CAPITOLO XVIII

CORRELAZIONE E COVARIANZA

CAPITOLO XVIII

CORRELAZIONE E COVARIANZA

18.1. LA CORRELAZIONE

La regressione lineare è finalizzata all'analisi della dipendenza tra due variabili, delle quali

- una (**Y**) è a priori definita come **dipendente o effetto**,
- l'altra (**X**) è individuata come **indipendente o causa**.

L'interesse della ricerca è rivolta essenzialmente all'**analisi delle cause** o allo **studio predittivo delle quantità medie di Y**, che si ottengono come risposta al variare di X.

Spesso anche nella ricerca ambientale, biologica e medica, la relazione di causa-effetto non ha una direzione logica o precisa: potrebbe essere ugualmente applicata nei due sensi, da una variabile all'altra. Le coppie di fidanzati o sposi di solito hanno altezza simile: la relazione di causa effetto può essere applicata sia dall'uomo alla donna che viceversa; coppie di gemelli hanno strutture fisiche simili e quella di uno può essere stimata sulla base dell'altro.

Altre volte, la causa può essere individuata in un terzo fattore, che agisce simultaneamente sui primi due, in modo diretto oppure indiretto, determinando i valori di entrambi e le loro variazioni, come la quantità di polveri sospese nell'aria e la concentrazione di benzene, entrambi dipendenti dall'intensità del traffico. In altre ancora, l'interesse può essere limitato a **misurare come due serie di dati variano congiuntamente**, per poi andare alla ricerca delle eventuali cause, se la risposta fosse statisticamente significativa.

In tutti questi casi, è corretto utilizzare la correlazione.

Più estesamente, è chiamato **coefficiente di correlazione prodotto-momento di Pearson** (***Pearson product-moment correlation coefficient***), perché nella sua espressione algebrica è stato presentato per la prima volta da Karl **Pearson** (1857-1936) in un lavoro del 1895. In modo più semplice, anche nel testo di **Fisher** è chiamato **coefficiente di correlazione** oppure **correlazione del prodotto dei momenti**. Il termine correlazione era già presente nella ricerca statistica del secolo scorso, anche se **Galton** (1822-1911) parlava di ***co-relation***. Sir **Galton** è stato il primo ad usare il simbolo **r (**chiamato ***reversion*****)**, ma per indicare il coefficiente angolare **b** nei suoi studi sull'ereditarietà. La pratica di indicare il coefficiente di correlazione con **r** diventa generale a partire dal 1920.

La scuola francese sovente utilizza la dizione "**coefficiente di correlazione di Bravais-Pearson"**, per ricordare il connazionale Bravais (1846), che aveva presentato alcuni concetti importanti di tale metodo cinquanta anni prima di Karl Pearson.

Per spiegare le differenze logiche nell'uso della regressione e della correlazione, vari testi di statistica ricorrono a esempi divertenti o paradossali. Uno di questi è quanto evidenziato da un ricercatore dei

paesi nordici. In un'ampia area rurale, per ogni comune durante il periodo invernale è stato contato il numero di cicogne e quello dei bambini nati. E' dimostrato che all'aumentare del primo cresce anche il secondo.

Ricorrere all'analisi della regressione su queste due variabili, indicando per ogni comune con X il numero di cicogne e con Y il numero di nati, implica una relazione di causa-effetto tra presenza di cicogne (X) e nascite di bambini (Y). Anche involontariamente si afferma che i bambini sono portati dalle cicogne; addirittura, stimando **b,** si arriva ad indicare quanti bambini sono portati mediamente da ogni cicogna.

In realtà durante i mesi invernali, nelle case in cui è presente un neonato, la temperatura viene mantenuta più alta della norma, passando indicativamente dai 16 ai 20 gradi centigradi. Soprattutto nei periodi più rigidi, le cicogne sono attratte dal maggior calore emesso dai camini e nidificano più facilmente su di essi o vi si soffermano più a lungo. Con la correlazione si afferma solamente che le due variabili cambiano in modo congiunto.

L'analisi della correlazione misura solo il grado di associazione spaziale o temporale dei due fenomeni; ma lascia liberi nella scelta della motivazione logica, nel rapporto logico tra i due fenomeni. Il coefficiente r è una misura dell'intensità dell'associazione tra le due variabili.

