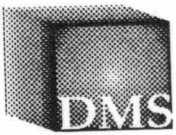
alberto battaglia

marco besozzi

STATISTICA PER MEDICI E BIOLOGI

con sviluppo mediante foglio elettronico su personal-computer



Editore: Data Management & Synthesis - Varese

I edizione - 1986
II edizione - 1987

alberto battaglia marco besozzi

STATISTICA PER MEDICI E BIOLOGI

con sviluppo mediante foglio elettronico su personal-computer

DMS

INDICE

1. INTRODUZIONE

Lo studio scientifico dei processi biologici, iniziato in sordina nel secolo scorso e letteralmente esploso nel corso degli ultimi decenni, ha comportato un sempre maggior ricorso a procedimenti di misura .

Cosi' oggi nessuno si meraviglia se i fisiologi misurano i potenziali generati dall'attivita' bioelettrica del cuore, se i biochimici misurano la costante di affinita' di un enzima per il suo substrato, se i farmacologi misurano la velocita' di assorbimento di un farmaco e il suo spazio di diffusione, e cosi' via.

Lo scopo dei ricercatori e', una volta ottenute le misure sperimentali, poter effettuare delle generalizzazioni dei dati ottenuti (delle inferenze), estendendo i risultati da un numero limitato di osservazioni (il campione), alla totalita' dei casi possibili (la popolazione).

Questo libro descrive appunto otto metodi di statistica inferenziale di impiego corrente nell'analisi di dati sperimentali.

Trattandosi di metodi che implicano sempre discrete difficolta' dal punto di vista computazionale, risulta ovvio il ricorso all'ausilio del personal-computer che oramai quasi tutti hanno, piu' o meno direttamente, a disposizione.

Tuttavia se statistica e personal-computer fossero i suoi unici ingredienti, questo libro non avrebbe forse ragione di essere: di libri di questo argomento ve ne sono gia' molti.

Ciò che lo rende veramente unico nel suo genere e' l'implementazione dei metodi statistici descritti su di un tabellone elettronico (spreadsheet), anzi su quello che attualmente puo' essere considerato un po' come il tabellone elettronico per eccellenza, cioe' MULTIPLAN.

A parte la rapidita', cosa che potrete ovviamente constatare di persona, con cui e' possibile implementare un metodo statistico anche abbastanza complesso su MULTIPLAN, gli altri sostanziali vantaggi offerti dall'utilizzo di un tabellone elettronico sono di tre tipi:

-possibilita' di avere globalmente una buona visione d'insieme dei dati e dei risultati intermedi dei calcoli

-possibilita' di utilizzare la logica del "cosa accade se.."

-possibilita' di elaborare dati provenienti da programmi di archivio (data-base) del commercio.

La prima possibilita' non richiede particolari commenti: semplicemente imparerete ben presto ad apprezzarla.

La seconda vi risultera' preziosa soprattutto in relazione alla possibilita' di ricalcolare rapidamente i risultati dei tests statistici, per esempio dopo l'eliminazione di un sospetto dato aberrante, o dopo l'aggiunta dei risultati di nuove osservazioni.

La terza rappresenta un po' un punto di forza. MULTIPLAN e' un tabellone elettronico in grado di accettare dati provenienti da vari programmi di archivio del commercio: potete quindi utilizzare i programmi statistici implementati su MULTIPLAN per elaborare direttamente dati provenienti da data-base compatibili, senza dover ogni volta reintrodurre da tastiera lunghe liste di valori numerici.

La descrizione dell'implementazione dei programmi viene fatta in questo libro utilizzando Macintosh, per le sue veramente eccellenti caratteristiche di presentazione grafica; tuttavia si

tenga presente che l'analoga versione di MULTIPLAN disponibile per PC-IBM e compatibili consente un rapido trasferimento dell'intero pacchetto su queste macchine.

Il contenuto di questo volume e' stato oggetto di una presentazione tenuta al First International Workshop in Occupational Health, organizzato da International Commission on Occupational Health e da Commission of the European Communities Joint Research Centre, Ispra Establishment, e tenutosi a Varese il 30 e 31 ottobre 1986.

2. IL CALCOLO DELLE STATISTICHE ELEMENTARI

Il problema

La raccolta sistematica di osservazioni e misure rappresenta la base dell'attivita' di ricerca: dall'analisi delle osservazioni e delle misure effettuate il ricercatore cerca di trarre conclusioni che egli spera abbiano una piu' ampia validita'.

Si consideri ora il caso piu' semplice di raccolta di dati, e cioe' quello della misura di una certa proprieta' in un insieme omogeneo di fenomeni od oggetti. Le domande che ci si pone sono di questo genere: le misure differiscono fra di loro 'tanto' o 'poco'? E' possibile definire dei descrittori che consentano una espressione oggettiva del 'tanto' e del 'poco' ? E' possibile in definitiva codificare una serie di descrittori in grado di darci una visione 'sintetica' ma sempre oggettiva dei risultati ottenuti, per non perdersi in elenchi di numeri che finirebbero con l'avere ben poco significato ? E i descrittori ottenuti analizzando un numero limitato di casi (un campione) possono essere considerati descrittori

attendibili anche dell'universo (la popolazione) da cui questo campione proviene ?

Le risposte a queste domande sono tutte dei 'si', con un 'se' concernente l'ultima, sulla quale e' necessario fermare per un momento l'attenzione : in effetti perche' i descrittori ottenuti dal campione possano essere generalizzati alla popolazione e' necessario che gli assunti posti alla base del modello matematico che viene impiegato per calcolare tali descrittori trovino riscontro nella popolazione stessa. In caso contrario la stima delle caratteristiche della popolazione effettuata a partire dal campione puo' risultare inattendibile. Ed e' quello che, nel caso delle statistiche elementari, puo' accadere se media e deviazione standard vengono impiegate per descrivere dei dati che non siano distribuiti in modo gaussiano, quando il modello che propone media e deviazione standard come descrittori attendibili assume come fondamento della propria validita' proprio il fatto di operare in presenza di una distribuzione gaussiana. Per questo motivo si sente fare riferimento, in contrapposizione alle tecniche

statistiche parametriche, che basano la loro validita' sull'assunto di una distribuzione gaussiana dei dati sperimentali (e degli errori), a tecniche statistiche non parametriche, che basano la loro validita' su assunti distribuzionali minimi, e che sono in grado di consentire descrizioni attendibili anche nel caso di distribuzioni non gaussiane.

La soluzione statistica

Una distribuzione gaussiana (o normale) e' completamente definita, da un punto di vista matematico, quando sono note due grandezze o 'parametri' (onde deriva il termine che ha dato il nome alla corrispondente classe di tecniche statistiche) della distribuzione : la media μ e la deviazione standard σ della popolazione.

In genere queste due grandezze non sono note, e vengono stimate da $\bar{x}$ e da s , rispettivamente la media e la deviazione standard di un campione estratto dalla popolazione, essendo, per un insieme comprendente un numero n di dati x_i

$$\bar{x} = \Sigma x_i / n \qquad (I)$$

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

ed essendo l'errore standard (cioe' la deviazione standard della media) espresso come

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

Al numeratore di **s**, nella (II), compare una grandezza, la devianza dei dati campionari, per calcolare la quale si ricorre in genere all'equivalenza algebrica, che consente uno svolgimento dei calcoli piu' pratico

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

Per esprimere il grado di scostamento della distribuzione dei dati trovata rispetto alla distribuzione teorica, si puo' ricorrere al coefficiente di asimmetria (g_1) e al coefficiente di curtosi (g_2), che indicano il tipo di scostamento dalla normalita' come qui di seguito specificato :

$g_1 < 0$: asimmetria negativa, cioe' coda sinistra della distribuzione eccessivamente lunga;

$g_1 > 0$: asimmetria positiva, cioe' coda destra della distribuzione eccessivamente lunga;

$g_2 < 0$: platicurtosi, cioe' distribuzione eccessivamente appiattita, con code troppo corte;

$g_2 > 0$: leptocurtosi, cioe' distribuzione eccessivamente alta, con code troppo lunghe.

Essendo allora

$$m_2 = \Sigma(x_i - \bar{x})^2 / n \quad \text{(IV)}$$

$$m_3 = \Sigma(x_i - \bar{x})^3 / n \quad \text{(V)}$$

$$m_4 = \Sigma(x_i - x)^4 / n \quad \text{(VI)}$$

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 e del coefficiente di curtosi sono calcolati come

$$g_1 = m_3 / (m_2 * \sqrt{m_2}) \quad \text{(VII)}$$

$$g_2 = m_4/m_2^2 - 3 \quad \text{(VIII)}$$

essendo il valore 3 sottratto nella (VIII) semplicemente per fare risultare uguale a zero il coefficiente di curtosi della curva normale (altrimenti pari a 3). Per il calcolo dei momenti secondo, terzo e quarto dalla media e' possibile utilizzare le seguenti equivalenze algebriche

$$\Sigma(x_i - x)^2/n = (\Sigma x_i^2 - n * x^2)/n \quad \text{(IX)}$$

$$\Sigma(x_i-\bar{x})^3/n = (\Sigma x_i^3 - 3*\bar{x}*\Sigma x_i^2+2*n*\bar{x}*\bar{x}^2) / n \quad (X)$$

$$\Sigma(x_i-\bar{x})^4/n = (\Sigma x_i^4-4*\bar{x}*\Sigma x_i^3+6*\bar{x}^2*\Sigma x_1^2-3*n*\bar{x}^2*\bar{x}^2)/n \quad (XI)$$

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

$$s_1 = \sqrt{6/n} \quad (XII)$$

$$s_2 = \sqrt{24/n} \quad (XIII)$$

Un test per la significativita' di g_1 (o g_2) ragionevolmente approssimato e' il seguente : il coefficiente di asimmetria (o quello di curtosi) viene considerato significativo se supera, in valore assoluto, di 2.6 volte la sua deviazione standard. In questo caso si rigetta l'ipotesi che i dati siano distribuiti

normalmente.

Qualora asimmetria e/o curtosi siano significative e' opportuno rinunciare a riportare la media e la deviazione standard, e conviene esprimere i risultati in termini di valore minimo osservato, valore massimo osservato, range. Una piu' ampia analisi non parametrica dei dati e' naturalmente possibile, ma esula dai limiti di questa trattazione.

L'applicazione su MULTIPLAN

Per impostare il programma su MULTIPLAN la prima cosa da fare e' selezionare un insieme di celle riservate all'introduzione dei dati : scegliete a questo scopo le celle della prima colonna, dalla riga 1 alla riga 200, cioe' le celle da R1C1 a R200C1, e definite questo insieme di celle con il nome

	1
1	
2	
3	
4	
5	
6	
7	
195	
196	
197	
198	
199	
200	

DATI, utilizzando il comando di MULTIPLAN 'definisci col nome'. In queste celle inserirete i valori dei dati campionari : il primo nella cella R1C1, il secondo nella cella R2C1, e via dicendo.

Per lo sviluppo dei successivi calcoli e' necessario disporre del quadrato, del cubo e della quarta potenza di ciascuno dei valori originali : incominciate a calcolarli per il primo dato, impostando rispettivamente

	2	3	4
1	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4
2			
3			
4			

-nella prima cella della colonna 2 (cella R1C2) l'espressione

=(RC[-1])^2

per l'elevazione al quadrato

-nella prima cella della colonna 3 (cella R1C3) l'espressione

=(RC[-2])^3

per l'elevazione al cubo

-nella prima cella della colonna 4 (cella R1C4) l'espressione

=(RC[-3])^4

per l'elevazione alla quarta potenza).

Mediante il comando di MULTIPLAN 'completa in basso' replicate tali espressioni fino alla riga 200 compresa , quindi

	2	3	4
1	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4
2	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4
3	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4
4	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4
5	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4
195	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4
196	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4
197	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4
198	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4
199	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4
200	=(RC[-1])^2	=(RC[-2])^3	=(RC[-3])^4

definite l'insieme delle 200 celle della colonna 2 (da R1C2 a R200C2) con il nome DATI2, l'insieme delle 200 celle della colonna 3 (da R1C3 a R200C3) con il nome DATI3, e l'insieme delle 200 celle della colonna 4 (da R1C4 a R200C4) con il nome DATI4.

E' necessario ora definire una serie di variabili necessarie al successivo sviluppo dei calcoli, e cioe' nell'ordine

-il numero dei dati mediante l'espressione

=COUNT(DATI)

(alla riga 202, definita con il nome NUM)

	1	2
202	NUM	=COUNT(DATI)
203	B	=SUM(DATI)
204	K	=SUM(DATI2)
205	D	=SUM(DATI3)
206	E	=SUM(DATI4)
207	H	=AVERAGE(DATI)
208	W	=H*H
209	IM	=(K-NUM*W)/NUM
210	LM	=(D-3*H*K+2*NUM*H*W)/NUM
211	MM	=(E-4*H*D+6*W*K-3*NUM*W*W)/NUM
212	AS	=LM/(IM*SQRT(IM))
213	SA	=SQRT(6/NUM)
214	CU	=MM/(IM*IM)-3
215	SC	=SQRT(24/NUM)

-la somma dei valori osservati con l'espressione

=SUM(DATI)

(alla riga 203, definita con il nome B)

-la somma dei quadrati dei valori osservati con l'espressione

=SUM(DATI2)

(alla riga 204, definita con il nome K)

-la somma dei cubi dei valori osservati con l'espressione

=SUM(DATI3)

(alla riga 205, definita con il nome D)

-la somma delle quarte potenze dei valori osservati con l'espressione

=SUM(DATI4)

(alla riga 206, definita con il nome E)

-la media dei valori osservati con l'espressione

=AVERAGE(DATI)

(alla riga 207, definita con il nome H)

-il quadrato della media dei valori osservati con l'espressione

=H*H

(alla riga 208, definita con il nome W)

-il momento di ordine secondo dalla media, calcolato come nella (IX), con l'espressione

=(K-NUM*W)/NUM

(alla riga 209, definita con il nome IM)

-il momento di ordine terzo dalla media, calcolato come nella (X), con l'espressione

=(D-3*H*K+2*NUM*H*W)/NUM

(alla riga 210, definita con il nome LM)

-il momento di ordine quarto dalla media, calcolato come nella (XI), con l'espressione

=(E-4*H*D+6*W*K-3*NUM*W*W)/NUM

(alla riga 211, definita con il nome MM)

-il coefficiente di asimmetria, calcolato come nella (VII), con l'espressione

=LM/(IM*SQRT(IM))

(alla riga 212, definita con il nome AS)

-la deviazione standard del coefficiente di asimmetria, calcolata come nella (XII), con l'espressione

=SQRT(6/NUM)

(alla riga 213, definita con il nome SA)

-il coefficiente di curtosi, calcolato come nella (VIII), con l'espressione

=MM/(IM*IM)-3

(alla riga 214, definita con il nome CU)

-la deviazione standard del coefficiente di curtosi, calcolata

come nella (XIII), con l'espressione

=SQRT(24/NUM)

(alla riga 215, definita con il nome SC)

Ecco che allora e' possibile riassumere in una videata finale il risultato dei calcoli effettuati (figura seguente), definendo rispettivamente

-il numero dei dati, gia' calcolato in precedenza alla riga 202, con l'espressione

=NUM

(alla riga 218)

	1	2
218	numero dati	=NUM
219	minimo	=MIN(DATI)
220	massimo	=MAX(DATI)
221	range	=MAX(DATI)-MIN(DATI)
222		
223	media	=AVERAGE(DATI)
224	deviazione standard	=STDEV(DATI)
225	errore standard	=STDEV(DATI)/SQRT(NUM)
227	coefficiente di asimmetria	=AS
228	ds del coeff.di asimmetria	=SA
229	rapporto	=ABS(AS/SA)
230	coefficiente di curtosi	=CU
231	ds del coeff.di curtosi	=SC
232	rapporto	=ABS(CU/SC)

-il valore minimo osservato con l'espressione, che utilizza un altro interessantissimo comando di MULTIPLAN,

=MIN(DATI)

(alla riga 219)

-il valore massimo osservato con l'espressione

=MAX(DATI)

(alla riga 220)

-il range dei valori osservati con l'espressione

=MAX(DATI)-MIN(DATI)

(alla riga 221)

-la media dei valori osservati con l'espressione che abbiamo gia' visto in precedenza, cioe'

=AVERAGE(DATI)

(alla riga 223)

-la deviazione standard dei valori osservati con l'espressione (ancora un'altro potente comando di MULTIPLAN)

=STDEV(DATI)

(alla riga 224)

-l'errore standard della media con l'espressione

=STDEV(DATI)/SQRT(NUM)

(alla riga 225)

-il coefficiente di asimmetria, gia' definito in precedenza alla riga 212, con l'espressione

=AS

(alla riga 227)

-la deviazione standard del coefficiente di asimmetria, gia' definita in precedenza alla riga 213, con l'espressione

=SA

(alla riga 228)

-il valore assoluto del rapporto fra il coefficiente di asimmetria e la sua deviazione standard con l'espressione

=ABS(AS/SA)

(alla riga 229)

-il coefficiente di curtosi, gia' definito in precedenza alla riga 214, con l'espressione

=CU

(alla riga 230)

-la deviazione standard del coefficiente di curtosi, gia' definita

in precedenza alla riga 215, con l'espressione

=SC

(alla riga 231)

-il valore assoluto del rapporto fra il coefficiente di curtosi e la sua deviazione standard con l'espressione

=ABS(CU/SC)

(alla riga 232).

