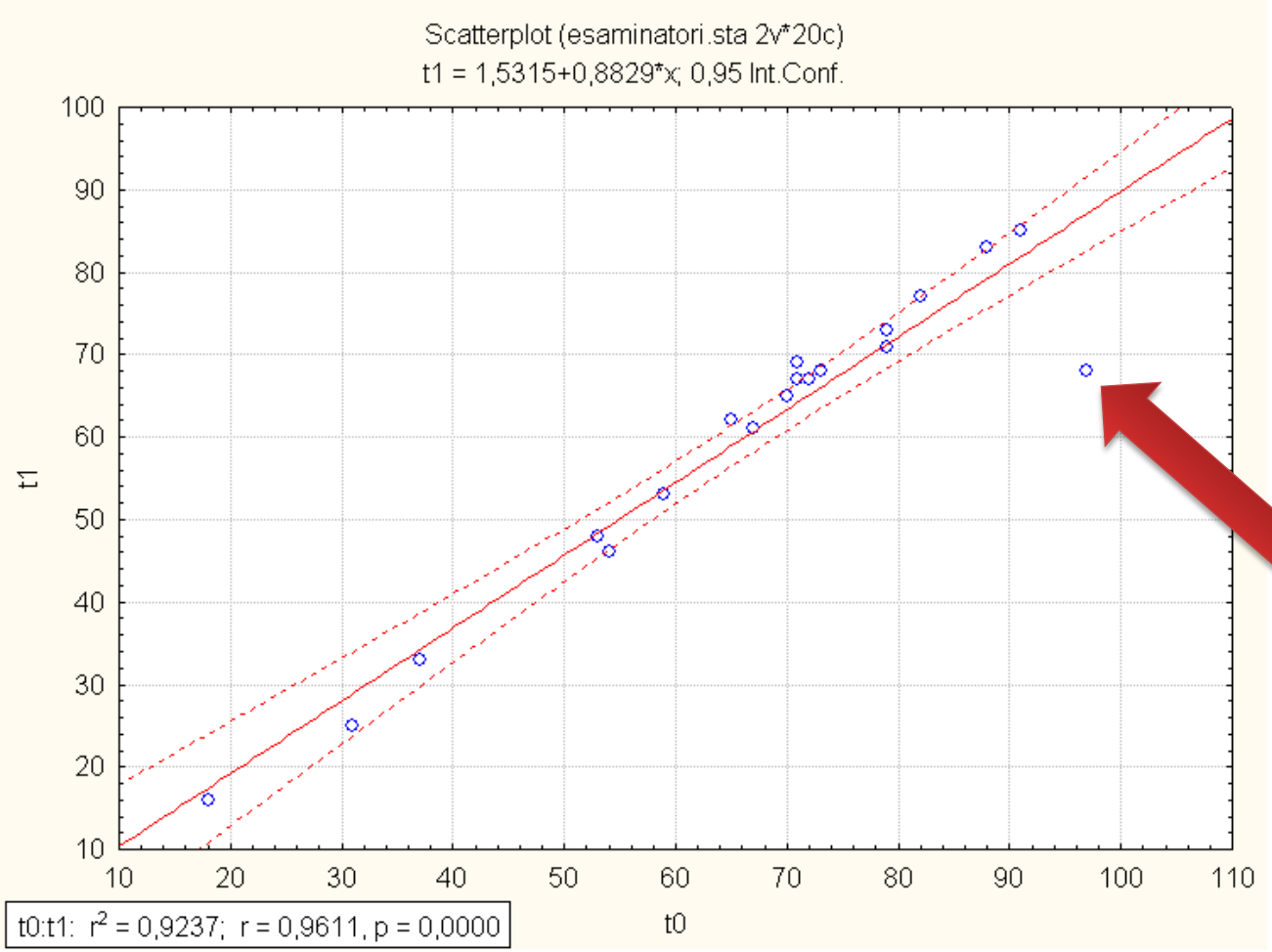
- Nota: (I dati sono contenuti nel file "OTI.xls")


Correlazioni (OTI.sta 13v*129c)
Keep Commitment
Negotiate Honestly
Avoid taking
oti/r

- Viene chiesto a due esaminatori di valutare su di una scala 0 – 100 l'efficienza di un servizio per il pubblico, secondo alcuni parametri noti.

- Si può dire che vi è concordanza tra gli esaminatori ?

- Nota: i dati sono contenuti nel file "esaminatori.sta"


Scatterplot (esaminatori.sta 2v*20c)
t1 = 1,5315+0,8829*x; 0,95 Int.Conf.
t1
t0
100
90
80
70
60
50
40
30
20
10
10
20
30
40
50
60
70
80
90
100
110
t0:t1: r² = 0,9237; r = 0,9611, p = 0,0000

E adesso …

- Su un campione di 1650 matricole della facoltà Psicologia sono state rilevate le seguenti variabili:
  - Voto all'Esame di Stato (0-100)
  - Voto ottenuto al Test di Ingresso all'Università (0-70)
  - Voto ottenuto nelle conoscenze di matematica, scienze umane, fisica e biologia, logica, comprensione di un brano.
- Commentare le relazioni tra le suddette variabili.

- Nota: I dati sono contenuti nel file "TestAccesso.xls"

SUGGERIMENTI
1. Costruire i grafici di dispersione per ciascuna coppia di variabili.
2. Osservare i grafici di dispersione.
3. Calcolare i coefficienti di correlazione lineare tra le variabili osservate.
4. Discutere i risultati ottenuti.

Confronti tra medie

- Il test t di Student è il metodo più comune per valutare la differenza tra le medie di due gruppi di osservazioni.

- Per utilizzare le varie tipologie di t – test:
  - Statistiche ➔ Statistiche di Base/Tabelle
- e scegliere il tipo di t-test desiderato

Tipo di t -test	Obiettivo del test	Verifica di ipotesi
test per campione singolo	Verificare se la media rilevata su un campione differisce rispetto a quella di una popolazione	Se il valore di probabilità osservato associato al test (p_{OSS}) è inferiore a un livello di probabilità fissato a priori (p_{CRIT}) si conclude che le media rilevata sul campione differisce significativamente da quella della popolazione.
test per campioni indipendenti	Verificare se le medie di 2 campioni indipendenti differiscono significativamente tra loro.	Se $p_{OSS} < p_{CRIT}$ si conclude che esiste differenza significativa tra le medie de due campioni
test per campioni appaiati	Verificare se le medie di una variabile rilevata 2 volte sullo stesso campione differiscono tra loro.	Se $p_{OSS} < p_{CRIT}$ si conclude che le medie rilevate nelle 2 occasioni differiscono significativamente tra loro.

