MINISTAT
GUIDA ALL'USO DEL SOFTWARE

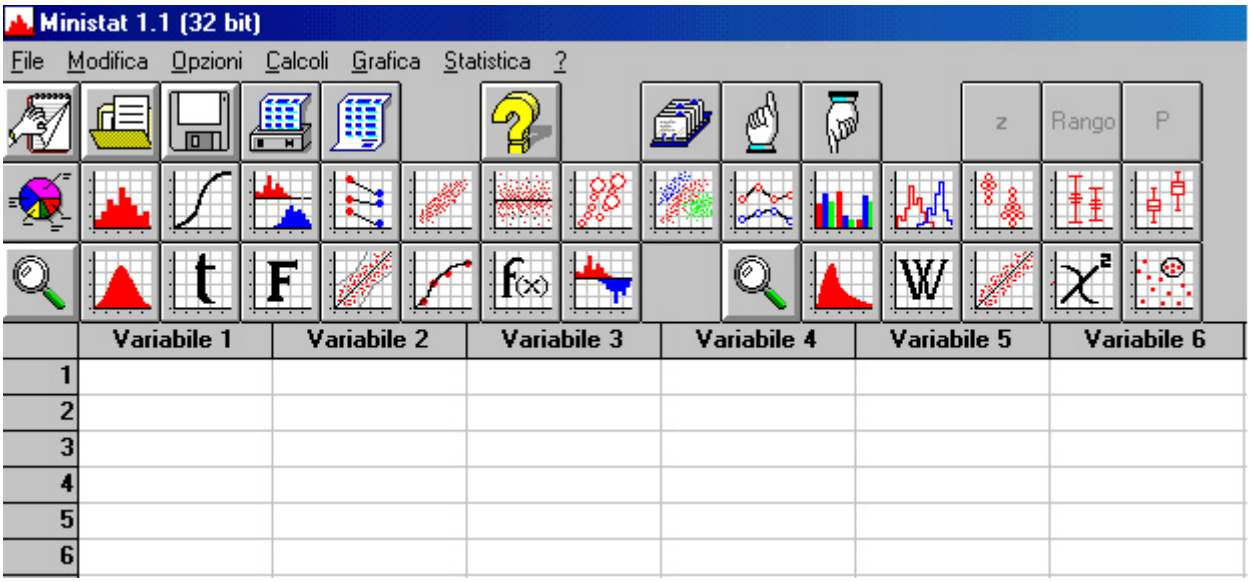
1.1. Accesso alla Guida on-line

Durante l'utilizzo di Ministat, è possibile richiamare la Guida in qualsiasi momento, selezionando Guida nel menù ? oppure facendo click sull'icona ? o ancora premendo il tasto di funzione F1.
Fare click sulle icone per una guida rapida, oppure utilizzare i collegamenti ipertestuali riportati subito dopo per accedere alla guida completa.

1.2. Caratteristiche di Ministat

1.2.1. Strutturazione dei dati

La tabella dati di Ministat, che compare nel menù principale,



costituisce la base di partenza per qualsiasi ulteriore operazione, come i calcoli, la rappresentazione grafica e l'analisi statistica.

- Ogni colonna rappresenta una variabile.
- Ogni cella contiene un dato (un valore di una variabile).
- È possibile gestire contemporaneamente fino a 10 variabili, comprendenti ciascuna fino a un massimo di 1000 dati.

Un caso particolare è costituito da due aspetti dell'Analisi della Varianza (l'analisi della varianza a due fattori e lo studio delle componenti della variabilità), e dal Test Chi-Quadrato, nei quali la strutturazione dei dati non segue la regola colonne/variabili e righe/casi. Si rimanda alle informazioni e agli esempi forniti sotto le voci specifiche per la comprensione (in realtà molto facile) di come strutturare i dati in questi casi particolari.

1.2.2. Immissione dei dati

I dati possono essere immessi nella tabella di Ministat da tastiera, essere letti da un file in formato Ministat (con l'opzione Apri del Menù File) nel quale siano stati salvati in precedenza, o essere importati da un database esterno mediante l'opzione Importa File (in formato Access® o dBase III®) del Menù File.

- Nelle celle dei dati possono essere immessi solamente caratteri di tipo numerico (inclusi i separatori virgola o punto).
- I dati devono essere limitati a 11 cifre significative, sei prima della virgola e cinque dopo la virgola (vedere l'opzione Numero di Decimali del Menù Opzioni).
- Si tratta di limiti solo apparentemente ristretti, se si tiene conto che le misure ottenute mediante i comuni dispositivi di misura comprendono in genere non più di due o tre cifre significative, rarissimamente quattro (per il concetto di numero di cifre significative in una misura, da non dare sempre per scontato, vedere Taylor[1].
- Il separatore dei decimali è quello configurato nelle Impostazioni Internazionali del Pannello di Controllo di Windows® (il punto [.] se si è scelta l'impostazione USA, la virgola [,] se si è scelta l'impostazione italiana).
- Possono essere immessi direttamente da tastiera solamente valori positivi: per i valori negativi fare doppio click sulla cella dopo avere immesso il valore.
- Per correggere i valori immessi è attivo il solo tasto di backspace.
- È possibile cancellare l'intero contenuto della cella anche selezionando Menù Modifica, quindi selezionando Cella e infine Vuota.

1.2.3. Salvataggio dei dati

I dati, che siano stati immessi da tastiera o importati da un file esterno, possono essere salvati solamente nel particolare formato utilizzato da Ministat.
Questo è in realtà un formato molto semplice (ASCII formato fisso), con 1001 righe di 140 caratteri ciascuna. Ogni riga comprende 10 campi di 14 caratteri, per un totale di 140 caratteri (alla mancanza di un carattere corrisponde a un carattere spazio). La prima riga contiene il nome delle variabili, le righe rimanenti (dalla 2 alla 1001) contengono i dati, esattamente così come appaiono nella tabella del menù principale di Ministat. Ecco, a titolo di esempio, come appaiono, quando siano visualizzati con un qualsiasi editor di testi ASCII, i primi 70 caratteri delle prime 6 righe del file IGA.MI1, un file dimostrativo in formato Ministat che si trova nella directory nella quale è stato installato il programma:

Normali	NCAH	CPH	CAH	AC
1,22	7,44	2,45	2,35	3,51
2,81	4,58	1,63	3,21	4,23
4,02	3,71	3,44	3,88	7,66
2,23	4,94	2,47	1,56	9,54
1,96	4,32	5,78	4,56	2,14

Il formato dati impiegato da Ministat per il salvataggio dei dati è talmente semplice che qualsiasi utente mediamente esperto è in grado di utilizzare direttamente i file di Ministat per esportare i

[1] *Taylor JR Introduzione all'analisi degli errori. Bologna: Zanichelli, 1990:223pp.*

dati verso altre applicazioni, ovvero di importare i dati dei suoi applicativi in Ministat (per questo sono sufficienti le informazioni qui fornite ed eventualmente un editor di testi ASCII). Tuttavia con Ministat viene fornito un ulteriore strumento per garantire la completa standardizzazione nel trasferimento dei dati. L'opzione Esporta File compresa nel Menù File consente infatti di esportare i dati con due modalità (file ASCII in formato fisso e file ASCII in formato delimitato) assolutamente standard e accessibili a qualsiasi altro programma (in pratica tutti i tabelloni elettronici, i database e i programmi di wordprocessing sia in ambiente Windows® che in ambiente DOS® sono forniti delle opzioni necessarie per importare dati indifferentemente da file ASCII dell'uno e dell'altro tipo).

1.2.4. Selezione dei dati da elaborare

Per selezionare i dati da elaborare è necessario evidenziare, nella tabella dati di Ministat, le/e colonna/e contenente/i i dati stessi.
Per selezionare una sola colonna (variabile) è sufficiente fare click con il tasto sinistro del mouse sulla cella contenente il nome della variabile stessa: tutta la colonna verrà evidenziata e continuerà a rimanere evidenziata finche non si deselezionerà la colonna facendo click su una cella qualsiasi. La selezione di una colonna determina l'inattivazione di tutte le voci dei menù che non risultano più pertinenti: in particolare risulteranno inattivate tutte le voci che corrispondono ad operazioni (calcoli, statistiche o elaborazioni grafiche) che possono essere eseguite solamente su più variabili.
Per selezionare più colonne (variabili) è possibile procedere in due modi:

- fare click con il tasto sinistro del mouse sulla cella contenente il nome della variabile corrispondente alla prima colonna da selezionare, tenere il tasto del mouse premuto, portarsi sulla cella contenente il nome della variabile corrispondente all'ultima colonna da selezionare, e rilasciare il tasto del mouse;
- fare click con il tasto sinistro del mouse sulla cella contenente il nome della variabile corrispondente alla prima colonna da selezionare, rilasciare il tasto del mouse (la colonna risulterà selezionata), premere il tasto SHIFT e, tenendolo premuto (il tasto SHIFT è quello che serve a generare le lettere maiuscole), premere il tasto FRECCIA A DESTRA tante volte quante sono quelle necessarie a selezionare tutte le colonne (variabili) aggiuntive desiderate.

La selezione di più colonne (variabili) determina l'inattivazione di tutte le voci dei menù che non risultano più pertinenti: in particolare risulteranno inattivate tutte le voci che corrispondono ad operazioni (calcoli, statistiche o elaborazioni grafiche) che possono essere eseguite solamente su una colonna (variabile). Qualsiasi sia il numero delle colonne selezionate, vengono comunque sempre inattivate tutte le voci del Menù File.
Per l'effettuazione delle rappresentazioni grafiche e dei test statistici, Ministat parte dalla prima riga (riga 1) e considera terminati i dati da elaborare al raggiungimento della prima cella vuota (eventuali dati successivi vengono ignorati). Per questo motivo l'inserimento di una cella vuota all'interno di una colonna (variabile) rappresenta un utile mezzo per limitare il numero dei dati da sottoporre ad elaborazione. L'inserimento di una riga vuota nella tabella dati determina un risultato analogo, esteso contemporaneamente a tutte le colonne (variabili) della tabella dati.
Per selezionare contemporaneamente tutte le colonne fare click con il tasto sinistro del mouse sulla cella nell'angolo in alto a sinistra.
Per selezionare una riga fare click con il tasto sinistro del mouse sulla cella contenente il numero della riga. È possibile selezionare solamente una singola riga per volta.

1.2.5. Limiti nell'elaborazione dei dati

Il primo limite nell'elaborazione dei dati è ovviamente quello derivante dal numero massimo di colonne e di righe contenute nella tabella del menù principale di Ministat: quindi 10 variabili con 1000 dati ciascuna.
Tutte la rappresentazioni grafiche e i calcoli e la maggior parte dei test statistici possono essere eseguiti sul numero massimo di dati previsti. Limiti più restrittivi possono determinarsi per alcuni particolari test statistici, che implicano l'utilizzo di risorse di memoria veramente imponenti. In particolare alcune opzioni del Menù Statistica, come la Regressione Multipla, e alcune opzioni della Statistica Non Parametrica, come l'Analisi della Somiglianza (cluster analysis) e la Regressione Lineare Non Parametrica, comportano l'impiego di matrici le cui dimensioni aumentano in ragione geometrica all'aumentare del numero dei dati da analizzare. In questi casi la quantità di memoria disponibile sul sistema utilizzato e l'ambiente nel quale è stato sviluppato Ministat possono drasticamente limitare il numero di dati elaborabili. In ogni caso viene tentata l'elaborazione dei dati, ed eventualmente viene segnalata l'impossibilità di portarla a termine.

1.2.6. Messaggi di errore

Molti errori possono essere intercettati dal sistema operativo: in questo caso i messaggi di errore sono generati da Windows®, la cui guida fornisce l'aiuto necessario per interpretare sia le cause sia i relativi rimedi. Tra gli errori intercettati dal sistema operativo vi sono tipicamente quelli determinati dalle periferiche: stampanti non in linea, dischetti non inseriti nel drive, dischetti protetti da scrittura o danneggiati, e così via.
Gli errori non intercettati dal sistema operativo sono gestiti da Ministat, che genera i relativi messaggi di errore. Spesso questi errori hanno origine banale, come quando per esempio si cerca di calcolare la regressione lineare su due variabili che non contengono lo stesso numero di dati. Occasionalmente tuttavia questi errori possono riflettere problemi più complessi. In particolare questo può accadere quando si utilizza la funzione che consente di importare dati da database esterni nella tabella di Ministat (vedere File: Importa File). Le difficoltà sono sostanzialmente riconducibili a tre situazioni: (1) i motivi più banali e più frequenti di difficoltà sono costituiti da file danneggiati, per danni al mezzo fisico (supporto magnetico) o per danni logici dei dati (tabella di allocazione dei file [FAT], file-system, settori del disco o dischetto). Si raccomanda come prima cosa di controllare il disc(hett)o mediante una utility del sistema operativo (come SCANDISK), che consente di identificare (e il più delle volte correggere) tali danni, e che oltretutto sarebbe buona norma utilizzare a scadenza regolare per assicurare la manutenzione del disco. Come seconda cosa utilizzare una utility per escludere la presenza di virus (sempre possibile); (2) in una tabella di database (ovvero in un database, se questo è strutturato come un'unica tabella), al fine di non consentire ambiguità, non vi possono essere due campi con lo stesso nome. Alcuni tabelloni elettronici consentono di manipolare i dati e di salvarli in formato dBase III® (file con estensione .DBF) senza un congruo controllo sui nomi dei campi: Ministat non consente di importare dati da file (anche se apparentemente nel formato corretto) con nomi di campi duplicati; (3) Windows® è un sistema operativo in continua evoluzione, e questo può creare problemi di compatibilità nella gestione di applicativi che utilizzano versioni diverse dei driver del sistema operativo. In particolare l'impossibilità di importare file in formato dBase III® potrebbe derivare da conflitti di questo genere. In questo caso, dopo avere accuratamente escluso tutte le possibili cause sopra elencate, mediante un editor di testi ASCII, accedere alla directory nella quale è stato installato Ministat, aprire il file MINISTAT.INI, togliere il punto e virgola (;) dalla riga in cui si trova

[Installable ISAMs]
dBASE III=C:\WINDOWS\SYSTEM\xbs110.dll
;dBASE III=C:\WINDOWS\SYSTEM\xbs200.dll

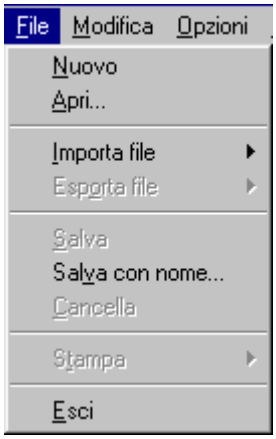
e aggiungerlo alla riga immediatamente precedente

[Installable ISAMs]
;dBASE III=C:\WINDOWS\SYSTEM\xbs110.dll
dBASE III=C:\WINDOWS\SYSTEM\xbs200.dll

In questo modo viene attivato un driver che consente di risolvere il problema.

1.3. Menù di Ministat

1.3.1. Menù File



Questo menù comprende tutte le opzioni necessarie per importare, salvare, esportare e stampare i file di dati. La selezione di una riga o di una o più colonne della tabella dati determina l'inattivazione di tutte le voci del Menù File.

- Le opzioni Esporta File, Salva, Cancella e Stampa risultano attive solamente dopo che i dati della tabella sono stati salvati come file in formato Ministat.

1.3.1.1. Nuovo

Consente di svuotare completamente la tabella dati di Ministat, al fine di preparare un nuovo file in formato Ministat. Impiegare questa opzione solamente quando si desidera immettere i dati da tastiera: quando si importano i dati da un database esterno, la tabella dati di Ministat viene svuotata automaticamente. In ogni caso viene richiesta la conferma, in quanto i dati preesistenti, se non salvati su disco in un file, sono persi. Le opzioni Esporta File, Salva, Cancella e Stampa vengono disattivate.

1.3.1.2. Apri

Apre un file in formato Ministat (file con estensione .MI1), e lo carica nella tabella dati di Ministat: i dati eventualmente preesistenti nella tabella sono persi se non sono stati in precedenza salvati su disco in un file (vedere File: Salva con Nome). Una finestra di dialogo standard di Windows® consente di selezionare il disco, la directory e infine il nome del file da aprire. In seguito alla apertura di un file in formato Ministat vengono attivate le opzioni Esporta File, Salva, Cancella e Stampa.

1.3.1.3. Importa file

Consente di importare i dati direttamente da un file/database in formato ASCII, in formato Access® 1.x oppure in formato dBase III®. Come prima cosa una finestra di dialogo standard di Windows® consente di selezionare il disco, la directory e infine il nome del file da importare (per i file ASCII viene assunta l'estensione .txt, per i file Access® 1.x viene assunta l'estensione .mdb, per i file dBase III® viene assunta l'estensione .dbf). Quindi la tabella dati di Ministat viene svuotata, previa richiesta di conferma: i dati della tabella non salvati su disco in un file (vedere: File: Salva con Nome) sono ovviamente persi. A questo punto le procedure differiscono.
Nel caso di file ASCII, per i quali non esiste un formato standard, il file viene caricato dal disco e presentato in una tabella, con un carattere per ciascuna cella della tabella. L'utilizzatore dovrà selezionare manualmente le colonne che contengono i caratteri che costituiscono un campo, quindi, dopo avere selezionato l'opzione Importa Dalle Colonne del Menù File, scegliere in quale colonna/variabile della tabella dati di Ministat importare i valori, selezionando prima tale colonna/variabile, e quindi selezionando l'opzione Importa Nella Colonna del Menù File. È possibile ripetere le operazioni sopra indicate, per importare altri campi, selezionando l'opzione Importa del Menù File, ovvero terminare l'importazione del file selezionando l'opzione Esci del Menù File.
Nel caso di un file/database in formato Access® 1.x, una ulteriore finestra consente di selezionare il nome di una delle tabelle che costituiscono il file/database: verranno quindi importati i dati della sola tabella selezionata.
Nel caso di un database in formato dBase III®, verranno importati tutti i dati del file selezionato, in quanto un file/database dBase III® corrisponde ad una singola tabella di database. I dati vengono importati così come sono: qualora si renda necessario, possono essere riportati al numero di cifre significative gestibile da Ministat (undici più il separatore, sei cifre per gli interi e cinque per i decimali) prima trasformandoli (mediante le opzioni Moltiplica e Dividi del Menù Calcoli) quindi arrotondandoli opportunamente (mediante l'opzione Numero di Decimali del Menù Opzioni).
Qualora si dovessero riscontrare problemi nell'importare i file, consultare la sezione dedicata ai Messaggi di Errore.

1.3.1.4. Esporta file

Questa opzione risulta attiva solamente se è stato aperto un file Ministat (vedere File: Apri), ovvero dopo che i dati della tabella sono stati salvati su disco come file in formato Ministat (vedere File: Salva con Nome). Una finestra di dialogo standard di Windows® consente di selezionare il disco, la directory e infine il nome del file in cui salvare i dati. L'estensione (.TXT) del file viene aggiunta automaticamente. Qualsiasi altra estensione eventualmente fornita non viene considerata, e viene sostituita dal valore predefinito.

È possibile esportare i dati in due tipi di formato ASCII: formato fisso e formato delimitato. Si tratta di due formati assolutamente standard e accessibili a qualsiasi altro programma (in pratica tutti i programmi in ambiente Windows® possono importare dati indifferentemente da file ASCII dell'uno e dell'altro tipo).
I file ASCII in formato fisso contengono 1000 righe di 140 caratteri ciascuna. Ogni riga comprende 10 campi di 14 caratteri. Le righe, dalla 1 alla 1000, contengono i dati, esattamente così come appaiono nella tabella del menù principale di Ministat. Ecco, a titolo di esempio, come appaiono, quando siano visualizzati con un qualsiasi editor di testi ASCII, i primi 70 caratteri delle prime 10 righe del file IGA.MI1, quando sia stato esportato come file ASCII in formato fisso (con un nome a piacere, ed estensione .TXT):

1,22	7,44	2,45	2,35	3,51
2,81	4,58	1,63	3,21	4,23
4,02	3,71	3,44	3,88	7,66
2,23	4,94	2,47	1,56	9,54
2,35	3,49	1,95	1,78	11,35
1,64	3,88	4,56	2,49	6,43
2,08	4,71	7,31	3,11	5,28
1,96	4,32	5,78	4,56	2,14
1,54	4,90	3,40	5,11	4,76

I file ASCII in formato delimitato contengono anch'essi 1000 righe. Le righe, dalla 1 alla 1000, contengono i dati, essendo ogni dato racchiuso entro doppi apici, e i dati separati tra di loro da un virgola. Ecco, a titolo di esempio, come appaiono, quando siano visualizzate con un qualsiasi editor di testi ASCII, le prime 10 righe del file IGA.MI1, quando sia stato esportato come file ASCII in formato delimitato (con un nome a piacere, ed estensione .TXT). Notare come in questo caso il numero di caratteri per riga sia variabile:

```
"1,22","7,44","2,45","2,35","3,51","","","","","","
"2,81","4,58","1,63","3,21","4,23","","","","","","
"4,02","3,71","3,44","3,88","7,66","","","","","","
"2,23","4,94","2,47","1,56","9,54","","","","","","
"2,35","3,49","1,95","1,78","11,35","","","","","","
"1,64","3,88","4,56","2,49","6,43","","","","","","
"2,08","4,71","7,31","3,11","5,28","","","","","","
"1,96","4,32","5,78","4,56","2,14","","","","","","
"1,54","4,90","3,40","5,11","4,76","","","","","","
"1,63","11,43","5,12","2,36","7,91","","","","","","
```

1.3.1.5. Salva

Questa opzione risulta attiva solamente se è stato aperto un file Ministat (vedere File: Apri), ovvero dopo che i dati della tabella sono stati salvati come file in formato Ministat (vedere File: Salva con Nome), e copia su disco il contenuto della tabella dati nel file in formato Ministat (file con estensione .MI1) in uso al momento del salvataggio.