Una presentazione chiara del**l'uso della correlazione** è fornita da **Fisher** stesso. Nonostante l'italiano del traduttore risenta del periodo, sulla base di una cultura biologica minima è possibile comprendere il ragionamento e la procedura che dovrebbero anche oggi caratterizzare il biologo. Si riafferma il concetto di non attribuire troppa importanza al puro aspetto statistico, se sganciato dal problema; è necessario utilizzare le due competenze congiuntamente. Nel caso particolare, un aspetto culturale importante è la presentazione dell'ereditarietà nell'uomo, tipica della cultura di **Fisher**, della sua scuola e del periodo storico a partire da **Galton.** In "***Metodi statistici ad uso dei ricercatori***, Torino 1948, Unione Tipografica Editrice Torinese (UTET), 326 p. traduzione di M Giorda, del testo ***Statistical Methods for Research Workers*** di R. A. **Fisher** 1945, nona edizione (la prima nel 1925) a pag. 163 si legge:

Nessuna quantità è più caratteristicamente impiegata in biometrica quanto il coefficiente di correlazione e nessun metodo è stato applicato a tanta varietà di dati quanto il metodo di correlazione. Specialmente nei casi in cui si può stabilire la presenza di varie cause possibili contribuenti a un fenomeno, ma non si può controllarle, i dati ricavati dall'osservazione hanno con questo mezzo assunto un'importanza assolutamente nuova. In un lavoro propriamente sperimentale, peraltro, la posizione del coefficiente di correlazione è molto meno centrale; esso, infatti, può risultare utile negli stadi iniziali d'una indagine, come quando due fattori che sono ritenuti indipendenti, risultano invece associati; ma è raro che, disponendo di condizioni sperimentali controllate, si intenda esprimere una conclusione nella forma di un coefficiente di correlazione.

Uno dei primi e più notevoli successi del metodo della correlazione si riscontrò nello studio biometrico dell'ereditarietà. In un tempo in cui nulla si conosceva del meccanismo dell'ereditarietà o della struttura della materia germinale, fu possibile, con questo metodo, dimostrare l'esistenza dell'ereditarietà e "misurarne l'intensità"; questo in un organismo nel quale non si potrebbero praticare allevamenti sperimentali, cioè nell'Uomo. Comparando i risultati ottenuti dalle misurazioni fisiche sull'uomo, con quelli ottenuti su altri organismi, si stabilì che la natura dell'uomo è governata dall'ereditarietà non meno di quella del resto del mondo animato. Lo scopo dell'analogia fu ulteriormente allargato dalla dimostrazione che coefficienti di correlazione della stessa grandezza si potevano ottenere tanto per le misurazioni fisiche, quanto per le qualità morali ed intellettuali dell'uomo.

Questi risultati rimangono di importanza fondamentale perché, non soltanto l'ereditarietà nell'uomo non è ancora suscettibile di studi sperimentali e gli attuali metodi di prova riguardanti l'intelletto sono, tuttora, inadatti ad analizzare le disposizioni intellettuali, ma perché, anche con organismi passibili di esperimenti e di misurazioni, è soltanto nel più favorevole dei casi che coll'ausilio dei metodi mendeliani possono essere determinati i diversi fattori causanti la variabilità incostante e studiati i loro effetti. Tale variabilità fluttuante, con una distribuzione pressoché normale, è caratteristica della maggioranza delle varietà più utili delle piante e degli animali domestici; e, quantunque, ci sia qui una forte ragione per ritenere che in tali casi l'ereditarietà è, in definitiva, mendeliana, il metodo biometrico di studio è, oggi giorno, il solo capace di alimentare le speranze di un reale progresso.

Questo metodo, che è anticamente basato sul coefficiente di correlazione, conferisce a questa quantità statistica un'effettiva importanza anche per coloro che preferiscono sviluppare la loro analisi con altri termini.

Nella correlazione, le due variabili vengono indicate con $\mathbf{X_1}$ e $\mathbf{X_2}$, non più con **X** (causa) e **Y** (effetto), per rendere evidente l'**assenza del concetto di dipendenza funzionale**. (Purtroppo, in vari lavori sono usati ugualmente X e Y, senza voler implicare il concetto della regressione).

L'indice statistico **(+r oppure –r) misura**
- **il tipo (con il segno + o -)**
- **e il grado (con il valore assoluto)**

di interdipendenza tra due variabili.

Il **segno** indica il **tipo di associazione**:
- **positivo**, quando le due variabili aumentano o diminuiscono insieme,
- **negativo**, quando all'aumento dell'una corrisponde una diminuzione dell'altra o viceversa.