Un esempio

Nell'uomo sono presenti, nel siero, varie molecole contenenti ciascuna piu' funzioni alcooliche (polioli) oltre al glucosio : il fruttosio, il mannosio, l'inositolo, il glucitolo, il mannitolo e altri. Per quanto riguarda per esempio il mannosio, si sa che la sua concentrazione nel siero risulta aumentata nella chetoacidosi diabetica e nella candidiasi invasiva, ed e' stato anche suggerito che esso possa rappresentare un nuovo marcatore metabolico dei carboidrati nei pazienti diabetici. La determinazione del mannosio del siero implica l'utilizzo di metodi analitici complessi quali la

cromatografia gas-liquido e la spettrometria di massa : a causa di cio' una valutazione preliminare dei valori normali del mannosio nel siero e' stata fatta limitandosi a 11 soggetti, e ottenendo i seguenti valori (espressi in milligrammi per litro di siero): 1.6, 3.8, 2.1, 2.7, 2.5, 2.7, 2.4, 3.3, 4.2, 3.7, 2.8.

Una volta introdotti i risultati nel programma su MULTIPLAN (figura accanto) calcolate i risultati mediante il comando 'calcola ora' , ed esaminate i dati ottenuti : il coefficiente di asimmetria e il coefficiente di curtosi non sono significativi (nessuno dei due supera di almeno 2.6 volte la rispettiva deviazione standard), quindi la distribuzione dei valori puo' essere assunta come normale, e la media e la deviazione standard possono essere utilizzate per descriverla. Tuttavia tenete presente che la numerosita' del

	1
1	1.6
2	3.8
3	2.1
4	2.7
5	2.5
6	2.7
7	2.4
8	3.3
9	4.2
10	3.7
11	2.8

	1	2
218	numero dati	11
219	minimo	1.6
220	massimo	4.2
221	range	2.6
222		
223	media	2.89091
224	deviazione standard	0.78289
225	errore standard	0.23605
227	coefficiente di asimmetria	0.16484
228	ds del coeff.di asimmetria	0.73855
229	rapporto	0.2232
230	coefficiente di curtosi	-0.854
231	ds del coeff.di curtosi	1.4771
232	rapporto	0.57813

campione e' assai ridotta : in casi di questo genere e' forse preferibile esprimere i risultati in termini di valore minimo e valore massimo osservati e di range. Lo stesso dicasi per i casi in cui, pure con campioni assai piu' numerosi, si evidenzia un significativo scostamento della distribuzione dalla ideale distribuzione gaussiana.

3. CONFRONTO FRA MEDIE: IL TEST T DI STUDENT PER CAMPIONI INDIPENDENTI

Il problema

Nel caso di dati quantitativi, il primo passo nell'analisi del campione e' rappresentato dalla espressione dei risultati ottenuti in forma sintetica, mediante opportuni descrittori: in particolare, come abbiamo visto nel capitolo dedicato al calcolo delle statistiche elementari, se si ritiene che un modello di distribuzione gaussiana consenta una descrizione soddisfacente dei risultati ottenuti, sono utilizzati come descrittori dei dati campionari (oltre ovviamente alla numerosita' campionaria) la media e la deviazione standard. Queste ultime sono a loro volta considerate stime attendibili della media e della deviazione standard della popolazione da cui deriva il campione, e nei riguardi della quale (questo e' in definitiva lo scopo) si desidera fare delle inferenze.

Esistono tuttavia numerose situazioni nelle quali interessa non solo ottenere le statistiche elementari di un campione, ma

anche valutarle in relazione a quelle di un altro campione: tipicamente la domanda che ci si pone e' 'le medie dei due campioni differiscono in modo significativo ? '.

Puo' trattarsi ad esempio di verificare se il valore medio della concentrazione del colesterolo nel siero di adulti sani di sesso maschile in una determinata fascia di eta' sia significativamente diverso da quello riscontrato in un analogo gruppo di soggetti di sesso femminile, oppure di controllare se la diminuzione media del peso in un gruppo di soggetti obesi sottoposti a un certo tipo di dieta sia significativamente diversa da quella osservata in un altro gruppo di soggetti obesi, sottoposto a una dieta differente, o ancora di stabilire se vi sia in media una differenza significativa fra la concentrazione del cromo presente nelle sorgenti di una certa zona rispetto ad un'altra: in ogni caso l'aspetto fondamentale, che condiziona il tipo di analisi statistica, e' rappresentato dal fatto che viene effettuato un confronto fra le medie di due campioni, ciscuno dei quali e' stato ottenuto in modo indipendente rispetto all'altro.

La soluzione statistica

Il test t di Student per il confronto fra le medie di campioni indipendenti, che viene qui presentato, e' basato su un principio semplice: a parita' di valore assoluto della differenza, a tale differenza viene dato tanto maggior peso quanto minore e' la dispersione dei dati campionari, e viceversa. Puo' quindi accadere che una ampia differenza fra le medie, alla quale saremmo propensi a dare importanza, risulti non significativa a causa della grande dispersione dei dati campionari, mentre per converso una piccola differenza puo' rivelarsi inaspettatamente significativa, qualora la dispersione dei dati campionari sia minima.

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 di t viene calcolato come

$$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s^2/n_1 + s^2/n_2)} \quad (1)$$

con $n_1 + n_2 - 2$ gradi di liberta'.

Al numeratore dell'espressione compare la differenza fra le medie, mentre al denominatore compare l'errore standard di questa differenza, la grandezza che, come si e' precedentemente accennato, consente dei dare il peso maggiore o minore alla differenza fra le medie.

La varianza s^2 e' la varianza combinata dei due campioni, e viene calcolata come rapporto fra la somma delle devianze dei due campioni e la somma dei loro gradi di liberta', cioe' come

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

Per entrambi i campioni il calcolo della devianza viene semplificato dall'uso della seguente equivalenza algebrica

$$\Sigma (x - \bar{x})^2 = \Sigma x^2 - (\Sigma x)^2 / n \quad \text{(III)}$$

Se il valore di t calcolato supera quello tabulato, per i gradi di liberta' e al livello di significativita' prescelto, si conclude che la differenza fra le medie e' significativa.

Un problema tuttavia complica lievemente la situazione: il test t di Student ordinario tende a fornire troppo pochi risultati significativi quando il campione piu' grande ha la varianza (cioe' la dispersione dei dati) maggiore, e troppi risultati significativi quando il campione piu' grande ha la varianza minore. Si consiglia percio' di 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 calcolano le varianze del primo e del secondo campione, rispettivamente come

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

con $n_1 - 1$ gradi di liberta', e come

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

con $n_2 - 1$ gradi di liberta', ed il valore del test **F** di omogeneita' fra le varianze come rapporto fra le varianza maggiore e la varianza minore, cioe' come

$$F = s_1^2/s_2^2 \qquad \text{se } s_1^2 > s_2^2 \quad (VI)$$

ovvero

$$F = s_2^2/s_1^2 \qquad \text{se } s_2^2 > s_1^2 \quad (VII)$$

Se il valore di **F** trovato supera quello riportato, per i gradi di liberta' e al livello di significativita' prescelto, si respinge l'ipotesi di uguaglianza delle due varianze.

Nel caso di varianze diverse o, come si dice di solito, non omogenee, si calcola la statistica

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

con gradi di liberta' pari a

$$G.L. = (v_1 + v_2)^2 / (v_1^2/(n_1-1) + v_2^2/(n_2-1)) \quad \text{(IX)}$$

essendo rispettivamente $v_1 = s_1^2/n_1$ e $v_2 = s_2^2/n_2$.

I gradi di liberta' calcolati mediante la (IX) sono arrotondati all'intero piu' vicino. Per il valore di t' viene utilizzata sempre la tabella del t di Student: se il valore trovato supera quello tabulato, per i gradi di liberta' e al livello di significativita' prescelto, la differenza fra le medie viene considerata significativa.

L'applicazione su MULTIPLAN

Selezionate ora due insiemi di celle, uno comprendente le

prime duecento celle della colonna 1 (celle da R1C1 a R200C1) per l'introduzione dei dati del primo campione (figura accanto), uno comprendente le prime duecento celle della colonna 2 (cella da R1C2 a R200C2) per l'introduzione dei dati del secondo campione (figura seguente) : definite tali insiemi di celle rispettivamente con i nomi DATI1 e DATI2 , utilizzando come al solito il comando di MULTIPLAN 'definisci col nome'.

Potete ora sfruttare ora la potenza dei comandi di MULTIPLAN per effettuare i calcoli necessari,

	1
1	
2	
3	
4	
5	
6	
195	
196	
197	
198	
199	
200	

	2
1	
2	
3	
4	
5	
6	
195	
196	
197	
198	
199	
200	

definendo (come nella figura che segue):

	1	2
202	NP	:=COUNT(DATI1)
203	NS	:=COUNT(DATI2)
204	MP	:=AVERAGE(DATI1)
205	VP	:=(STDEV(DATI1))^2
206	SP	:=STDEV(DATI1)
207	MS	:=AVERAGE(DATI2)
208	VS	:=(STDEV(DATI2))^2
209	SS	:=STDEV(DATI2)

-il numero dei dati del primo campione con l'espressione

=COUNT(DATI1)

(variabile definita col nome NP, alla riga 202)

-il numero dei dati del secondo campione con l'espressione

=COUNT(DATI2)

(variabile definita col nome NS, alla riga 203)

-la media dei dati del primo campione con l'espressione

=AVERAGE(DATI1)

(variabile definita col nome MP, alla riga 204)

-la varianza del primo campione come quadrato della sua deviazione standard , cioe' come

=(STDEV(DATI1))^2

(variabile definita col nome VP, alla riga 205)

-la deviazione standard del primo campione con l'espressione

=STDEV(DATI1)

(variabile definita col nome SP, alla riga 206)

-la media dei dati del secondo campione con l'espressione

=AVERAGE(DATI2)

(variabile definita col nome MS, alla riga 207)

-la varianza del secondo campione come quadrato della sua deviazione standard , cioe' come

=(STDEV(DATI2))^2

(variabile definita col nome VS, alla riga 208)

-la deviazione standard del secondo campione con l'espressione

=STDEV(DATI2)

(variabile definita col nome SS, alla riga 209).

Potete ora effettuare il calcolo delle statistiche, a partire dal rapporto fra varianze F, calcolato, come abbiamo detto in precedenza, come rapporto fra la varianza maggiore e la

varianza minore, utilizzando un particolare comando di MULTIPLAN, il comando

=IF(condizione, valore se vera, valore se falsa)

	1	2
210	F	:=IF(VP>VS,VP/VS,VS/VP)
211	LN	:=IF(VP>VS,NP-1,NS-1)
212	LD	:=IF(VP>VS,NS-1,NP-1)
213	I1	:=(VP*(NP-1)+VS*(NS-1))/(NP+NS-2)
214	T1	:=(MP-MS)/SQRT(I1/NP+I1/NS)
215	N1	:=NP+NS-2
216	Q	:=VP/NP
217	J	:=VS/NS
218	T2	:=(MP-MS)/SQRT(Q+J)
219	N	:=((Q+J)*(Q+J)/(Q*Q/(NP-1)+J*J/(NS-1)))
220	N2	:=INT(N+0.5)

Questo comando consente di assegnare ad una variabile (la variabile F nel nostro caso) uno di due valori, che dipende dal verificarsi (valore se vera) o dal non verificarsi (valore se falsa) di una condizione. In pratica nel nostro caso se la varianza del primo campione e' maggiore della varianza del secondo campione (VP>VS) alla variabile F (riga 210) viene assegnato il 'valore se vera' (cioe' VP/VS), mentre se la varianza del primo campione non e' maggiore della varianza

del secondo campione alla variabile F viene assegnato il 'valore se falsa' (cioe' VS/VP), il che viene facilmente espresso come indicato alla riga 210

=IF(VP>VS,VP/VS,VS/VP)

Lo stesso procedimento viene utilizzato per calcolare i gradi di liberta' del numeratore (variabile LN) e i gradi di liberta' del denominatore (variabile LD) che definirete rispettivemente come

=IF(VP>VS,NP-1,NS-1)

(variabile LN, alla riga 211) e come

=IF(VP>VS,NS-1,NP-1)

(variabile LD, alla riga 212) .

Alla riga 213 potete calcolare la varianza combinata dei due campioni come

=(VP*(NP-1)+VS*(NS-1))/(NP+NS-2)

assegnando tale valore alla variabile I1.

Alla riga 214 ecco finalmente il valore del test t di Student per varianze omogenee, calcolato secondo la (I) come

=(MP-MS)/SQRT(I1/NP+I1/NS)

e definito col nome T1, cui corrispondono i gradi di liberta' calcolati come

=NP+NS-2

alla riga 215, definiti con il nome N1.

Una volta definita la variabile Q (alla riga 216) come

=VP/NP

cioe' come rapporto fra varianza e numero dei dati del primo campione, e la variabile J (alla riga 217) come

=VS/NS

ovvero come rapporto fra varianza e numero dei dati del secondo campione, e' possibile calcolare il valore del test t per varianze non omogenee con l'espressione

=(MP-MS)/SQRT(Q+J)

in accordo con la (VIII) e assegnargli il nome T2 (alla riga 218).

Mediante l'espressione (IX) e' allora possibile calcolare i gradi di liberta' del test t per varianze non omogenee (definiti con il nome N alla riga 219) come

=((Q+J)*(Q+J)/(Q*Q/(NP-1)+J*J/(NS-1)))

arrotondati poi all'intero piu' vicino (variabile definita col

nome N2 alla riga 220) mediante l'espressione

=INT(N+0.5)

Ecco che allora potete facilmente strutturare una 'videata' che riassume i risultati ottenuti, riportando via via

	1	2	3
224		primo campione	secondo campione
225	numero delle osservazioni	=NP	=NS
226	media	=MP	=MS
227	deviazione standard	=SP	=SS
228			
229	rapporto F	=F	
230	gradi liberta' numeratore	=LN	
231	gradi liberta' denominatore	=LD	
232			
233	t per varianze omogenee	=T1	
234	gradi di libertà	=N1	
235			
236	t per varianze non omogenee	=T2	
237	gradi di libertà	=N2	

-il numero delle osservazioni, la media e la deviazione standard dei dati del primo campione alle righe da 225 a 227, nella colonna 2, definiti rispettivamente come =NP, =MP e =SP

-il numero delle osservazioni, la media e la deviazione standard dei dati del secondo campione alle righe da 225 a

227, nella colonna 3, definiti rispettivamente come =NS, =MS e =SS

-il valore del rapporto F definito come

=F

alla riga 229 e i corrispondenti gradi di liberta' del numeratore, definiti come

=LN

e i gradi di liberta' del denominatore definiti come

=LD

alle righe 230 e 231 rispettivamente

-il valore del t per varianze omogenee, definito come

=T1

alla riga 233, con i gradi di liberta' corrispondenti definiti alla riga 234 come

=N1

-il valore del t per varianze non omogenee, definito come

=T2

alla riga 236, con i gradi di liberta' corrispondenti definiti alla

riga 237 come

=N2

Un esempio

La riboflavina, o vitamina B_2 , entra come e' noto a far parte del flavin-adenin-dinucleotide (FAD) e del flavin-mononucleotide (FMN), che giocano un ruolo fondamentale nel trasporto dell'idrogeno nella catena respiratoria e nella biosintesi degli acidi grassi. I fabbisogni di riboflavina variano in funzione dell'eta' e del peso corporeo : aumentano per esempio in corso di gravidanza e allattamento. Allo scopo di verificare quali siano gli alimenti piu' consigliabili nel caso di aumentata richiesta di riboflavina, viene effettuato un confronto fra il muscolo e il fegato di bue. Una serie di determinazioni del contenuto di riboflavina nel muscolo di bue da' i seguenti risultati (in milligrammi di riboflavina per 100 grammi di carne) : 0.22, 0.18, 0.46, 0.86, 0.64, 0.28, 0.33, 0.35, 0.42. Nel caso del fegato si ottengono invece i seguenti valori : 0.95, 2.18, 1.12, 1.86.