1.3.1.6. Salva con nome

Consente di salvare su disco, in un file in formato Ministat (file con estensione .MI1), i dati della tabella. Una finestra di dialogo standard di Windows® consente di selezionare il disco, la directory e infine il nome del file in cui salvare i dati. L'estensione (.MI1) del file viene aggiunta automaticamente. Qualsiasi altra estensione eventualmente fornita non viene considerata, e viene sostituita dal valore predefinito. Il salvataggio del file in formato Ministat attiva le opzioni Esporta File, Salva, Cancella e Stampa.

1.3.1.7. Cancella

Questa opzione risulta attiva solamente se è stato aperto un file Ministat (vedere File: Apri), ovvero dopo che i dati della tabella sono stati salvati come file in formato Ministat (vedere File: Salva con Nome). Dopo una richiesta di conferma, la tabella dati di Ministat viene vuotata e il file Ministat in uso viene cancellato dal disco: i dati sono irrimediabilmente persi.

1.3.1.8. Stampa


Questa opzione, che risulta attiva solamente se è stato aperto un file Ministat (vedere File: Apri), ovvero dopo che i dati della tabella sono stati salvati come file in formato Ministat (vedere File: Salva con Nome), consente di stampare la tabella dati di Ministat (le opzioni per la stampa dei risultati dell'elaborazione grafica e dell'elaborazione statistica sono contenuti all'interno dei menù delle specifiche elaborazioni).

- È possibile stampare i dati della tabella in due parti: la prima comprendente le colonne (variabili) da 1 a 5, la seconda comprendente le colonne (variabili) da 6 a 10.

In entrambi i casi la stampa parte dalla prima riga (riga 1) e si interrompe quando viene incontrata la prima riga completamente vuota. Per questo motivo l'inserimento di una riga vuota all'interno della tabella dati di Ministat (vedere l'opzione Riga del Menù Modifica) può rappresentare un utile mezzo per limitare il numero dei dati da stampare.

1.3.1.9. Esci

Termina l'esecuzione di Ministat. Viene richiesta conferma in quanto i dati non salvati vengono persi.

1.3.2. Menù Modifica

Comprende una serie di opzioni che conferiscono una notevole efficienza nella gestione della tabella dei dati, consentendo di spostare le variabili all'interno della tabella, di ordinare i dati in ordine numerico crescente o decrescente, di aggiungere e togliere colonne, righe o anche singole celle.

1.3.2.1. Taglia colonna

Taglia la colonna (variabile) selezionata. La differenza rispetto all'opzione Copia Colonna è rappresentata dal fatto che la colonna originaria viene vuotata. I dati potranno poi essere trasferiti in un'altra colonna con l'opzione Incolla Colonna. È possibile utilizzare questa opzione anche per accorpare in un'unica tabella variabili contenute in file differenti, nel seguente modo:

- aprire un file Ministat o importare un file dBase III® o Access®;
- tagliare la colonna (variabile) che interessa;
- aprire un nuovo file Ministat; (4) incollare la colonna (variabile) nel nuovo file;
- salvare il file; aprire il primo file e tagliare una nuova colonna;
- aprire il secondo file e incollare la colonna;
- salvare il file;
- ripetere i tre punti precedenti, tagliando via via le varie colonne dal primo file e incollandole nel secondo.

1.3.2.2. Copia colonna

Copia la colonna (variabile) selezionata. La differenza rispetto all'opzione Taglia Colonna è rappresentata dal fatto che la colonna originaria rimane immodificata. I dati potranno poi essere trasferiti in un'altra colonna con l'opzione Incolla Colonna. È possibile utilizzare questa opzione anche per accorpare in un'unica tabella variabili contenute in file differenti, nel seguente modo:

- aprire un file Ministat o importare un file dBase III® o Access®;
- copiare la colonna (variabile) che interessa;
- aprire un nuovo file Ministat;
- incollare la colonna (variabile) nel nuovo file;
- salvare il file;
- aprire il primo file e copiare una nuova colonna;
- aprire il secondo file e incollare la colonna;
- salvare il file;

- ripetere i tre punti precedenti, copiando via via le varie colonne dal primo file e incollandole nel secondo.

1.3.2.3. Incolla colonna

Incolla la colonna (variabile) precedentemente tagliata (vedere Taglia colonna) o copiata (vedere Copia colonna) sulla colonna selezionata come destinazione. Eventuali dati presenti nella colonna destinazione verranno persi, sovrascritti dai dati della colonna incollata.

1.3.2.4. Annulla incolla

Dopo l'utilizzo delle opzioni Taglia Colonna e Copia Colonna, i dati della colonna (variabile) tagliata o copiata restano in attesa di una destinazione, per cui la successiva selezione di una colonna viene interpretata da Ministat come indicazione della destinazione dei dati tagliati o copiati, con il risultato di bloccare qualsiasi altra operazione. L'opzione Annulla Incolla consente di rinunciare all'operazione Incolla, riabilitando le opzioni dei menù altrimenti rese inaccessibili.

1.3.2.5. Cella

L'opzione Cella risulta attiva solamente se non sono state selezionate né una riga né una o più colonne, e agisce sulla cella selezionata dal puntatore (cella attiva).

- L'opzione Vuota consente di vuotare la cella, lasciando inalterato il resto della tabella dati.
- L'opzione Elimina consente di eliminare la cella: la cella immediatamente successiva viene spostata al posto della cella eliminata, e così via fino all'ultima cella della colonna.
- L'opzione Inserisci consente di inserire una nuova cella, vuota, al posto della cella selezionata. La cella selezionata viene spostata di una posizione in giù, per fare posto alla nuova cella, e così via fino all'ultima cella della colonna, il cui contenuto viene ovviamente perso.

Poiché Ministat, per l'effettuazione delle rappresentazioni grafiche e dei test statistici, parte dalla prima riga (riga 1) della tabella e considera terminati i dati al raggiungimento della prima cella vuota (eventuali dati successivi vengono ignorati), l'inserimento di una cella vuota all'interno di una colonna (variabile) rappresenta un utile mezzo per limitare il numero dei dati da elaborare per quella colonna (variabile).

1.3.2.6. Colonna

L'opzione Colonna risulta attiva solamente se è stata selezionata una colonna.

- L'opzione Vuota consente di vuotare la colonna, lasciando inalterato il resto della tabella dati.
- L'opzione Elimina consente di eliminare una colonna: la colonna immediatamente successiva viene spostata al posto della colonna eliminata, e così via fino all'ultima colonna.
- L'opzione Inserisci consente di inserire una nuova colonna, vuota, al posto della colonna selezionata. La colonna selezionata viene spostata di una posizione a destra, per fare posto alla nuova colonna, e così via fino all'ultima colonna, il cui contenuto viene ovviamente perso.

1.3.2.7. Riga

L'opzione Riga risulta attiva solamente se è stata selezionata una riga.

- L'opzione Vuota consente di vuotare la riga, lasciando inalterato il resto della tabella dati.
- L'opzione Elimina consente di eliminare la riga: la riga immediatamente successiva viene spostata al posto della riga eliminata, e così via fino all'ultima riga.

- L'opzione Inserisci consente di inserire una nuova riga, vuota, al posto della riga selezionata. La riga selezionata viene spostata di una posizione in giù, per fare posto alla nuova riga, e così via fino all'ultima riga, il cui contenuto viene ovviamente perso.

Poiché Ministat, per l'effettuazione delle rappresentazioni grafiche e dei test statistici, parte dalla prima riga (riga 1) e considera terminati i dati di ciascuna colonna (variabile) al raggiungimento della prima cella vuota (eventuali dati successivi vengono ignorati), l'inserimento di una riga vuota nella tabella dati rappresenta un utile mezzo per limitare contemporaneamente su tutte le colonne (variabili) il numero dei dati da elaborare.

1.3.2.8. Unisci dati

Consente di unire i dati di due colonne (variabili).

- È necessario selezionare due colonne adiacenti.

I dati delle due colonne sono uniti in una terza colonna, che viene inserita automaticamente immediatamente sulla sinistra delle due colonne selezionate, e che contiene, in immediata successione e nelle sequenze originali, i dati della prima e quelli della seconda colonna. Per effetto dell'inserimento della nuova colonna il contenuto della colonna 10 viene perso.

1.3.2.9. Scambia colonne

Consente di scambiare tra di loro due colonne (variabili).

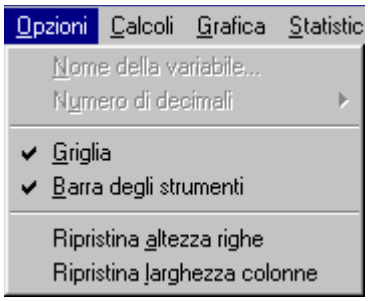
- È necessario selezionare due colonne adiacenti.

Al termine dell'operazione le due colonne saranno state scambiate di posizione tra di loro: quella di sinistra si troverà nella posizione di quella di destra e viceversa.

1.3.2.10. Ordina dati


Consente di ordinare In Ordine Crescente oppure In Ordine Decrescente i dati della colonna (variabile) selezionata. Al fine di mantenere l'allineamento dei dati appartenenti ad una stessa riga (che viene quindi intesa come un record), i dati delle altre colonne (variabili) della tabella di Ministat seguono l'ordinamento della colonna selezionata. In questo modo i dati appaiati continuano a rimanere tali anche in seguito al/i processo/i di ordinamento. Prima di ordinare i dati è consigliabile numerare le righe, al fine di potere successivamente ripristinare l'ordine originario.

1.3.3. Menù Opzioni



Contiene le opzioni necessarie per gestire il nome delle variabili e il numero di decimali con cui rappresentare i dati, e le opzioni necessarie per gestire e ripristinare l'aspetto della tabella dati di Ministat.

1.3.3.1. Nome della variabile

Consente di immettere e/o correggere il nome della/e variabile/i corrispondente alla colonna (variabile) selezionata. Nel caso di dati importati da database esterni, come nome di ciascuna variabile viene assunto il nome del campo importato, eventualmente troncato a 12 caratteri. Nel caso di un nuovo file Ministat (vedere l'opzione Nuovo del Menù File), le colonne (variabili) sono denominate automaticamente con i valori predefiniti di Variabile 1 (prima colonna), Variabile 2 (seconda colonna), e così via, che possono poi ovviamente essere modificati a piacere. Il nome della variabile non può superare i 12 caratteri di lunghezza.

1.3.3.2. Numero di decimali

Consente di arrotondare al numero di decimali desiderato i valori della colonna (variabile) selezionata. Questa operazione si rende sempre necessaria quando si importano da database esterni dati rappresentati con un numero di decimali superiore a cinque (questo avviene tipicamente per campi di database che contengono valori derivati, come per esempio il rapporto tra i valori di due campi). In questi casi i dati sono importati così come sono, e devono essere ricondotti al numero di cifre significative gestibili da Ministat (undici più il separatore, sei cifre per gli interi e cinque per i decimali) arrotondandoli. L'arrotondamento comporta la perdita delle cifre in eccesso. Per questo motivo può rendersi necessario effettuare preliminarmente una riconversione delle scale. Per esempio, se viene importato il dato 0,0015437, al fine di conservare tutte le cinque cifre (ammesso che siano tutte e cinque significative) mantenendo un numero limitato di decimali, può essere opportuno moltiplicare il dato per 1000. Il Menù Calcoli contiene le opzioni Moltiplica e Dividi, che consentono di trasformare i valori delle variabili importate al fine di rappresentarli in modo congruo. Per il problema del numero delle cifre significative con cui rappresentare un dato vedere quanto riportato al capitolo 3 e anche Taylor[2].

1.3.3.3. Griglia

Consente di attivare o disattivare la visualizzazione della griglia della tabella dati. Il valore predefinito comporta la visualizzazione della griglia.

1.3.3.4. Barra degli strumenti

Consente di attivare o disattivare la visualizzazione della barra degli strumenti. Il valore predefinito comporta la visualizzazione della barra degli strumenti.

[2] *Taylor JR Introduzione all'analisi degli errori. Bologna: Zanichelli, 1990:223pp.*

1.3.3.5. Ripristina altezza righe

Posizionando il puntatore del mouse nelle celle al margine sinistro della tabella (quelle che contengono il numero delle righe), è possibile ridimensionare l'altezza della riga. Per fare questo posizionarsi esattamente sulla linea che separa due righe contigue: il puntatore del mouse cambierà di forma. Fare click sul tasto sinistro e, mantenendo la pressione sul tasto, spostarsi in su o in giù fino a conferire alla riga l'altezza desiderata, quindi rilasciare il tasto del mouse. L'opzione Ripristina Altezza Righe consente di riportare automaticamente tutte le righe all'altezza predefinita. L'opzione Ripristina Larghezza Colonne consente di riportare automaticamente alla larghezza predefinita tutte le colonne.

1.3.3.6. Ripristina larghezza colonne

Posizionando il puntatore del mouse nelle celle al margine superiore della tabella (quelle che contengono il nome della colonna (variabile)), è possibile ridimensionare la larghezza della colonna. Per fare questo posizionarsi esattamente sulla linea che separa due colonne contigue: il puntatore del mouse cambierà di forma. Fare click sul tasto sinistro e, mantenendo la pressione sul tasto, spostarsi a sinistra o a destra fino a conferire alla colonna la larghezza desiderata, quindi rilasciare il tasto del mouse. L'opzione Ripristina Larghezza Colonne consente di riportare automaticamente tutte le colonne alla larghezza predefinita. L'opzione Ripristina Altezza Righe consente di riportare automaticamente all'altezza predefinita tutte le righe.

1.3.4. Menù Calcoli

Le opzioni di questo menù consentono di effettuare una serie di utili trasformazioni numeriche sui valori delle colonne (variabili) della tabella dati. A seconda dei casi, è necessario selezionare una o due colonne. In tutti i casi i valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della/e colonna/e selezionata/e: questo consente di salvaguardare i dati originari. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi: un apposita finestra di dialogo avverte l'utilizzatore e gli consente di rinunciare eventualmente all'operazione. A causa

dell'inserimento della nuova colonna, le opzioni del menu Calcoli non possono ovviamente essere effettuate oltre la colonna 9. Infine le trasformazioni richieste sono applicate solamente alle celle contenenti dati.

1.3.4.1. Somma

Consente di sommare ai dati della colonna (variabile) selezionata un numero positivo.

- Questa opzione risulta attiva quando viene selezionata una sola colonna.

I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.2. Sottrai

Consente di sottrarre ai dati della colonna (variabile) selezionata un numero positivo.

- Questa opzione risulta attiva quando viene selezionata una sola colonna.

I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.3. Moltiplica

Consente di moltiplicare i dati della colonna (variabile) selezionata per un numero positivo.

- Questa opzione risulta attiva quando viene selezionata una sola colonna.

I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.4. Dividi

Consente di dividere i dati della colonna (variabile) selezionata per un numero positivo. Questa opzione risulta attiva quando viene selezionata una sola colonna. I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.5. Quadrato

Calcola il quadrato dei dati della colonna (variabile) selezionata.

- Questa opzione risulta attiva quando viene selezionata una sola colonna.

I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.6. Radice

Calcola la radice quadrata dei dati della colonna (variabile) selezionata.

- Questa opzione risulta attiva quando viene selezionata una sola colonna.

I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.7. Logaritmo

Calcola il logaritmo naturale (logaritmo neperiano, in base *e* con *e* = 2,71828) oppure il logaritmo decimale (logaritmo in base 10) dei dati della colonna (variabile) selezionata.

- Questa opzione risulta attiva quando viene selezionata una sola colonna.

I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.8. Differenza

Calcola la differenza tra i dati della prima e quelli della seconda di due colonne (variabili) selezionate.

- Questa opzione risulta attiva quando vengono selezionate due colonne.

I valori calcolati sono riportati in una colonna inserita ex-novo alla sinistra della prima delle due colonne selezionate. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.9. Media

Calcola la media aritmetica (somma diviso due) o geometrica (radice quadrata del prodotto) dei dati della prima e quelli della seconda di due colonne (variabili) selezionate.

- Questa opzione risulta attiva quando vengono selezionate due colonne.

I valori calcolati sono riportati in una colonna inserita ex-novo alla sinistra della prima delle due colonne selezionate. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.10. Rapporto

Calcola il rapporto tra i dati della prima e quelli della seconda di due colonne (variabili) selezionate.

- Questa opzione risulta attiva quando vengono selezionate due colonne.

I valori calcolati sono riportati in una colonna inserita ex-novo alla sinistra della prima delle due colonne selezionate. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.11. Cambia il segno

Cambia il segno dei dati della colonna (variabile) selezionata. Per cambiare di segno il contenuto di una singola cella (un singolo dato) fare doppio click sulla cella.

- Questa opzione risulta attiva quando viene selezionata una sola colonna.

I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.12. Valore assoluto

Toglie il segno ai dati della colonna (variabile) selezionata.

- Questa opzione risulta attiva quando viene selezionata una sola colonna.

I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.13. Numera

Numera le righe, riportandone il numero di posizione.

➠ Questa opzione risulta attiva quando viene selezionata una sola colonna.

Il numero di posizione di ogni riga viene riportato in una colonna inserita ex-novo alla sinistra della colonna selezionata, e denominata "Numero riga". Con l'inserimento della nuova colonna i dati della colonna 10 sono persi. Eseguire questa opzione prima di ordinare i dati. Infatti riordinandoli successivamente in base al numero di riga della colonna "Numero riga" sarà possibile ripristinare l'ordine originario .

1.3.4.14. Deviata z

Calcola la deviata normale standardizzata *z* dei dati della colonna selezionata. La deviata normale standardizzata *z* viene calcolata per ogni singolo dato come

z = (valore del dato - media dei dati) / deviazione standard dei dati

➠ Questa opzione risulta attiva quando viene selezionata una sola colonna.

Il valore della deviata normale standardizzata viene riportato in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.15. Rango

Calcola il rango, cioè il numero di posizione dei singoli dati nella lista dei dati ordinati in ordine numerico crescente. Quando due dati sono uguali, a ciascuno viene assegnato come numero di posizione la media dei numeri di posizione dei dati che risultano uguali. Così, per esempio, se il quinto e il sesto dato, in una lista ordinata, sono tutti e due uguali, e pari a 223, il rango del quinto e del sesto dato sarà uguale, e pari a 5,5. Il procedimento viene esteso anche al caso in cui più di due dati siano uguali. Per il metodo vedere Snedecor[3].

➠ Questa opzione risulta attiva quando viene selezionata una sola colonna.

Il rango viene riportato in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.4.16. Percentili

Calcola il percentile, cioè la posizione del dato all'interno della lista dei dati ordinati in ordine numerico crescente, espressa come frazione percentuale. Essendo *n* il numero dei dati, il percentile *P* (p-esimo) viene calcolato come

$$P = 100 \cdot rango / (n + 1)$$

[3] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:135-148.*

Per il metodo vedere Snedecor[4].

➨ Questa opzione risulta attiva quando viene selezionata una sola colonna.

I percentili sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

1.3.5. Menù Grafica

Le opzioni di questo menù consentono di effettuare una serie di interessanti e utili rappresentazioni grafiche dei dati, in aggiunta a quelle già disponibili all'interno delle opzioni del Menù Statistica. Per l'importanza della rappresentazione grafica vedere Bossi[5], Lantieri[6] e Campbell[7].