Alcune note importanti (povera Statistica):

- dal punto di vista teorico il t test può essere utilizzato solo se la variabile oggetto di studio è distribuita normalmente.
- nel caso di t-test per campioni indipendenti è necessario che le varianze dei due gruppi siano tra loro omogenee.
  - Per valutare l'omegeneità della varianza può essere utilizzta l'opzione "Test di Levene" (se tale test risulta significativo l'omogeneità delle varianze non può essere accettata).
- nei casi di ridotta numerosità campionaria ($n < 30$) il test-t non è da considerarsi statisticamente robusto.
- se le ipotesi per l'applicabiltà del t-test non sono verificate, e/o nei casi caratterizzati da ridotta numerosità campionaria è consigliabile utilizzare metodi non parametrici.

Test t, campione singolo

- Esegue un confronto tra una media osservata e una media nota (un valore atteso per la popolazione), ad esempio il confronto fra il salario medio di una ditta e il salario medio nazionale.
  - Selezionare su "Variabili" la/le variabile/i che si vuole confrontare con la media nota.
  - Scrivere il valore della media nota su "Valori di riferimento": "Testa ogni media rispetto".
  - Per avere una rappresentazione grafica, cliccare su "Box & whisker".

Test t, indipendenti, per gruppi

- Esegue il t-test per campioni indipendenti.
- Si usa quando si vogliono confrontare le medie di due gruppi di soggetti se, per ogni variabile, i dati dei due gruppi sono impostati in un'unica colonna e ci si serve di una colonna aggiuntiva con i numeri (codici) assegnati ai gruppi (ad es. la variabile "GRUPPO" in cui 1=maschi e 2=femmine);
  - Selezionare in "Variabili" la variabile dipendente che contiene le medie da confrontare e la variabile di gruppo che contiene i codici per i gruppi (che verranno automaticamente inseriti nel campo sottostante).
  - Cliccare su "Riepilogo: Test t ".
  - Per avere una rappresentazione grafica, cliccare su "Box & whisker".

Test t, indipendenti, per variabili

- Esegue il t-test per campioni indipendenti.
- Si usa quando si vogliono confrontare le medie di due gruppi di soggetti se i dati di ciascun gruppo sono impostati in due colonne separate (nella pratica avviene di rado).
  - Selezionare in "Variabili" le due colonne da confrontare.
  - Cliccare su "Riepilogo: Test t ".
  - Per avere una rappresentazione grafica, cliccare su "Box & whisker".

Test t per campioni appaiati

- Test t, campioni dipendenti
- Esegue il t-test per campioni dipendenti; si usa quando si vogliono confrontare due medie rilevate sullo stesso gruppo di soggetti (ad es., baseline vs. stimolo, oppure i risultati ottenuti prima e dopo un trattamento).
  - Selezionare su "Variabili" le variabili da confrontare (anche più di 2, ma il confronto sarà sempre effettuato a due a due).
  - Nota: se si inseriscono 2 variabili nella prima lista STATISTICA effettuerà i t-test tra tutte le variabili, due a due (quindi anche di ogni variabile con se stessa), se si inserisce una variabile nella prima lista e una nella seconda, effettuerà il t-test solamente tra queste due.
  - Cliccare su "Avanzate", Mostra "Risultati dettagliati", e quindi "Riepilogo: Test t".
  - Per avere una rappresentazione grafica, cliccare su "Box & whisker".

- È noto in letteratura che la lunghezza delle pannocchie di grano è distribuita normalmente con media pari a 25 cm. Un contadino, appassionato di statistica, è dell'opinione che le pannocchie da lui prodotte quest'anno abbiano una lunghezza diversa rispetto alla media generale.
- Per valutare tale ipotesi, il contadino ha:
  - 1) selezionato casualmente un campione di 40 pannocchie, tra quelle da lui prodotte;
  - 2) misurato ciascuna pannocchia;
  - 3) costruito un dataset contenente i dati rilevati.

- Verificare ad un livello di significatività del 5% ($\alpha = 0.05$) l'ipotesi che le pannocchie del contadino abbiano una lunghezza media diversa rispetto alla lunghezza media generale.
- Discutere i risultati ottenuti.
- Nota: i dati sono contenuti nel file "pannocchie.csv".