Introducete i valori nel programma di MULTIPLAN ed effettuate i calcoli lanciando il comando 'calcola ora' .

	1	2
1	0.22	0.95
2	0.18	2.18
3	0.46	1.12
4	0.86	1.86
5	0.64	
6	0.28	
7	0.33	
8	0.35	
9	0.42	

I risultati ottenuti indicano che il contenuto medio di riboflavina del fegato e' superiore (1.5275 contro 0.4156 mg/100 grammi) a quello del muscolo, ma tale la differenza

	1	2	3
224		primo campione	secondo campione
225	numero delle osservazioni	9	4
226	media	0.4156	1.5275
227	deviazione standard	0.2161	0.5876
228			
229	rapporto F	7.3934	
230	gradi liberta' numeratore	3	
231	gradi liberta' denominatore	8	
232			
233	t per varianze omogenee	-5.1692	
234	gradi di libertà	11.0000	
235			
236	t per varianze non omogenee	-3.6757	
237	gradi di libertà	3.0000	

puo' essere considerata significativa ?

Per sapere quale forma del test t scegliere esaminate il valore del rapporto F : il valore ottenuto (7.3934) supera il valore tabulato (4.07) per 3 gradi di liberta' del numeratore e 8 del denominatore al livello di significativita' del 5%, quindi le varianze dei campioni differiscono significativamente, ed e' opportuno valutare i risultati mediante il test t per varianze non omogenee. Quest'ultimo (3.6757 ignorando il segno) supera a sua volta il valore tabulato (3.18) per 3 gradi di liberta', al livello di significativita' del 5% ($0.01<p<0.05$) : la differenza fra le medie risulta pertanto statisticamente significativa.

Si noti che il test t per varianze omogenee (5.1692 ignorando come al solito il segno) avrebbe comportato (erroneamente) un livello di significativita' ben maggiore, superando addirittura il valore (4.44) tabulato per 11 gradi di liberta' al livello di significativita' dello 0.1% ($p<0.001$).

4. CONFRONTO FRA MEDIE : IL TEST T DI STUDENT PER DATI APPAIATI

Il problema

Abbiamo visto nel capitolo precedente come confrontare fra di loro, mediante il test t di Student, le medie di campioni indipendenti, ciascuno dei quali cioe' tratto in modo del tutto autonomo dalla rispettiva popolazione.

In molti casi quella dei campioni indipendenti rappresenta una strada obbligata : si pensi per esempio al confronto della concentrazione del colesterolo totale nel siero fra un gruppo di soggetti di sesso maschile ed un gruppo di soggetto di sesso femminile omogeneo per eta', abitudini alimentari, estrazione socio-economica e cosi' via. Non vi e' alcuna sostanziale possibilita' di mettere in relazione i singoli elementi di un campione con quelli dell'altro : tuttavia e' chiaro che per quanti sforzi si facciano e' pur sempre possibile che i due campioni non risultino in tutto e per tutto omogenei, come invece si desidererebbe poiche' lo scopo e' di mettere in luce (se

esiste) la sola differenza dovuta al sesso. A questo punto e' inevitabile porsi una domanda : esiste un qualche modo per rendere ottimale l'omogeneita' dei campioni ? La risposta e' affermativa, e puo' essere bene illustrata da un altro esempio. Si pensi di volere confrontare il valore medio della pressione arteriosa sistolica a riposo con quello dopo una prova da sforzo standard, in adulti sani di sesso maschile, compresi in una certa fascia di eta'. Sarebbe possibile effettuare il confronto misurando la pressione sistolica a riposo in un campione magari anche ampio di soggetti, e quindi misurando la pressione sistolica dopo la prova da sforzo standard in un campione di soggetti piu' limitato, omogeneo con il primo. Ma perche' invece non misurare la pressione sistolica nello stesso gruppo di soggetti, prima e dopo la prova da sforzo ?. E' chiaro che in questo caso il problema dell'omogeneita' dei campioni viene superato : i due campioni, formati dagli stessi individui, risultano omogenei in tutto tranne che nel fattore che il ricercatore ha deliberatamente introdotto, cioe'

nella prova da sforzo.

Un opportuno disegno sperimentale consente in definitiva di superare il problema dell'omogeneita' dei campioni posti a confronto.

E' questo il semplice concetto che sta alla base dell'appaiamento dei dati, e quindi del test t di Student per dati appaiati.

La soluzione statistica

Nel test **t** di Student per dati appaiati la differenza fra le coppie di osservazioni diventa la variabile in esame. Dato un numero **n** di coppie 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

$$d_i = x_i - y_i \quad (1)$$

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

$$\bar{d} = \Sigma d_i / n \quad \text{(II)}$$

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

Si rammenta ancora una volta, a proposito del calcolo della devianza, l'equivalenza algebrica

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

Il valore del test t di Student e' calcolato come rapporto fra la differenza media e osservata e il suo errore standard, cioe'

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

con $n - 1$ gradi di liberta'. Se il valore di t calcolato supera quello tabulato, per i gradi di liberta' e al livello di significativita' prescelto, si conclude che la differenza media osservata e' significativa.

L'applicazione su MULTIPLAN

Iniziate l'impostazione del programma su MULTIPLAN selezionando le celle che utilizzerete per l'introduzione dei dati : le celle della prima colonna, dalla riga 1 alla riga 200, per l'introduzione del primo valore della coppia, le celle della seconda colonna, sempre dalla riga 1 alla riga 200, per l'introduzione del secondo valore di ciascuna coppia (figura della pagina seguente). Definite quindi con il nome DATI1 le celle da R1C1 a R200C1, e con il nome DATI2 le celle da R1C2 a R200C2.

	1
1	
2	
3	
4	
5	
6	
7	
195	
196	
197	
198	
199	
200	

Per calcolare la differenza fra il primo e il secondo valore di ciascuna coppia impostate nella prima cella della colonna 3 l'espressione

=RC[-2]-RC[-1]

e per calcolare il quadrato di tale differenza impostate nella prima cella della colonna 4 l'espressione

=RC[-1]^2

e quindi riportate queste due espressioni nelle 199 celle sottostanti rispettivamente di colonna 3 e di colonna 4 mediante il comando 'completa in basso': assegnate infine alle celle da R1C3 a R200C3 il nome DIFF e alle celle da R1C4 a R200C4 il nome DIFF2.

	2
1	
2	
3	
4	
5	
6	
7	
195	
196	
197	
198	
199	
200	

	3	4
1	=RC[-2]-RC[-1]	=RC[-1]^2
2	=RC[-2]-RC[-1]	=RC[-1]^2
3	=RC[-2]-RC[-1]	=RC[-1]^2
4	=RC[-2]-RC[-1]	=RC[-1]^2
5	=RC[-2]-RC[-1]	=RC[-1]^2
6	=RC[-2]-RC[-1]	=RC[-1]^2
7	=RC[-2]-RC[-1]	=RC[-1]^2
193	=RC[-2]-RC[-1]	=RC[-1]^2
194	=RC[-2]-RC[-1]	=RC[-1]^2
195	=RC[-2]-RC[-1]	=RC[-1]^2
196	=RC[-2]-RC[-1]	=RC[-1]^2
197	=RC[-2]-RC[-1]	=RC[-1]^2
198	=RC[-2]-RC[-1]	=RC[-1]^2
199	=RC[-2]-RC[-1]	=RC[-1]^2
200	=RC[-2]-RC[-1]	=RC[-1]^2

Definite poi (alla riga 202) con il nome NUM l'espressione

=COUNT(DATI1)

che rappresenta il numero delle coppie di valori introdotte, e alla riga 203 calcolate la varianza delle differenze, in base alla (III) e alla (IV) combinate, mediante l'espressione

=(SUM(DIFF2)-(SUM(DIFF))^2/NUM)/(NUM-1)

cui assegnerete il nome VAR.

	1	2
202	NUM	:=COUNT(DATI1)
203	VAR	:=(SUM(DIFF2)-(SUM(DIFF))^2/NUM)/(NUM-1)

I calcoli molto semplici fin qui effettuati possono allora essere riassunti in una tabella finale quale quella presentata, nella quale definirete

	1	2
204		:
205	numero delle osservazioni	:=NUM
206	media del primo campione	:=AVERAGE(DATI1)
207	media del secondo campione	:=AVERAGE(DATI2)
208	differenza media	:=AVERAGE(DATI1)-AVERAGE(DATI2)
209		:
210	valore di t	:=(SUM(DIFF)/NUM)/SQRT(VAR/NUM)
211	gradi di liberta'	:=NUM-1

-il numero delle osservazioni mediante l'espressione

=NUM

alla riga 205

-la media del primo valore di ciascuna coppia (media del primo campione) mediante l'espressione

=AVERAGE(DATI1)

alla riga 206

-la media del secondo valore di ciascuna coppia (media del secondo campione) mediante l'espressione

=AVERAGE(DATI2)

alla riga 207

-la differenza media osservata mediante l'espressione

=AVERAGE(DATI1)-AVERAGE(DATI2)

alla riga 208

-il valore di t calcolato come nella (V) mediante l'espressione

=(SUM(DIFF)/NUM)/SQRT(VAR/NUM)

alla riga 210

-i gradi di liberta' corrispondenti calcolati mediante l'espressione

=NUM-1

alla riga 211

Un esempio

La determinazione quantitativa dell'emoglobina A_2 rappresenta

il test diagnostico per lo stato di portatore eterozigote di ß-talassemia. Nel corso di un programma di screening della ß-talassemia risulta necessario, per ragioni organizzative, concentrare le analisi in giorni determinati : cio' comporta che alcuni campioni vengano analizzati immediatamente dopo il prelievo, mentre altri vengono analizzati a distanza di 2 o tre giorni da esso. Per verificare se il rinvio delle analisi comporti qualche modifica dei livelli di emoglobina A_2, questa viene determinata su un certo numero di campioni entro 2 ore dal prelievo e dopo 72 ore, ottenendo le seguenti coppie di valori (emoglobina A_2 come percentuale dell'emoglobina totale, valore dopo 2 ore/valore dopo 72 ore) : 1.8/1.8, 2.6/2.5, 4.8/4.5, 3.2/3.2, 3.1/3.2, 5.2/5.3, 4.5/4.5, 4.3/4.1, 1.9/2.1, 2.5/2.4.

Una volta introdotti i dati nel programma su MULTIPLAN, potete ottenere rapidamente i

	1	2
1	1.8	1.8
2	2.6	2.5
3	4.8	4.5
4	3.2	3.2
5	3.1	3.2
6	5.2	5.3
7	4.5	4.5
8	4.3	4.1
9	1.9	2.1
10	2.5	2.4

risultati lanciando la procedura di calcolo con il solito comando 'calcola ora'.

	1	2
204		
205	numero delle osservazioni	10
206	media del primo campione	3.39
207	media del secondo campione	3.36
208	differenza media	0.03
209		
210	valore di t	0.63481
211	gradi di liberta'	9

La differenza media osservata risulta molto bassa (0.03) : il valore di t calcolato (0.63481) si colloca fra 0.70 e 0.13, i valori tabulati per 9 gradi di liberta', ai livelli di probabilita' di 0.5 e 0.9 rispettivamente. La probabilita' di osservare per caso una differenza della stessa entita' di quella effettivamente osservata risulta quindi superiore al 50%: una probabilta' troppo elevata, inaccettabile, che viene respinta, concludendo che le medie non differiscono significativamente.

5. REGRESSIONE LINEARE : IL MODELLO STANDARD

Il problema

Abbiamo finora considerato problemi che implicano una singola misura su ciascun elemento del campione : vediamo ora il caso in cui interessi ricostruire la relazione di funzione che lega due variabili, la variabile y (variabile dipendente) alla variabile x (variabile indipendente).

Se si ritiene che la relazione esistente fra le due variabili possa essere convenientemente descritta mediante una retta, l'equazione di tale retta puo' essere calcolata mediante la tecnica statistica nota come regressione lineare.

La denominazione di regressione lineare deriva dagli studi sull'ereditarieta' dei caratteri condotti da F. Galton sul finire dell'800. Nel corso di questi studi vennero, fra le altre cose, registrate le altezze dei componenti di piu' di 1000 gruppi familiari. Ponendo su un sistema di assi cartesiani in ascisse le altezze dei padri e in ordinate le altezze dei figli, si nota un fatto : sebbene in genere padri piu' alti avessero figli piu' alti

(come era del resto da attendersi), padri che erano di 16 centimetri circa piu' alti della media dei padri, avevano figli che erano solamente 8 centimetri piu' alti della media dei figli. In altre parole sembrava che vi fosse un 'tornare indietro', una regressione delle altezze dei figli rispetto a quelle dei padri : e il termine che descriveva il risultato di questa iniziale applicazione, fini' con l'essere impiegato per indicare la tecnica statistica, ed e' rimasto ancora oggi .

La soluzione statistica

Una volta accettato che l'equazione di una retta sia adatta a rappresentare la relazione di funzione che intercorre fra due variabili, il problema che si pone consiste nello stimare nel migliore modo possibile la regressione vera $y = \alpha + \beta x$ a partire da una serie di osservazioni costituite da coppie di valori sperimentali (x_i, y_i). Per fare cio' bisogna disporre di un metodo che consenta un buon adattamento della retta ai dati sperimentali, intendendosi con cio' un adattamento tale

da rendere piccolo l'errore totale. Il metodo dei minimi quadrati risponde a questo scopo, essendo possibile dimostrare che esso, fra gli stimatori lineari corretti di β (o α), e' quello che presenta la varianza minima (cioe' appunto l'errore totale piu'piccolo).

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

$$y = a + b\,x$$

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

$$b = \Sigma(x_i - \bar{x})*(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 \quad (I)$$

$$a = \bar{y} - b*\bar{x} \quad (II)$$

Il coefficiente di correlazione **r** viene calcolato come

$$r = \Sigma(x_i - \bar{x})*(y_i - \bar{y}) / \sqrt{\Sigma(x_i - \bar{x})^2 * \Sigma(y_i - \bar{y})^2} \quad \text{(III)}$$

La varianza residua attorno alla regressione viene calcolata come

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

essendo

$$s_1^2 = (\Sigma(x_i - \bar{x})*(y_i - \bar{y}))^2 / \Sigma(x_i - \bar{x})^2 \quad \text{(V)}$$

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** e di **b**, sono calcolati come

$$s_{yx} = \sqrt{s_0^2} \quad \text{(VI)}$$

$$s_b = s_{yx} * \sqrt{1 / \Sigma(x_i - \bar{x})^2} \quad \text{(VII)}$$

$$s_a = s_b * \sqrt{\Sigma x_i^2 / n} \quad \text{(VIII)}$$

Si tenga presente che per il calcolo della devianza delle x, della devianza delle y e della codevianza e' possibile utilizzare, per semplicita', le seguenti equivalenze algebriche

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

$$\Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n \quad \text{(X)}$$

$$\Sigma(x_i - \bar{x})*(y_i - \bar{y}) = \Sigma x_i * y_i - (\Sigma x_i)*(\Sigma y_i) / n \quad \text{(XI)}$$

Il fatto importante e' che tutti i calcoli possono essere sviluppati a partire dalle seguenti grandezze fondamentali : numero dei dati (n), sommatoria delle x_i (Σx_i), sommatoria delle y_i

(Σy_i), sommatoria dei quadrati delle x_i (Σx_i^2), sommatoria dei quadrati delle y_i (Σy_i^2), sommatoria dei prodotti delle x_i per le y_i ($\Sigma x_i * y_i$).