1.3.5.1. Torta

Consente di rappresentare la distribuzione dei valori di una riga di dati, relativi ad un numero compreso tra due e dieci variabili (colonne), sotto forma di un diagramma a torta.

➨ Questa opzione risulta attiva quando viene selezionata una sola riga.

---

[4] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:135-148.*

[5] *Bossi A, Cortinovis I, Duca P, Marubini E. Introduzione alla statistica medica. Roma: La Nuova Italia Scientifica, 1994:37-68.*

[6] *Lantieri PB, Risso D, Rovida S, Ravera G. Statistica medica ed elementi di informatica. Milano: McGraw-Hill, 1994:71-101.*

[7] *Campbell MJ, Machin D. Medical statistics. A commonsense approach. Chichester: John Wiley & Sons, 1993:44-59.*

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III® del Menù File). Eliminare la prima colonna, che contiene l'età dei soggetti. Selezionare la prima riga, e osservare la ripartizione dei lipidi totale nel siero nelle quattro principali classi che li compongono, colesterolo totale, colesterolo HDL, trigliceridi e colesterolo LDL, nel soggetto in questione..

- L'opzione Diagramma a Torta consente di scegliere tra una rappresentazione a due dimensioni e una a tre dimensioni del diagramma. In entrambi i casi viene esploso lo spicchio corrispondente alla variabile che compare nella prima colonna.
- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende non è prevista.

1.3.5.2. Istogramma

Consente di rappresentare la distribuzione dei valori di una variabile singola sotto forma di istogramma.

- Questa opzione risulta attiva quando viene selezionata una sola colonna (variabile).
- Per visualizzare oltre all'istogramma anche le statistiche elementari parametriche utilizzare l'opzione Statistiche Elementari del Menù Statistica.

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III® del Menù File). Selezionare la colonna COLEST, e osservare la distribuzione del colesterolo totale (in mg/dL, in ascisse) in un gruppo di soggetti non selezionati.

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.5.3. Cumulativa

Consente di rappresentare la distribuzione dei valori di una variabile singola sotto forma di curva cumulativa.

- Questa opzione risulta attiva quando viene selezionata una sola colonna (variabile).

- Per visualizzare oltre alla distribuzione cumulativa anche le statistiche elementari non parametriche utilizzare l'opzione Statistica Non Parametrica del Menù Statistica e selezionare le Statistiche Elementari.

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III® del Menù File). Selezionare la colonna COLEST, e osservare la distribuzione del colesterolo totale (in mg/dL, in ascisse) in un gruppo di soggetti non selezionati. È possibile rappresentare la distribuzione cumulativa sia nella forma tradizionale che nella forma ripiegata proposta da Krouwer[8].

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.5.4. Confronto

Consente di rappresentare, in uno stesso grafico, la distribuzione dei valori di due variabili, sotto forma di due istogrammi.

- Questa opzione risulta attiva quando vengono selezionate due colonne (variabili). La prima delle due colonne selezionate corrisponde all'istogramma riportato superiormente, la seconda corrisponde all'istogramma riportato inferiormente.
- La rappresentazione viene effettuata anche se le due colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File). Selezionare le colonne COLEST e HDL, e osservare la distribuzione del colesterolo totale e del colesterolo HDL (entrambi in mg/dL) in un gruppo di soggetti non selezionati.

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

[8] *Krouwer JS, Monti KL. A simple, graphical method to evaluate laboratory assays. Eur J Clin Chem Clin Biochem 1995;33:525-7.*

1.3.5.5. Appaiamento

Consente di rappresentare i valori di due variabili appaiate (paired-dots plot). Tipicamente questo si realizza quando lo stesso fenomeno viene misurato in condizioni che differiscono tra di loro per una sola caratteristica, introdotta dal ricercatore mediante il disegno sperimentale.

- Questa opzione risulta attiva quando vengono selezionate due colonne (variabili).
- I valori della prima delle due colonne selezionate sono riportati sulla sinistra, quelli della seconda sulla destra.
- La rappresentazione viene effettuata solamente se le due colonne (variabili) contengono lo stesso numero di dati.

Utilizzare la tabella AST del file ESEMPI.MDB: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Contiene il valore della AST (aspartato amminotransferasi) determinata su sieri freschi (colonna SUBITO), e rideterminata sugli stessi sieri dopo conservazione a +4 °C per 24 ore (colonna DOPO24ORE), al fine di stabilire se tale conservazione sia idonea a mantenere l'attività dell'enzima. Selezionare entrambe le colonne. L'andamento generale delle coppie di valori (in U/L), valutato alla luce dell'imprecisione analitica che caratterizza un metodo per la determinazione dell'AST, sembra suggerire che mediamente non vi siano differenze tra i due valori. Provare a verificare tale conclusione utilizzando il Confronto tra Medie.

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.5.6. Dispersione

Consente di rappresentare in una sistema cartesiano una serie di punti di coordinate (x,y).

- Questa opzione risulta attiva quando vengono selezionate due colonne (variabili). La prima delle due colonne selezionate corrisponde ai valori delle ascisse, la seconda corrisponde ai valori delle ordinate dei singoli punti.
- La rappresentazione viene effettuata solamente se le due colonne (variabili) contengono lo stesso numero di dati.

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File). Selezionare le colonne ETA (manca l'accento, che non può essere incluso nel nome di un campo) e COLEST, e osservare la relazione che esiste tra colesterolo totale (in mg/dL, in ordinate) ed età (in anni, in ascisse) in un gruppo di soggetti non selezionati.

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.5.7. Altman

Consente di rappresentare in una sistema cartesiano i dati di due variabili appaiate, riportando sull'asse delle ascisse la media dei valori (cioè (valore della prima variabile + valore della seconda variabile) / 2), e sull'asse delle ordinate la loro differenze (cioè valore della prima variabile - valore della seconda variabile).

- Questa opzione risulta attiva quando vengono selezionate due colonne (variabili). La prima delle due colonne selezionate corrisponde ai valori della prima variabile, la seconda corrisponde ai valori della seconda variabile.
- La rappresentazione viene effettuata solamente se le due colonne (variabili) contengono lo stesso numero di dati.

Utilizzare la tabella AST del file ESEMPI.MDB: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Contiene il valore della AST (aspartato amminotransferasi) determinata su sieri freschi (colonna SUBITO), e rideterminata sugli stessi sieri dopo conservazione a +4 °C per 24 ore (colonna DOPO24ORE), al fine di stabilire se tale conservazione sia idonea a mantenere l'attività dell'enzima. Selezionare entrambe le colonne. L'andamento generale delle differenze (in U/L), sembra suggerire che mediamente le differenze non cambino al variare della concentrazione, e che globalmente la differenza media sia praticamente uguale a zero. Provare a verificare tale conclusione utilizzando il Confronto tra Medie. Il metodo è descritto da Altman[9].

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende

1.3.5.8. Associazione

Consente di rappresentare in una sistema cartesiano una serie di punti di coordinate (x,y), e vi associa il valore di una terza variabile.

[9] *Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307-10.*

- Questa opzione risulta attiva quando vengono selezionate tre colonne (variabili). La prima delle tre colonne selezionate corrisponde ai valori delle ascisse, la seconda corrisponde ai valori delle ordinate dei singoli punti. In corrispondenza di ciascun punto viene quindi tracciato un cerchio il cui diametro risulta proporzionale al valore della terza variabile.
- La rappresentazione viene effettuata solamente se le tre colonne (variabili) contengono lo stesso numero di dati.

Utilizzare il file ESEMPI.MDB: caricare la tabella COLEST_F nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Selezionare le colonne ETA (manca l'accento, che non può essere incluso nel nome di un campo), COLEST e HDL, e osservare la relazione che esiste tra colesterolo totale (in mg/dL, in ordinate), età (in anni, in ascisse), e colesterolo HDL (in mg/dL, proporzionale al diametro del cerchio) in un gruppo di soggetti non selezionati, di sesso femminile.

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.5.9. Sovrapposizione

Consente di rappresentare in una sistema cartesiano fino a quattro serie di punti di coordinate (x,y).

- Questa opzione risulta attiva quando vengono selezionate da tre a cinque colonne (variabili). La prima delle colonne selezionate corrisponde ai valori delle ascisse, le successive (fino a un massimo di quattro) corrispondono ai valori delle ordinate dei singoli punti delle diverse serie.
- La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file ESEMPI.MDB: caricare la tabella COLEST_M nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Selezionare le colonne ETA (manca l'accento, che non può essere incluso nel nome di un campo), COLEST, HDL, TRIGLI e LDL, e osservare la distribuzione di colesterolo totale, colesterolo HDL, trigliceridi e colesterolo LDL (età in anni in ascisse, il resto in mg/dL in ordinate) in un gruppo di soggetti non selezionati, di sesso maschile.

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.5.10. Linee spezzate

Consente di rappresentare in una sistema cartesiano fino a quattro serie di punti di coordinate (x,y), unendoli mediante segmenti di retta (linee spezzate).

- Questa opzione risulta attiva quando vengono selezionate da due a cinque colonne (variabili). La prima delle colonne selezionate corrisponde ai valori delle ascisse, le successive (fino a un massimo di quattro) corrispondono ai valori delle ordinate dei singoli punti delle diverse serie.
- La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file VARBIOL.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Contiene la concentrazione del colesterolo totale determinata ripetutamente in alcuni volontari in un periodo di 25 giorni, selezionare le colonne (variabili) da GIORNO a SOGGETTO_D, e osservare l'andamento del colesterolo totale (in mg/dL, in ordinate) nei quattro soggetti, nell'arco di tempo (in giorni, in ascisse) indicato.

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.5.11. Barre verticali

Consente di rappresentare in una sistema cartesiano fino a quattro serie di punti di coordinate (x,y) mediante barre verticali.

- Questa opzione risulta attiva quando vengono selezionate da due a cinque colonne (variabili). La prima delle colonne selezionate corrisponde ai valori delle ascisse, le successive (fino a un massimo di quattro) corrispondono ai valori delle ordinate dei singoli punti delle diverse serie.
- La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Quindi numerare le righe, inserendo il numero d'ordine dei dati nella prima colonna. Selezionare infine questa colonna e le successive quattro colonne denominate Controlli, NCAH, CPH e CAH, che contengono i valori di IgA nel siero in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), e infine in un gruppo di soggetti con epatite cronica attiva (CAH). Osservare il diagramma a barre verticali (tutti i valori sono in g/L).

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.5.12. Istogrammi multipli

Consente di rappresentare la distribuzione dei valori di più variabili (fino a quattro contemporaneamente) sotto forma di istogrammi.

- Questa opzione risulta attiva quando vengono selezionate da due a quattro colonne (variabili).
- La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.
- Per analizzare una a una le singole variabili utilizzare l'opzione Statistiche Elementari del Menù Statistica.

Il file BAYES.DBF contiene i risultati ottenuti in un gruppo di soggetti sani (colonna SANI) e in un gruppo di soggetti malati (colonna MALATI): caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Selezionare entrambe le colonne ed effettuarne la rappresentazione.

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.5.13. Distribuzione

Consente di rappresentare sotto forma di un diagramma a pallini (dot-plot) la distribuzione dei dati di una o più colonne (variabili). Contrariamente al diagramma Quartili e Range, in questo diagramma viene conservata la rappresentazione dei singoli dati.

- Questa opzione risulta attiva quando vengono selezionate da una a quattro colonne (variabili).
- La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Selezionare le colonne Controlli, NCAH, CPH e CAH: contengono i valori di IgA nel siero in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), e infine in un gruppo di soggetti con epatite cronica attiva (CAH). Osservare il diagramma quartili e

range (tutti i valori sono in g/L). Il gruppo dei soggetti con epatite alcolica non cirrogena, pure presentando una distribuzione notevolmente asimmetrica verso i valori alti, è anche quello che presenta in nucleo centrale più omogeneo, e che più si discosta dal gruppo di controllo.

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.5.14. Limiti di confidenza

Consente di rappresentare in forma grafica la media, i limiti di confidenza al 95% della media e i limiti di confidenza al 95% della distribuzione campionaria. La tacca orizzontale inferiore corrisponde al limite di confidenza inferiore della distribuzione campionaria, la tacca subito al di sotto di quella centrale corrisponde al limite di confidenza inferiore della media, la tacca centrale alla media, la tacca subito al di sopra di quella centrale corrisponde al limite di confidenza superiore della media, e infine la tacca orizzontale superiore corrisponde al limite di confidenza superiore della distribuzione campionaria.

- Questa opzione risulta attiva quando vengono selezionate da una a quattro colonne (variabili).
- La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Selezionare le colonne Controlli, NCAH, CPH e CAH: contengono i valori di IgA nel siero in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), e infine in un gruppo di soggetti con epatite cronica attiva (CAH). Osservare il diagramma dei limiti di confidenza (tutti i valori sono in g/L).

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.5.15. Quartili e range

Consente di rappresentare sotto forma di una scatola con i baffi (box-whisker plot) il valore minimo, il primo quartile, la mediana, il terzo quartile e il valore massimo osservati in una o più

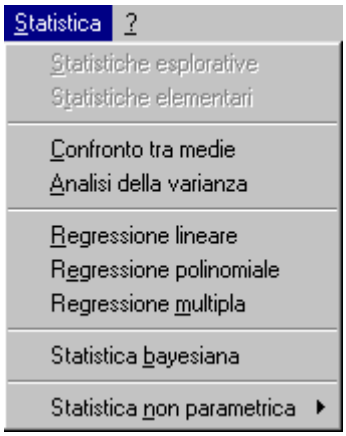
colonne (variabili). La tacca orizzontale inferiore (baffo inferiore) corrisponde al valore minimo, il margine inferiore della scatola al primo quartile, il segmento orizzontale all'interno della scatola alla mediana, il margine superiore della scatola al terzo quartile, la tacca orizzontale superiore (baffo superiore) al valore massimo.

- Questa opzione risulta attiva quando vengono selezionate da una a quattro colonne (variabili).
- La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Selezionare le colonne Controlli, NCAH, CPH e CAH: contengono i valori di IgA nel siero in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), e infine in un gruppo di soggetti con epatite cronica attiva (CAH). Osservare il diagramma quartili e range (tutti i valori sono in g/L). Il gruppo dei soggetti con epatite alcolica non cirrogena, pure presentando una coda notevolmente asimmetrica verso i valori alti, è quello più omogeneo, e che più si discosta dal gruppo di controllo.

- È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
- Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

1.3.6. Menù Statistica



Le opzioni di questo menù consentono di effettuare le elaborazioni statistiche desiderate. Le principali rappresentazioni grafiche (istogramma, distribuzione cumulativa, grafico dei dati e della retta e/o curva sovrapposta, eccetera) sono incluse ciascuna nell'opzione pertinente. Ulteriori interessanti e utili elaborazioni grafiche dei dati possono essere ottenute impiegando le opzioni del Menù Grafica, mentre le opzioni del Menù Calcoli consentono di effettuare sui dati eventuali trasformazioni preliminari. Tutte le opzioni del Menù Statistica presentano i risultati dell'elaborazione in una finestra di testo. È possibile selezionare il testo desiderato e copiarlo negli Appunti

1.3.6.1. Statistiche esplorative

Questa opzione consente di calcolare la media, la deviazione standard, il coefficiente di variazione percentuale (CV), l'errore standard della media e i quartili della distribuzione. Queste statistiche rappresentano il modo in cui esprimere i risultati dell'analisi elementare di dati campionari che siano distribuiti in modo gaussiano. Il metodo è riportato su tutti i testi di statistica, vedere per esempio Armitage[10].
È infine possibile valutare lo scostamento della distribuzione dei dati trovata rispetto alla distribuzione gaussiana teorica ricorrendo al coefficiente di asimmetria g_1 e al coefficiente di curtosi g_2, che indicano il tipo di scostamento come qui di seguito specificato:
$g_1 < 0$: asimmetria negativa, cioè coda sinistra della distribuzione eccessivamente lunga;
$g_1 > 0$: asimmetria positiva, cioè coda destra della distribuzione eccessivamente lunga;
$g_2 < 0$: platicurtosi, cioè distribuzione eccessivamente appiattita, con code troppo corte;
$g_2 > 0$: leptocurtosi, cioè distribuzione eccessivamente alta, con code troppo lunghe.
Un test per la significatività di g_1 e g_2 sufficientemente approssimato per le esigenze pratiche è il seguente: il coefficiente di asimmetria (o quello di curtosi) viene considerato significativo se il rapporto tra esso e il suo errore standard è superiore a 2,6: in questo caso si rigetta l'ipotesi che i dati siano distribuiti in modo gaussiano, e si impiegano le statistiche non parametriche. Il metodo è tratto da Snedecor[11], mentre il test approssimato per la significatività è quello suggerito da Solberg[12].

- L'opzione risulta attiva quando non è stata selezionata alcuna colonna. I calcoli vengono eseguiti automaticamente su tutte le colonne contenenti dati. viene selezionata una sola colonna (variabile).

Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Le colonne Controlli, NCAH, CPH e CAH contengono i valori di IgA nel siero in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), in un gruppo di soggetti con epatite cronica attiva (CAH) e infine in un gruppo di soggetti con cirrosi epatica alcolica (AC).

1.3.6.2. Statistiche elementari

Questa opzione consente di calcolare media, deviazione standard, coefficiente di variazione percentuale (CV) ed errore standard della media, i quartili della distribuzione, una tabella dei percentili più importanti, e una tabella della suddivisione dei dati in classi. È possibile rappresentare anche l'istogramma della distribuzione. Queste statistiche rappresentano il modo in

[10] *Armitage P. Statistica medica. Milano: Feltrinelli Editore, 1979:13-52.*
[11] *Snedecor GW, Cochran WG: Statistical methods. VII Edition. Ames:The Iowa State University Press, Ames, 1980:78-80.*
[12] *Solberg HE. The theory of reference values. Part 5. Statistical treatment of collected reference values - Determination of reference limits. J Clin Chem Clin Biochem 1983;21:749-60.*

cui esprimere i risultati dell'analisi elementare di dati campionari che siano distribuiti in modo gaussiano. Il metodo è riportato su tutti i testi di statistica, vedere per esempio Armitage[13].
Lo scostamento della distribuzione dei dati trovata rispetto alla distribuzione gaussiana teorica viene valutato ricorrendo ancora al coefficiente di asimmetria g_1 e al coefficiente di curtosi g_2.
Il metodo è tratto da Snedecor[14], mentre il test approssimato per la significatività di g_1 e di g_2 è quello suggerito da Solberg[15].

- L'opzione risulta attiva quando viene selezionata una sola colonna (variabile).

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Il calcolo delle Statistiche Elementari sul valore dei trigliceridi (per l'elaborazione selezionare la colonna TRIGLI) porta ad una tabella dei percentili nella quale il 2,5-esimo percentile parametrico e il 97,5-esimo percentile parametrico (rispettivamente la media meno 1,96 volte la deviazione standard e la media più 1,96 volte la deviazione standard) sono pari a -45,3 e 318,3. In base a questo dovremmo affermare che nel 95% dei soggetti esaminati i trigliceridi sono compresi tra -45,3 e 318,3 mg/dL: un'affermazione che, a dispetto del fatto che i risultati sono stati ottenuti con un metodo statistico rigoroso, nessuno si sentirebbe di sostenere. Le cose vanno meglio nel caso delle statistiche non parametriche: il valore non parametrico comparativo riportato nella stessa tabella per il 2,5-esimo e per il 97,5-esimo percentile è pari rispettivamente 44,7 e 382,0 mg/dL. Quello illustrato è, ancorché basato su dati reali, un caso limite, nel quale la contraddizione in termini biologici rappresentata da un valore di concentrazione negativo evidenzia l'errore nell'applicazione della metodologia statistica.

- L'opzione Statistica Non Parametrica del Menù Statistica offre una serie di test statistici non parametrici che consentono di risolvere i principali problemi in tal senso, tra i quali il calcolo delle Statistiche Elementari con metodo non parametrico.
- È possibile stampare sia i risultati dell'elaborazione statistica che l'istogramma mediante l'opzione Stampa del Menù File.
- Per la stampa dell'istogramma si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire la scala delle ordinate più opportuna, di personalizzare le legende e di cambiare il numero delle classi in cui sono suddivisi i dati.
- Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

[13] *Armitage P. Statistica medica. Milano: Feltrinelli Editore, 1979:13-52.*
[14] *Snedecor GW, Cochran WG: Statistical methods. VII Edition. Ames:The Iowa State University Press, Ames, 1980: 78-80.*
[15] *Solberg HE. The theory of reference values. Part 5. Statistical treatment of collected reference values - Determination of reference limits. J Clin Chem Clin Biochem 1983;21:749-60.*

1.3.6.3. Confronto tra medie

Questa opzione consente di effettuare il confronto tra medie sia per dati appaiati che per campioni indipendenti. Il confronto tra medie viene effettuato mediante un test statistico parametrico: il test *t* di Student.