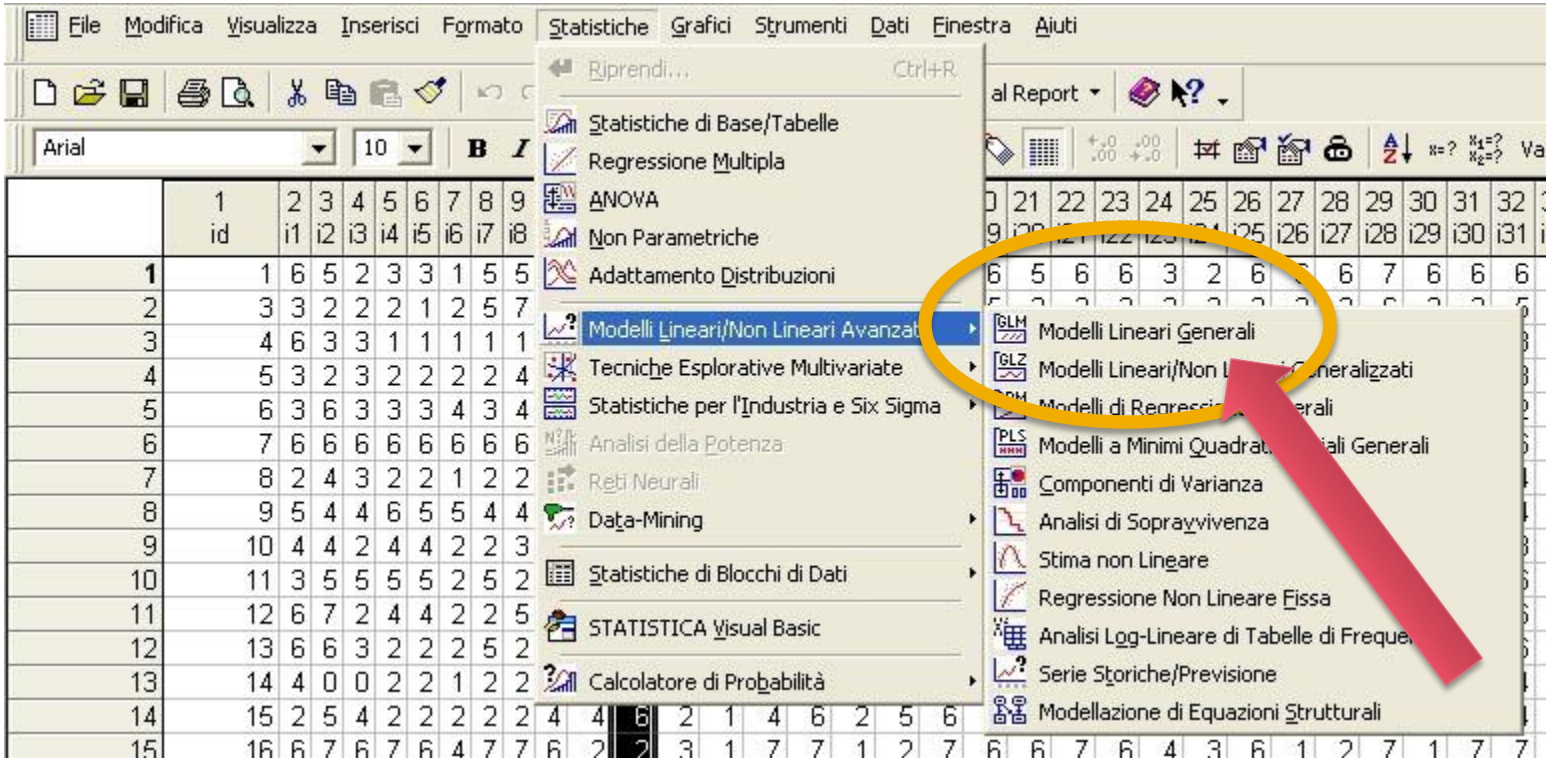
- Un ospedale vuole confrontare l'efficacia di 2 trattamenti relativi alla cura della claustrofobia. Per fare ciò, 50 pazienti vengono casualmente assegnati a 2 gruppi (gruppo A = 25 pazienti , gruppo B = 25 pazienti).
- Ai membri del gruppo A viene somministrato il trattamento A e a quelli del gruppo B il trattamento B. Alla fine dei due trattamenti, i soggetti vengono sottoposti a una serie di prove in ambienti chiusi e il loro comportamento viene videoregistrato.
- A 3 psicologi clinici viene richiesto di visionare i video e di valutare in maniera indipendente ogni soggetto su una scala da 1 (poco claustrofobico) a 10 (molto claustrofobico). A ciascun soggetto viene attribuito un giudizio complessivo derivante dalla media dei tre giudizi.
- Valutare se esiste differenza significativa nell'efficacia dei due trattamenti.
- Nota: i dati sono contenuti nel file "claustro.csv".

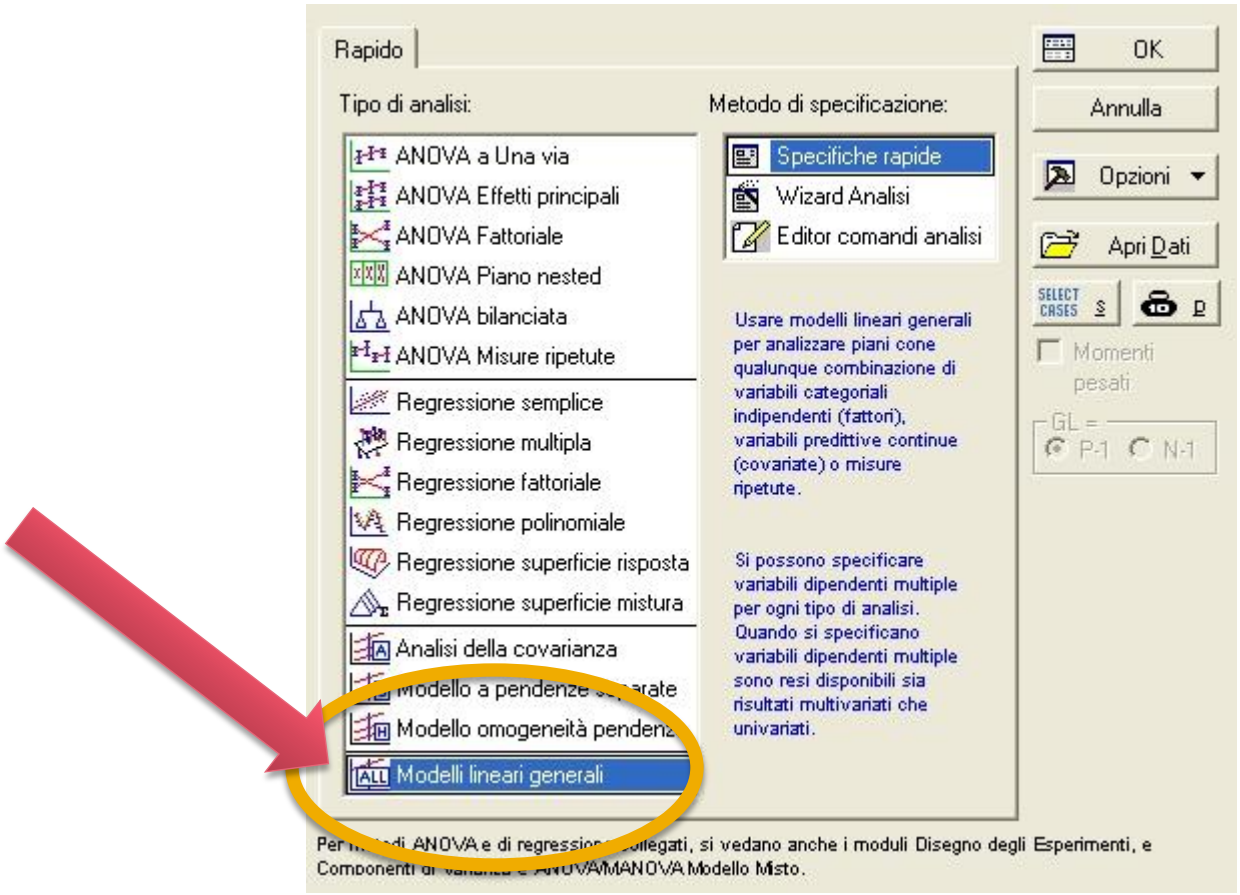
- Un istruttore di palestra vuole valutare se il suo corso di fitness ha degli effetti sul peso dei suoi allievi. Per fare ciò rileva il peso di 20 nuovi iscritti all'inizio del corso e alla fine del corso (dopo 2 mesi).
- L'istruttore è dell'idea che il corso che lui propone è in grado, tra le altre di accelerare il metabolismo delle persone, facendo loro ridurre la quantità di grasso corporeo.
- Valutare sei il peso degli allievi è cambiato dopo i due mesi di corso.
- Nota: i dati sono contenuti nel file "fitness.csv"