L'applicazione su MULTIPLAN

La prima cosa da fare, come al solito, e' stabilire quali siano le righe e le colonne nelle quali introdurremo i dati. Selezionate quindi le prime duecento celle della colonna 1 (celle da R1C1 a R200C1) per l'introduzione dei valori della x , e definite l'insieme di queste duecento celle con il nome DATI1.

Selezionate poi le prime duecento celle della colonna 2 (celle da R1C2 a R200C2) per l'introduzione dei valori della y, e definite l'insieme di queste duecento celle con il nome DATI2.

	1
1	.....
2	.....
3	.....
4	.....
5	.....
195	.....
196	.....
197	.....
198	.....
199	.....
200	

In questo modo ogni riga a partire dalla R1 conterra' le

coordinate di un punto, il valore della x nella colonna 1, e il corrispondente valore della y nella colonna 2. La necessita' di disporre, per il successivo sviluppo dei calcoli, dei quadrati delle x, dei quadrati delle y, e dei prodotti di ciascun valore della x per il corrispondente valore della y, viene risolta aprendo tre nuove colonne, dalla colonna tre alla colonna cinque. Nella prima riga della terza colonna (cella R1C3) riportate l'espressione

	2
1	.....
2	.....
3	.....
4	.....
5	
195	.....
196	.....
197	.....
198	.....
199	.....
200	

=RC[-2]^2

che consente di elevare al quadrato il valore contenuto nella cella della medesima riga e che la precede di due colonne, e quindi in definitiva di elevare al quadrato il contenuto della corrispondente cella della colonna 1, che e' il valore della x (figura alla pagina seguente).

Nella prima riga della quarta colonna (cella R1C4) riportate

	3	4	5
1	=RC[-2]^2	:=RC[-2]^2	:=RC[-4]*RC[-3]

ancora l'espressione

=RC[-2]^2

che consente di elevare al quadrato il valore contenuto nella cella della medesima riga e che la precede di due colonne, e quindi questa volta di elevare al quadrato il contenuto della corrispondente cella della colonna 2, che e' il valore della y.

Nella prima riga della quinta colonna (R1C5) riportate infine l'espressione

=RC[-4]*RC[-3]

che moltiplicando il contenuto della cella della colonna 1 per quello della cella della colonna 2 consente di ottenere il valore del prodotto x*y.

Ad evitare di dovere ripetere l'operazione per tutte le duecento righe riservate all'introduzione dei dati ci soccorre ancora una volta la potenza dei comandi di MULTIPLAN che, con l'istruzione 'completa in basso' (figura alla pagina seguente), consente di ottenere in pochi istanti la necessarie ripetizioni

	3	4	5
1	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
2	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
3	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
4	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
5	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
195	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
196	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
197	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
198	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
199	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
200	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]

delle formule.

Non resta ora che dare un nome al contenuto delle nuove colonne, e precisamente:

-il nome DATI1Q al contenuto della colonna 3

-il nome DATI2Q al contenuto della colonna 4

-il nome PROD al contenuto della colonna 5

Ecco che si puo' allora procedere allo sviluppo dei calcoli, definendo nell'ordine

-il numero dei dati con l'espressione

=COUNT(DATI1)

(alla riga 202, e che definirete con il nome NUM)

	1	2
202	NUM	:=COUNT(DATI1)
203	B	:=SUM(DATI1)
204	A	:=SUM(DATI2)
205	D	:=SUM(DATI1Q)
206	E	:=SUM(DATI2Q)
207	F	:=SUM(PROD)
208	G	:=D-B*B/NUM
209	H	:=E-A*A/NUM
210	I	:=F-B*A/NUM
211	L	:=I*I/G

-la sommatoria dei valori delle x con l'espressione

=SUM(DATI1)

(alla riga 203, definita con il nome B)

-la sommatoria dei valori delle y con l'espressione

=SUM(DATI2)

(alla riga 204, definita con il nome A)

-la sommatoria dei quadrati delle x con l'espressione

=SUM(DATI1Q)

(alla riga 205, definita con il nome D)

-la sommatoria dei quadrati delle y con l'espressione

=SUM(DATI2Q)

(alla riga 206, definita con il nome E)

-la sommatoria dei prodotti dei valori di ciascuna x per il corrispondente valore della y con l'espressione

=SUM(PROD)

(alla riga 207, definita con il nome F)

-la devianza delle x, calcolata secondo la (IX), con l'espressione

=D-B*B/NUM

(alla riga 208, definita con il nome G)

-la devianza delle y, calcolata secondo la (X), con l'espressione

=E-A*A/NUM

(alla riga 209, definita con il nome H)

-la codevianza, calcolata secondo la (XI), con l'espressione

=F-B*A/NUM

(alla riga 210, definita con il nome I)

-la varianza spiegata dalla regressione, calcolata secondo la (V), con l'espressione

=I*I/G

(alla riga 211, definita con il nome L)

e ancora definendo (figura alla pagina seguente)

	1	2
212	M	:=(H-L)/(NUM-2)
213	SL	:=I/G
214	IN	:=A/NUM-SL*B/NUM
215	CC	:=I/SQRT(G*H)
216	DR	:=SQRT(M)
217	DS	:=SQRT(M)*SQRT(1/G)
218	DI	:=SQRT(M)*SQRT(1/G)*SQRT(D/NUM)

-la varianza residua attorno alla regressione, calcolata secondo la (IV), con l'espressione

=(H-L)/(NUM-2)

(alla riga 212, definita con il nome M)

-il coefficiente angolare della retta di regressione calcolato in base alla (I), con l'espressione

=I/G

(alla riga 213, definita con il nome SL)

-l'intercetta della retta di regressione calcolata in base alla (II), con l'espressione

=A/NUM-SL*B/NUM

(alla riga 214, definita con il nome IN)

-il coefficiente di correlazione, calcolato in base alla (III), con l'espressione

=I/SQRT(G*H)

(alla riga 215, definito con il nome CC)

-la deviazione standard residua, o errore standard della stima, calcolata in base alla (VI), con l'espressione

=SQRT(M)

(alla riga 216, definita con il nome DR)

-la deviazione standard del coefficiente angolare b, calcolata in base alla (VII), con l'espressione

=SQRT(M)*SQRT(1/G)

(alla riga 217, definita con il nome DS)

-la deviazione standard dell'intercetta a, calcolata in base alla (VIII), con l'espressione

=SQRT(M)*SQRT(1/G)*SQRT(D/NUM)

(alla riga 218, definita con il nome DI)

Non rimane ora che riassumere i risultati ottenuti in modo chiaro, riportando via via nelle righe da 220 in avanti il valore del coefficiente angolare b (=SL, alla riga 220) e dell'intercetta a (=IN, alla riga 221) della retta di regressione, il valore del

coefficiente di correlazione r (=CC, alla riga 222), e quindi i

	1	2
220	coefficiente angolare b	=SL
221	intercetta a	=IN
222	coefficiente di correlazione r	=CC
223		
224	deviazione standard di b	=DS
225	deviazione standard di a	=DI
226	errore standard della stima	=DR

valori della deviazione standard del coefficiente angolare (=DS, alla riga 224), della deviazione standard dell'intercetta (=DI, alla riga 225) e dell'errore standard della stima (=DR, alla riga 226). Oltre a questo, vale la pena di completare il lavoro fatto riservando alcune celle allo svolgimento delle interpolazioni, quella della x data la y, con l'espressione

=(Y-IN)/SL

e quella della y data la x mediante l'espressione

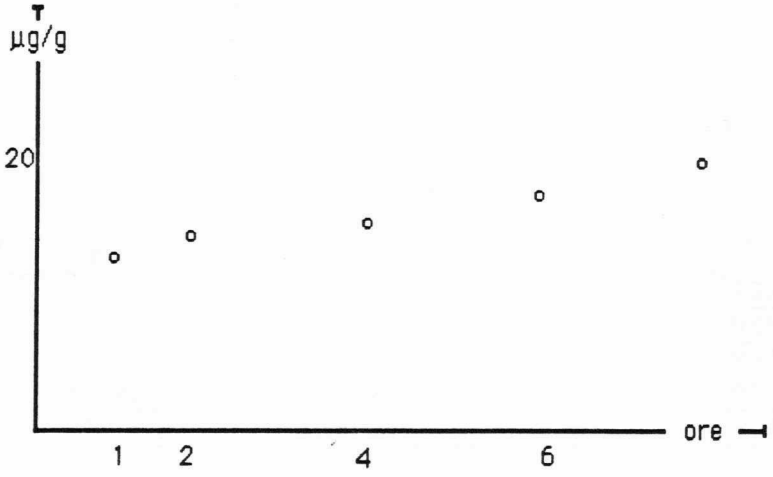
=IN+SL*X

	1	2	3	4
227				
228				
229				
230	Y=		X -->	=(Y-IN)/SL
231	X=		Y -->	=IN+SL*X

che riporterete entrambe nella colonna 4, rispettivamente alla riga 230 e alla riga 231. Riservate quindi le corrispondenti celle della colonna 2 all'introduzione dei dati : la cella R230C2 all'introduzione del valore della y da interpolare, definendola con il nome Y, e la cella R231C2 all'introduzione del valore della x da interpolare, definendola con il nome X.

Un esempio

Nell'ambito di uno studio sulla esposizione al cromo in ambienti lavorativi, si vuole verificare l'andamento della cromuria in funzione del tempo di esposizione a



verniciatura con vernice contenente cromo quale inibitore di corrosione. Per questo il cromo eliminato con le urine (espresso in microgrammi per grammo di creatinina) viene determinato in uno stesso soggetto dopo 1, 2, 4, 6 e 8 ore di

	1	2
1	1	8.2
2	2	10.4
3	4	12.6
4	6	16.1
5	8	20

esposizione, ottenendo le seguenti coppie di valori (tempo/cromuria) : 1/8.2, 2/10.4, 4/12.6, 6/16.1, 8/20.

L'ispezione dei dati mostra che essi possono essere ragionevolmente approssimati mediante una retta.

Una volta introdotti i dati nel programma di MULTIPLAN , potete facilmente calcolare i risultati, e tracciare quindi sul diagramma di distribuzione dei dati sperimentali la retta calcolati, che sembra fornire una piu' che ragionevole rappresentazione dell'andamento medio della cromuria in funzione del tempo di esposizione.


µg/g
20
ore
1
2
4
6

6. REGRESSIONE LINEARE: LA COMPONENTE PRINCIPALE STANDARDIZZATA

Il problema

Alla base del modello classico di regressione lineare stanno tre fondamentali assunti riguardanti la relazione tra y e x:

- per ogni valore della x vi e' una serie di valori della y distribuiti normalmente, dai quali il valore campionario e' tratto a caso (cioe' la x viene assunta come misurata senza errore, la y come affetta da un errore casuale distribuito normalmente)
- la popolazione dei valori di y che corrisponde a un dato valore della x ha una media che giace sulla retta $\mu = \alpha + \beta x$
- in ciascuna delle popolazioni dei valori di y la deviazione standard di y attorno alla media $\alpha + \beta x$ viene assunta come costante al variare della x.

Se tali assunti non vengono rispettati e' possibile che il modello fornisca una stima dei valori del coefficiente angolare e dell'intercetta anche grossolanamente scorretti.

Esistono delle situazioni nelle quali viene violato il primo

assunto, quello che prevede che la x non sia affetta da alcun errore di misura : e' allora possibile utilizzare modelli di regressione lineare alternativi al modello standard, che tengono conto del fatto che entrambe le variabili sono affette da errore di misura. La componente principale standardizzata rappresenta una possibile alternativa al modello standard di regressione lineare in casi di questo genere.

La soluzione statistica

Il modello standard di regressione lineare comporta (a meno che i punti siano perfettamente allineati su una retta) il calcolo di due differenti equazioni della retta di regressione : una e' quella (che viene normalmente calcolata, implementata anche su calcolatrici tascabili) ottenuta minimizzando la sommatoria dei quadrati delle differenze residue nella direzione dell'asse delle y, corrispondente all'ipotesi che sia la x la variabile indipendente (x misurata, come abbiamo detto, senza errore) ; l'altra corrisponde all'ipotesi che sia la y ad essere misurata senza errore, e comporta la minimizzazione della sommatoria

dei quadrati delle differenze residue nella direzione dell'asse delle x (e non viene in effetti mai utilizzata). Nei casi in cui entrambe le variabili siano soggette ad errore di misura, e quindi possa sorgere il dubbio su quale debba essere considerata come variabile indipendente, la componente principale standardizzata fornisce una stima intermedia del valore del coefficiente angolare b (e dell'intercetta a).

Si consideri il caso di **n** punti aventi coordinate cartesiane note (x_i , y_i), essendo $\bar{x}$ la media dei valori delle x_i , e $\bar{y}$ la media dei valori delle y_i . Essendo allora b_{yx} il coefficiente angolare della regressione x variabile indipendente, e b_{xy} il coefficiente angolare della regressione y variabile indipendente, il coefficiente angolare della componente principale standardizzata b_{cps} risulta pari alla media geometrica dei due, ovvero

$$b_{cps} = \sqrt{b_{xy} * b_{yx}} \quad (1)$$

Dato che

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

e che

$$b_{xy} = \Sigma (y_i - \bar{y})^2 / \Sigma (x_i - \bar{x})^*(y_i - \bar{y}) \quad \text{(III)}$$

avremo allora che, per la (I)

$$b_{cps} = \sqrt{\Sigma(y_i - \bar{y})^2 / \Sigma(x_i - \bar{x})^2} \quad \text{(IV)}$$

L'intercetta **a** viene calcolata sempre come

$$a = \bar{y} - b\bar{x} \quad \text{(V)}$$

mentre per il calcolo della devianza delle x e della devianza delle y si rammentano le gia' citate equivalenze algebriche

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

$$\Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2/n \quad \text{(VII)}$$

Il calcolo della codevianza, che viene effettuato con la seconda delle espressioni algebricamente equivalenti

$$\Sigma(x_i - \bar{x})*(y_i - \bar{y}) = \Sigma x_i*y_i - (\Sigma x_i)*(\Sigma y_i)/n \quad \text{(VIII)}$$

si rende necessario per ottenere il valore del coefficiente di correlazione **r**

$$r = \Sigma(x_i - \bar{x})*(y_i - \bar{y}) / \sqrt{\Sigma(x_i - \bar{x})^2*\Sigma(y_i - \bar{y})^2} \quad \text{(IX)}$$

che e' pertanto uguale (come vi era da attendersi) a quello della regressione standard.

L'applicazione su MULTIPLAN

L'implementazione su MULTIPLAN della componente principale standardizzata ricalca pressoche' completamente quanto fatto per la regressione lineare. Selezionate quindi le

prime duecento celle della colonna 1 (celle da R1C1 a R200C1) per l'introduzione dei valori della x , e definite l'insieme di queste duecento celle con il nome DATI1.

	1	2
1		
2		
3		
4		
5		
195		
196		
197		
198		
199		
200		

Selezionate poi le prime duecento celle della colonna 2 (celle da R1C2 a R200C2) per l'introduzione dei valori della y, e definite l'insieme di queste duecento celle con il nome DATI2.

In questo modo ogni riga a partire dalla R1 conterra' le coordinate di un punto, il valore della x nella colonna 1, e il corrispondente valore della y nella colonna 2.

La necessita' di disporre, per il successivo sviluppo dei calcoli, dei quadrati delle x, dei quadrati delle y, e dei prodotti di

ciascun valore della x per il corrispondente valore della y, viene risolta aprendo tre nuove colonne, dalla colonna tre alla colonna cinque. Nella prima riga della terza colonna (cella R1C3) riportate l'espressione

=RC[-2]^2

	3	4	5
1	=RC[-2]^2	:=RC[-2]^2	:=RC[-4]*RC[-3]

che consente di elevare al quadrato il valore contenuto nella cella della medesima riga e che la precede di due colonne, e quindi in definitiva di elevare al quadrato il contenuto della corrispondente cella della colonna 1, che e' il valore della x .