- L'opzione risulta attiva quando vengono selezionate due colonne (variabili).
- Il calcolo del test *t* di Student per dati appaiati è possibile solamente se le due colonne (variabili) contengono lo stesso numero di dati, quello per campioni indipendenti è possibile sempre.

Il confronto fra medie per dati appaiati è un caso particolare del confronto fra medie: infatti l'appaiamento dei dati consente di ottimizzare l'omogeneità fra i due campioni posti a confronto, che differiranno fra di loro solamente per il fattore che lo sperimentatore ha deliberatamente introdotto. Il test *t* di Student per dati appaiati viene in genere considerato significativo quando *p* è inferiore al 5% ($p < 0,05$). Il metodo è tratto da Snedecor[16].

Utilizzare come esempio la tabella AST del file ESEMPI.MDB: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Contiene il valore della AST (aspartato amminotransferasi) determinata su sieri freschi (colonna SUBITO), e rideterminata sugli stessi sieri dopo conservazione a +4 C per 24 ore (colonna DOPO24ORE), al fine di stabilire se tale conservazione sia idonea a mantenere l'attività dell'enzima. Selezionare entrambe le colonne. I risultati del test *t* di Student sembrano suggerire che mediamente non vi siano differenza tra i due valori.

Tuttavia è forse più frequente incontrare situazioni nelle quali i campioni sono indipendenti l'uno dall'altro: si pensi al confronto fra i valori di concentrazione di un qualsiasi analita in maschi e femmine, in adulti e in soggetti in età pediatrica, e via dicendo. Il test *t* di Student per il confronto fra le medie di campioni indipendenti è basato su un principio semplice, e tutto sommato intuitivo: a parità di valore assoluto della differenza tra le medie, a tale differenza viene data tanto maggior peso quanto minore è la dispersione dei dati campionari, e viceversa. Può quindi accadere che una ampia differenza fra le medie, alla quale si sarebbe propensi a dare importanza, risulti non significativa a causa della grande dispersione dei dati campionari, mentre per converso una piccola differenza può rivelarsi inaspettatamente significativa, qualora la dispersione dei dati campionari sia minima. Un problema tuttavia complica lievemente la situazione: il test *t* di Student ordinario tende a fornire troppo pochi risultati significativi quando il campione più grande ha la varianza (cioè la dispersione dei dati) maggiore, e troppi risultati significativi quando il campione più grande ha la varianza minore. Si consiglia perciò di utilizzare, nei casi in cui la varianza dei due campioni differisca in maniera significativa, una forma del test che preveda una opportuna correzione. Se il rapporto tra le varianze dei due campioni indica varianze significativamente diverse (test *F* con $p < 0,05$) utilizzare i risultati del test *t* per varianze non omogenee. Si fa notare che i risultati ottenuti con le due forme del test *t* di Student sono tanto più diversi quanto più differiscono tra di loro le varianze dei due campioni, mentre forniscono lo stesso identico risultato quando le varianze dei due campioni sono uguali. Infine il test *t* di Student per campioni indipendenti viene in genere considerato significativo quando *p* è inferiore al 5% ($p < 0,05$). Il metodo è tratto da Snedecor[17].

[16] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:83-89.*

Il file RIBOFLAV.MI1 (file in formato Ministat: caricarlo nella tabella dati con l'opzione Apri del Menù File) contiene la concentrazione di riboflavina nel fegato (colonna FEGATO) e nel muscolo di bue (colonna MUSCOLO), come determinata in una apposita sperimentazione. Selezionare entrambe le colonne. Il test *F* evidenzia varianze significativamente diverse nei due casi. Il test *t* di Student per varianze non omogenee conferma una differenza nel contenuto di riboflavina dei due tessuti, la cui significatività è pero meno rilevante di quanto apparirebbe a prima vista (*p* solo di poco inferiore a 0,05).

- È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.
- Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

1.3.6.4. Analisi della varianza

Questa opzione consente di effettuare l'analisi della varianza secondo diversi disegni sperimentali.

- Questa opzione risulta attiva quando vengono selezionate due o più colonne (variabili), che devono contenere lo stesso numero di dati.
- Se vengono selezionate due colonne, risulta abilitato il solo calcolo del rapporto tra varianze.
- Se vengono selezionate più di due colonne, il calcolo del rapporto tra varianze risulta disabilitato, mentre vengono abilitati il calcolo dell'analisi della varianza a un fattore, quello dell'analisi della varianza a due fattori e quello delle componenti della variabilità.

Il rapporto tra varianze riportato in questa opzione è identico a quello impiegato per verificare l'omogeneita delle varianza nel confronto tra medie: un test *F* con $p < 0,05$ indica varianze significativamente diverse.

Il file RIBOFLAV.MI1 (file in formato Ministat: caricarlo nella tabella dati con l'opzione Apri del Menù File) contiene la concentrazione di riboflavina nel fegato (colonna FEGATO) e nel muscolo (colonna MUSCOLO) di bue, come determinata in una apposita sperimentazione. Selezionare le due colonne (variabili) ed effettuare i calcoli: il test *F* evidenzia varianze significativamente diverse nei risultati ottenuti sui due tessuti.

Con le varie forme del test *t* di Student (test parametrico) e del test di Wilcoxon (test non parametrico) è possibile confrontare fra di loro due medie: l'analisi della varianza o ANOVA (da ANalysis Of VAriance) a un fattore consente di estendere il confronto a più di due medie.

Il file DIETA.DBF contiene la riduzione dei trigliceridi nel siero osservata in quattro gruppi di soggetti sottoposti a differenti diete. Selezionare le colonne da DIETA_A a DIETA_D: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. L'analisi della varianza a un fattore documenta una differenza significativa tra le riduzioni della concentrazione dei trigliceridi ottenute con le quattro diete. L'ispezione dei dati consente di

[17] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:89-99.*

attribuire alla DIETA_D la differenza significativa. Il metodo è trattato in tutti i testi di statistica; molto chiaro, fra gli altri Wonnacott[18].

Mentre l'ANOVA un fattore (detta anche "a un criterio di classificazione") consente di verificare se vi siano (in media) differenza significative fra gli elementi appartenenti alle colonne (variabili) della tabella in cui sono stati ordinatamente raccolte le osservazioni, l'ANOVA a due fattori consente di verificare contemporaneamente se vi siano (in media) differenza significative fra gli elementi appartenenti alle righe della tabella. Che questo sia significativo dipende esclusivamente dal disegno sperimentale.

Si consideri la tabella TECNICI del file ESEMPI.MDB: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Nella tabella sono stati registrati, per cinque giorni consecutivi, i risultati del controllo di qualità di un metodo analitico eseguito in modo semi-automatico. I risultati sono stati registrati separatamente per i tre tecnici che eseguono detto metodo. Il disegno sperimentale è stato strutturato in questo modo in quanto così ci si aspetta di poter rilevare dalla variabilità tra le colonne la variabilità dovuta alla differente manualità del personale, e dalla variabilità tra le righe la variabilità tra giorni dovuta al metodo analitico stesso. Selezionare le colonne da TECNICO_A a TECNICO_C. L'ANOVA a due fattori rivela una variabilità non significativa tra i tecnici (variabilità tra le colonne), ed una variabilità significativa tra giorni (variabilità tra le righe). Per il metodo vedere ancora Wonnacott[19].

La variabilità entro duplicati consente di calcolare la variabilità presente tra misure duplicate dello stesso evento. Tipicamente viene utilizzata per determinare l'imprecisione di un metodo analitico quando si hanno a disposizione misure in doppio effettuate su campioni della routine. Ovviamente il significato dell'imprecisione misurata è condizionato in modo determinante dal disegno sperimentale adottato: così se le misure in doppio sono ottenute nella stessa serie analitica l'imprecisione misurata rappresenterà una stima dell'imprecisione entro la serie, mentre se le misure in doppio sono ottenute in giorni e quindi in serie analitiche diverse l'imprecisione misurata rappresenterà un stima dell'imprecisione tra le serie.

Il file CLORURO.MI1 (file in formato Ministat: caricarlo nella tabella dati con l'opzione Apri del Menù File) contiene venti coppie di valori, ottenute determinando la concentrazione del cloruro in doppio su altrettanti sieri della routine, nella stessa serie analitica. Il metodo sembra avere una buona precisione. L'esempio è tratto dal Manuale del Controllo di Qualità a cura del CROSP della Regione Lombardia[20].

L'analisi delle componenti della variabilità consente di decomporre la variabilità totale in due componenti, quella tra le colonne (variabili) e quella entro le colonne. Viene utilizzata l'analisi della varianza nella forma modificata proposta da Krouwer[21], e molto simile a quella consigliata anche dal National Comittee for Cinical Laboratory Standards[22].

---

[18] *Wonnacott TH, Wonnacott RJ. Introduzione alla statistica. Milano:Franco Angeli, 1980:237-54.*
[19] *Wonnacott TH, Wonnacott RJ. Introduzione alla statistica. Milano:Franco Angeli, 1980:254-63.*
[20] *Giunta Regionale della Lombardia, Assessorato alla Sanità. Manuale del Controllo di Qualità nel Laboratorio di Patologia Clinica. Milano: Comitato Regionale per l'Ordinamento dei Servizi di Patologia, 1978;63.*
[21] *Krouwer JS, Rabinowitz R. How to improve estimates of imprecision. Clin Chem 1984;30:2902.*
[22] *NCCLS. Proposed Standard PSEP-3. Protocol for establishing performance claims for clinical chemical methods. Replication experiment. NCCLS, 771 E. Lancaster Ave., Villanova, PA 19085.*

La tabella POTASSIO del file ESEMPI.MDB contiene i risultati della determinazione del potassio su un singolo campione, effettuata per cinque giorni, con cinque replicati per giorno: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) Apri del Menù File. Lo scopo è di determinare l'imprecisione totale del metodo e di decomporla in due componenti: quella entro giorni e quella tra giorni. Selezionare le colonne da GIORNO_1 a GIORNO_5 ed effettuare il calcolo delle componenti della variabilità. Come atteso la variabilità tra giorni (tra le colonne) risulta superiore a quella entro giorni (entro le colonne).

Interessante anche il file VARBIOL.DBF. Contiene i valori di colesterolo totale, determinati in una serie di giorni successivi in cinque volontari. Lo scopo è quello dei decomporre la variabilità biologica totale in due componenti: variabilità biologica intra-individuale (stimata dalla variabilità entro le colonne) e variabilità biologica inter-individuale (stimata dalla variabilità tra le colonne). Selezionare le colonne da SOGGETTO_A a SOGGETTO_E ed effettuare il calcolo delle componenti della variabilità. Come atteso la variabilità biologica intra-individuale è inferiore a quella tra i diversi individui (inter-individuale).

- È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.
- Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

1.3.6.5. Regressione lineare

Questa opzione consente di effettuare il calcolo dell'equazione della retta di regressione con tre diversi modelli.

- Questa opzione risulta attiva quando vengono selezionate due colonne (variabili)
- Le due colonne devono contenere lo stesso numero di dati.
- La prima delle due colonne selezionate corrisponde ai valori delle ascisse, la seconda corrisponde ai valori delle ordinate dei singoli punti.

Per adattare la retta ai dati sperimentali viene impiegato il metodo dei minimi quadrati, una tecnica di approssimazione ben nota, che consente di minimizzare la somma dei quadrati delle differenze che residuano fra i punti sperimentali e la retta.
Il metodo dei minimi quadrati viene impiegato con tre differenti modelli: la regressione *x* variabile indipendente (regressione lineare standard), la regressione *y* variabile indipendente e la componente principale standardizzata.
A meno che non vi siano motivi particolari, che l'utilizzatore dovrà attentamente valutare, impiegare sempre per risultati della regressione lineare *x* variabile indipendente. Il modello matematico impiegato presuppone che la *x*, cioè la variabile indipendente, sia misurata senza errore, e che l'errore con cui si misura la *y* sia distribuito normalmente e con una varianza che non cambia al variare del valore della *x*. Tra le varie statistiche fornite, particolarmente interessanti sono la significatività della differenza del coefficiente angolare da 0 (zero) e da 1 (uno), e la significatività della differenza dell'intercetta da zero (per $p < 0,05$ le differenze citate possono essere considerate significative). Inoltre con il grafico della retta di regressione vengono tracciati i

limiti di confidenza al 90% della regressione. Il metodo è trattato su tutti i testi di statistica, fra i quali si ricordano Armitage[23] e Snedecor[24].
Per l'interpretazione dei risultati della regressione lineare in funzione dei diversi tipi di applicazione di questa tecnica statistica nel campo della chimica clinica, vedere Davis[25].
La regressione lineare *y* variabile indipendente è semplicemente l'immagine speculare della precedente. Deve essere utilizzata solamente a scopo esplorativo dei dati. Tanto più differisce dalla regressione *x* variabile indipendente, in presenza di errori nella misura delle *x*, tanto più si dovrà considerare di esprimere i risultati della regressione in termini di componente principale standardizzata. Infatti la regressione lineare standard non è raccomandata per l'analisi statistica di dati nei quali la variabile indipendente sia affetta da un errore di misura: in questo caso, impiegando la regressione lineare standard, si ottengono, a seconda di quale variabile sia posta in ascisse, equazioni della retta di regressione diverse: e questo fatto può portare a conclusioni contraddittorie. Per i casi di questo genere (tipico il confronto tra due metodi per la determinazione dello stesso analita) si consiglia di impiegare, in alternativa alla regressione lineare standard, la componente principale standardizzata, per la quale vedere Feldman[26].

La tabella CALCIO del file ESEMPI.MDB contiene i risultati di due ipotetici metodi per la determinazione del calcio nel siero: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Sui primi tre sieri i due metodi hanno fornito identici risultati. Sui rimanenti quattro sieri hanno fornito risultati speculari. La situazione appare caratterizzata da una assoluta indecidibilità. Selezionare entrambe le colonne: i valore della prima colonna selezionata (METODO_A) saranno riportati in ascisse, quelli della seconda colonna selezionata (METODO_B) saranno riportati in ordinate. La regressione x variabile indipendente e la regressione y variabile indipendente forniscono in effetti risultati speculari (utilizzare la rappresentazione delle rispettive rette di regressione per visualizzare le rette). La componente principale standardizzata più salomonicamente fa propendere per il fatto che, con le informazioni disponibili, possiamo solo concludere che i due metodi forniscono sostanzialmente gli stessi risultati.

Dati reali, con cui sperimentare la regressione lineare, sono quelli relativi alla determinazione dell'urea (valori in mg/dL) effettuata, su una serie di sieri freschi della routine, con cinque strumenti diversi.

I dati sono contenuti nel file UREA.DBF, che può essere importato con l'opzione Importa File (in formato dBase III®) del Menù File. Da notare come in questo caso la scelta di una adeguata ampiezza della dispersione dei valori di concentrazione dell'urea consente di ottenere, con la regressione *x* variabile indipendente e con la regressione *y* variabile indipendente, risultati molto simili. In ogni caso il risultato da utilizzare dovrebbe essere quello fornito dalla componente principale standardizzata. Si fa notare infine che i risultati della componente principale standardizzata sono validi solamente quando i coefficienti angolari della regressione lineare *x* variabile indipendente e della regressione lineare *y* variabile indipendente sono entrambi positivi.

---

[23] *Armitage P. Statistica medica. Milano: Feltrinelli Editore, 1979:151-167 e 264-266.*
[24] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:149-174.*
[25] *Davis RD, Thompson JE, Pardue HL. Characteristics of statistical parameters used to interpret least-squares results. Clin Chem 1978;24:611-20.*
[26] *Feldman U, Schneider B, Klinkers H. A multivariate approach for the biometric comparison on analytical methods in clinical chemistry. J Clin Chem Clin Biochem 1981;19:131-7.*

- È possibile stampare sia i risultati dell'elaborazione statistica che i grafici mediante l'opzione Stampa del Menù File.
- Per la stampa dei grafici si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune, di personalizzare le legende, di calcolare il valore della y corrispondente a un dato valore della x, e di calcolare il valore della x corrispondente a un dato valore della y.
- Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.
-

1.3.6.6. Regressione polinomiale

Questa opzione consente di adattare a una serie di punti, con il metodo dei minimi quadrati una retta (polinomio di primo grado) oppure, in alternativa, un polinomio di secondo grado o un polinomio di terzo grado.

- Questa opzione risulta attiva quando vengono selezionate due colonne (variabili)
- Le due colonne devono contenere lo stesso numero di dati.
- La prima delle due colonne selezionate corrisponde ai valori delle ascisse, la seconda corrisponde ai valori delle ordinate dei singoli punti.

Per il polinomio di primo grado viene utilizzata la regressione lineare *x* variabile indipendente (regressione lineare standard): con l'aggiunta però di un test per la linearità, che consente di verificare se la retta descrive adeguatamente la relazione di funzione *y = f(x)* che lega i valori in ordinate a quelli in ascisse come indicato in Burnett[27].
Per valori di $p < 0,05$ si può considerare che tale relazione non sia adeguatamente descritta da una retta, ma sia piuttosto curvilinea.

Il file TESTLIN.DBF contiene i valori di assorbanza (in ordinate, ASSORBANZA) ottenuti per una serie di concentrazioni note di un ipotetico analita (in ascisse, CONCENTRAZ): caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Selezionare entrambe le colonne. Il test per la linearità (*p* di gran lunga inferiore a 0,05) indica che la relazione tra assorbanza e concentrazione si discosta di molto dalla linearità: la relazione viene meglio descritta da un polinomio di secondo grado.

Anche nel caso della regressione polinomiale di secondo grado il metodo dei minimi quadrati viene utilizzato per adattare la corrispondente funzione polinomiale ai dati sperimentali in modo che risulti minimizzata la somma dei quadrati delle differenze residue fra i dati sperimentali e la funzione stessa. Il sistema di equazioni che consente di calcolare termine noto e coefficienti del polinomio di secondo grado può essere facilmente risolto in forma matriciale, mediante la regola di Cramer (il matematico svizzero vissuto fra il 1704 e il 1752). Vedere Batschelet[28] e Spiegel[29].

---

[27] *Burnett RW. Quantitative evaluation of linearity. Clin Chem 1980;26:644-6.*
[28] *Batschelet E. Introduzione alla matematica per biologi. Padova: Piccin, 1979:511-8.*
[29] *Spiegel MS. Statistica. Milano: Gruppo Editoriale Fabbri-Bompiani, 1976:221.*

La tabella APTOGLOB del file ESEMPI.MDB contiene i valori di assorbanza (in ordinate, ASSORBANZA) e di concentrazione (in ascisse, MG_DL) di una curva di calibrazione immunoturbidimetrica: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File). Selezionare entrambe le colonne: un polinomio di secondo grado descrive adeguatamente la curva.

Infine anche nel caso della regressione polinomiale di terzo grado il metodo dei minimi quadrati viene utilizzato per adattare la corrispondente funzione polinomiale ai dati sperimentali in modo che risulti minimizzata la somma dei quadrati delle differenze residue fra i dati sperimentali e la funzione stessa. Il sistema di equazioni che consente di calcolare termine noto e coefficienti del polinomio di terzo grado può anche in questo caso essere facilmente risolto in forma matriciale mediante la regola di Cramer (il matematico svizzero vissuto fra il 1704 e il 1752). Vedere Batschelet[30] e Scheid[31].

La tabella CINETICA del file ESEMPI.MDB contiene i valori di assorbanza (in ordinate, ASSORBANZA) e di tempo (in ascisse, SECONDI) registrati per una cinetica enzimatica: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File). Il polinomio di terzo grado appare l'unico ad essere in grado di descrivere adeguatamente l'andamento della cinetica nell'intervallo dei dati sperimentali.

- È possibile stampare sia i risultati dell'elaborazione statistica che i grafici mediante l'opzione Stampa del Menù File.
- Per la stampa dei grafici si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune, di personalizzare le legende, di calcolare il valore della y corrispondente a un dato valore della x, e di calcolare il valore della x corrispondente a un dato valore della y.
- Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

1.3.6.7. Regressione multipla

Questa opzione consente di definire la migliore relazione lineare che lega un numero di variabili superiore a quello che ne consente la rappresentazione in un piano cartesiano (quindi un numero di variabili superiore a due). Per il metodo vedere Snedecor[32].