Analisi della Varianza

Applicazioni di modelli di analisi della varianza

- Per poter costruire dei modelli di analisi della varianza (univariata, multivariata, a misure ripetute):
  - Statistiche → Modelli Lineari / Non Lineari Avanzati → Modelli Lineari Generali → Modelli Lineari Generali


File Modifica Visualizza Inserisci Formato Statistiche Grafici Strumenti Dati Finestra Aiuti
Riprendi... Ctrl+R
Statistiche di Base/Tabelle
Regressione Multipla
ANOVA
Non Parametriche
Adattamento Distribuzioni
Modelli Lineari/Non Lineari Avanzati
Tecniche Esplorative Multivariate
Statistiche per l'Industria e Six Sigma
Analisi della Potenza
Reti Neurali
Data-Mining
Statistiche di Blocchi di Dati
STATISTICA Visual Basic
Calcolatore di Probabilità
Modelli Lineari Generali
Componenti di Varianza
Analisi di Sopravvivenza
Stima non Lineare
Regressione Non Lineare Fissa
Serie Storiche/Previsione
Modellazione di Equazioni Strutturali
Arial
10
id


Rapido
Tipo di analisi:
Metodo di specificazione:
ANOVA a Una via
ANOVA Effetti principali
ANOVA Fattoriale
ANOVA Piano nested
ANOVA bilanciata
ANOVA Misure ripetute
Regressione semplice
Regressione multipla
Regressione fattoriale
Regressione polinomiale
Regressione superficie risposta
Regressione superficie mistura
Analisi della covarianza
Modello a pendenze
Modello omogeneità pendenz
Modelli lineari generali
Specifiche rapide
Wizard Analisi
Editor comandi analisi
Usare modelli lineari generali per analizzare piani cone qualunque combinazione di variabili categoriali indipendenti (fattori), variabili predittive continue (covariate) o misure ripetute.
Si possono specificare variabili dipendenti multiple per ogni tipo di analisi. Quando si specificano variabili dipendenti multiple sono resi disponibili sia risultati multivariati che univariati.
OK
Annulla
Opzioni
Apri Dati
SELECT CASES
Momenti pesati
GL =
P-1
N-1
ANOVA/MANOVA Modello Misto.

Analisi della Varianza

- E una statistica molto più potente del t-test e si applica a disegni molto più complessi (confronto tra medie di più gruppi e più condizioni).
- Si può testare l'effetto di un fattore tenendo sotto controllo gli altri e si accede alla verifica delle interazioni tra fattori.
- Se si stanno confrontando solo due medie tuttavia ANOVA fornirà gli stessi (identici) risultati del test t.

  - Per quanto concerne gli aspetti tecnici e di implementazione si approfondisca l'argomento con la dispensa allegata preparata dalla dr.ssa Silvia Poli, *Uso del programma STATISTICA 6.1*, pag. 25-36.
  - Oppure **http://www.statsoft.com/textbook/stathome.html**

- Il termine "analisi della varianza" deriva dal fatto che, pur basandosi su una analisi delle medie, la tecnica statistica utilizzata si basa sulla "scomposizione" della variabilità totale dei dati osservati in due parti:
  - **variabilità sperimentale** (varianza sperimentale o spiegata o tra gruppi (between groups) detta anche Mean Square Effect, Media del Quadrato degli Effetti, o MSeffetto) che e dovuta alle variabili introdotte e studiate dal disegno di ricerca e cioè alla manipolazione della variabile indipendente.
  - **variabilità residua** o accidentale (varianza non spiegata, o di errore, o entro i gruppi (within groups) detta anche Mean Square Error, Media del Quadrato dell'Errore o MSerrore) che e dovuta a tutte le condizioni o variabili non controllabili o non controllate dal disegno stesso.

- Ipotesi sperimentali
- H_0: non vi sono differenza tra le medie dei gruppi nella popolazione
  - ci si può aspettare che la varianza stimata sulla base della variabilità tra i gruppi (dovuta alla manipolazione della VI) è all'incirca pari a quella dovuta alla variabilità entro gruppi (variabilità accidentale).

- Queste due dimensioni di varianza possono essere confrontate tramite il test F.
  - F = varianza tra i gruppi / varianza entro i gruppi

- Il valore di F è tanto più grande quanto più è grande la varianza tra i gruppi e piccola quella entro i gruppi.
- Per valutare se esso è abbastanza grande per rigettare l'ipotesi nulla si confronta la probabilità associata (p-value) con il livello di significatività fissato (solitamente 0.05).