Nella prima riga della quarta colonna (cella R1C4) riportate ancora l'espressione

=RC[-2]^2

che consente di elevare al quadrato il valore contenuto nella cella della medesima riga e che la precede di due colonne, e quindi questa volta di elevare il quadrato il contenuto della corrispondente cella della colonna 2, che e' il valore della y.

Nella prima riga della quinta colonna (R1C5) riportate infine l'espressione

=RC[-4]*RC[-3]

che moltiplicando il contenuto della cella della colonna 1 per quello della cella della colonna 2 consente di ottenere il valore del prodotto x*y.

Ad evitare di dovere ripetere l'operazione per tutte le duecento righe riservate all'introduzione dei dati ci soccorre ancora una volta la potenza dei comandi di MULTIPLAN che, con l'istruzione 'completa in basso', consente di ottenere in pochi istanti la necessarie ripetizioni delle formule.

	3	4	5
1	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
2	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
3	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
4	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
5	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
6	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
195	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
196	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
197	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
198	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
199	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]
200	=RC[-2]^2	=RC[-2]^2	=RC[-4]*RC[-3]

Non resta ora che dare un nome al contenuto delle nuove colonne, e precisamente:

-il nome DATI1Q al contenuto della colonna 3

-il nome DATI2Q al contenuto della colonna 4

-il nome PROD al contenuto della colonna 5

Ecco che si puo' allora procedere allo sviluppo dei calcoli, definendo nell'ordine:

-il numero dei dati con l'espressione

=COUNT(DATI1)

(alla riga 202, e che definirete con il nome NUM)

-la sommatoria dei valori delle x con l'espressione

=SUM(DATI1)

(alla riga 203, definita con il nome B)

	1	2
202	NUM	:=COUNT(DATI1)
203	B	:=SUM(DATI1)
204	A	:=SUM(DATI2)
205	D	:=SUM(DATI1Q)
206	E	:=SUM(DATI2Q)
207	F	:=SUM(PROD)
208	G	:=D-B*B/NUM
209	H	:=E-A*A/NUM
210	I	:=F-B*A/NUM
211	SL	:=SQRT(H/G)
212	IN	:=A/NUM-SL*B/NUM
213	CC	:=I/SQRT(G*H)

-la sommatoria dei valori delle y con l'espressione

=SUM(DATI2)

(alla riga 204, definita con il nome A)

-la sommatoria dei quadrati delle x con l'espressione

=SUM(DATI1Q)

(alla riga 205, definita con il nome D)

-la sommatoria dei quadrati delle y con l'espressione

=SUM(DATI2Q)

(alla riga 206, definita con il nome E)

-la sommatoria dei prodotti dei valori di ciascuna x per il corrispondente valore della y con l'espressione

=SUM(PROD)

(alla riga 207, definita con il nome F)

-la devianza delle x, calcolata secondo la (VI), con l'espressione

=D-B*B/NUM

(alla riga 208, definita con il nome G)

-la devianza delle y, calcolata secondo la (VII), con l'espressione

=E-A*A/NUM

(alla riga 209, definita con il nome H)

-la codevianza, calcolata secondo la (VIII), con l'espressione

=F-B*A/NUM

(alla riga 210, definita con il nome I)

-il coefficiente angolare della retta di regressione calcolato in

base alla (IV), con l'espressione

=SQRT(H/G)

(alla riga 211, definita con il nome SL)

-l'intercetta della retta di regressione calcolata in base alla (V), con l'espressione

=A/NUM-SL*B/NUM

(alla riga 212, definita con il nome IN)

-il coefficiente di correlazione, calcolato in base alla (IX), con l'espressione

=I/SQRT(G*H)

(alla riga 213, definito con il nome CC)

Non rimane ora che riassumere i risultati ottenuti in modo chiaro, riportando via via nelle righe da 215 a 217 il valore del coefficiente angolare b (=SL, alla riga 215), dell'intercetta a (=IN, alla riga 216) della retta di regressione e il valore del coefficiente di correlazione r (=CC, alla riga 217).

Oltre a questo, vale la pena di completare il lavoro fatto riservando alcune celle allo svolgimento delle

	1	2
215	coefficiente angolare b	=SL
216	intercetta a	=IN
217	coefficiente di correlazione r	=CC

interpolazioni, quella della x data la y, con l'espressione

=(Y-IN)/SL

e quella della y data la x mediante l'espressione

	1	2	3	4
221	Y=		X -->	=(Y-IN)/SL
222	X=		Y -->	=IN+SL*X
223				

=IN+SL*X

che riporterete nella colonna 4, rispettivamente alla riga 221 e alla riga 222.

Riservate quindi le corrispondenti celle della colonna 2 all'introduzione dei dati : la cella R221C2 all'introduzione del valore della y da interpolare, definendola con il nome Y, e la cella R222C2 all'introduzione del valore della x da interpolare, definendola con il nome X.

Un esempio

L'antigene carcinoembrionario (CEA) e' un antigene oncofetale prodotto da vari tipi di tumori che, pur non essendo utilizzabile a scopo di screening e/o di diagnosi precoce, puo'

rappresentare un marcatore precoce di eventuali recidive (in loco o metastatiche) di tumori asportati chirurgicamente. Per la determinazione del CEA nel siero vengono utilizzate tecniche immunologiche, che prevedono l'utilizzo di anticorpi anti-CEA marcati con radioisotopi, con enzimi o con composti fluorescenti. Allo scopo di migliorare l'affidabilita' del test, anche nel caso del CEA vengono proposti dosaggi impieganti anticorpi monoclonali : un confronto preliminare fra il metodo in uso che impiega un anticorpo policlonale e un nuovo metodo basato sull'utilizzo di un anticorpo monoclonale fornisce i seguenti risultati

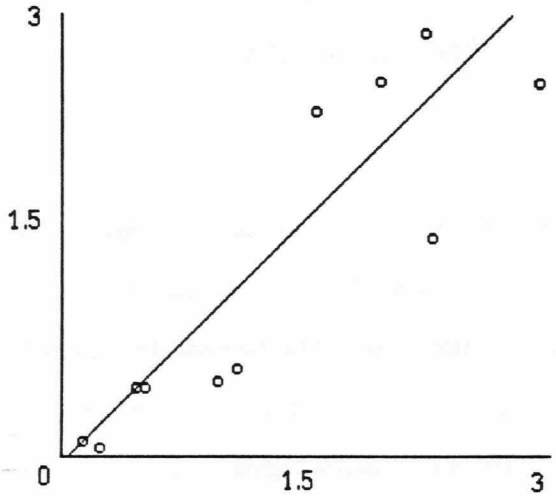
	1	2
1	0.2	0.4
2	0.8	0.8
3	0.6	0.8
4	2.2	1.4
5	1.9	2.4
6	3.1	2.4
7	1.6	1.9
8	2.4	2.9
9	0.9	0.6
10	1.1	0.7
11	0.3	0.1

(vecchio metodo/nuovo metodo, in nanogrammi per millilitro di siero) : 0.2/0.4, 0.8/0.8, 0.6/0.8, 2.2/1.4, 1.9/2.4, 3.1/2.4, 1.6/1.9, 2.4/2.9, 0.9/0.6, 1.1/0.7, 0.3/0.1.

Una volta introdotti i valori nel programma di MULTIPLAN e lanciata l'esecuzione dei calcoli , potete visualizzare i risultati, che mostrano un'equazione della retta

	1	2	3	4
215	coefficiente angolare b	1.00833		
216	intercetta a	-0.0751		
217	coefficiente di correlazione	0.8848		
218				
219				
220				
221	Y=	0	X -->	0.07
222	X=	0	Y -->	-0.08
223				

di regressione molto prossima alla retta y = x che rappresenta il caso ideale, quello in cui entrambi i metodi forniscono il medesimo risultato. Tenete presente che in questo caso la regressione lineare standard (provate a verificarlo!) fornisce un coefficiente angolare pari a 0.89217, che sembrerebbe



suggerire una conclusione ben diversa !

7. ANALISI DI FREQUENZE : IL TEST CHI-QUADRATO PER TABELLE DI CONTINGENZA

Il problema

I fenomeni biologici sono associati a un grado piu' o meno elevato di variabilita' : l'approccio allo studio su base scientifica di tale variabilita' prevede in genere l'utilizzo di descrizioni di tipo quantitativo, cioe' di misure dei fenomeni oggetto dell'indagine. Esistono tuttavia delle situazioni nelle quali la descrizione dei fenomeni puo' essere effettuata solamente in termini qualitativi o, nella migliore delle ipotesi, semiquantitativi : fumatore - non fumatore, sano - ammalato, HBsAg positivo - HBsAg negativo, soggetto sovrappeso - soggetto in peso ideale - soggetto sottopeso, sono solamente alcuni esempi presi a caso di situazioni nelle quali ci si imbatte quotidianamente.

Quando non sono possibili valutazioni di tipo quantitativo, interessa spesso avere la possibilita' di verificare la frequenza di incidenza dei fenomeni. Cosi' puo' interessare controllare la

frequenza di due differenti forme morbose in soggetti fumatori e non fumatori, verificare la frequenza di positivita' dell'antigene di superficie dell'epatite B (HBsAg) in differenti strati della popolazione, e via dicendo. La cosa importante e' evidentemente quella di disporre, una volta che i dati siano stati organizzati secondo un ordine logico che rispecchia il problema che ci si pone, di un indice statistico che consenta di rispondere alla domanda : le frequenze osservate nei diversi gruppi sono uguali? Chi-quadrato e' l'indice statistico di dispersione che consente di rispondere a tale domanda.

La soluzione statistica

Il test chi-quadrato viene qui proposto nella sua forma piu' generale, nella quale e' applicato a tabelle di contingenza.

Perche' cio' possa avvenire e' 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 e **C**

colonne, contenente **n = R * C** celle.

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

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

Il valore di **f** per ciascuna cella e' noto, essendo come detto **f** la frequenza osservata. Il valore di **F**, cioe' la frequenza attesa non e' noto, ma puo' essere specificato qualora si operi alla luce di una ben definita ipotesi riguardante i dati. Nel caso particolare dell'ipotesi " non vi e' differenza fra le frequenze osservate", detta '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 il totale dei casi osservati, ovvero

F =(totale della riga)*(totale della colonna)/n (II)

Per illustrare meglio il concetto si consideri il caso piu' semplice, quello di una tabella 2 x 2 in cui sono :

- le frequenze osservate per ciascuna delle quattro celle indicate rispettivamente con f_1 , f_2 , f_3 , ed f_4
- il totale della riga 1 indicato con R_1 ($R_1 = f_1 + f_2$) e quello della riga 2 indicato con R_2 ($R_2 = f_3 + f_4$)
- il totale della colonna 1 indicato con C_1 ($C_1 = f_1 + f_3$) e quello della colonna 2 indicato con C_2 ($C_2 = f_2 + f_4$)
- il totale dei casi indicato con **n** ($n = f_1 + f_2 + f_3 + f_4$)

Data allora la seguente tabella dei valori osservati

f_1	f_2
f_3	f_4

i quattro valori di **F** attesi corrispondenti ai valori di **f** osservati sono allora

$$F_1 = C_1 * R_1/n \qquad F_2 = C_2 * R_1/n$$

$$F_3 = C_1 * R_2/n \qquad F_4 = C_2 * R_2/n$$

essendo ovviamente ancora $F_1 + F_2 + F_3 + F_4 = n$, mentre il valore di chi-quadrato sara' in base alla (I) 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) * (C - 1)** gradi di liberta'. Facile estendere i calcoli dalla tabella 2 x 2 a tabelle di maggiore estensione.

Il valore di chi-quadrato ottenuto va confrontato con quello riportato nelle apposite tabelle : se esso supera il valore tabulato per i corrispondenti gradi di liberta' e al livello di significativita' prescelto, si conclude che esiste una differenza significativa fra i valori delle celle, e quindi fra le frequenze osservate.

L'applicazione su MULTIPLAN

Incominciate con il selezionare un'insieme di celle comprendente le prime 10 righe e le prime 10 colonne di MULTIPLAN (figura seguente), al quale assegnerete il nome CHIQUADRO e in cui inserirete le vostre tabelle dei valori

osservati con l'accortezza di iniziare l'inserimento sempre a

	1	8	9	10
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

partire dalla prima cella in alto (cella R1C1, della prima riga e prima colonna).

Assegnate ora il nome RIGHE all'insieme di celle da R1C1 a R10C1,

	1
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

e il nome COLONNE all'insieme di celle da R1C1 a R1C10 , che serviranno fra poco per calcolare numero di righe e colonne della tabella del chi-quadrato.

	1	9	10
1			

Avete visto in precedenza come le grandezze necessarie per il calcolo di chi-quadrato siano solamente tre, e cioe' : totale delle osservazioni, somma dei valori di ciscuna riga e somma dei valori di ciascuna colonna. Iniziamo da queste ultime, riportando nella prima cella di riga 11 (cella R11C1) l'espressione

=SUM(R[-1]C:R[-10]C)

che consente di assegnare a tale cella un valore pari alla somma dei valori delle 10 celle sovrastanti (cioe' delle celle da R1C1 a R10C1).

E' ora necessario ripetere l'operazione per le celle da R11C2 a R11C10, cioe' per le celle della riga 11, dalla colonna 2 alla colonna 10 compresa : per evitare di riscrivere altre nove volte

l'espressione ricorrete al comando di MULTIPLAN 'completa a destra'.

	1	2	10
9			
10			
11	=SUM(R[-1]C:R[-10]C)	=SUM(R[-1]C:	=SUM(R[-1]C:R[-10]C)

Procedete in modo analogo per calcolare la somma dei valori di ciascuna riga , riportando nella prima cella di colonna 11 (cella R1C11) l'espressione

=SUM(RC[-1]:RC[-10])

che consente di assegnare a tale cella un valore pari alla somma dei valori delle 10 celle a sinistra (cioe' delle celle da R1C1 a R1C10).

E' ora necessario ripetere l'operazione per le celle da R2C11 a R10C11, cioe' per le celle della colonna 11, dalla riga 2 alla riga 10 compresa : per evitare di riscrivere altre nove volte l'espressione ricorrete questa volta al comando di MULTIPLAN 'completa in basso' (figura alla pagina seguente).

Per semplificare le espressioni che dovremo fra poco impiegare, assegnate alle celle di colonna 11, da riga 1 a riga 10,

i nomi rispettivamente di SOR1 (al contenuto della cella R1C11), SOR2 (al contenuto della cella R2C11), e cosi' via, (figura successiva).

	11
1	=SUM(RC[-1]:RC[-10])
2	=SUM(RC[-1]:RC[-10])
3	=SUM(RC[-1]:RC[-10])
4	=SUM(RC[-1]:RC[-10])
5	=SUM(RC[-1]:RC[-10])
6	=SUM(RC[-1]:RC[-10])
7	=SUM(RC[-1]:RC[-10])
8	=SUM(RC[-1]:RC[-10])
9	=SUM(RC[-1]:RC[-10])
10	=SUM(RC[-1]:RC[-10])

	11	12
1	=SUM(RC[-1]:RC[-10])	'SOR1
2	=SUM(RC[-1]:RC[-10])	'SOR2
3	=SUM(RC[-1]:RC[-10])	'SOR3
4	=SUM(RC[-1]:RC[-10])	'SOR4
5	=SUM(RC[-1]:RC[-10])	'SOR5
6	=SUM(RC[-1]:RC[-10])	'SOR6
7	=SUM(RC[-1]:RC[-10])	'SOR7
8	=SUM(RC[-1]:RC[-10])	'SOR8
9	=SUM(RC[-1]:RC[-10])	'SOR9
10	=SUM(RC[-1]:RC[-10])	'SOR10

Potete adesso calcolare

- la somma che rappresenta il totale delle osservazioni della tabella di chi-quadrato, mediante l'espressione

=SUM(CHIQUADRO)

	1	2
13	TOT	:=SUM(CHIQUADRO)
14	RI	:=COUNT(RIGHE)
15	CO	:=COUNT(COLONNE)

(alla riga 13, definita con il nome TOT)

- il numero di righe della tabella mediante l'espressione

=COUNT(RIGHE)

(alla riga 14, definita con il nome RI)

- il numero delle colonne della tabella mediante l'espressione

=COUNT(COLONNE)

(alla riga 15, definita con il nome CO).