- Questa opzione risulta attiva quando vengono selezionate da 2 a 10 colonne (variabili)

---

[30] *Batschelet E. Introduzione alla matematica per biologi. Padova: Piccin, 1979:511-8.*
[31] *Scheid F. Analisi numerica. Milano: Gruppo Editoriale Fabbri-Bompiani, Sonzogno, 1975:235-66.*
[32] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:334-64.*

- Le due colonne devono contenere tutte lo stesso numero di dati.
- L'ultima colonna selezionata deve contenere la variabile dipendente (*y*), le colonne selezionate a sinistra di questa sono assunte contenere le variabili indipendenti.

Un esempio può chiarire meglio di tutto le potenzialità di impiego di questo test statistico.

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Limitare a 30 i dati da analizzare: per questo è sufficiente vuotare la riga 31 (selezionare la riga 31 facendo click sulla cella contenente il numero della riga, dal Menu Modifica scegliere Riga e quindi Vuota). Selezionare le colonne COLEST, HDL, TRIGLI e LDL. Calcolare la regressione multipla. Il colesterolo LDL è la variabile dipendente (LDL, in mg/dL): le variabili indipendenti sono il colesterolo totale (COLEST, in mg/dL), il colesterolo HDL (HDL, in mg/dL) e i trigliceridi (TRIGLI, in mg/dL). Osservare i coefficienti della regressione multipla: forniscono il risultato atteso ? [Risposta: si, i coefficienti della regressione forniscono con una notevole accuratezza la formula di Friedewald, LDL = COLEST - HDL - (TRIGLI/5) con la quale è stato calcolato il colesterolo LDL].

- È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.
- Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

1.3.6.8. Statistica bayesiana

Questa opzione consente di calcolare, per un dato test di laboratorio, la sensibilità, la specificità, il valore predittivo del test positivo, il valore predittivo del test negativo, l'efficienza diagnostica e la curva ROC. Il valore predittivo del test positivo e il valore predittivo del test negativo possono essere ricalcolati facendo variare la prevalenza della malattia.

- L'opzione risulta attiva quando vengono selezionate due colonne (variabili)
- Le due colonne possono contenere un diverso numero di dati.
- I risultati del test nei soggetti sani devono trovarsi nella prima delle due colonne selezionate (colonna di sinistra), i risultati del test nei soggetti malati devono trovarsi nella seconda delle colonne selezionate (colonna di destra).

Ovviamente i soggetti devono essere stati classificati nei due gruppi (sani e malati) con un criterio indipendente dal test di laboratorio di cui si vogliono determinare le caratteristiche. Tutte le grandezze citate possono essere rappresentate graficamente, sovraimposte agli istogrammi della distribuzione dei risultati del test ottenuti nel un gruppo di soggetti sani e nel gruppo di soggetti malati.

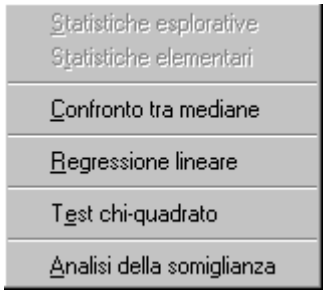
Il file BAYES.DBF contiene i risultati ottenuti in un gruppo di soggetti sani (colonna SANI) e in un gruppo di soggetti malati (colonna MALATI): caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Selezionare entrambe le colonne ed effettuare i calcoli e le rappresentazioni previste dall'opzione. Si ricordi che la scelta del livello decisionale dipende dalle considerazioni di merito relative agli obbiettivi che ci si propone. Per un test con un costo contenuto, e che rivela una malattia curabile, la scelta più ovvia sarà quella di fissare il livello decisionale al massimo della sensibilità (100%), anche se ciò andrà a scapito della specificità (aumento dei falsi positivi, cioè dei falsi allarmi). Ma se i falsi positivi dovessero innescare processi diagnostici per l'esclusione della malattia complessi, molto costosi, e magari potenzialmente

pericolosi, ecco magari la necessità di bilanciare meglio, anche in funzione delle risorse disponibili, sensibilità e specificità.

Per una trattazione dettagliata della statistica bayesiana illustrata da eccellenti esempi clinici delle strategie da seguire, vedere Galen e Gambino[33] e Gherardt[34]. Per una trattazione delle curve ROC vedere Robertson[35].

- È possibile stampare sia i risultati dell'elaborazione statistica che i grafici mediante l'opzione Stampa del Menù File.
- Per la stampa dei grafici si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune, di personalizzare le legende, e di ridefinire il valore della prevalenza della malattia (in quest'ultimo caso la sensibilità, la specificità, il valore predittivo del test positivo, il valore predittivo del test negativo, l'efficienza diagnostica e la curva ROC sono ricalcolati automaticamente).
- Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

1.3.6.9. Statistica non parametrica



Questa opzione comprende un vero e proprio menù, e come tale viene presentata. Per un approccio generale al problema delle statistiche non parametriche vedere Maritz[36] e Siegel[37].

1.3.6.9.1. Statistiche esplorative

Questa opzione consente di calcolare la mediana, il range e i quartili della distribuzione. Queste statistiche rappresentano il modo in cui esprimere i risultati dell'analisi elementare di dati campionari che non siano distribuiti in modo gaussiano.

---

[33] *Galen RS, Gambino SR. Oltre il concetto di normalità. Padova: Piccin, 1980:237pp.*
[34] *Gherardt W, Keller H. Evaluation of test data from clinical studies. Scand J Clin Lab Invest;1986:46(Supplement 181),74pp.*
[35] *Robertson EA, Zweig MH. Use of receiver operating chracteristic curves to evaluate the clinical performance of analytical systems. Clin Chem 1981;27:1569-74.*
[36] *Maritz JS. Distribution-free statistical methods. London: Chapman and Hall Editor, 1984:264pp.*
[37] *Siegel S, Castellan NJ Jr. Statistica non parametrica. Milano: McGraw-Hill, 1988:477pp.*

- L’opzione risulta attiva quando non è stata selezionata alcuna colonna.

Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Le colonne Controlli, NCAH, CPH e CAH contengono i valori di IgA nel siero ottenuti rispettivamente in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), in un gruppo di soggetti con epatite cronica attiva (CAH) e infine in un gruppo di soggetti con cirrosi epatica alcolica (AC).

1.3.6.9.2. Statistiche elementari

Questa opzione consente di calcolare la mediana, il range, i quartili della distribuzione, e una tabella dei percentili più importanti. È possibile rappresentare anche la distribuzione cumulativa dei dati, nella forma standard e nella forma di distribuzione cumulativa ripiegata proposta da Krouwer[38].
Queste statistiche rappresentano il modo in cui esprimere i risultati dell'analisi elementare di dati campionari che non siano distribuiti in modo gaussiano. Come anche nel caso delle Statistiche Elementari determinate con metodo parametrico, sono riportati i limiti di confidenza al 90% dei percentili. Per i metodi vedere La Rocca[39].

- L'opzione risulta attiva quando viene selezionata una sola colonna (variabile).

Utilizzare il file COOLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Il calcolo delle Statistiche Elementari sul valore dei trigliceridi (per l'elaborazione selezionare la colonna TRIGLI) con metodo parametrico (scegliere l'opzione Statistiche Elementari dal Menù Statistica) porta ad una tabella dei percentili nella quale il 2,5-esimo percentile parametrico e il 97,5-esimo percentile parametrico (rispettivamente la media meno 1,96 volte la deviazione standard e la media più 1,96 volte la deviazione standard) sono pari a -40,3 e 318,3. In base a questo dovremmo affermare che nel 95% dei soggetti esaminati i trigliceridi sono compresi tra -40,3 e 318,3 mg/dL: un'affermazione che, e dispetto del fatto che i risultati sono stati ottenuti con un metodo statistico rigoroso, nessuno si sentirebbe di sostenere. Le cose vanno meglio nel caso delle statistiche non parametriche con metodo non parametrico, con la presente opzione (provare).

- È possibile stampare sia i risultati dell'elaborazione statistica che i grafici mediante l'opzione Stampa del Menù File.
- Per la stampa dei grafici si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

---

[38] *Krouwer JS, Monti KL. A simple, graphical method to evaluate laboratory assays. Eur J Clin Chem Clin Biochem 1995;33:525-7.*
[39] *La Rocca S, Milani S, Oriani G. I valori di riferimento. Principi teorici e metodologia di produzione. Biochim Clin 1989;13:168-79.*

- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune, di personalizzare le legende, di calcolare il percentile corrispondente a un dato valore della x, e di calcolare il valore della x corrispondente a un dato percentile.
- Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

1.3.6.9.3. Confronto tra mediane

Questa opzione consente di effettuare il confronto tra le mediane nel caso di campioni indipendenti e di dati appaiati.

- L'opzione risulta attiva quando vengono selezionate due colonne (variabili).
- Il calcolo del test di Wilcoxon per dati appaiati è possibile solamente se le due colonne (variabili) contengono lo stesso numero di dati, quello per campioni indipendenti è possibile sempre.

Il test di Wilcoxon per dati appaiati è l'equivalente non parametrico del test *t* di Student per dati appaiati, e va utilizzato in luogo di questo quando i dati non siano distribuiti in modo gaussiano. La soluzione utilizzata per il calcolo della significatività è sufficientemente accurata per n > 16. Il metodo è tratto da Snedecor[40].

Il file ARTRITER.DBF contiene i valori delle immunoglobuline, determinati in pazienti trattati con due diversi farmaci (FARMACO_A e FARMACO_B): caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Il disegno sperimentale era strutturato in modo da determinare l'appaiamento dei pazienti. Si sapeva inoltre che i valori delle immunoglobuline non sono distribuiti in modo gaussiano. Importare il file e selezionare entrambe le colonne. Calcolare innanzitutto le differenze tra le coppie di valori mediante l'opzione Differenza del Menù Calcoli. Le differenze hanno per lo più segno negativo, e indicano che i valori delle immunoglobuline rilevati dopo trattamento con il farmaco A sono tendenzialmente inferiori a quelli rilevati dopo trattamento con il farmaco B. Il test di Wilcoxon per dati appaiati conferma la significatività delle differenze legate ai due tipi di trattamento. L'esempio è tratto da Bossi[41].

Sebbene spesso chiamato test di Mann-Whitney, l'equivalente non parametrico del test *t* di Student per campioni indipendenti è dovuto anch'esso a Wilcoxon, come ricorda lo Snedecor: e qui si è voluto adottare l'eponimo che sembra essere storicamente più corretto. La soluzione utilizzata per il calcolo della significatività è sufficientemente accurata per n > 16. Il metodo è tratto da Snedecor[42].

Il file RIBOFLAV.MI1 (file in formato Ministat: caricarlo nella tabella dati con l'opzione Apri del Menù File) contiene la concentrazione di riboflavina nel fegato (colonna FEGATO) e nel muscolo (colonna MUSCOLO) di bue, come determinata in una apposita sperimentazione. Selezionare le

---

[40] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:141-3.*

[41] *Bossi A, Cortinovis I, Duca P, Marubini E. Introduzione alla statistica medica. Roma: La Nuova Italia Scientifica, 1994:341.*

[42] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:144-5.*

due colonne (variabili) ed effettuare i calcoli: il test di Wilcoxon per campioni indipendenti evidenzia una differenza significativa tra i risultati ottenuti sui due tessuti.

- È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.
- Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.
-

1.3.6.9.4. Regressione lineare

Questa opzione consente di calcolare la regressione *x* variabile indipendente, la regressione *y* variabile indipendente, e l'equivalente della componente principale standardizzata in modo non parametrico. Come equivalente non parametrico della componente principale standardizzata viene utilizzato il modello di regressione lineare non parametrica proposto da Passing e Bablok[43,44]. Per un articolo riassuntivo e una bibliografia abbastanza completa riguardo il problema del confronto tra metodi vedere Besozzi[45].

- L'opzione risulta attiva quando vengono selezionate due colonne (variabili)
- Le due colonne devono contenere lo stesso numero di dati.

La tabella CALCIO del file ESEMPI.MDB contiene i risultati di due ipotetici metodi per la determinazione del calcio nel siero: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Sui primi tre sieri i due metodi hanno fornito identici risultati. Sui rimanenti quattro sieri hanno fornito risultati speculari. La situazione appare caratterizzata da una assoluta indecidibilità. Selezionare entrambe le colonne: i valore della prima colonna selezionata (METODO_A) saranno riportati in ascisse, quelli della seconda colonna selezionata (METODO_B) saranno riportati in ordinate. La regressione x variabile indipendente e la regressione y variabile indipendente forniscono in effetti risultati speculari (utilizzare la rappresentazione delle rispettive rette di regressione per visualizzare le rette). La regressione lineare di Passing e Bablok più salomonicamente fa propendere per il fatto che, con le informazioni disponibili, possiamo solo concludere che i due metodi forniscono sostanzialmente gli stessi risultati.

Dati reali, con cui sperimentare la regressione lineare, sono quelli relativi alla determinazione dell'urea (valori in mg/dL) effettuata, su una serie di sieri freschi della routine, con cinque strumenti diversi.

I dati sono contenuti nel file UREA.DBF, che può essere importato con l'opzione Importa File (in formato dBase III®) del Menù File. Da notare come in questo caso la scelta di una adeguata

[43] *Passing H, Bablok W. A new biometric procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. J Clin Chem Clin Biochem 1983;21:709-20.*
[44] *Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample size. Application of linear regression procedures for method comparison studies in clinical chemistry, Part II. J Clin Chem Clin Biochem 1984;22:431-45.*
[45] *Besozzi M, Franzini C. Impiego della regressione lineare nel confronto fra metodi: nuove tecniche statistiche parametriche e non parametriche (con in appendice un programma in BASIC). Giorn It Chim Clin 1986;11:29-41.*

ampiezza della dispersione dei valori di concentrazione dell'urea consente di ottenere, con la regressione *x* variabile indipendente e con la regressione *y* variabile indipendente, risultati molto simili. In ogni caso il risultato da utilizzare dovrebbe essere quello fornito dalla regressione di Passing e Bablok. Si fa notare infine che i risultati della regressione lineare di Passing e Bablok sono validi solamente quando i coefficienti angolari della regressione lineare *x* variabile indipendente e della regressione lineare *y* variabile indipendente sono entrambi positivi.

- È possibile stampare sia i risultati dell'elaborazione statistica che i grafici mediante l'opzione Stampa del Menù File.
- Per la stampa dei grafici si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
- L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune, di personalizzare le legende, di calcolare il valore della y corrispondente a un dato valore della x, e di calcolare il valore della x corrispondente a un dato valore della y.
- Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

1.3.6.9.5. Test chi quadrato

Questa opzione consente di calcolare il test chi-quadrato nella forma generalizzata per tabelle di contingenza, che è quella impiegata da Ministat. Per questo è necessario che ciascun elemento del campione in esame possa essere classificato per una caratteristica in un numero R di classi, e per una seconda caratteristica in un numero C di classi, in modo tale che i dati possano essere organizzati in una tabella di R righe per C colonne. Il metodo è trattato in tutti i testi di statistica, vedere per esempio Ingelfinger[46].

- L'opzione risulta attiva quando vengono selezionate almeno due colonne (variabili)
- Le colonne devono contenere lo stesso numero di dati.

Utilizzare il file FUMO.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Selezionare le colonne VIVI e DECEDUTI (notare come non vengano importati i contenuti dei campi non numerici. Comunque la riga 1 contiene i risultati ottenuti nel gruppo dei non fumatori, la riga 2 contiene i risultati ottenuti nel gruppo di fumatori). Il file contiene i risultati di un esperimento effettuato allo scopo di verificare se il fumo di pipa determini un significativo aumento della mortalità. Per questo un gruppo di non fumatori e uno di fumatori di pipa furono seguiti per sei anni e, al termine del periodo di osservazione, venne determinato il numero di soggetti deceduti in ciascuno dei due gruppi, con i seguenti risultati: tra i non fumatori, 950 soggetti erano ancora vivi, mentre 117 erano deceduti (riga 1). Tra i fumatori di pipa, 348 erano ancora vivi, mentre 54 erano deceduti (riga 2). Il test chi-quadrato evidenzia una differenza non significativa tra la mortalità nei due gruppi. Conclusione: fumare la pipa non fa male. L'esempio, famoso per le sue conclusioni, è tratto da Snedecor[47].

[46] *Ingelfinger JA, Mosteller FA, Thibodeau LA, Ware JH. Biostatistica in medicina. Milano: Raffaello Cortina, 1986:195-210.*

Una considerazione in merito al disegno sperimentale: l'osservazione per periodi di tempo più prolungati potrebbe portare a conclusioni differenti ?.

Un'ultima osservazione: la distribuzione chi-quadrato è un'approssimazione tanto più valida quanto più grandi sono le frequenze attese. Se anche una sola frequenza attesa risulta inferiore a 1, ovvero se più del 20% delle frequenze attese risulta inferiore a 5, i risultati del test chi-quadrato, e in particolare la sua significatività, devono essere valutati con estrema prudenza.

- È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.
- Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

1.3.6.9.6. Analisi della somiglianza

Questa opzione consente di effettuare l'analisi della somiglianza (cluster analysis) di un insieme di oggetti omogenei.

- L'opzione risulta attiva qualsiasi sia il numero delle colonne (variabili) selezionate.
- Le colonne devono contenere lo stesso numero di dati.

Il file CALCOLI.DBF contiene i dati di composizione di 10 formazioni calcolose delle vie urinarie: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Per ognuno dei calcoli è nota la composizione in calcio, fosfato, ossalato e magnesio. Selezionare le colonne CALCIO, FOSFATO, OSSALATO e MAGNESIO. La matrice dei cluster contiene, per ciascuno dei casi in esame, il/i cluster nei quali il caso confluisce. I cluster sono numerati in ordine progressivo di formazione, da sinistra verso destra. Ogni colonna della matrice dei cluster rappresenta in questo modo un livello crescente di aggregazione dei casi in base alla loro somiglianza. Per ogni colonna è riportata la distanza euclidea (espressa come deviata normale standardizzata *z*) alla quale si è formato l'ultimo cluster.
Il primo cluster, comprende i calcoli numero 6 e numero 1, che hanno una distanza euclidea di 0.94. Quindi i calcoli 1 e 6 sono i più simili tra loro in assoluto. Ma anche i calcolo 2 e 7 sono molto simili tra di loro, avendo una distanza euclidea di 0.98. Notare come all'ottavo passaggio si siano formati due cluster, il primo comprendente i calcoli 10, 5 4 e 9, il secondo comprendente i calcoli 6, 1, 2, 7, 8 e 3: perché questi due cluster confluiscano si arriva ad una distanza euclidea di 8.60. Quindi questi due gruppi di calcoli sembrano rappresentare due famiglie ben distinte per composizione.

Il metodo è tratto da Massart[48]. Per una discussione critica delle applicazioni della cluster analysis a problemi del laboratorio clinico vedere Vogt[49].

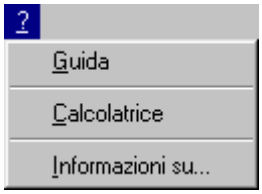
- È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.

---

[47] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:124.*
[48] *Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kaufman L. Chemometrics: a textbook. Amsterdam: Elsevier Science Publishers BV, 1988:371-84.*
[49] *Vogt W, Nagel D. Cluster analysis in diagnosis. Clin Chem 1992;38:182-98.*

- Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®. Nel caso di elaborazioni di oltre 40-50 dati, la finestra di testo potrebbe non riuscire a contenere completamente la matrice dei cluster: ciò si può evincere dal fatto che non compare il caratteristico messaggio di *fine* elaborazione. Tuttavia anche in questi casi, se l'analisi della somiglianza termina senza segnalazioni di errore, la stampa dei risultati riporterà la matrice dei cluster nella sua interezza.