- H_0
  - Se non possiamo rigettare l'ipotesi nulla
  - possiamo concludere che i campioni provengano dalla stessa popolazione e quindi la varianza tra-i-gruppi e la varianza entro-i-gruppi sono due stime indipendenti della stessa varianza della popolazione.
- H_1
  - se la varianza tra-i-gruppi è significativamente più grande di quella entro-i-gruppi,
  - possiamo concludere che la variabilità osservata nella variabile dipendente è riconducibile alla manipolazione della variabile indipendente.
  - Esiste una differenza tra le medie dei gruppi riconducibile alla variabile indipendente.

- Riassumendo
  - Se il risultato del test F non è significativo **è inutile** procedere all'esame delle differenze tra medie particolari, perche vi è il rischio reale che un certo numero di confronti sia dato come significativo mentre la maggior parte di essi è dovuto solo alla variabilità casuale.
  - Se invece il risultato del test F è statisticamente significativo vuol dire che **almeno una media** risulta essere diversa dalle altre.

- Per individuare quale gruppo o quali gruppi differiscono si può procedere invece in due modi:
  - confronti a priori o contrasti **pianificati** prima della raccolta dati, in quanto aventi "a priori" un particolare interesse.
  - confronti a posteriori o **post-hoc** (definiti dopo aver raccolto i dati ed esaminato le medie, tipicamente tutti i confronti a coppie possibili)

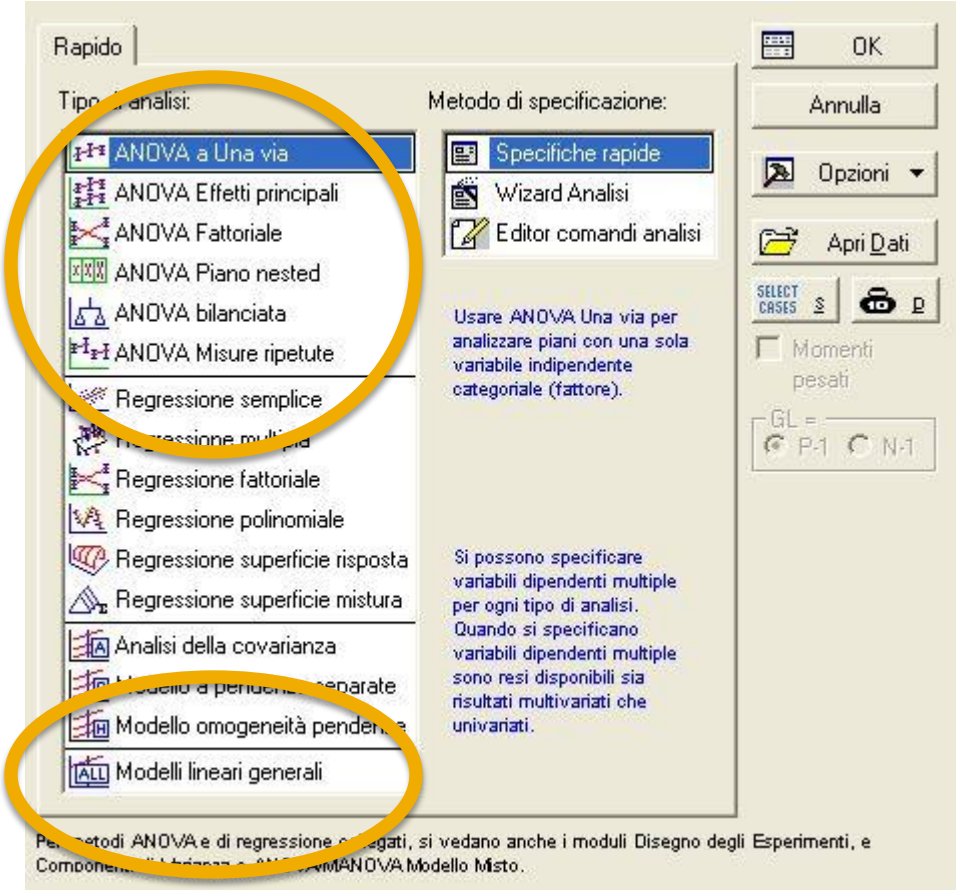
Nota bene:

- L'attendibilita del test F nell'analisi della varianza si basa sulla soddisfazione dei seguenti **assunti**:
  - normalita della distribuzione della variabile dipendente.
    - Questa si verifica con i test di normalità di Kolmogorov-Smirnof o di Shapiro-Wilk;
  - estrazione casuale dei campioni della popolazione;
  - omogeneita delle varianze dei gruppi.
    - Si verifica con il test di Levene.

- A seconda del numero di Variabili Indipendenti avremo:
  - analisi della varianza **univariata** a una via se si ha una sola VI
  - analisi **fattoriale** se si hanno più variabili indipendenti

- A seconda del numero delle Variabili Dipendenti oggetto di analisi potremmo avere:
  - analisi della varianza **univariata** (ANOVA) se è indagata una sola VD
  - disegni a **misure ripetute** se la VD è misurata più volte
  - analisi della varianza multivariata (MANOVA) se sono indagate diverse VD

- Il modulo ANOVA in STATISTICA è un sottoinsieme del modulo Modelli Lineari Generali (GLM)
- Può eseguire analisi della varianza univariate (ANOVA) e multivariate (MANOVA), di piani fattoriali con o senza una misura ripetuta.



Alcuni esempi

- Si supponga di aver somministrato un test sulla memoria ad un campione di soggetti appartenenti a tre fasce d'età (A: 20-29 anni, B: 30-49 anni, C: 50 anni e oltre).
- Si vuole valutare se l'età ha un effetto sulla memoria.
- Come procedere:
  - Formulare le ipotesi sperimentali
  - Caricare il dataset
  - Analisi descrittive
  - Effettuare il test statistico
  - Commentare i risultati
- Nota: i dati sono contenuti nel file "memoria.csv"

ANOVA una via

- Selezionare le variabili dipendenti e il predittore categoriale

- Assunti ➔ Test di Levene
  - Per verificare l'assunto di omogeneità delle varianze (verificato se p > 0.05)
- Rapido ➔ Tutti gli effetti / Grafici
- Post Hoc ➔ HSD di Tukey