I valori di frequenza F attesi, cosi' come definiti nella (II), corrispondenti ai valori di frequenza f osservati, vengono allora calcolati in una tabella che corrisponde, cella per cella, alla tabella CHIQUADRATO precedentemente definita. Per costruire tale tabella, fra le righe 1 e 10 e le colonne 14 e 23,

- nella prima cella di colonna 14 (cella R1C14) riportate l'espressione

=(SOR1/TOT)*R[+10]C[-13]

- nella seconda cella di colonna 14 (cella R2C14) riportate

l'espressione

=(SOR2/TOT)*R[+9]C[-13]

- nella terza cella di colonna 14 (cella R3C14) riportate

	14
1	=(SOR1/TOT)*R[+10]C[-13]
2	=(SOR2/TOT)*R[+9]C[-13]
3	=(SOR3/TOT)*R[+8]C[-13]
4	=(SOR4/TOT)*R[+7]C[-13]
5	=(SOR5/TOT)*R[+6]C[-13]
6	=(SOR6/TOT)*R[+5]C[-13]
7	=(SOR7/TOT)*R[+4]C[-13]
8	=(SOR8/TOT)*R[+3]C[-13]
9	=(SOR9/TOT)*R[+2]C[-13]
10	=(SOR10/TOT)*R[+1]C[-13]

l'espressione

=(SOR3/TOT)*R[+8]C[-13]

- nella quarta cella di colonna 14 (cella R4C14) riportate l'espressione

=(SOR4/TOT)*R[+7]C[-13]

- nella quinta cella di colonna 14 (cella R5C14) riportate l'espressione

=(SOR5/TOT)*R[+6]C[-13]

- nella sesta cella di colonna 14 (cella R6C14) riportate l'espressione

=(SOR6/TOT)*R[+5]C[-13]

- nella settima cella di colonna 14 (cella R7C14) riportate l'espressione

=(SOR7/TOT)*R[+4]C[-13]

- nell'ottava cella di colonna 14 (cella R8C14) riportate l'espressione

=(SOR8/TOT)*R[+3]C[-13]

- nella nona cella di riga 14 (cella R9C14) riportate l'espressione

=(SOR9/TOT)*R[+2]C[-13]

- nella decima cella di colonna 14 (cella R10C14) riportate l'espressione

=(SOR10/TOT)*R[+1]C[-13]

Ora mediante il comando di MULTIPLAN 'completa a destra', che qui mostra veramente tutta la sua potenza, riportate le espressioni della colonna 14 nelle colonne da 15 a 23, ottenendo cosi' il completamento della tabella dei valori delle

frequenze attese F.

	14	23
1	=(SOR1/TOT)*R[+10]C[-13]	=(SOR1/TOT)*R[+10]C[-13]
2	=(SOR2/TOT)*R[+9]C[-13]	=(SOR2/TOT)*R[+9]C[-13]
3	=(SOR3/TOT)*R[+8]C[-13]	=(SOR3/TOT)*R[+8]C[-13]
4	=(SOR4/TOT)*R[+7]C[-13]	=(SOR4/TOT)*R[+7]C[-13]
5	=(SOR5/TOT)*R[+6]C[-13]	=(SOR5/TOT)*R[+6]C[-13]
6	=(SOR6/TOT)*R[+5]C[-13]	=(SOR6/TOT)*R[+5]C[-13]
7	=(SOR7/TOT)*R[+4]C[-13]	=(SOR7/TOT)*R[+4]C[-13]
8	=(SOR8/TOT)*R[+3]C[-13]	=(SOR8/TOT)*R[+3]C[-13]
9	=(SOR9/TOT)*R[+2]C[-13]	=(SOR9/TOT)*R[+2]C[-13]
10	=(SOR10/TOT)*R[+1]C[-13]	=(SOR10/TOT)*R[+1]C[-13]

Per il calcolo del valore di chi-quadrato, in base alla (I), costruiamo ora una terza tabella di 10 righe per 10 colonne, che corrisponde cella per cella alle prime due, e riportiamo in essa via via i valori dei rapporti fra quadrato della differenza valore osservato meno valore atteso (al numeratore) e valore atteso (al denominatore). Per questo riportiamo nella prima cella della colonna 26 (punto di inizio della nuova tabella) l'espressione

=IF(RC[-12]=0,0,(RC[-25]-RC[-12])^2/RC[-12])

che riporteremo anche nelle righe sottostanti, fina alla riga 10

compresa, mediante il comando 'completa in basso'. Quindi riportiamo tutto il contenuto della colonna 26 nelle colonne da 27 a 35 mediante il comando 'completa a destra' per ottenera la tabella completa, alla quale infine assegnamo il nome SCARTI mediante il comando 'definisci col nome' secondo le modalita' che ben conoscete.

	26	35
1	=IF(RC[-12]=0,0,(RC[-25]-RC[-12])^2/RC[-12])	]=0,0,(RC[-25]-
2	=IF(RC[-12]=0,0,(RC[-25]-RC[-12])^2/RC[-12])	]=0,0,(RC[-25]-
3	=IF(RC[-12]=0,0,(RC[-25]-RC[-12])^2/RC[-12])	]=0,0,(RC[-25]-
4	=IF(RC[-12]=0,0,(RC[-25]-RC[-12])^2/RC[-12])	]=0,0,(RC[-25]-
5	=IF(RC[-12]=0,0,(RC[-25]-RC[-12])^2/RC[-12])	]=0,0,(RC[-25]-
6	=IF(RC[-12]=0,0,(RC[-25]-RC[-12])^2/RC[-12])	]=0,0,(RC[-25]-
7	=IF(RC[-12]=0,0,(RC[-25]-RC[-12])^2/RC[-12])	]=0,0,(RC[-25]-
8	=IF(RC[-12]=0,0,(RC[-25]-RC[-12])^2/RC[-12])	]=0,0,(RC[-25]-
9	=IF(RC[-12]=0,0,(RC[-25]-RC[-12])^2/RC[-12])	]=0,0,(RC[-25]-
10	=IF(RC[-12]=0,0,(RC[-25]-RC[-12])^2/RC[-12])	]=0,0,(RC[-25]-

Come conseguenza dell'utilizzo

=IF(condizione, valore se vera, valore se falsa)

nella terza tabella vengono riportati, cella per cella, dei valori che sono rispettivamente uno zero per le celle che nella tabella delle frequenze attese non contenevano alcun valore (la

condizione RC[-12]=0 risulta vera), e il valore calcolato come detto in precedenza per le celle che nella tabella delle frequenze attese contengono un valore qualsiasi (la condizione RC[-12]=0 risulta falsa) : in tal modo vengono riconosciute automaticamente dal programma le celle utilizzate rispetto a quelle non utilizzate (vuote).

Potete ora riassumere in tre sole righe i risultati del lavoro fatto, riportando il numero totale delle osservazioni mediante l'espressione

=TOT

alla riga 17, calcolando il valore di chi-quadrato mediante l'espressione

=SUM(SCARTI)

alla riga 19, e i relativi gradi di liberta' come

=(RI-1)*(CO-1)

alla riga 20.

	1	2
17	totale osservazioni	:=TOT
18		:
19	chiquadro	:=SUM(SCARTI)
20	gradi di liberta'	:=(RI-1)*(CO-1)

Un esempio

In uno studio sulle possibili interrelazioni fra malattie e antigeni gruppoematici, un gruppo di pazienti affetti da ulcera peptica, un gruppo di pazienti affetti da cancro allo stomaco e un gruppo di pazienti di controllo sono stati classificati in base al sistema ABO in soggetti di gruppo O, di gruppo A, di gruppo B e di gruppo AB. I dati ottenuti sono riassunti nella tabella che segue

Gruppo ABO	Ulcera peptica	Ca gastrico	Controlli
Gruppo O	983	383	2892
Gruppo A	679	416	2625
Gruppo B	134	84	570
Gruppo AB	22	16	71

L'idea e' quella di verificare se la distribuzione dei gruppi sia la stessa in tutti e tre i campioni. Utilizzando il programma su

MULTIPLAN introducete i dati a partire dalla prima riga e prima colonna nello stesso ordine in cui sono presentati nella tabella, e lanciate la procedura di calcolo dei risultati.

	1	2	3
1	983	383	2892
2	679	416	2625
3	134	84	570
4	22	16	71

Dopo pochi istanti potete verificare il valore di chi-quadrato, che risulta pari a 43.04 con 6 gradi di liberta'.

	1	2
16		
17	totale osservazioni	8875
18		
19	chiquadro	43.0431
20	gradi di liberta'	6

Si tratta di un valore molto elevato, che supera di gran lunga quello tabulato (22.5) per 6 gradi di liberta' al livello di probabilita' dello 0.1% : la probabilita' di osservare per caso un valore di chi-quadrato dell'ordine di grandezza di quello

effettivamente osservato risulta di gran lunga inferiore allo 0.1% ($p<0.001$), per cui si conclude che la distribuzione dei gruppi non e' la stessa nei tre campioni (in effetti nel gruppo dei pazienti con ulcera gastrica vi e' un eccesso di gruppi 0).

8. ANALISI DELLA VARIANZA A UN FATTORE

Il problema

Abbiamo visto in precedenza come effettuare il confronto fra due medie utilizzando il test t di Student, tanto nel caso di campioni indipendenti quanto in quello di dati appaiati.

Vediamo ora come sia possibile estendere il confronto a piu' di due medie, utilizzando l'analisi della varianza a un fattore. Per motivi di semplicita' ci limiteremo a rispondere alla piu' elementare delle domande che ci si puo' porre quando si vogliono confrontare contemporaneamente piu' di due medie, e cioe' : le medie sono realmente differenti?.

Questa domanda corrisponde, come al solito, a stabilire se le medie (campionarie) osservate differiscono fra di loro a causa di una differenza fra le medie delle popolazioni da cui provengono, oppure se le differenze (fra le medie osservate) possono essere ragionevolmente attribuite solamente a fluttuazioni casuali.

La soluzione statistica

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

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

nella quale quindi le **r*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*n** osservazioni. Allora la variabilita' totale (S_t) osservata viene calcolata come

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

con (**r*n - 1**) gradi di liberta'.

La variabilita' S_s spiegata dalle differenze fra le medie x_i delle righe viene calcolata come

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

con (**r - 1**) gradi di liberta', mentre la variabilita' casuale, non spiegata (S_n), detta anche "residua", viene calcolata come

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

con $r^*(n-1)$ gradi di liberta' tenendo presente che, per semplicita', essa puo' essere calcolata anche per differenza, come

$$S_n = S_t - S_s \quad \text{(IV)}$$

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

$$V_s = S_s / (r - 1) \quad \text{(V)}$$

$$V_n = S_n / (r^*(n - 1)) \quad \text{(VI)}$$

vengono allora impiegate per calcolare finalmente il rapporto fra varianze **F**

$$F = V_s / V_n \quad \text{(VII)}$$

con **r - 1** gradi di liberta' al numeratore e **r*(n - 1)** gradi di liberta' al denominatore: se il valore di **F** calcolato supera il valore tabulato, per i gradi di liberta' e al livello di significativita' prescelto, si conclude che esiste una differenza significativa fra le medie $\bar{x}_i$ delle **r** righe: sara' compito di una ulteriore analisi dei dati, che peraltro esula dai limiti di questa semplice introduzione, l'identificazione della media (o delle medie) cui puo' presumibilmente essere attribuita la differenza osservata.

L'applicazione su MULTIPLAN

Definite, a partire dalla cella di riga 1 e colonna 1 (R1C1) un insieme di celle (figura alla pagina seguente) comprendente le prime 10 righe e le prime 20 colonne e quindi le celle da R1C1 (angolo estremo in alto a sinistra) a R10C20 (angolo

	1	20
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

estremo in basso a destra), e assegnate a tale insieme di celle il nome ANOVA1.

Nella tabella cosi' definita introdurrete via via i valori delle osservazioni: quelli relativi al campione 1 nella riga 1 (il replicato 1 nella colonna 1, il replicato 2 nella colonna 2 e via dicendo), quelli relativi al campione 2 nella riga 2 (il replicato 1 nella colonna 1, il replicato 2 nella colonna 2, e via dicendo), proseguendo fino alla completa introduzione dei dati. Definite anco-

	1
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

	1	2	20
1		:	

ra con il nome RIGHE l'insieme delle celle da R1C1 a R10C1 , e con il nome NUMREP (che sta per 'numero dei replicati') le celle da R1C1 a R1C20.

Riportate ora

- alla riga 12 l'espressione

	1	2
12	RI	:=COUNT(RIGHE)
13	NU	:=COUNT(NUMREP)
14	SOT	:=SUM(ANOVA1)
15	ME	:=SOT/(RI*NU)

=COUNT(RIGHE)

che consente di calcolare il numero delle righe della tabella ANOVA1, e assegnatele il nome RI

- alla riga 13 l'espressione

=COUNT(NUMREP)

che consente di calcolare il numero delle osservazioni (o numero dei replicati) per ciascuna riga della tabella ANOVA1, e assegnatele il nome NU

- alla riga 14 l'espressione

=SUM(ANOVA1)

che consente di calcolare la somma totale dei valori delle osservazioni della tabella ANOVA1, e assegnatele il nome SOT

- alla riga 15 l'espressione

=SOT/(RI*NU)

che consente di calcolare la media generale delle RI*NU osservazioni della tabella ANOVA1, e alla quale assegnerete il nome ME.

Riservate la colonna 21 al calcolo delle medie di ciascuna riga, riportando innanzitutto nella prima cella di questa colonna (cella R1C21) l'espressione

=(SUM(RC[-1]:RC[-20]))/NU

e replicandola con il comando di MULTIPLAN 'completa in basso' per tutte le nove celle sottostanti, fino alla 10 compresa.

	21
1	=(SUM(RC[-1]:RC[-20]))/NU
2	=(SUM(RC[-1]:RC[-20]))/NU
3	=(SUM(RC[-1]:RC[-20]))/NU
4	=(SUM(RC[-1]:RC[-20]))/NU
5	=(SUM(RC[-1]:RC[-20]))/NU
6	=(SUM(RC[-1]:RC[-20]))/NU
7	=(SUM(RC[-1]:RC[-20]))/NU
8	=(SUM(RC[-1]:RC[-20]))/NU
9	=(SUM(RC[-1]:RC[-20]))/NU
10	=(SUM(RC[-1]:RC[-20]))/NU

Nella prima riga della colonna 22 riportate ora l'espressione

=(RC[-1]-ME)^2

che consente di elevare al quadrato il valore della differenza fra la media degli NU valori della riga 1 e la media generale ME delle osservazioni : ancora una volta, mediante il comando 'completa in basso' riportate l'espressione nelle 9 celle sottostanti, fino alla cella R10C22 compresa

Definite l'insieme delle celle di colonna 22, da R1C22 a R10C22, con il nome SCARTIRI.

	22
1	=(RC[-1]-ME)^2
2	=(RC[-1]-ME)^2
3	=(RC[-1]-ME)^2
4	=(RC[-1]-ME)^2
5	=(RC[-1]-ME)^2
6	=(RC[-1]-ME)^2
7	=(RC[-1]-ME)^2
8	=(RC[-1]-ME)^2
9	=(RC[-1]-ME)^2
10	=(RC[-1]-ME)^2

Costruite ora fra le colonne 23 e 42 e le righe 1 e 10 una tabella che corrisponda alla prima cella per cella, nella quale siano calcolati via via i quadrati delle differenze fra ciascun elemento della tabella ANOVA1 e la media generale ME. Per

questo scrivete nella prima riga della colonna 23 (cella R1C23) l'espressione

=(RC[-22]-ME)^2

e riportate poi con i comandi 'completa a destra' e 'completa in basso' tale espressione nelle 20*10 celle di questa seconda tabella, cui assegnerete il nome ANOVABIS.