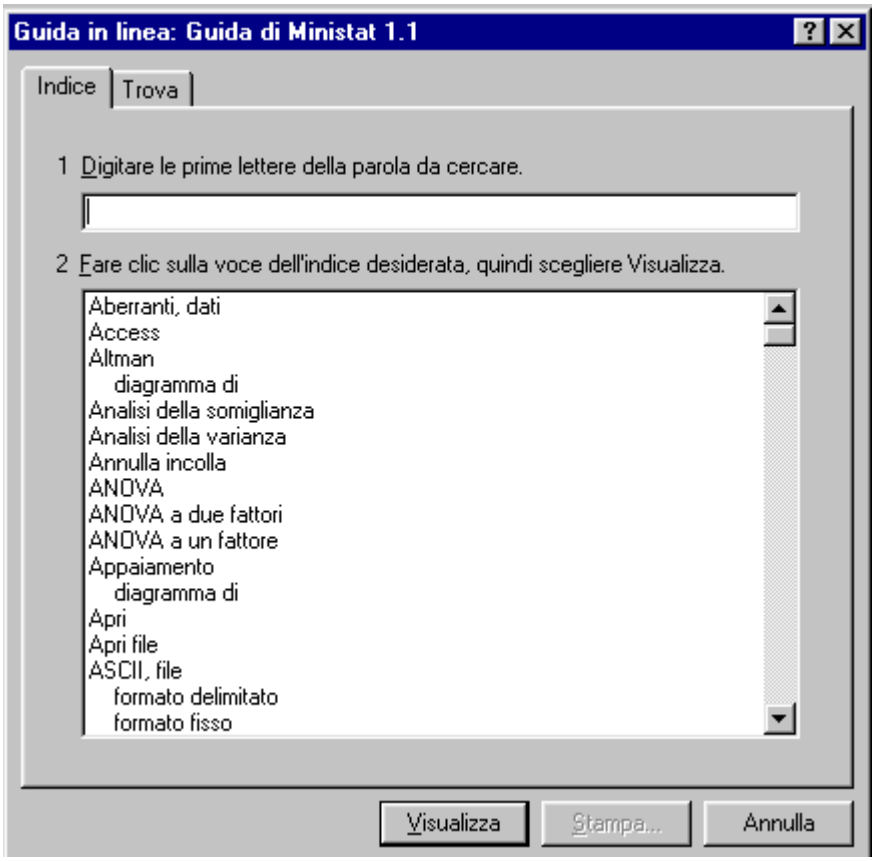
1.3.7. Menù Aiuto



1.3.7.1. Guida

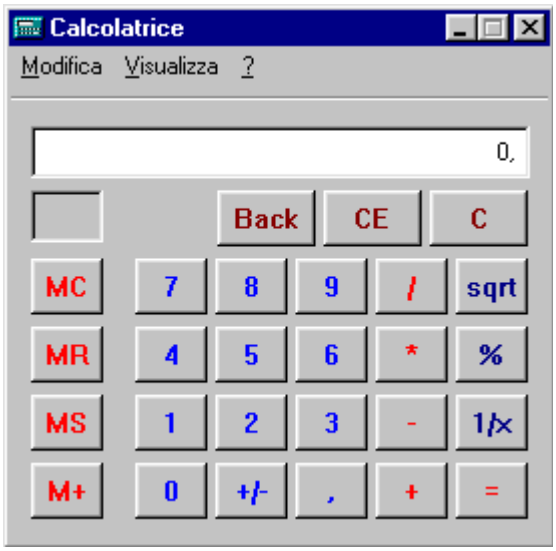
Contiene la versione elettronica on-line di questo manuale, che comprende in pratica tutto il contenuto del capitolo 5 del volume (sono quindi esclusi i quattro capitoli introduttivi e il sesto capitolo, riservato alle formule e agli algoritmi di calcolo). Si fa notare che nell'aggiornamento alla versione a 32 bit, come del resto specificato nella guida on-line, la nuova funzione che consente di tracciare diagrammi a torta è documentata solo nel testo. Il menù principale di Ministat riportato nella guida on-line e qui sotto, differisce da quello effettivo per la mancanza dell'icona corrispondente al diagramma a torta, e per una lieve differenza nella disposizione delle altre icone, che peraltro sono rimaste immodificate sia nell'aspetto che nelle funzioni.

Mediante l'opzione Cerca è possibile accedere a un Indice delle parole e delle espressioni per le quali è disponibile una ricerca rapida.



1.3.7.2. Calcolatrice

Consente di utilizzare la calcolatrice di Windows® per effettuare calcoli estemporanei.



Da notare che l'aspetto della calcolatrice dipende dalla versione di Windows® e dalle opzioni relative.

1.3.7.3. Informazioni su…

Consente di visualizzare una finestra contenente le informazioni sulla versione del programma.

ADDENDUM
FORMULE E ALGORITMI DI CALCOLO

Sono riportati qui di seguito le formule e gli algoritmi più importanti, in quanto utili sia per comprendere a fondo il significato dei vari test statistici, sia per consentire ai lettori più pazienti e più intraprendenti di implementare qualcuna delle tecniche statistiche più semplici all'interno di propri applicativi, realizzabili con gli strumenti standard disponibili su PC (tipicamente i tabelloni elettronici, come Excel®).

Nell'interpretazione delle formule si tenga presente che, laddove non compaiono le parentesi a determinarle, le priorità degli operatori aritmetici sono le seguenti: prima devono essere eseguite le operazioni di sommatoria, poi quelle di elevamento a potenza e di radice, poi le moltiplicazioni e le divisioni, e infine le operazioni di somma e di sottrazione.

Per le voci bibliografiche dalle quali sono stati tratti formule e algoritmi, si rimanda alle sezioni precedenti.

2.1. Asimmetria e curtosi

Per un insieme comprendente un numero n di dati x_i, è possibile esprimere lo scostamento della distribuzione dei dati trovata rispetto alla distribuzione gaussiana teorica ricorrendo al coefficiente di asimmetria (g_1) e al coefficiente di curtosi (g_2), che indicano il tipo di scostamento dalla normalità come qui di seguito specificato:

$\Rightarrow$ $g_1 < 0$: asimmetria negativa, cioè coda sinistra della distribuzione eccessivamente lunga;
$\Rightarrow$ $g_1 > 0$: asimmetria positiva, cioè coda destra della distribuzione eccessivamente lunga;
$\Rightarrow$ $g_2 < 0$: platicurtosi, cioè distribuzione eccessivamente appiattita, con code troppo corte;
$\Rightarrow$ $g_2 > 0$: leptocurtosi, cioè distribuzione eccessivamente alta, con code troppo lunghe.

Essendo allora

$$m_2 = \Sigma (x_i - \bar{x})^2 / n$$
$$m_3 = \Sigma (x_i - \bar{x})^3 / n$$
$$m_4 = \Sigma (x_i - \bar{x})^4 / n$$

rispettivamente il momento di ordine secondo (m_2), il momento di ordine terzo (m_3) e il momento di ordine quarto (m_4) dalla media, i valori del coefficiente di asimmetria g_1 e del coefficiente di curtosi g_2 sono calcolati come

$$g_1 = m_3 / (m_2 \cdot \sqrt{m_2})$$
$$g_2 = m_4 / m_2^2 - 3$$

La deviazione standard s_1 del coefficiente di asimmetria e la deviazione standard s_2 del coefficiente di curtosi sono calcolate rispettivamente come

$$s_1 = \sqrt{(6 / n)}$$
$$s_2 = \sqrt{(24 / n)}$$

Il coefficiente di asimmetria g_1 viene considerato significativo se supera (in valore assoluto) di 2,6 volte la sua deviazione standard s_1, cioè se

$$|g_1 / s_1| > 2,6$$

In questo caso si rigetta l'ipotesi che l'asimmetria osservata sia compatibile con quella di una distribuzione gaussiana.

Il coefficiente di curtosi g_2 viene considerato significativo se supera (in valore assoluto) di 2,6 volte la sua deviazione standard s_2, cioè se

$$|g_2 / s_2| > 2,6$$

In questo caso si rigetta l'ipotesi che la curtosi osservata sia compatibile con quella di una distribuzione gaussiana.

Qualora si arrivi a concludere che la distribuzione campionaria osservata si discosta significativamente da quella della gaussiana teorica, sarà necessario ricorrere a tecniche statistiche che non basano la loro validità (quindi la validità delle conclusioni che mediante esse è possibile trarre) su assunti distribuzionali di gaussianità, e cioè a tecniche statistiche non-parametriche.

2.2. Test di Kolmogorov-Smirnov

In alternativa ai test di asimmetria e curtosi, è possibile verificare se la distribuzione osservata si discosta significativamente dalla distribuzione gaussiana teorica ricorrendo al test di Kolmogorov-Smirnov, che può essere calcolato procedendo come segue:

⇒ ordinare i dati in ordine numerico crescente;
⇒ calcolare la media $\bar{x}$ e la deviazione standard s dei dati;
⇒ per ciascun dato calcolare la deviata normale standardizzata (*DNS*) come

$$DNS = (x - \bar{x}) / s$$

⇒ calcolare un fattore di correzione (*FC*) come metà del rapporto fra la differenza minima misurabile per due campioni (*DMM*) e la deviazione standard calcolata (*s*), cioè

$$FC = 0,5 \cdot DMM / s$$

⇒ per ciascuno dei dati calcolare il valore della deviata normale standardizzata corretta (*DNSC*) come

$$DNSC = DNS + FC$$

⇒ per ciascuno dei dati calcolare il valore della funzione di distribuzione cumulativa normale *(FDCN*) corrispondente alla *DNSC* ;
⇒ per ciascuno dei dati calcolare il valore della distribuzione empirica (*FDE*) come rapporto fra il numero progressivo del dato (numero d'ordine nella lista ordinata dei dati) e il numero totale dei dati ;
⇒ per ciascuno dei dati calcolare il valore assoluto della differenza fra il valore della funzione di distribuzione cumulativa normale e quello della funzione di distribuzione empirica, cioè

$$|FDCN - FDE|$$

e chiamare *KS* la maggiore di tale differenze.

Essendo *n* il numero dei dati, un test sufficientemente approssimato per valutare la gaussianità di una distribuzione è basato sul calcolo dei valori critici del test (*KS*) ai livelli di probabilità del 5% e dell'1% rispettivamente come

$$KS_{0,05} = 0,886 / \sqrt{n}$$
$$KS_{0,01} = 1,031 / \sqrt{n}$$

Se il valore ottenuto supera quello previsto, al livello di probabilità prescelto, si conclude per un significativo scostamento della distribuzione dei dati trovata rispetto alla distribuzione gaussiana teorica .

Il test di Kolmogorov-Smirnov è qui riportato per completezza, non essendo incluso nell'attuale versione di Ministat. In realtà questo test fornisce (come atteso) gli stessi risultati dei test di asimmetria e di curtosi, con lo svantaggio di riassumere l'informazione in un'unica statistica, che non consente di separare la componente di asimmetria da quella di curtosi, e quindi non consente di identificare il contributo che ciascuna delle due fornisce alla non-gaussianità (se presente) della distribuzione osservata.

2.3. Statistiche elementari parametriche

Per una distribuzione gaussiana data, il valore μ (la *media della popolazione*) che rappresenta la misura di posizione della distribuzione gaussiana, e il valore σ (la *deviazione standard della popolazione*) che rappresenta la misura di dispersione della distribuzione gaussiana, possono essere stimati rispettivamente mediante la media campionaria $\bar{x}$ e la deviazione standard campionaria *s*.

Dato quindi un campione che include *n* dati (osservazioni) x_i, la media campionaria $\bar{x}$ e la deviazione standard campionaria *s* sono calcolate rispettivamente come

$$\bar{x} = \Sigma x_i / n$$
$$s = \sqrt{s^2}$$

essendo la varianza s^2 dei dati calcolata come

$$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$$

E' possibile semplificare il calcolo della varianza s^2 ricordando che

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$

L'errore standard della media *es* viene calcolato come rapporto tra la deviazione standard *s* e la radice quadrata del numero *n* di osservazioni

$$es = s / \sqrt{n}$$

Il coefficiente di variazione CV viene calcolato come rapporto, espresso in percentuale, tra la deviazione standard s e la media $\bar{x}$, cioè come

$$CV = 100 \cdot (s \ / \ \bar{x})$$

Fatti uguale 100 il numero di dati di una distribuzione, il percentile è il valore che lascia alla sua sinistra la percentuale di dati desiderata (e corrispondentemente alla destra il complemento a 100). Così ad esempio il 2,5-esimo percentile è il valore che lascia alla sua sinistra il 2,5% dei dati osservati (e corrispondentemente lascia alla sua destra il 97,5% dei dati osservati), il 50-esimo percentile (che nel caso delle statistiche parametriche è la media campionaria $\bar{x}$) è il valore che lascia alla sua sinistra il 50% dei dati osservati (e corrispondentemente lascia alla sua destra il 50% dei dati osservati).
Per calcolare i percentili parametrici di una distribuzione ricorrere alla tabella della deviata normale standardizzata *z* riportata in tutti i testi di statistica. Alcuni percentili di frequente utilizzo sono riportati nella seguente tabella (ove $\bar{x}$ e *s* sono rispettivamente la media e la deviazione standard dei dati):

Percentile	Deviata normale standardizzata z	Valore del percentile
2,5	1,96	$\bar{x}$ - 1,96 *s*
5,0	1,645	$\bar{x}$ - 1,645 *s*
10	1,282	$\bar{x}$ - 1,282 *s*
20	0,842	$\bar{x}$ - 0,842 *s*
30	0,525	$\bar{x}$ - 0,525 *s*
40	0,253	$\bar{x}$ - 0,253 *s*
50	0	$\bar{x}$
60	0,253	$\bar{x}$ + 0,253 *s*
70	0,525	$\bar{x}$ + 0,525 *s*
80	0,842	$\bar{x}$ + 0,842 *s*
90	1,282	$\bar{x}$ + 1,282 *s*
95	1,645	$\bar{x}$ + 1,645 *s*
97,5	1,96	$\bar{x}$ + 1,96 *s*

2.4. Statistiche elementari non parametriche

Le statistiche elementari non parametriche comprendono in genere, oltre alla *mediana*, il *valore minimo* osservato, il *valore massimo* osservato, il *range* (cioè la differenza tra valore massimo e valore minimo) e spesso i *quartili*. In base al principio elementare per cui un segmento di retta può essere suddiviso in quattro parti mediante tre tacche equidistanti tra di loro e dagli estremi del segmento, i quartili sono tre. Il primo quartile è il valore che lascia alla sua sinistra il 25% dei dati (e che lascia alla sua destra il 75% dei dati). Il secondo quartile è il valore che lascia alla sua sinistra il 50% dei dati (e che lascia alla sua destra il 50% dei dati): il secondo quartile è quindi la mediana. Il terzo quartile è il valore che lascia alla sua sinistra il 75% dei dati (e che lascia alla sua destra il 25% dei dati).

Anche nel caso delle statistiche non parametriche, fatti uguale 100 il numero di dati di una distribuzione, il percentile è il valore che lascia alla sua sinistra la percentuale di dati desiderata (e

corrispondentemente alla destra il complemento a 100). Così ad esempio il 2,5-esimo percentile è il valore che lascia alla sua sinistra il 2,5% dei dati osservati (e corrispondentemente lascia alla sua destra il 97,5% dei dati osservati), il 50-esimo percentile (che nel caso delle statistiche non parametriche è la mediana) è il valore che lascia alla sua sinistra il 50% dei dati osservati (e corrispondentemente lascia alla sua destra il 50% dei dati osservati). Si noti che il primo quartile corrisponde al 25-esimo percentile non parametrico, il secondo quartile corrisponde al 50-esimo percentile non parametrico (cioè alla mediana), il terzo quartile corrisponde al 75-esimo percentile non parametrico.

2.4.1. Calcolo del percentile corrispondente a un dato

Per calcolare il percentile corrispondente a un dato, percentile che indica la posizione del dato all'interno della distribuzione osservata:

⇒ calcolare il rango del dato. Il rango è il numero di posizione dei singoli dati nella lista dei dati ordinati in ordine numerico crescente. Quando due dati sono uguali, a ciascuno viene assegnato come numero di posizione la media dei numeri di posizione dei dati che risultano uguali. Così, per esempio, se il quinto e il sesto dato, in una lista ordinata, sono tutti e due uguali, e pari a 223, il rango del quinto e del sesto dato sarà uguale, e pari a 5,5. Il procedimento viene esteso anche al caso in cui più di due dati siano uguali;

⇒ essendo n il numero dei dati, il percentile P (p-esimo) non parametrico, che indica la posizione del dato all'interno della distribuzione osservata, viene calcolato come

$$P = 100 \cdot rango / (n + 1)$$

2.4.2. Calcolo del dato corrispondente a un percentile

In un insieme di n dati, il P-esimo percentile non parametrico corrisponde al dato il cui numero C (numero progressivo, non necessariamente intero, nella serie dei dati ordinati in ordine numerico crescente) è

$$C = P \cdot (n + 1) / 100$$

Qualora C non sia un intero (accade quasi sempre) viene effettuata una interpolazione lineare fra il valore corrispondente all'intero che immediatamente precede C e il valore corrispondente all'intero che immediatamente segue C.

2.5. Rapporto tra varianze

Il test F applicato al rapporto tra due varianze campionarie, consente di verificare se le varianze di due campioni sono tra di loro omogenee (cioè sostanzialmente uguali).

Indicati con x_1 i dati del primo campione, con n_1 il loro numero e con $\bar{x}_1$ la loro media, ed indicati con x_2 i dati del secondo campione, con n_2 il loro numero e con $\bar{x}_2$ la loro media, la varianza del primo campione viene calcolato come

$$s_1^2 = \Sigma (x_1 - \bar{x}_1)^2 / (n_1 - 1)$$

con gradi di libertà pari a

$$n_1 - 1$$

la varianza del secondo campione viene calcolata come

$$s_2^2 = \Sigma (x_2 - \bar{x}_2)^2 / (n_2 - 1)$$

con gradi di libertà pari a

$$n_2 - 1$$

E' possibile semplificare il calcolo della varianza s_1^2 e della varianza s_2^2 ricordando che

$$\Sigma (x_1 - \bar{x}_1)^2 = \Sigma x_1^2 - (\Sigma x_1)^2 / n_1$$
$$\Sigma (x_2 - \bar{x}_2)^2 = \Sigma x_2^2 - (\Sigma x_2)^2 / n_2$$

Il valore del test *F* di omogeneità fra le varianze viene calcolato allora come rapporto fra la varianza maggiore e la varianza minore, cioè

$$F = s_1^2 / s_2^2 \qquad se \;\; s_1^2 > s_2^2$$

ovvero

$$F = s_2^2 / s_1^2 \qquad se \;\; s_2^2 > s_1^2$$

Il valore di *p* corrispondente alla statistica *F* rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude che le varianze sono significativamente diverse. Varianze diverse stanno ad indicare che i due campioni sono stati tratti da popolazioni diverse, aventi medie diverse. In questo senso il rapporto tra varianze corrisponde a un confronto tra medie.

2.6. Test t di Student per dati appaiati

Nel confronto tra medie mediante il test *t* di Student per dati appaiati la differenza fra le coppie di osservazioni diventa la variabile in esame. Dato un numero *n* di dati x_i , y_i sia d_i la differenza (presa con il segno) fra il valore del primo e il valore del secondo elemento della coppia, ovvero

$$d_i = x_i - y_i$$

Allora la differenza media $\bar{d}$ e la varianza s_d^2 delle differenza sono calcolate rispettivamente come

$$\bar{d} = \Sigma d_i / n$$
$$s_d^2 = \Sigma (d_i - \bar{d})^2 / (n - 1)$$

Si ricorda che anche in questo caso è possibile semplificare il calcolo della varianza s_d^2 ricordando che

$$\Sigma (d_i - \bar{d})^2 = \Sigma d_i^2 - (\Sigma d_i)^2 / n$$

Il valore del test t di Student è calcolato come rapporto fra la differenza media osservata e il suo errore standard, cioè

$$t = \bar{d} / \sqrt{(s_d^2 / n)}$$

con $n - 1$ gradi di libertà.

Il valore di p corrispondente alla statistica t rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza.

2.7. Test t di Student per campioni indipendenti

Il test t di Student per il confronto fra le medie di campioni indipendenti è basato su un principio semplice e intuitivo: a parità di valore assoluto della differenza fra le medie, a tale differenza viene data tanto maggior peso quanto minore è la dispersione dei dati campionari (cioè in definitiva quanto meno le due distribuzioni campionarie si sovrappongono), e viceversa.