ANOVA una via

- Test di Levene per verificare l'omogeneità delle varianze

Test di Levene di Omogeneità delle Varianze (memoria.sta)
Effetto: s
Gradi di libertà per tutte le F: 2, 55

	MS Effetto	MS Errore	F	p
m	135,1959	57,60678	2,346875	0,105179

ANOVA una via

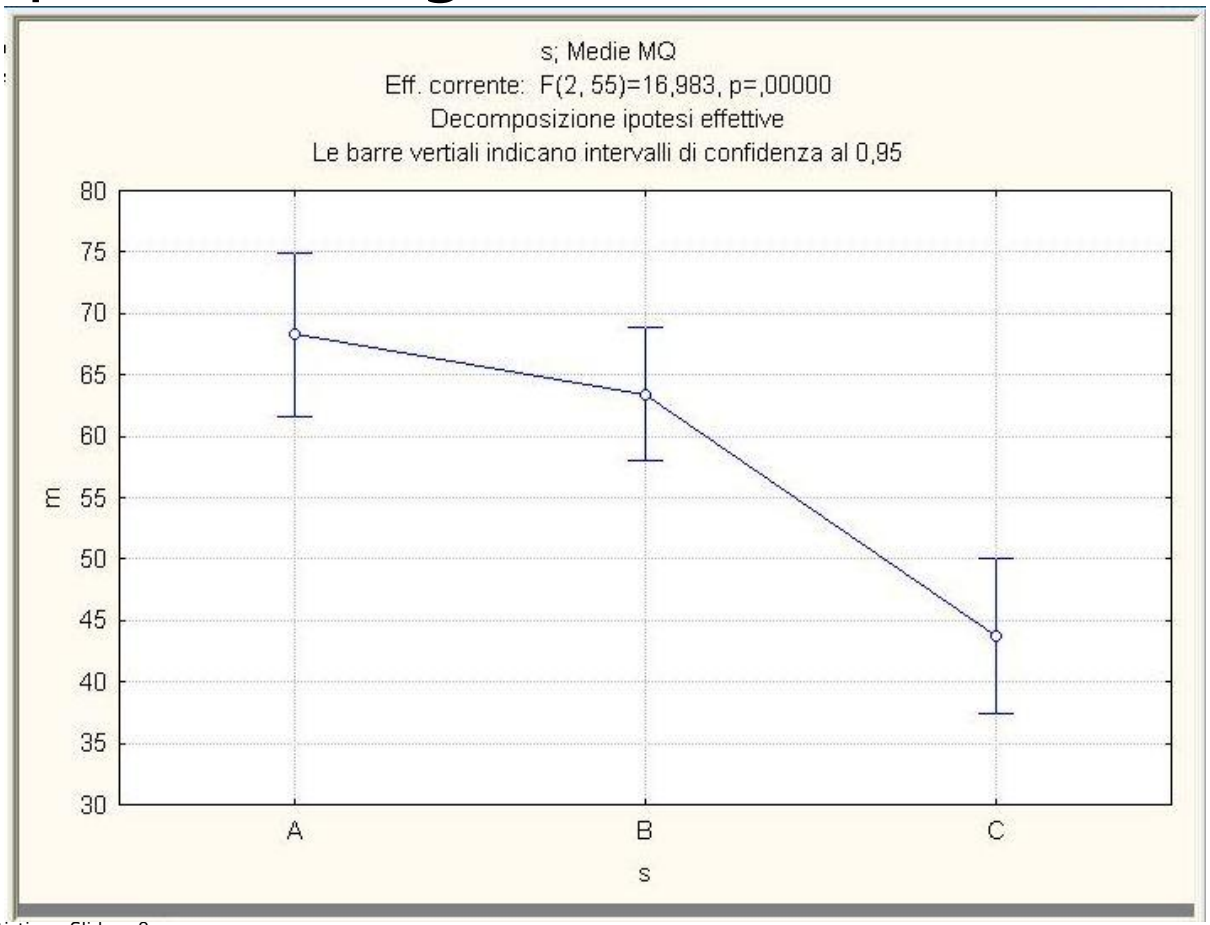
- Analisi della varianza univariata a una via

Effetto	Test Univariati di Significatività per m (memoria.sta) Parametrizzazione sigma-ristretta Decomposizione ipotesi effettive SS	Gradi di Libertà	MS	F	p
Intercetta	192851,6	1	192851,6	1088,061	0,000000
s	6020,4	2	3010,2	16,983	0,000002
Errore	9748,4	55	177,2		

- L'ANOVA ad una via ha mostrato come vi siano differenze significativa nelle medie osservate attribuibili al fattore s ($F_{2,55}$=19.98, p < 0.0001)

ANOVA una via

- Interpretazione grafica



ANOVA una via

- Analisi Post-hoc

Cella N.	s	{1} 68,313	{2} 63,417	{3} 43,778
1	A		0,494333	0,000124
2	B	0,494333		0,000161
3	C	0,000124	0,000161	

Test HSD Tukey;variabile m (memoria.sta)
Probabilità Approssimate per Test Post Hoc
Errore: MS Tra grp= 177,24, gl = 55,000

N. Cella	s	{1} 68,313	{2} 63,417	{3} 43,778
1	A		0,778420	0,000005
2	B	0,778420		0,000048
3	C	0,000005	0,000048	

Test Bonferroni;variabile m (memoria.sta)
Probabilità per Test Post Hoc
Errore: MS Tra-grp= 177,24, gl = 55,000

ANOVA fattoriale

- Si supponga di voler studiare gli effetti del fumo da sigaretta su alcuni tipi di prestazione. A tale scopo è stato selezionato un campione i cui soggetti sono stati suddivisi in tre gruppi rispetto al fumo:
  - non fumatori (NS),
  - fumatori ma non prima-durante la prova (DS),
  - fumatori attivi prima-durante la prova (AS).
- In maniera casuale all'interno di ciascun gruppo un terzo dei soggetti ha fatto
  - un compito di pattern recognition (PR),
  - un compito di tipo cognitivo (C)
  - una simulazione di guida con un video game (VG).
- In ogni caso la variabile dipendente è il numero di errori commessi.