	23	24	42
1	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2
2	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2
3	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2
4	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2
5	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2
6	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2
7	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2
8	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2
9	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2
10	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2	=(RC[-22]-ME)^2

Potete ora calcolare (figura alla pagina successiva)

- la variabilita' totale, cosi' come indicata nella (I), mediante l'espressione

=(SUM(ANOVABIS))-(ME*ME*(200-RI*NU))

(alla riga 16, definita col nome ST)

	1	2
16	ST	=(SUM(ANOVABIS))-(ME*ME*(200-RI*NU))
17	SS	=NU*(SUM(SCARTIRI)-(ME*ME*(10-RI)))
18	SN	=ST-SS

- la variabilita'spiegata dalle differenze fra le medie delle righe, cosi' come indicata nella (II), mediante l'espressione

=NU*(SUM(SCARTIRI)-(ME*ME*(10-RI)))

(alla riga 17, definita col nome SS)

- la variabilita' non spiegata (o residua), cosi' come indicata nella (IV), mediante l'espressione

=ST-SS

(alla riga 18, definita col nome SN).

Tenete presente che i fattori ME*ME*(200-RI*NU) e ME*ME*(10-RI) vengono sottratti nel calcolo rispettivamente della variabilita' totale (ST) e della variabilita' spiegata dalle differenze fra le medie delle righe (SS) al fine di correggere per

	1	2	3	4	5
20	variabilita'	somma dei quadrati	gradi liberta'	varianza	F
21					
22	fra righe	=SS	=RI-1	=SS/(RI-1)	=VS/VN
23	residua	=SN	=RI*(NU-1)	=SN/(RI*(NU-1))	
24	totale	=ST	=NU*RI-1		

le celle della tabella dell'analisi della varianza non utilizzate.

E finalmente potete riassumere i risultati ottenuti in una classica tabella dell'analisi della varianza, riportando alla riga 22

- nella seconda colonna l'espressione

=SS

che rappresenta la variabilita' spiegata calcolata come abbiamo visto in base alla (II)

- nella terza colonna l'espressione

=RI-1

che rappresenta i gradi di liberta' della variabilita' spiegata

-nella terza colonna l'espressione

=SS/(RI-1)

che rappresenta la varianza spiegata, calcolata in base alla (V), alla quale assegnerete il nome VS, quindi riportando alla riga 23

-nella colonna 2 l'espressione

=SN

che rappresenta la variabilita' non spiegata

-nella colonna 3 l'espressione

=RI*(NU-1)

che rappresenta i gradi di liberta' della variabilita' non spiegata

-nella colonna 4 l'espressione

=SN/(RI*(NU-1))

che consente di calcolare la varianza non spiegata in base alla (VI), e alla quale assegnerete il nome VN, quindi riportando alla riga 24

-nella colonna 2 la variabilita' totale gia' precedentemente calcolata mediante l'espressione

=ST

-nella colonna 3 i gradi di liberta' corrispondenti alla variabilita' totale come

=NU*RI-1

Finalmente potrete riportare nella cella R22C5 il valore del rapporto fra varianze F calcolato come nella (VII) mediante l'espressione

=VS/VN

Un esempio

Nell'intento di verificare se diete contenenti diverse proporzioni di glicidi, lipidi e proteine siano in grado di influenzare la velocita' di crescita dei pulcini, quattro gruppi comprendenti ciascuno cinque animali vengono sottoposti ad alimentazione con quattro diverse diete (diete A,B,C e D), e al termine del periodo stabilito di osservazione viene misurato l'aumento del peso corporeo (in grammi) che si e' verificato, ottenendosi i risultati riportati nella seguente tabella

	Replicato				
Dieta	1	2	3	4	5
A	55	49	42	21	52
B	61	112	30	89	63
C	42	97	81	95	92
D	169	137	169	85	154

che riportati nello stesso ordine nel programma su MULTIPLAN

	1	2	3	4	5
1	55	49	42	21	52
2	61	112	30	89	63
3	42	97	81	95	92
4	169	137	169	85	154

consentono un rapido calcolo dell'analisi della varianza.

	1	2	3	4	5
19					
20	variabilita'	somma dei quadrati	gradi liberta'	varianza	F
21					
22	fra righe	26234.95	3	8744.98	12.105
23	residua	11558.8	16	722.425	
24	totale	37793.75	19		

Il valore del rapporto fra varianze F (12.1) supera di gran lunga il valore tabulato (5.28) per tre gradi di liberta' del numeratore e per 16 gradi di liberta' del denominatore al livello di probabilita' dell'1%, e consente di concludere che le diverse diete comportano aumenti di peso in media differenti : in effetti la dieta D comporta un aumento medio di peso di 142.8 grammi, contro un aumento medio di 43.8, 71 e 81.4 grammi delle diete A, B e C rispettivamente.

9. ANALISI DELLA VARIANZA A DUE FATTORI

Il problema

L'espressione "analisi della varianza" viene impiegata in riferimento non a una, bensi' a un insieme di tecniche statistiche per il confronto contemporaneo di piu' medie: quello che in sostanza determina le caratteristiche di ciascuna tecnica e' il problema che ci si trova a dovere risolvere, il quale a sua volta deve essere ovviamente riflesso nel disegno sperimentale in base al quale vengono ottenuti i dati.

L'analisi della varianza a un fattore e' utilizzata per lo studio di dati per i quali esiste un solo criterio di classificazione (con osservazioni replicate in ciascuno dei gruppi): l'analisi della varianza a due fattori viene impiegata per lo studio di dati per i quali esistono due criteri di classificazione (e viene qui per semplicita' limitata al caso in cui non vi siano replicati).

La soluzione statistica

Si consideri la seguente tabella dell'analisi della varianza a due

fattori

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

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 ancora $\bar{x}_{i.}$ la media di una generica riga, $\bar{x}_{.j}$ la media di una generica colonna, e $\bar{x}_{g}$ la media generale degli **r*c** dati.

E' facile notare che, rispetto alla tabella dell'analisi della varianza a un fattore, la novita' sostanziale consiste nell'avere introdotto un elemento di classificazione a livello delle colonne, e quindi le corrispondenti medie $\bar{x}_{.j}$.

La variabilita' totale (S_t) osservata viene calcolata come al solito come

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

con ($r^*c - 1$) gradi di liberta'.

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

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

con (r - 1) gradi di liberta.

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

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

con c - 1 gradi di liberta'.

Infine la variabilita' casuale, non spiegata (S_n), detta anche

"residua", viene calcolata come

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

con $(r - 1)*(c - 1)$ gradi di liberta', tenendo presente che, per semplicita', essa puo' essere calcolata anche come

$$S_n = S_t - S_r - S_c \quad \text{(V)}$$

La varianza spiegata dalle differenze fra le medie $\bar{x}_{i.}$ delle righe (V_r), quella spiegata dalle differenze fra le medie $\bar{x}_{.j}$

delle colonne (V_c) e la varianza non spiegata (V_n) sono allora calcolate rispettivamente come

$$V_r = S_r / (r - 1) \quad \text{(VI)}$$

$$V_c = S_c / (c - 1) \quad \text{(VII)}$$

$$V_n = S_n / ((r - 1)*(c-1)) \quad \text{(VIII)}$$

Il rapporto fra varianze **F**

$$F = V_r / V_n \quad \text{(IX)}$$

con **r - 1** gradi di liberta' al numeratore e **(r - 1)*(c - 1)** gradi di liberta' al denominatore viene allora impiegato per

verificare l'esistenza di una differenza significativa fra le medie delle righe ($\bar{x}_{i.}$), e il rapporto fra varianze

$$F = V_c / V_n \quad (x)$$

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

Se il valore di **F** calcolato supera quello tabulato, per i gradi di liberta' e al livello di significativita' prescelto, la differenza fra le medie (delle righe o delle colonne, secondo i casi) viene considerata significativa : come gia' detto a proposito dell'analisi della varianza a un fattore sara' compito di una ulteriore analisi dei dati l'identificazione della media (o delle medie) cui puo' essere presumibilmente attribuita la differenza osservata.

L'applicazione su MULTIPLAN

Iniziate, come nel caso dell'analisi della varianza a un fattore, definendo, a partire dalla cella di riga 1 e colonna 1 (R1C1), un

	1	2	3		19	20
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

insieme di celle comprendente le prime 10 righe e le prime 20 colonne e quindi le celle da R1C1 (angolo estremo in alto a sinistra) a R10C20 (angolo estremo in basso a destra), e assegnate a tale insieme di celle il nome ANOVA2.

	1
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Nella tabella cosi' definita introdurrete

	1	2	20
1			

via via i valori delle osservazioni, a partire come al solito dalla prima riga e prima colonna, proseguendo fino alla completa introduzione dei dati. Definite ora con il nome RIGHE l'insieme comprendente le celle da R1C1 a R10C1, e definite con il nome COLONNE l'insieme comprendente le celle da R1C1 a R20C1.

Riportate ora

- alla riga 14 l'espressione

=COUNT(RIGHE)

che consente di calcolare il numero delle righe della tabella ANOVA2, e assegnatele il nome RI

14	RI	=COUNT(RIGHE)
15	CO	=COUNT(COLONNE)
16	SO	=SUM(ANOVA2)
17	ME	=SO/(RI*CO)
18		

- alla riga 15 l'espressione

=COUNT(COLONNE)

che consente di calcolare il numero delle colonne della tabella ANOVA2, e assegnatele il nome CO

- alla riga 16 l'espressione

=SUM(ANOVA2)

che consente di calcolare la somma totale dei valori delle osservazioni della tabella ANOVA2, e assegnatele il nome SO

- alla riga 17 l'espressione

SO/(RI*CO)

che consente di calcolare la media generale delle RI*CO osservazioni della tabella ANOVA2, e assegnatele il nome ME.

Riservate la colonna 21 al calcolo delle medie di ciascuna riga, riportando innanzitutto nella prima cella di questa colonna (cella R1C21) l'espressione

=SUM(RC[-1]:RC[-20])/CO

e replicandola con il comando MULTIPLAN 'completa in basso' per tutte le nove celle sottostanti, fino alla 10 compresa (figura alla pagina seguente).

Nella prima riga della colonna 22 riportate ora l'espressione

=(RC[-1]-ME)^2

che consente di elevare al quadrato il valore della differenza fra la media dei CO valori della riga 1 e la media generale ME delle osservazioni: ancora una volta, mediante il comando 'completa in basso' riportate l'espressione nelle 9 celle sottostanti, fino alla cella R10C22 compresa.

	21
1	=SUM(RC[-1]:RC[-20])/CO
2	=SUM(RC[-1]:RC[-20])/CO
3	=SUM(RC[-1]:RC[-20])/CO
4	=SUM(RC[-1]:RC[-20])/CO
5	=SUM(RC[-1]:RC[-20])/CO
6	=SUM(RC[-1]:RC[-20])/CO
7	=SUM(RC[-1]:RC[-20])/CO
8	=SUM(RC[-1]:RC[-20])/CO
9	=SUM(RC[-1]:RC[-20])/CO
10	=SUM(RC[-1]:RC[-20])/CO

	22
1	=(RC[-1]-ME)^2
2	=(RC[-1]-ME)^2
3	=(RC[-1]-ME)^2
4	=(RC[-1]-ME)^2
5	=(RC[-1]-ME)^2
6	=(RC[-1]-ME)^2
7	=(RC[-1]-ME)^2
8	=(RC[-1]-ME)^2
9	=(RC[-1]-ME)^2
10	=(RC[-1]-ME)^2

Assegnate all'insieme delle celle della colonna 22, da R1C22 a R10C22, il nome SCARTIRI.

Ripetete ora per le colonne le operazioni fin qui eseguite sulle righe : riservate la riga 11 al calcolo delle medie di ciascuna colonna, riportando innanzitutto nella prima cella di questa riga (cella R11C1) l'espressione

=SUM(R[-1]C:R[-10]C)/RI

e replicandola con il comando MULTIPLAN 'completa a destra' per tutte le diciannove colonne seguenti, fino alla cella R11C20 compresa.

	1	2	20
11	=SUM(R[-1]C:R[-10]C)/RI	:=SUM(R[-1]C:R	=SUM(R[-1]C:R[-10]C)

Nella prima colonna della riga 12 riportate ora l'espressione

=(R[-1]C-ME)^2

che consente di elevare al quadrato il valore della differenza fra la media dei RI valori della colonna 1 e la media generale ME delle osservazioni : ancora una volta, mediante il comando 'completa a destra' riportate l'espressione nelle 19 colonne seguenti, fino alla cella R12C20 compresa.

	1	2	20
12	=(R[-1]C-ME)^2	:=(R[-1]C-ME)^2	=(R[-1]C-ME)^2

Definite con il nome SCARTICO l'insieme delle celle di riga 12, da R12C1 a R12C20 .

Costruite ora fra le colonne 23 e 42 e le righe 1 e 10 una

tabella che corrisponda alla prima cella per cella, nella quale sono calcolati via via i quadrati delle differenze fra ciascun elemento della tabella ANOVA2 e la media generale ME. Per questo scrivete nella prima riga della colonna 23 (cella R1C23) l'espressione

=(RC[-22]-ME)^2

e riportate poi con i comandi 'completa a destra' e 'completa in

	23
1	=(RC[-22]-ME)^2

basso' tale espressione nelle 20*10 celle di questa seconda tabella, cui assegnerete il nome ANOVA2BIS.

	23	24	42
1	=(RC[-22]-ME)^2	=(RC[-22]-ME	=(RC[-22]-ME)^2
2	=(RC[-22]-ME)^2	=(RC[-22]-ME	=(RC[-22]-ME)^2
3	=(RC[-22]-ME)^2	=(RC[-22]-ME	=(RC[-22]-ME)^2
4	=(RC[-22]-ME)^2	=(RC[-22]-ME	=(RC[-22]-ME)^2
5	=(RC[-22]-ME)^2	=(RC[-22]-ME	=(RC[-22]-ME)^2
6	=(RC[-22]-ME)^2	=(RC[-22]-ME	=(RC[-22]-ME)^2
7	=(RC[-22]-ME)^2	=(RC[-22]-ME	=(RC[-22]-ME)^2
8	=(RC[-22]-ME)^2	=(RC[-22]-ME	=(RC[-22]-ME)^2
9	=(RC[-22]-ME)^2	=(RC[-22]-ME	=(RC[-22]-ME)^2
10	=(RC[-22]-ME)^2	=(RC[-22]-ME	=(RC[-22]-ME)^2

Riportate ora

- alla riga 18 l'espressione

=SUM(ANOVA2BIS)-ME*ME*(200-RI*CO)

(assegnatele il nome ST), che consente di calcolare la varianza

	1	2
18	ST	:=SUM(ANOVA2BIS)-ME*ME*(200-RI*CO)
19	SR	:=(SUM(SCARTIRI)-ME*ME*(10-RI))*CO
20	SC	:=(SUM(SCARTICO)-ME*ME*(20-CO))*RI
21	SN	:=ST-SR-SC

totale secondo la (I), essendo il termine ME*ME*(200-RI*CO) sottratto al fine di correggere per i valori corrsipondenti alle celle della tabella ANOVA2BIS non utilizzate

- alla riga 19 l'espressione

=(SUM(SCARTIRI)-ME*ME*(10-RI))*CO

(assegnatele il nome SR) : tale espressione consente di calcolare la somma dei valori presenti nelle 10 celle di colonna 22 , che moltiplicata per il numero delle colonne, fornisce la variabilita' spiegata dalle differenza fra le righe secondo la (II) essendo il termine (ME*ME*(10-RI) sottratto al fine di correggere per i valori corrispondenti alle righe della tabella ANOVA2 non

utilizzate

- alla riga 20 l'espressione

=(SUM(SCARTICO)-ME*ME*(20-CO))*RI

(assegnatele il nome SC) : tale espresssione consente di calcolare la somma dei valori presenti nelle 20 celle di riga 12 sopra definite, che moltiplicata per il numero delle righe, fornisce la variabilita' spiegata dalle differenza fra le colonne secondo la (III) essendo il termine ME*ME*(20-CO) sottratto al fine di correggere per i valori corrispondenti alle colonne della tabella ANOVA2 non utilizzate

- alla riga 21 l'espressione

=ST-SR-SC

(assegnatele il nome SN) che consente di calcolare la variabilita' non spiegata (o residua) secondo la (V).