Dati allora due campioni, il primo comprendente n_1 osservazioni x_1, aventi media $\bar{x}_1$, e il secondo comprendente n_2 osservazioni x_2, aventi media $\bar{x}_2$, il valore t viene calcolato come

$$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s^2 / n_1 + s^2 / n_2)}$$

con gradi di libertà pari a

$$n_1 + n_2 - 2$$

Al numeratore compare la differenza fra le medie, mentre al denominatore compare l'errore standard di tale differenza, la grandezza che, come si è precedentemente accennato, consente di dare il peso maggiore o minore alla differenza fra le medie.
La varianza s^2 è la varianza combinata dei due campioni, e viene calcolata come rapporto fra la somma delle devianza dei due campioni e la somma dei loro gradi di libertà, cioè come

$$s^2 = (\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2) / (n_1 + n_2 - 2)$$

E' possibile semplificare il calcolo della varianza s_1^2 e della varianza s_2^2 ricordando che

$$\Sigma (x_1 - \bar{x}_1)^2 = \Sigma x_1^2 - (\Sigma x_1)^2 / n_1$$
$$\Sigma (x_2 - \bar{x}_2)^2 = \Sigma x_2^2 - (\Sigma x_2)^2 / n_2$$

Il valore di p corrispondente alla statistica t rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza fra le medie.

Un problema tuttavia complica lievemente la situazione: il test t di Student ordinario tende a fornire troppo pochi risultati significativi quando il campione più grande ha la varianza (cioè la dispersione dei dati) maggiore, e troppi risultati significativi quando il campione più grande ha la varianza minore. E' perciò necessario utilizzare, nei casi in cui la varianza dei due campioni differisca in maniera significativa, una forma del test che preveda una opportuna correzione. Per fare questo si calcola il rapporto tra varianze.

La varianza del primo campione viene calcolata come

$$s_1^2 = \Sigma (x_1 - \bar{x}_1)^2 / (n_1 - 1)$$

con gradi di libertà pari a

$$n_1 - 1$$

la varianza del secondo campione viene calcolata come

$$s_2^2 = \Sigma (x_2 - \bar{x}_2)^2 / (n_2 - 1)$$

con gradi di libertà pari a

$$n_2 - 1$$

Il valore del test *F* di omogeneità fra le varianze viene calcolato allora come rapporto fra la varianza maggiore e la varianza minore, cioè

$$F = s_1^2 / s_2^2 \qquad se \;\; s_1^2 > s_2^2$$

ovvero

$$F = s_2^2 / s_1^2 \qquad se \;\; s_2^2 > s_1^2$$

Il valore di *p* corrispondente alla statistica *F* rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una significatività della diversità fra le varianze.

Nel caso di varianze significativamente diverse (varianze non omogenee) si calcola la statistica

$$t' = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s_1^2 / n_1 + s_2^2 / n_2)}$$

con gradi di libertà uguali a

$$(v_1 + v_2)^2 / (v_1^2 / (n_1 - 1) + v_2^2 / (n_2 - 1))$$

essendo rispettivamente

$$v_1 = s_1^2 / n_1$$
$$v_2 = s_2^2 / n_2$$

Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza fra le medie.

2.8. Test t di Student per una media teorica

Dato un campione che include *n* dati (osservazioni) x_i, la media campionaria $\bar{x}$ e varianza campionaria s^2 sono calcolate rispettivamente come

$$\bar{x} = \Sigma x_i / n$$
$$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$$

E' possibile semplificare il calcolo della varianza s^2 ricordando che

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$

Essendo allora μ la media teorica attesa, il test *t* nella forma

$$t = (\bar{x} - \mu) / \sqrt{(s^2 / n)}$$

consente di verificare se la media osservata $\bar{x}$ differisce dalla media teorica μ.

Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza fra le medie.

Come è possibile vedere più avanti, nella parte dedicata al calcolo della regressione lineare, questa forma del test t di Student è molto importante in quanto consente di effettuare i test di significatività sui valori dell'intercetta *a* e del coefficiente angolare *b* dell'equazione della retta di regressione *x* variabile indipendente calcolata con il metodo dei minimi quadrati.

2.9. Test di Wilcoxon per dati appaiati

Il test di Wilcoxon per dati appaiati è l'equivalente non parametrico del test t di Student per dati appaiati, e va utilizzato in luogo di questo quando i dati non siano distribuiti in modo gaussiano.

Per i calcoli si procede in questo modo :

- $\Rightarrow$ determinare le differenze (con il segno) fra le *n* coppie di valori;
- $\Rightarrow$ stabilire, per ciascuna differenza, il numero di posizione nella lista delle differenze ordinate (questa volta ignorando il segno) in ordine numerico crescente: la più piccola differenza osservata avrà numero di posizione 1, e via dicendo;
- $\Rightarrow$ quando due o più differenze sono uguali, assegnare a ciascuna di esse la media dei numeri di posizione che esse dovrebbero avere; così, per esempio, se la quinta e la sesta differenza sono uguali, assegnare come numero di posizione nella lista il valore 5.5 a entrambe;
- $\Rightarrow$ riassegnare il segno ai numeri di posizione nella lista (se la differenza avente quel numero di posizione era negativa, assegnare il segno meno, se era positiva assegnare il segno più);
- $\Rightarrow$ calcolare il totale per i numeri di posizione con segno negativo e per i numeri di posizione con segno positivo, e chiamare *T* il più piccolo di questi due totali;
- $\Rightarrow$ calcolare la deviata normale standardizzata *Z* come

$$Z = (\mu - T - 0{,}5) / s$$

essendo

$$\mu = n \cdot (n + 1) / 4$$
$$s = \sqrt{((2n + 1) \cdot \mu / 6)}$$

La statistica *Z* così calcolata corrisponde a sottoporre al test la mediana delle differenze.

Il valore di *p* corrispondente alla statistica *Z* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza fra le mediane. Questa soluzione è sufficientemente accurata per n > 16.

2.10. Test di Wilcoxon per campioni indipendenti

Sebbene spesso chiamato test di Mann-Whitney, l'equivalente non parametrico del test t di Student per dati appaiati è dovuto anch'esso a Wilcoxon. Rappresenta l'equivalente non parametrico del test t di Student per campioni indipendenti, e va utilizzato in luogo di questo quando i dati non siano distribuiti in modo gaussiano.

Per i calcoli si procede in questo modo:

⇒ mettere i dati dei due campioni in una singola lista, badando ad etichettarli in modo che poi possano essere successivamente di nuovo distinti;

⇒ stabilire, per ciascun dato, il numero di posizione nella lista ordinata in ordine numerico crescente: il dato più piccolo avrà numero di posizione 1, e via dicendo;

⇒ quando due o più dati sono uguali, assegnare a ciascuno di essi la media dei numeri di posizione che esse dovrebbero avere; così, per esempio, se i dati dal primo al sesto sono uguali, assegnare come numero di posizione nella lista a tutti il valore 3,5 (media dei numeri da 1 a 6);

⇒ se i campioni hanno lo stesso numero di dati, calcolare il totale per i numeri di posizione del primo e per i numeri di posizione del secondo campione, e chiamare T il più piccolo di questi due totali;

⇒ se i due campioni hanno diverso numero di dati chiamare T_1 il totale per il campione che ha il numero minore di dati, diciamo n_1 , e quindi, essendo n_2 il numero di dati del secondo campione, calcolare

$$T_2 = n_1 \cdot (n_1 + n_2 + 1) - T_1$$

⇒ chiamare T il più piccolo fra i valori T_1 e T_2

⇒ calcolare la deviata normale standardizzata Z come

$$Z = (| \mu - T | - 0{,}5) / s$$

essendo

$$\mu = n_1 \cdot (n_1 + n_2 + 1) / 2$$
$$s = \sqrt{(n_2 \cdot \mu / 6)}$$

La statistica Z così calcolata corrisponde a sottoporre al test la mediana delle differenze.

Il valore di p corrispondente alla statistica Z rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza mediana osservata. Questa soluzione è sufficientemente accurata per n > 16.

2.11. Analisi della varianza a un fattore

Con le varie forme del test t di Student e del test di Wilcoxon è possibile confrontare fra di loro due medie: l'analisi della varianza o ANOVA (da ANalysis Of VAriance) consente di estendere il confronto a più di due medie.

Sia i (con $i = 1,2,3,....r$) un generico campione, sia j (con $j = 1,2,3,....n$) un generico replicato, e quindi $x_{i,j}$ un generico valore della tabella

Campione	Replicato					Media
	j=1	j=2	j=3	. . . . .	j=n	$\bar{x}_i$
i=1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	. . . . .	$x_{1,n}$	$\bar{x}_1$
i=2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	. . . . .	$x_{2,n}$	$\bar{x}_2$
. . . .	. . . .	. . . .	. . . .	. . . . .	. . . .	. . . .
i=r	$x_{r,1}$	$x_{r,2}$	$x_{r,3}$	. . . . .	$x_{r,n}$	$\bar{x}_r$

nella quale le $r \cdot n$ osservazioni corrispondenti a r campioni, per ciascuno dei quali sono disponibili n dati, sono riportate in modo ordinato.

Siano ancora $\bar{x}_i$ la media di un generico campione, e $\bar{x}_g$ la media generale di tutte le $r \cdot n$ osservazioni. Allora la variabilità totale (S_t) osservata viene calcolata come

$$S_t = \sum_{i=1}^{i=r} \sum_{j=1}^{j=n} (x_{i,j} - \bar{x}_g)^2$$

con $r \cdot n - 1$ gradi di libertà .

La variabilità S_s spiegata dalle differenza fra le medie $\bar{x}_i$ delle righe viene calcolata come

$$S_s = n \cdot \sum_{i=1}^{i=r} (\bar{x}_i - \bar{x}_g)^2$$

con $r - 1$ gradi di libertà, mentre la variabilità casuale, non spiegata, detta anche "residua" (S_n), viene calcolata come

$$S_n = \sum_{i=1}^{i=r} \sum_{j=1}^{j=n} (x_{i,j} - \bar{x}_i)^2$$

con $r \cdot (n - 1)$ gradi di libertà, tenendo presente che, per semplicità, essa può essere calcolata anche per differenza come

$$S_n = S_t - S_s$$

La varianza spiegata (V_s) e la varianza non spiegata (V_n), calcolate rispettivamente come

$$V_s = S_s / (r - 1)$$
$$V_n = S_n / (r \cdot (n - 1))$$

vengono allora impiegate per calcolare finalmente il rapporto fra varianze F

$$F = V_s / V_n$$

con $r - 1$ gradi di libertà al numeratore e $r \cdot (n - 1)$ gradi di libertà al denominatore.

Il valore di p corrispondente alla statistica F rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una significatività della diversità fra le varianze, e conseguentemente che esistono delle differenze significative fra le medie $\bar{x}_i$ delle r righe.

2.12. Analisi della varianza a due fattori

Se l'analisi della varianza a 1 fattore (detta anche "a un criterio di classificazione") consente di verificare se vi siano(in media) differenze significative fra gli elementi appartenenti alle righe della tabella in cui sono stati ordinatamente raccolte le nostre osservazioni, l'ANOVA a 2 fattori consente di verificare contemporaneamente se vi siano (in media) differenze significative fra gli elementi appartenenti alle colonne della tabella, che si presenta così

Riga	Colonna					Media
	j=1	j=2	j=3	. . . . .	j=c	$\bar{x}_{i.}$
i=1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	. . . . .	$x_{1,c}$	$\bar{x}_{1.}$
i=2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	. . . . .	$x_{2,c}$	$\bar{x}_{2.}$
. . . .	. . . .	. . . .	. . . .	. . . . .	. . . .	. . . .
i=r	$x_{r,1}$	$x_{r,2}$	$x_{r,3}$	. . . . .	$x_{r,c}$	$\bar{x}_{r.}$
Media $\bar{x}_{.j}$	$\bar{x}_{.1}$	$\bar{x}_{.2}$	$\bar{x}_{.3}$	. . . . .	$\bar{x}_{.c}$	

Ciascuno dei dati $x_{i,j}$ risulta pertanto assegnato in base a una caratteristica a una delle i ($i = 1,2,3,....r$) righe e per un'altra caratteristica a una delle j ($j = 1,2,3,....c$) colonne.

Siano allora $\bar{x}_{i.}$ la media di una generica riga, $\bar{x}_{.j}$ la media di una generica colonna, e $\bar{x}_g$ la media generale degli $r \cdot c$ dati. È facile notare che, rispetto alla tabella dell'analisi della varianza a 1 fattore, la novità consiste nell'avere introdotto un elemento di classificazione a livello delle colonne, e quindi le corrispondenti medie $\bar{x}_{.j}$.

La variabilità totale (S_t) osservata viene calcolata come

$$S_t = \sum_{i=1}^{i=r} \sum_{j=1}^{j=c} (x_{i,j} - \bar{x}_g)^2$$

con $r \cdot c - 1$ gradi di libertà.

Tuttavia, essendo stato introdotto un nuovo criterio di classificazione dei dati, essa verrà scomposta non più in due, bensì in tre componenti. La prima di esse, la variabilità S_r spiegata dalle differenze fra le medie $\bar{x}_{i.}$, cioè dalle differenze fra le medie delle righe, viene calcolata come

$$S_r = c \cdot \sum_{i=1}^{i=r} (\bar{x}_{i.} - \bar{x}_g)^2$$

con $r - 1$ gradi di libertà.

La variabilità S_c spiegata dalle differenze fra le medie $\bar{x}_{.j}$, cioè dalle differenze fra le medie delle colonne, viene calcolata come

$$S_c = r \cdot \sum_{j=1}^{j=c} (\bar{x}_{.j} - \bar{x}_g)^2$$

con $c - 1$ gradi di libertà.

Infine la variabilità casuale, non spiegata (S_n), detta anche "residua", viene calcolata come

$$S_n = \sum_{i=1}^{i=r} \sum_{j=1}^{j=c} (x_{i,j} - \bar{x}_{i.} - \bar{x}_{.j} - \bar{x}_g)^2$$

con $(r - 1) \cdot (c - 1)$ gradi di libertà, tenendo presente che può più semplicemente essere calcolata per differenza come

$$S_n = S_t - S_r - S_c$$

La varianza spiegata dalle differenze fra le medie $\bar{x}_{i.}$ delle righe (V_r), quella spiegata dalle differenze fra le medie $x_{.j}$ delle colonne (V_c) e la varianza non spiegata (V_n) sono allora calcolate rispettivamente come

$$V_r = S_r / (r - 1)$$
$$V_c = S_c / (c - 1)$$
$$V_n = S_n / ((r - 1) \cdot (c - 1))$$

Il rapporto fra varianze F

$$F = V_r / V_n$$

con $r - 1$ gradi di libertà al numeratore e $(r - 1) \cdot (c - 1)$ gradi di libertà al denominatore viene allora impiegato per verificare l'esistenza di una differenza significativa fra le medie delle righe ($\bar{x}_{i.}$).

Il valore di p corrispondente alla statistica F rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una significatività della diversità fra le varianze, e conseguentemente che esistono delle differenze significative fra le medie $\bar{x}_i$ delle r righe.

Il rapporto fra varianze

$$F = V_c / V_n$$

con $c - 1$ gradi di libertà al numeratore e $(r - 1) \cdot (c - 1)$ gradi di libertà al denominatore viene impiegato per verificare l'esistenza di una differenza significativa fra le medie delle colonne ($\bar{x}_j$).

Il valore di p corrispondente alla statistica F rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una significatività della diversità fra le varianze, e conseguentemente che esistono delle differenze significative fra le medie $\bar{x}_j$ delle c colonne.

2.13. Regressione lineare parametrica

Per adattare la retta ai dati sperimentali viene impiegato il metodo dei minimi quadrati, una tecnica di approssimazione ben nota, che consente di minimizzare la somma dei quadrati delle differenze che residuano fra i punti sperimentali e la retta.

2.13.1. Regressione lineare x variabile indipendente

Il modello matematico impiegato presuppone che la x, cioè la variabile indipendente, sia misurata senza errore, e che l'errore con cui si misura la y sia distribuito normalmente e con una varianza che non cambia al variare del valore della x.

Sia allora n il numero dei punti aventi coordinate note (x_i,y_i), e siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i: il coefficiente angolare b_{yx} e l'intercetta a_{yx} dell'equazione della retta di regressione x variabile indipendente

$$y = a_{yx} + b_{yx} \cdot x$$

che meglio approssima (in termini di minimi quadrati) i dati vengono calcolati rispettivamente come

$$b_{yx} = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \Sigma (x_i - \bar{x})2$$

$$a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}$$

Il coefficiente di correlazione r viene calcolato come

$$r = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2)}$$

E' possibile semplificare i calcoli ricordando che

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma (y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$

La varianza residua attorno alla regressione viene calcolata come

$$s_0^2 = (\Sigma (y_i - \bar{y})^2 - s_1^2) / (n - 2)$$

essendo

$$s_1^2 = (\Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}))^2 / \Sigma (x_i - \bar{x})^2$$

Infine l'errore standard della stima s_{yx} e le deviazioni standard del coefficiente angolare (s_b) e dell'intercetta (s_a), che forniscono una misura rispettivamente della dispersione dei dati attorno alla retta calcolata, e del grado di incertezza connesso con i valori ottenuti di a_{yx} e di b_{yx}, sono calcolati come

$$s_{yx} = \sqrt{(s_0^2)}$$
$$s_b = s_{yx} \cdot \sqrt{(1 / \Sigma (x_i - \bar{x})^2)}$$
$$s_a = s_b \cdot \sqrt{(\Sigma x_i^2 / n)}$$

Si consideri che la retta di regressione campionaria

$$y = a_{yx} + b_{yx} \cdot x$$

rappresenta la migliore stima possibile della retta di regressione della popolazione

$$y = \alpha + \beta \cdot x$$

Si consideri che il test t di Student per una media teorica nella forma già vista

$$t = (\bar{x} - \mu) / \sqrt{(s^2 / n)}$$

può essere riscritto tenendo conto delle seguenti identità

$$\bar{x} = a_{yx}$$
$$\mu = \alpha$$
$$\sqrt{(s^2 / n)} = s_a$$

assumendo quindi la forma

$$t = (a_{yx} - \alpha) / s_a$$

Questo consente di sottoporre a test la differenza dell'intercetta *a* rispetto a un valore atteso (per esempio rispetto a 0, cioè all'intercetta di una retta passante per l'origine). Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza dell'intercetta rispetto al valore atteso.

Si consideri che il test t di Student per una media teorica può anche essere riscritto tenendo conto delle seguenti identità

$$\bar{x} = b_{yx}$$
$$\mu = \beta$$
$$\sqrt{(s^2 / n)} = s_b$$

assumendo quindi la forma

$$t = (b_{yx} - \beta) / s_b$$

Questo consente di sottoporre a test la differenza del coefficiente angolare *b* rispetto a un valore atteso (per esempio rispetto a 0, cioè al coefficiente angolare di una retta orizzontale, oppure rispetto a 1, cioè al coefficiente angolare di una retta a 45 gradi). Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza del coefficiente angolare rispetto al valore atteso.

2.13.2. Regressione lineare y variabile indipendente

Il modello matematico impiegato presuppone che la *y*, cioè la variabile indipendente, sia misurata senza errore, e che l'errore con cui si misura la *x* sia distribuito normalmente e con una varianza che non cambia al variare del valore della *y*. Si noti che in questo caso inizialmente la *y* (variabile indipendente) viene posta in ascisse e la *x* (variabile dipendente) viene posta in ordinate.

Sia allora *n* il numero dei punti aventi coordinate note (x_i,y_i), e siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i: il coefficiente angolare b_{xy} e l'intercetta a_{xy} dell'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

che meglio approssima (in termini di minimi quadrati) i dati vengono calcolati rispettivamente come

$$b_{xy} = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \Sigma (y_i - \bar{y})^2$$

$$a_{xy} = \bar{x} - b_{xy} \cdot \bar{y}$$

Il coefficiente di correlazione *r* viene calcolato come

$$r = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2)}$$

(si noti che, come atteso, esso risulta identico a quello calcolato mediante la regressione *x* variabile indipendente).

E' possibile semplificare i calcoli ricordando che

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma (y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$

Per riportare i dati sullo stesso sistema di coordinate cartesiane utilizzato per la regressione *x* variabile indipendente, si esplicita l'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

rispetto alla *y*, ottenendo

$$x - a_{xy} = b_{xy} \cdot y$$

e quindi, dividendo entrambi i membri per b_{xy}

$$y = - a_{xy} / b_{xy} + 1 / b_{xy} \cdot x$$

Quindi l'intercetta *a* e il coefficiente angolare *b* che consentono di rappresentare la regressione *y* variabile indipendente sullo stesso sistema di coordinate cartesiane della regressione *x* variabile indipendente saranno rispettivamente uguali a

$$a = - a_{xy} / b_{xy}$$
$$b = 1 / b_{xy}$$

2.13.3. Componente principale standardizzata

Il modello matematico impiegato presuppone tanto la *x* quanto la *y* siano affette da un errore di misura equivalente.