- Le domande di ricerca riguardano la valutazione dell'effetto del fumo, dell'effetto del tipo di compito, e dell'eventuale interazione tra fumo e compito sulle performance dei soggetti.
- Nota: i dati sono contenuti nel file "smoking.csv"

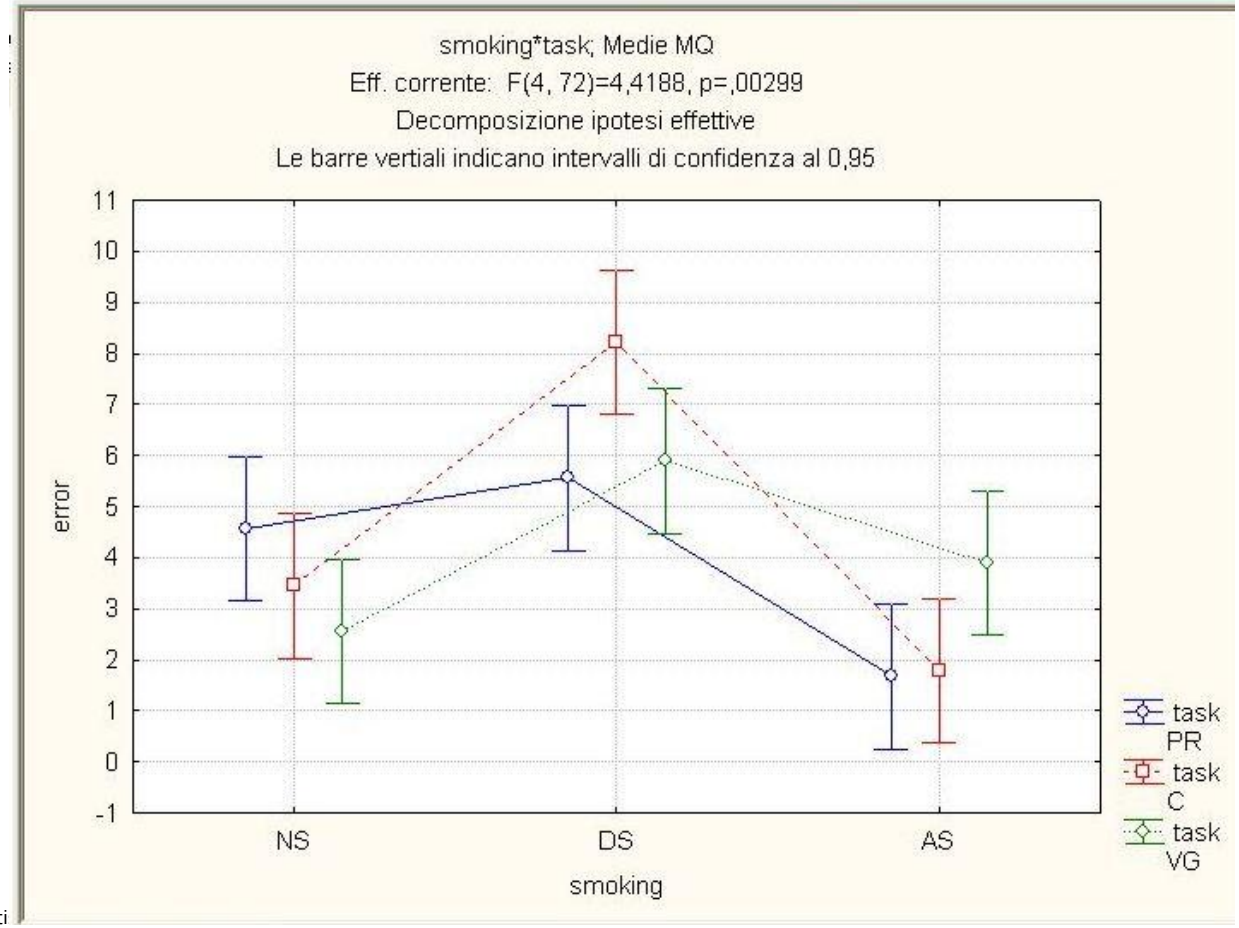
ANOVA fattoriale

- Analisi della varianza univariata 3X3

Effetto	Risultati Univariati per Ogni VD (smoking.sta) Parametrizzazione sigma-ristretta Decomposizione ipotesi effettive **Gradi di Libertà**	error SS	error MS	error F	error p
Intercetta	1	1410,420	1410,420	311,7162	0,000000
smoking	2	245,506	122,753	27,1296	0,000000
task	2	4,321	2,160	0,4775	0,622289
smoking*task	4	79,975	19,994	4,4188	0,002992
Errore	72	325,778	4,525		
Totale	80	655,580			