E' allora finalmente possibile riassumere i risultati ottenuti in una classica tabella dell'analisi della varianza (figura alla pagina successiva), in cui alla riga 25 riportate

- nella colonna 2 l'espressione

=SR

	1	2	3	4	5
23	variabilita'	somma dei quadrati	gradi libertà	varianza	F
24					
25	fra righe	=SR	=RI-1	=SR/(RI-1)	=VR/VN
26	fra colonne	=SC	=CO-1	=SC/(CO-1)	=VC/VN
27	residua	=SN	=(RI-1)*(CO-1)	=SN/((RI-1)*(CO-1))	
28	totale	=ST	=RI*CO-1		

che rappresenta la variabilita' spiegata dalle differenze fra le medie delle righe

- nella colonna 3 l'espressione

=RI-1

che rappresenta i gradi di liberta' corrispondenti

- nella colonna 4 l'espressione

=SR/(RI-1)

che rappresenta la varianza spiegata dalle differenza fra le medie delle righe, calcolata in base alla (VI), cui assegnerete il nome VR, mentre alla riga 26 riportate

- nella colonna 2 l'espressione

=SC

che rappresenta la variabilita' spiegata dalle differenze fra le medie delle colonne

- nella colonna 3 l'espressione

=CO-1

che rappresenta i gradi di liberta' corrispondenti

- nella colonna 4 l'espressione

=SC/(CO-1)

che rappresenta la varianza spiegata dalle differenza fra le medie delle colonne, calcolata in base alla (VII), cui assegnerete il nome VC, e ancora alla riga 27 riportate

- nella colonna 2 l'espressione

=SN

che rappresenta come gia' visto la variabilita' residua, e a cui assegnerete il nome SN

- nella colonna 3 l'espressione

=(RI-1)*(CO-1)

che rappresenta i gradi di liberta' della variabilita' non spiegata

- nella colonna 4 l'espressione

=SN/((RI-1)*(CO-1))

che consente di calcolare la varianza non spiegata in base alla

(VIII), e a cui assegnerete il nome VN, e finalmente alla riga 28 riportate

- nella colonna 2 la variabilita' totale gia' precedentemente calcolata mediante l'espressione

=ST

- nella colonna 3 i gradi di liberta' corrispondenti alla variabilita' totale come

=RI*CO-1

Ecco che nella colonna 5 potrete allora riportare

- alla riga 25 il valore del rapporto fra varianze F calcolato come nella (IX)

=VR/VN

che serve per testare la significativita' della variabilita' osservata fra le righe

- e alla riga 26 il valore del rapporto fra varianze F calcolato come nella (X)

=VC/VN

che serve per testare la significativita' della variabilita' osservata fra le colonne.

Un esempio

Il valore del rapporto fra superficie delle foglie e loro peso secco (in centimetri quadrati per grammo) viene determinato in tre varieta' di piante appartenenti alla stessa specie, in tre diverse condizioni di soleggiamento, ottenendo in definitiva la seguente tabella di valori

Soleggiamento	Varieta'		
	A	B	C
Pieno sole	112	90	123
Mezza ombra	86	73	89
Ombra	80	62	81

Le domande che ci si pone sono due : esiste in media una differenza legata alle differenti condizioni di esposizione ? Ed esiste in media una qualche differenza fra le varieta' esaminate?

L'analisi della varianza a due fattori ci consente di rispondere contemporaneamente ad entrambe le domande.

Introducete i dati nel programma su MULTIPLAN, nello stesso ordine in cui compaiono nella tabella, e calcolate i risultati.

	1	2	3
1	112	90	123
2	86	73	89
3	80	62	81

Il rapporto fra varianze F risulta in entrambi i casi significativo, superando nel primo caso (F=43.26) di molto e nel secondo caso (F=19.53) di poco il valore tabulato per 2 gradi di liberta' del numeratore e 4 gradi di liberta' del denominatore (pari a 18 per il livello di probabilita' dell'1%).

Si conclude che esiste una grande differenza indotta dalle

	1	2	3	4	5
23	variabilita'	somma dei quadrati	gradi libertà	varianza	F
24					
25	fra righe	1884.2222222	2	942.111	43.2602
26	fra colonne	850.88888892	2	425.444	19.5357
27	residua	87.111112361	4	21.7778	
28	totale	2822.2222235	8		

diverse condizioni di soleggiamento (fra righe) : in effetti le medie risultano essere pari a 108.3 (pieno sole), 82.7 (mezza ombra) e 74.3 (ombra). Esiste contemporaneamente una significativa differenza fra le medie delle tre varieta', pari a 92.7, 75 e 97.7 per le tre varieta' A, B e C rispettivamente : tale differenza sembra imputabile al basso valore osservato per la varieta' B.

APPENDICE A . Tavola del t di Student

Gradi di liberta'	Probabilita' (p)					
	0.90	0.50	0.10	0.05	0.01	0.001
1	0.16	1.00	6.31	12.7	63.7	636.6
2	0.14	0.82	2.92	4.30	9.93	31.6
3	0.14	0.77	2.35	3.18	5.84	12.9
4	0.13	0.74	2.13	2.78	4.60	8.61
5	0.13	0.73	2.02	2.57	4.03	6.86
6	0.13	0.72	1.94	2.45	3.71	5.96
7	0.13	0.71	1.90	2.37	3.50	5.41
8	0.13	0.71	1.86	2.31	3.36	5.04
9	0.13	0.70	1.83	2.26	3.25	4.78
10	0.13	0.70	1.81	2.23	3.17	4.59
11	0.13	0.70	1.80	2.20	3.11	4.44
12	0.13	0.70	1.78	2.18	3.06	4.32
13	0.13	0.69	1.77	2.16	3.01	4.22
14	0.13	0.69	1.76	2.15	2.98	4.14
15	0.13	0.69	1.75	2.13	2.95	4.07
16	0.13	0.69	1.75	2.12	2.92	4.02
17	0.13	0.69	1.74	2.11	2.90	3.97
18	0.13	0.69	1.73	2.10	2.88	3.92
19	0.13	0.69	1.73	2.09	2.86	3.88
20	0.13	0.69	1.73	2.09	2.85	3.85
30	0.13	0.68	1.70	2.04	2.75	3.65
40	0.13	0.68	1.68	2.02	2.70	3.55
50	0.13	0.68	1.68	2.01	2.68	3.50
60	0.13	0.68	1.67	2.00	2.66	3.46
70	0.13	0.68	1.67	1.99	2.65	3.44
80	0.13	0.68	1.67	1.99	2.64	3.42
90	0.13	0.68	1.66	1.99	2.63	3.40
100	0.13	0.68	1.65	1.98	2.62	3.39
∞	0.13	0.67	1.64	1.96	2.58	3.29

La probabilita' riportata e' la probabilita' di osservare per caso una differenza della stessa entita' di quella effettivamente osservata : se tale probabilita' risulta sufficientemente piccola (tipicamente inferiore a 0.05) la differenza viene considerata significativa.

Esempio : un t pari a 3.16 con 19 gradi di liberta' si colloca fra 2.86 e 3.88 , valori corrispondenti rispettivamente alle probabilita' p=0.01 e p=0.001. Cio' significa che la probabilita' di osservare per caso una differenza della stessa entita' di quella effettivamente osservata e' compresa fra l'1% e lo 0.1% , una probabilita' abbastanza piccola : la conclusione e' che la differenza osservata e' presumibilmente non casuale, e quindi significativa.

APPENDICE B . Tavola del chi-quadrato

Gradi di liberta'	Probabilita' (p)					
	0.90	0.50	0.10	0.05	0.01	0.001
1	0.02	0.45	2.71	3.84	6.63	10.8
2	0.21	1.39	4.61	5.99	9.21	13.8
3	0.58	2.37	6.25	7.81	11.3	16.3
4	1.06	3.36	7.78	9.49	13.3	18.5
5	1.61	4.35	9.24	11.1	15.1	20.5
6	2.20	5.35	10.6	12.6	16.8	22.5
7	2.83	6.35	12.0	14.1	18.5	24.3
8	3.49	7.34	13.4	15.5	20.1	26.1
9	4.17	8.34	14.7	16.9	21.7	27.8
10	4.87	9.34	16.0	18.3	23.2	29.6
11	5.58	10.3	17.3	19.7	24.7	31.3
12	6.30	11.3	18.6	21.0	26.2	32.9
13	7.04	12.3	19.8	22.4	27.7	34.5
14	7.79	13.3	21.1	23.7	29.1	36.1
15	8.55	14.3	22.3	25.0	30.6	37.7
16	9.31	15.3	23.5	26.3	32.0	39.3
17	10.1	16.3	24.8	27.6	33.4	40.8
18	10.9	17.3	26.0	28.9	34.8	42.3
19	11.7	18.3	27.2	30.1	36.2	43.8
20	12.4	19.3	28.4	31.4	37.6	45.3
30	20.6	29.3	40.3	43.8	50.9	59.7
40	29.1	39.3	51.8	55.8	63.7	73.4
50	37.7	49.3	63.2	67.5	76.2	86.7
60	46.5	59.3	74.4	79.1	88.4	99.6
70	55.3	69.3	85.5	90.5	100.4	112.3
80	64.3	79.3	96.6	101.9	112.3	124.8
90	73.3	89.3	107.6	113.1	124.1	137.2
100	82.4	99.3	118.5	124.3	135.8	149.5

La probabilita' riportata e' la probabilita' di osservare per caso una differenza della stessa entita' di quella effettivamente osservata : se tale probabilita' risulta sufficientemente piccola (tipicamente inferiore a 0.05) la differenza viene considerata significativa.

Esempio : un chi-quadrato pari a 8.7 con 12 gradi di liberta' si colloca fra 6.30 e 11.3 , valori corrispondenti rispettivamente alle probabilita' p=0.9 e p=0.5 . Cio' significa che la probabilita' di osservare per caso una differenza della stessa entita' di quella effettivamente osservata e' compresa fra il 50% e il 90% , una probabilita' molto grande : la conclusione e' che la differenza osservata e' presumibilmente casuale, e quindi non significativa.

APPENDICE C. Tavola dei valori di F (n = gradi di liberta' del numeratore, d = gradi di liberta' del denominatore) per i livelli di probabilita p=0.05 (carattere normale) e p=0.01 (in corsivo)

	n									
d	1	2	3	4	5	6	8	12	24	∞
1	161	200	216	225	230	234	239	244	249	254
	4052	*4999*	*5403*	*5625*	*5764*	*5859*	*5981*	*6106*	*6234*	*6366*
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.5	19.5
	98.5	*99.0*	*99.2*	*99.3*	*99.3*	*99.3*	*99.4*	*99.4*	*99.5*	*99.5*
3	10.1	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
	34.1	*30.8*	*29.5*	*28.7*	*28.2*	*27.9*	*27.5*	*27.1*	*26.6*	*26.1*
4	7.71	6.94	6.58	6.39	6.26	6.16	6.04	5.91	5.77	5.63
	21.3	*18.0*	*16.7*	*16.0*	*15.5*	*15.2*	*14.8*	*14.4*	*13.9*	*13.5*
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
	16.3	*13.3*	*12.1*	*11.4*	*11.0*	*10.7*	*10.3*	*9.89*	*9.47*	*9.02*
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
	13.7	*10.9*	*9.78*	*9.15*	*8.75*	*8.47*	*8.10*	*7.72*	*7.31*	*6.88*
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
	12.3	*9.55*	*8.45*	*7.85*	*7.45*	*7.19*	*6.84*	*6.47*	*6.07*	*5.65*
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
	11.3	*8.65*	*7.59*	*7.01*	*6.63*	*6.37*	*6.03*	*5.67*	*5.28*	*4.86*
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
	10.6	*8.02*	*6.99*	*6.42*	*6.06*	*5.80*	*5.47*	*5.11*	*4.73*	*4.31*

	n									
d	1	2	3	4	5	6	8	12	24	∞
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
	10.0	*7.56*	*6.55*	*5.99*	*5.64*	*5.39*	*5.06*	*4.71*	*4.33*	*3.91*
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
	9.65	*7.20*	*6.22*	*5.67*	*5.32*	*5.07*	*4.74*	*4.40*	*4.02*	*3.60*
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
	9.33	*6.93*	*5.95*	*5.41*	*5.06*	*4.82*	*4.50*	*4.16*	*3.78*	*3.36*
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
	9.07	*6.70*	*5.74*	*5.20*	*4.86*	*4.62*	*4.30*	*3.96*	*3.59*	*3.16*
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
	8.86	*6.51*	*5.56*	*5.03*	*4.69*	*4.46*	*4.14*	*3.80*	*3.43*	*3.00*
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
	8.68	*6.36*	*5.42*	*4.89*	*4.56*	*4.32*	*4.00*	*3.67*	*3.29*	*2.87*
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
	8.53	*6.23*	*5.28*	*4.77*	*4.44*	*4.20*	*3.89*	*3.55*	*3.18*	*2.75*
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
	8.40	*6.11*	*5.18*	*4.67*	*4.34*	*4.10*	*3.79*	*3.45*	*3.08*	*2.65*
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
	8.208	*6.01*	*5.09*	*4.58*	*4.25*	*4.01*	*3.71*	*3.37*	*3.00*	*2.57*
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
	8.18	*5.93*	*5.01*	*4.50*	*4.17*	*3.94*	*3.63*	*3.30*	*2.92*	*2.49*

	n									
d	1	2	3	4	5	6	8	12	24	∞
20	4.36	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
	8.10	*5.85*	*4.94*	*4.43*	*4.10*	*3.87*	*3.56*	*3.23*	*2.86*	*2.42*
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
	7.56	*5.39*	*4.51*	*4.02*	*3.70*	*3.47*	*3.17*	*2.84*	*2.47*	*2.01*
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
	7.31	*5.18*	*4.31*	*3.83*	*3.51*	*3.29*	*2.99*	*2.66*	*2.29*	*1.80*
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
	7.08	*4.98*	*4.13*	*3.65*	*3.34*	*3.12*	*2.82*	*2.50*	*2.12*	*1.60*
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00
	6.64	*4.60*	*3.78*	*3.32*	*3.02*	*2.80*	*2.51*	*2.18*	*1.79*	*1.00*

La probabilita' riportata e' la probabilita' di osservare per caso una differenza della stessa entita' di quella effettivamente osservata : se tale probabilita' risulta sufficientemente piccola (tipicamente inferiore a 0.05) la differenza viene considerata significativa.

Esempio : un F pari a 4.48 con 3 gradi di liberta' al numeratore e 19 gradi di liberta' al denominatore si colloca fra

3.13 e 5.01 , valori corrispondenti rispettivamente alle probabilita' p=0.05 e p=0.01. Cio' significa che la probabilita' di osservare per caso una differenza della stessa entita' di quella effettivamente osservata e' compresa fra il 5% e l'1% , una probabilita' abbastanza piccola : la conclusione e' che la differenza osservata e' presumibilmente non casuale, e quindi significativa.

finito di stampare dalla Data Management & Synthesis

nel marzo 1987

Il mercato del software offre un gran numero di programmi di statistica, spesso pero' anche i piu' completi e i piu' sofisticati sono dotati di uno scarso grado di compatibilita' con i data base. Questo fatto obbliga il ricercatore a dover trasferire da tastiera grandi quantita' di dati dagli archivi ai programmi di elaborazione statistica.

Gli autori di questo volume, due medici che lavorano nell'ambito della ricerca scientifica, hanno risolto il problema della compatibilita' implementando i metodi statistici di uso piu' frequente sugli spreadsheets, pacchetti notoriamente compatibili con gli archivi elettronici.

La presentazione fornita in questo volume prevede l'uso di Multiplan e di Macintosh, il primo scelto per la sua grande diffusione ed il secondo per le sue eccellenti caratteristiche di presentazione grafica. L'utente pero' sfogliando le prime pagine si accorgera' subito di come sia agevole il trasferimento dell'intero pacchetto su IBM compatibili o su altri spreadsheets.

Questo lavoro e' stato presentato al First International Workshop in Occupational Health organizzato dalla International Commission on Occupational Health e da Commission of the European Communities Joint Research Centre e tenutosi a Varese il 30 e 31 ottobre 1986.