Sia allora *n* il numero dei punti aventi coordinate note (x_i,y_i), siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i, sia b_{yx} il coefficiente angolare dell'equazione della retta di regressione *x* variabile indipendente, e sia b_{xy} il coefficiente angolare dell'equazione della retta di regressione *y* variabile indipendente.

Il coefficiente angolare b_{cps} dell'equazione della retta di regressione calcolata come componente principale standardizzata è allora uguale a

$$b_{cps} = \sqrt{(b_{yx} \cdot b_{xy})}$$

cioè alla media geometrica tra il coefficiente angolare b_{yx} della regressione *x* variabile indipendente e il coefficiente angolare b_{xy} della regressione *y* variabile indipendente, cioè

$$b_{xy} = \sqrt{(\Sigma (y_i - \bar{y})^2 / \Sigma (x_i - \bar{x})^2)}$$

mentre l'intercetta a_{cps} dell'equazione della retta di regressione calcolata come componente principale standardizzata è uguale a

$$a_{cps} = \bar{y} - b_{cps} \cdot \bar{x}$$

Infine il coefficiente di correlazione *r* viene calcolato come

$$r = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2)}$$

(si noti che, come atteso, esso risulta identico sia a quello calcolato mediante la regressione *x* variabile indipendente sia a quello calcolato mediante la regressione *y* variabile indipendente).

E' possibile semplificare i calcoli ricordando che

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma (y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$

2.14. Regressione lineare non parametrica

Per adattare la retta ai dati sperimentali viene impiegato un metodo che non fa ricorso ad assunti preliminari riguardo la distribuzione dei dati e degli errori ad essi associati.

2.14.1. Regressione lineare x variabile indipendente

Essendo *n* il numero dei punti aventi coordinate note (x_i,y_i), esistono $n \cdot (n - 1) / 2$ modi di connettere due punti qualsiasi con una retta, cioè esistono $n \cdot (n - 1) / 2$ coefficienti angolari

$$b = (y_i - y_j) / (x_i - x_j)$$

con *i* e *j* che variano fra *1* e *n* (con $i < j$).

Allora il coefficiente angolare b_{yx} e l'intercetta a_{yx} dell'equazione della retta di regressione *x* variabile indipendente

$$y = a_{yx} + b_{yx} \cdot x$$

che meglio approssima i dati, vengono calcolati nel seguente modo:

⇒ si calcolano i coefficienti angolari delle $n \cdot (n - 1) / 2$ rette che passano per tutte le coppie possibili di punti;

⇒ si ordinano gli $n \cdot (n - 1) / 2$ coefficienti angolari così calcolati in ordine numerico crescente;

⇒ si calcola il coefficiente angolare b_{yx} dell'equazione della retta di regressione come mediana degli $N = n \cdot (n - 1) / 2$ valori di cui al punto precedente, cioè come

$$b_{yx} = b_{((N + 1)/2)} \quad \text{se } N \text{ è dispari}$$
$$b_{yx} = (b_{(N/2)} + b_{(N/2+1)}) / 2 \quad \text{se } N \text{ è pari}$$

⇒ si calcolano allora gli *n* valori possibili per l'intercetta *a* come

$$a_i = y_i - b_{yx} \cdot x_i$$

⇒ si ordinano gli *n* valori di intercetta così calcolati in ordine numerico crescente;

⇒ si calcola l'intercetta a_{yx} dell'equazione della retta di regressione come mediana dei valori di cui al punto precedente, cioè come

$$a_{yx} = a_{((n+1)/2)} \quad \text{se } n \text{ è dispari}$$
$$a_{yx} = (a_{(n/2)} + a_{(n/2+1)}) / 2 \quad \text{se } n \text{ è pari}$$

2.14.2. Regressione lineare y variabile indipendente

Si noti che in questo caso inizialmente la *y* (variabile indipendente) viene posta in ascisse e la *x* (variabile dipendente) viene posta in ordinate.

Essendo allora *n* il numero dei punti aventi coordinate note (x_i,y_i), esistono $n \cdot (n - 1) / 2$ modi di connettere due punti qualsiasi con una retta, cioè esistono $n \cdot (n - 1) / 2$ coefficienti angolari

$$b = (x_i - x_j) / (y_i - y_j)$$

con *i* e *j* che variano fra *1* e *n* (con $i < j$).

Allora il coefficiente angolare b_{xy} e l'intercetta a_{xy} dell'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

che meglio approssima i dati vengono calcolati nel seguente modo:

⇒ si calcolano i coefficienti angolari delle $n \cdot (n - 1) / 2$ rette che passano per tutte le coppie possibili di punti;

⇒ si ordinano gli $n \cdot (n - 1) / 2$ coefficienti angolari così calcolati in ordine numerico crescente;

⇒ si calcola il coefficiente angolare b_{xy} dell'equazione della retta di regressione come mediana degli $N = n \cdot (n - 1) / 2$ valori di cui al punto precedente, cioè come

$$b_{xy} = b_{((N + 1)/2)} \quad \text{se } N \text{ è dispari}$$
$$b_{xy} = (b_{(N/2)} + b_{(N/2+1)}) / 2 \quad \text{se } N \text{ è pari}$$

⇒ si calcolano allora gli *n* valori possibili per l'intercetta *a* come

$$a_i = x_i - b_{xy} \cdot y_i$$

⇒ si ordinano gli *n* valori di intercetta così calcolati in ordine numerico crescente;

⇒ si calcola l'intercetta a_{xy} dell'equazione della retta di regressione come mediana dei valori di cui al punto precedente, cioè come

$$a_{xy} = a_{((n+1)/2)} \quad \text{se } n \text{ è dispari}$$
$$a_{xy} = (a_{(n/2)} + a_{(n/2+1)}) / 2 \quad \text{se } n \text{ è pari}$$

Per riportare i dati sullo stesso sistema di coordinate cartesiane utilizzato per la regressione *x* variabile indipendente, si esplicita l'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

rispetto alla *y*, ottenendo

$$x - a_{xy} = b_{xy} \cdot y$$

e quindi, dividendo entrambi i membri per b_{xy}

$$y = - a_{xy} / b_{xy} + 1 / b_{xy} \cdot y$$

Quindi l'intercetta *a* e il coefficiente angolare *b* che consentono di rappresentare la regressione *y* variabile indipendente sullo stesso sistema di coordinate cartesiane della regressione *x* variabile indipendente saranno rispettivamente uguali a

$$a = - a_{xy} / b_{xy}$$
$$b = 1 / b_{xy}$$

2.14.3. Regressione lineare di Passing e Bablok

Essendo *n* il numero dei punti aventi coordinate note (x_i,y_i), esistono $n \cdot (n - 1) / 2$ modi di connettere due punti qualsiasi con una retta, cioè esistono $n \cdot (n - 1) / 2$ coefficienti angolari

$$b = (y_i - y_j) / (x_i - x_j)$$

con *i* e *j* che variano fra *1* e *n* (con $i < j$).

Nel caso in cui sia

$$x_i = x_j \text{ e } y_i = y_j$$

il valore del coefficiente angolare non è definito, e quindi viene scartato dai calcoli successivi; ugualmente vengono scartati tutti i valori del coefficiente angolare uguali a -1.

I rimanenti *N* valori vengono allora ordinati in ordine numerico crescente; essendo *K* il numero dei valori del coefficiente angolare inferiori a -1, il coefficiente angolare b_{pb} della retta di regressione calcolata con il metodo di Passing e Bablok

$$y = a_{pb} + b_{pb} \cdot x$$

viene calcolato come

$$b_{pb} = b_{((N + 1)/2 + K)} \quad \text{se } N \text{ è dispari}$$
$$b_{pb} = (b_{(N/2+K)} + b_{(N/2+1+K)}) / 2 \quad \text{se } N \text{ è pari}$$

mentre l'intercetta a_{pb} viene calcolata come mediana degli *n* valori

$$a_i = y_i + b_{pb} \cdot x_i$$

e cioè, nella lista dei valori di *a* così calcolati e ordinati in ordine numerico crescente, come

$$a_{pb} = a_{((n+1)/2)} \quad \text{se } n \text{ è dispari}$$
$$a_{pb} = (a_{(n/2)} + a_{(n/2+1)}) / 2 \quad \text{se } n \text{ è pari}$$

2.15. Regressione polinomiale di secondo grado

Il metodo dei minimi quadrati consente di adattare un funzione $y = f(x)$ ad una serie di dati sperimentali in modo che risulti minimizzata la somma dei quadrati delle differenze residue fra i dati sperimentali e la funzione stessa.

In questo caso il metodo dei minimi quadrati viene impiegato per adattare ai dati un polinomio di secondo grado, nella forma

$$y = a + b \cdot x + c \cdot x^2$$

Essendo *n* il numero dei dati aventi coordinate (x_i,y_i), il termina noto *a*, e i coefficienti *b* e *c* del polinomio di secondo grado possono essere determinati risolvendo contemporaneamente le equazioni

$$a \cdot n + b \cdot \Sigma x_i + c \cdot \Sigma x_i^2 = \Sigma y_i$$
$$a \cdot \Sigma x_i + b \cdot \Sigma x_i^2 + c \cdot \Sigma x_i^3 = \Sigma x_i \cdot y_i$$
$$a \cdot \Sigma x_i^2 + b \cdot \Sigma x_i^3 + c \cdot \Sigma x_i^4 = \Sigma x_i^2 \cdot y_i$$

Il sistema può essere facilmente risolto, utilizzando qualsiasi programma che includa la soluzione dei sistemi di equazioni in forma matriciale (per esempio Excel®).

2.16. Regressione polinomiale di terzo grado

Il metodo dei minimi quadrati consente di adattare una funzione $y = f(x)$ ad una serie di dati sperimentali in modo che risulti minimizzata la somma dei quadrati delle differenze residue fra i dati sperimentali e la funzione stessa.

In questo caso il metodo dei minimi quadrati viene impiegato per adattare ai dati un polinomio di secondo grado, nella forma

$$y = a + b \cdot x + c \cdot x^2 + d \cdot x^3$$

Essendo *n* il numero dei dati aventi coordinate (x_i,y_i), il termina noto *a*, e i coefficienti *b*, *c* e *d* del polinomio di terzo grado possono essere determinati risolvendo contemporaneamente le equazioni

$$a \cdot n + b \cdot \Sigma x_i + c \cdot \Sigma x_i^2 + d \cdot \Sigma x_i^3 = \Sigma y_i$$
$$a \cdot \Sigma x_i + b \cdot \Sigma x_i^2 + c \cdot \Sigma x_i^{3+} d \cdot \Sigma x_i^4 = \Sigma x_i \cdot y_i$$
$$a \cdot \Sigma x_i^2 + b \cdot \Sigma x_i^3 + c \cdot \Sigma x_i^{4+} d \cdot \Sigma x_i^5 = \Sigma x_i^2 \cdot y_i$$
$$a \cdot \Sigma x_i^3 + b \cdot \Sigma x_i^4 + c \cdot \Sigma x_i^{5+} d \cdot \Sigma x_i^6 = \Sigma x_i^3 \cdot y_i$$

Il sistema può essere facilmente risolto, utilizzando qualsiasi programma che includa la soluzione dei sistemi di equazioni in forma matriciale (per esempio Excel®).

2.17. Test chi-quadrato per tabelle di contingenza

Per applicare il test chi-quadrato (χ^2) nella forma generalizzata per tabelle di contingenza, che è quella qui impiegata, è necessario che ciascun elemento del campione in esame possa essere classificato per una caratteristica in un numero *R* di classi, e per una seconda caratteristica in un numero *C* di classi, in modo tale che i dati possano essere organizzati in una tabella di *R* righe per *C* colonne, comprendente quindi un totale di $N = R \cdot C$ celle.

Essendo allora *f* la frequenza osservata in una data cella e *F* la frequenza attesa per la stessa cella, il test χ^2 viene calcolato come somma dei rapporti $(F - f)^2 / F$ per tutte le celle della tabella, cioè come

$$\chi^2 = \Sigma (f - F)^2 / F$$

Il valore di *f* per ciascuna cella è noto, essendo come detto *f* la frequenza osservata. Il valore di *F*, cioè la frequenza attesa, non è noto, ma può essere specificato qualora si operi alla luce di una ben definita ipotesi riguardante i dati. Nel caso particolare dell'ipotesi "non vi è differenza fra le frequenze osservate", cioè di quella che gli statistici chiamano la "ipotesi nulla", e che viene qui impiegata, le frequenze attese *F* possono essere stimate, e assumono ciascuna un valore pari al prodotto del totale della riga per il totale della colonna cui la cella appartiene diviso per il totale *n* dei casi osservato, ovvero

$$F = (\text{totale della riga}) \cdot (\text{totale della colonna}) / n$$

Per illustrare le modalità di sviluppo dei calcoli si consideri il caso più semplice, quello della seguente tabella 2 x 2

$$\begin{matrix} f_1 & f_2 \\ f_3 & f_4 \end{matrix}$$

in cui le frequenze osservate per ciascuna delle quattro celle sono indicate rispettivamente con f_1, f_2, f_3 e f_4.

Indicando allora:

$\Rightarrow$ il totale della riga 1 con R_1 ($R_1 = f_1 + f_2$);
$\Rightarrow$ il totale della riga 2 con R_2 ($R_2 = f_3 + f_4$);
$\Rightarrow$ il totale della colonna 1 con C_1 ($C_1 = f_1 + f_3$);
$\Rightarrow$ il totale della colonna 2 con C_2 ($C_2 = f_2 + f_4$);
$\Rightarrow$ il totale dei casi con n ($n = f_1 + f_2 + f_3 + f_4$);

i quattro valori di *F* attesi

$$\begin{matrix} F_1 & F_2 \\ F_3 & F_4 \end{matrix}$$

corrispondenti ai valori di *f* osservati saranno per definizione uguali rispettivamente a

$$\begin{matrix} F_1 = C_1 \cdot R_1 / n & F_2 = C_2 \cdot R_1 / n \\ F_3 = C_1 \cdot R_2 / n & F_4 = C_2 \cdot R_2 / n \end{matrix}$$

essendo ancora ovviamente $F_1 + F_2 + F_3 + F_4 = n$, mentre il valore di χ^2 sarà pari a

$$\chi^2 = (f_1 - F_1)^2 / F_1 + (f_2 - F_2)^2 / F_2 + (f_3 - F_3)^2 / F_3 + (f_4 - F_4)^2 / F_4$$

con $(R - 1) \cdot (C - 1)$ gradi di libertà.

E' quindi facile estendere i calcoli dalla tabella 2 x 2 a tabelle di qualsiasi estensione.

Il valore di *p* corrispondente alla statistica χ^2 rappresenta la probabilità di osservare per caso una differenza tra frequenze osservate e frequenze attese della grandezza di quella effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una differenza significativa di incidenza nei diversi gruppi del fattore in esame.

2.18. Analisi della somiglianza (cluster analysis)

I problemi di classificazione rivestono un ruolo centrale nella scienza. E classificare gli oggetti in base a criteri oggettivi, basati su quantità misurabili, è lo scopo fondamentale della cluster analysis.

Si consideri il seguente esempio, relativo a 10 calcoli delle vie urinarie, analizzati per la loro composizione in calcio, fosfato, ossalato e magnesio (dati simulati e valori espressi in unità di misura arbitrarie).

CALCOLO	CALCIO	FOSFATO	OSSALATO	MAGNESIO
1	99	81	69	61
2	78	65	53	43
3	81	66	38	54
4	45	23	19	16
5	44	18	24	19
6	102	83	72	66
7	83	68	49	45
8	74	71	41	57
9	38	19	22	14
10	48	14	21	12

All'inizio del processo di classificazione (clustering) ad ogni oggetto corrisponde un cluster (e viceversa). In questo stadio tutti gli oggetti sono considerati dissimili (diversi) tra di loro. Al passaggio successivo i due oggetti più simili sono raggruppati in un unico cluster. Il numero dei cluster risulta quindi pari al numero di oggetti - 1. Il procedimento viene ripetuto ciclicamente, fino ad ottenere (all'ultimo passaggio) un unico cluster.

Stabilire a quale livello di aggregazione degli oggetti fermarsi, e quindi quali conclusioni trarre, dipende esclusivamente dai giudizio di merito dell'utilizzatore. Per questo la cluster analysis riveste un ruolo centrale, in statistica, limitatamente alla analisi esplorativa dei dati (in effetti il livello a cui fermarsi trarre le conclusioni a questo punto non è più quantitativo, e quindi non è più oggettivo).

La prima cosa che si fa nell'analisi della somiglianza è quella di calcolare le distanze euclidee. La distanza può essere calcolata in vari modi: quello qui seguito prevede di calcolare la distanza tra tutte le possibili coppie di punti. Nel caso di due variabili (come per esempio sarebbe stato se si fossero avute, per ciascun calcolo, le sole misure di calcio e fosfato) mediante il teorema di Pitagora. Che peraltro può essere esteso dal piano cartesiano, bidimensionale, ad uno spazio tridimensionale (nel caso di tre misure sullo stesso campione) e a uno spazio n-dimensionale (nel caso di n misure sullo stesso campione).

Nel caso dell'esempio illustrato la matrice delle distanze è la seguente:

Caso	1	2	3	4	5	6	7	8	9	10
1	0,00	4,72	5,56	13,50	13,21	0,94	4,51	5,44	14,05	13,87
2	4,72	0,00	2,79	8,85	8,60	5,63	0,98	2,81	9,39	9,28
3	5,56	2,79	0,00	8,85	8,74	6,32	2,12	1,20	9,58	9,44
4	13,50	8,85	8,85	0,00	1,03	14,38	9,12	9,14	1,12	1,21
5	13,21	8,60	8,74	1,03	0,00	14,07	8,95	8,99	1,08	1,27
6	0,94	5,63	6,32	14,38	14,07	0,00	5,41	6,17	14,92	14,75
7	4,51	0,98	2,12	9,12	8,95	5,41	0,00	2,39	9,75	9,58
8	5,44	2,81	1,20	9,14	8,99	6,17	2,39	0,00	9,79	9,81
9	14,05	9,39	9,58	1,12	1,08	14,92	9,75	9,79	0,00	1,41
10	13,87	9,28	9,44	1,21	1,27	14,75	9,58	9,81	1,41	0,00

La matrice delle distanze è formata da un numero di righe e di colonne uguale, e pari al numero di casi in esame (i casi sono numerati in ordine progressivo: il caso 1 corrisponde alla prima riga di dati, il caso 2 alla seconda, e così via). La matrice contiene le distanze tra tutte le possibili coppie di casi. Notare come essa sia simmetrica rispetto alla diagonale. Le distanze sono espresse come deviata normale standardizzata (*z*). Si noti che la matrice è simmetrica, e che la diagonale assume valori uguali a zero (la distanza euclidea tra un calcolo e sé stesso è ovviamente nulla).

La corrispondente matrice dei cluster appare così

Caso	Cluster								
10	0	0	0	0	0	6	6	6	9
5	0	0	3	4	4	6	6	6	9
4	0	0	3	4	4	6	6	6	9
9	0	0	0	4	4	6	6	6	9
6	1	1	1	1	1	1	1	8	9
1	1	1	1	1	1	1	1	8	9
2	0	2	2	2	2	2	7	8	9
7	0	2	2	2	2	2	7	8	9
8	0	0	0	0	5	5	7	8	9
3	0	0	0	0	5	5	7	8	9

0.94	0.98	1.03	1.08	1.20	1.21	2.12	4.51	8.60
Distanza euclidea								

La matrice dei cluster contiene, per ciascuno dei casi in esame (righe) il/i cluster nei quali il caso confluisce. I cluster sono numerati in ordine progressivo di formazione, da sinistra verso destra. Ogni colonna rappresenta in questo modo un livello crescente di aggregazione dei casi in base alla loro somiglianza. Per ogni colonna è riportata la distanza (standardizzata) alla quale si è formato l'ultimo cluster.
Il primo cluster, comprende gli oggetti numero 6 e numero 1, che hanno una distanza euclidea di 0.94 (controllare anche la matrice delle distanze). Quindi gli oggetti 1 e 6 sono i più simili tra loro in assoluto. Ma anche gli oggetti2 e 7 sono molto simili tra di loro, avendo una distanza euclidea di 0.98. Notare come all'ottavo passaggio si siano formati due cluster, il primo comprendente gli oggetti 10, 5 4 e 9, il secondo comprendente gli oggetti 6, 1, 2, 7, 8 e 3: perché questi due cluster confluiscano si arriva ad una distanza euclidea di 8.60. Quindi questi due gruppi di oggetti sembrano rappresentare due famiglie ben distinte per composizione.