ANOVA fattoriale

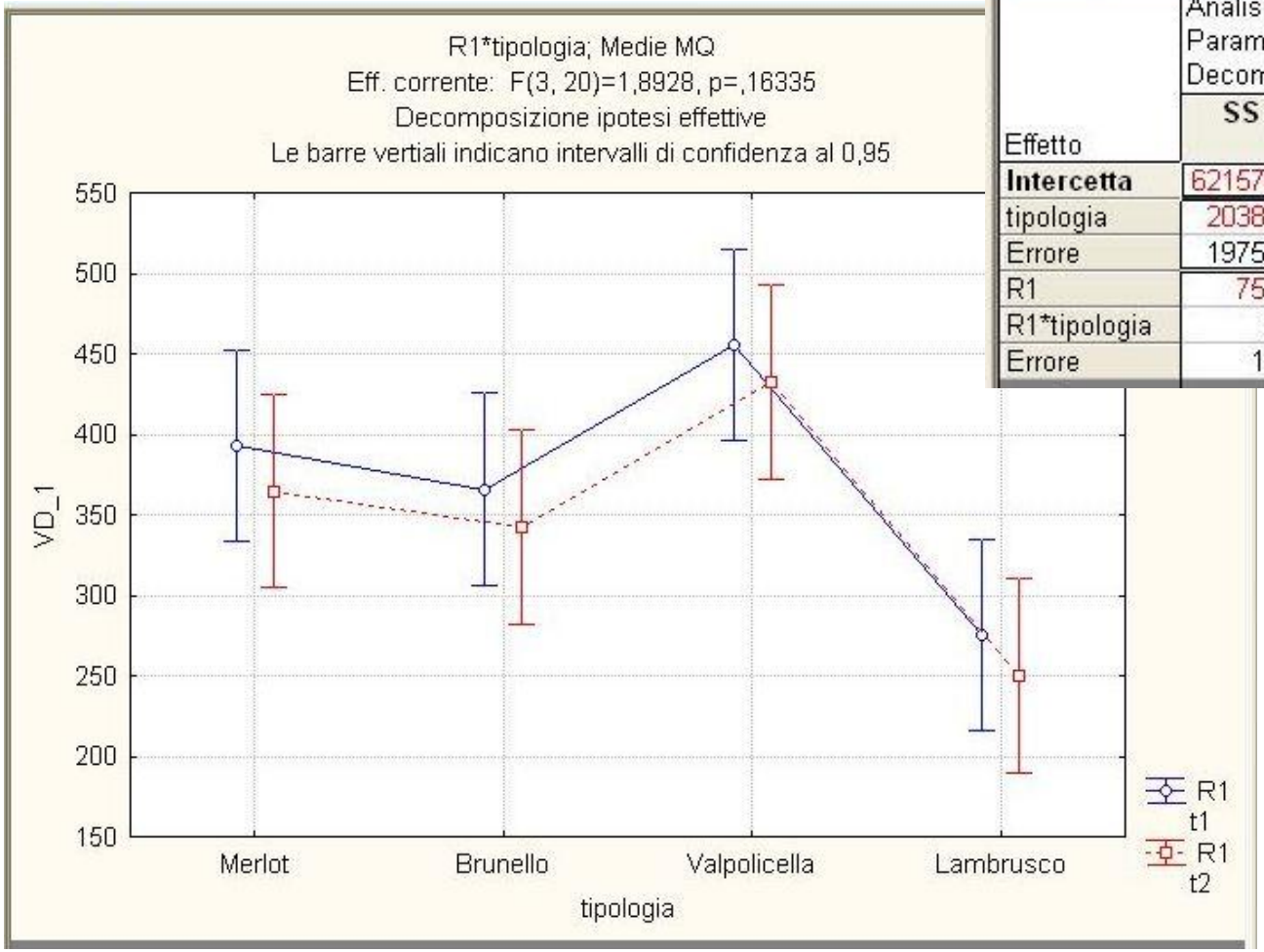
- Rappresentazione grafica dell'interazione



ANOVA misure ripetute

- Si supponga di voler studiare l'effetto di 4 diversi tipi di vino sui tempi di reazione ad una particolare prova di abilità.
- Nella conduzione dell'esperimento un tempo sufficiente viene fatto trascorrere tra una prova e l'altra, in modo da minimizzare gli effetti della "somministrazione" di un tipo di vino sui tempi di reazione legati alla "successiva somministrazione" (Winer, 1971).
- Nota: i dati sono contenuti nel file "vini.csv".

ANOVA misure ripetute



Analisi della Varianza Misure Ripetute (vini)
Parametrizzazione sigma-ristretta
Decomposizione ipotesi effettive

Effetto	SS	Gradi di Libertà	MS	F	p
Intercetta	6215761	1	6215761	629,1699	0,000000
tipologia	203888	3	67963	6,8793	0,002292
Errore	197586	20	9879		
R1	7525	1	7525	790,3742	0,000000
R1*tipologia	54	3	18	1,8928	0,163349
Errore	190	20	10		

ANOVA Fattoriale

- A due gruppi, uno sottoposto a una condizione stressante (gruppo sperimentale) ed uno sottoposto ad una condizione neutra (gruppo di controllo), vengono letti tre brani di crescente difficoltà. Dopo la lettura di ciascun brano vengono poste ai soggetti 10 domande di comprensione del testo e viene rilevato il numero di risposte corrette.
- Si vogliono studiare i seguenti aspetti:
  - la difficoltà dei brani ha un effetto sul numero di risposte corrette?
  - il gruppo sottoposto ad una condizione di stress risponde complessivamente in maniera diversa rispetto al gruppo di controllo?
  - esiste un'interazione tra la difficoltà dei brani ed il livello di stress (le differenze tra i due gruppi sono costanti per i tre livelli di difficoltà dei brani) ?
- Nota: i dati sono contenuti nel file "stress.csv".

ANOVA Fattoriale

- In un esperimento di percezione viene studiata la capacità degli utenti di sentire il contatto con un corpo morbido. Vengono utilizzati più dispositivi per la misurazione dei dati (device), superfici di differente morbidezza (stiffness), e differenti velocità di contatto con le superfici (speed). Ciascun soggetto ripete l'esperimento più volte (trial).
- Si intende studiare l'effetto delle variabili indicate nel definire la forza di contatto esercitata
- Nota: I dati sono contenuti nel file "stiffness.csv".

Statistiche non parametriche

- Come comportarsi quando gli assunti di normalità non vengono rispettati, o quando la numerosità è esigua, o quando i dati sono su scala ordinale o categoriale ?

	Analisi parametrica	Analisi non parametrica	
		Variabile dipendente almeno ordinale	**Variabile dipendente dicotomica (scala nominale)**
2 campioni indipendenti	Test t di Student per campioni indipendenti	Test U di Mann-Whitney	Test esatto di Fisher
N campioni indipendenti	ANOVA fattoriale o a una via	Test H di Krusal-Wallis	Test χ^2
2 campioni appaiati	Test t di Student per campioni appaiati	Test T di Wilcoxon	Test di McNemar
N campioni dipendenti	ANOVA a misure ripetute	Test di Friedman	Test Q di Cochran
Associazione fra 2 variabili	Correlazione di Pearson e regressione lineare	Correlazione dei ranghi (test di Spearman)	Coefficiente di contingenza (test K di Kendal)

E quindi uscimmo …

Ringraziamenti

- Questi appunti devono molto al lavoro precedentemente svolto dal dott. Gian Marco Altoè e dalla dott.ssa Silvia Poli.

- Il corso è stato finanziato nell'ambito del progetto FSE "L'analisi dei dati nella ricerca psicosociale" (2105/1/7/1017/2008).