Il **valore assoluto** varia da 0 a 1:

- **è massimo** (uguale a 1) quando c'è una **perfetta corrispondenza lineare** tra X_1 e X_2;
- tende a ridursi al diminuire della corrispondenza ed è zero quando essa è nulla.

L'indicatore della correlazione r è fondato sulla Codevianza e la Covarianza delle due variabili.

La **Codevianza** e la **Covarianza** tra $\mathbf{X_1}$ e $\mathbf{X_2}$ ($\mathbf{Cod_{X1/X2}}$ e $\mathbf{Cov_{X1/X2}}$) hanno la proprietà vantaggiosa di contenere queste due informazioni sul tipo (segno) ed sul grado (valore) di associazione; ma presentano anche lo svantaggio della regressione, poiché il loro valore risente in modo determinante della scala con la quale le due variabili $\mathbf{X_1}$ e $\mathbf{X_2}$ sono misurate.

Quantificando il peso in chilogrammi oppure in grammi e l'altezza in metri oppure in centimetri, si ottengono valori assoluti di Codevianza con dimensioni diverse, appunto perché fondati sugli scarti dalle medie ($x_i - \bar{x}$):

$$\mathrm{Cod}_{X1/X2} = \sum \left(x_{1i} - \bar{x}_1 \right) \cdot \left(x_{2i} - \bar{x}_2 \right)$$

E' possibile **pervenire a valori direttamente comparabili,** qualunque sia la dimensione dei due fenomeni, cioè ottenere valori **adimensionali**, solo ricorrendo ad unità standard, quale appunto la variazione tra – 1 e +1 . Si perviene ad essa,

- **mediante il rapporto tra la codevianza e la media geometrica delle devianze di $\mathbf{X_1}$ e $\mathbf{X_2}$**:

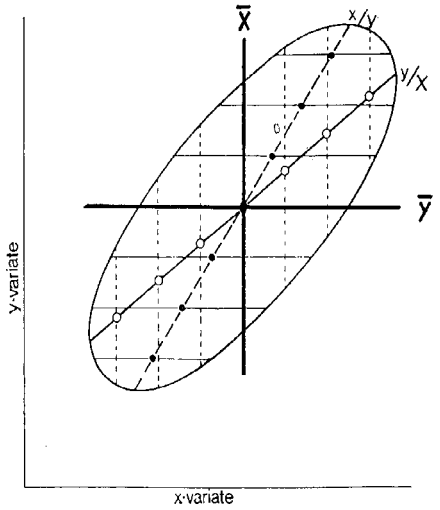
$$r = \frac{\sum \left(X_{1i} - \bar{X}_1 \right) \cdot \left(X_{2i} - \bar{X}_2 \right)}{\sqrt{\sum \left(X_{1i} - \bar{X}_1 \right)^2 \cdot \sum \left(X_{2i} - \bar{X}_2 \right)^2}}$$

In realtà la definizione è basata sulla covarianza e le due varianze: **la stima della correlazione è il rapporto tra la covarianza e la media geometrica delle due varianze.** Tuttavia, dato che le varianze sono ottenute dividendo le devianze per **n** (oppure i gradi di libertà in caso di campioni come sempre usato nelle formule presentate), anche **nel testo di Fisher si afferma che conviene basare il calcolo sulla codevianza e devianze**

Per comprendere il significato geometrico dell'indice **r** di correlazione e derivarne la formula, un approccio semplice è il **confronto tra le due rette di regressione**, calcolate dai valori di $\mathbf{X_1}$ e $\mathbf{X_2}$:

- la prima calcolata con $\mathbf{X_1}$ usata come variabile dipendente e $\mathbf{X_2}$ come variabile indipendente;
- la seconda scambiando le variabili, quindi utilizzando $\mathbf{X_2}$ come dipendente e $\mathbf{X_1}$ come indipendente.

(Per meglio distinguere le due rette, anche se errato è qui conveniente utilizzare X e Y per le due variabili)



Nella figura precedente, l'ellisse (la superficie contenuta nella figura piana chiusa) descrive la distribuzione di una nuvola di punti.

Quando si calcola la retta di regressione classica, presentata nel capitolo relativo,

$$\mathbf{Y = a + bX}$$

si ottiene la retta indicata con i punti vuoti (o bianchi).

Se si scambiano le variabili e si stima

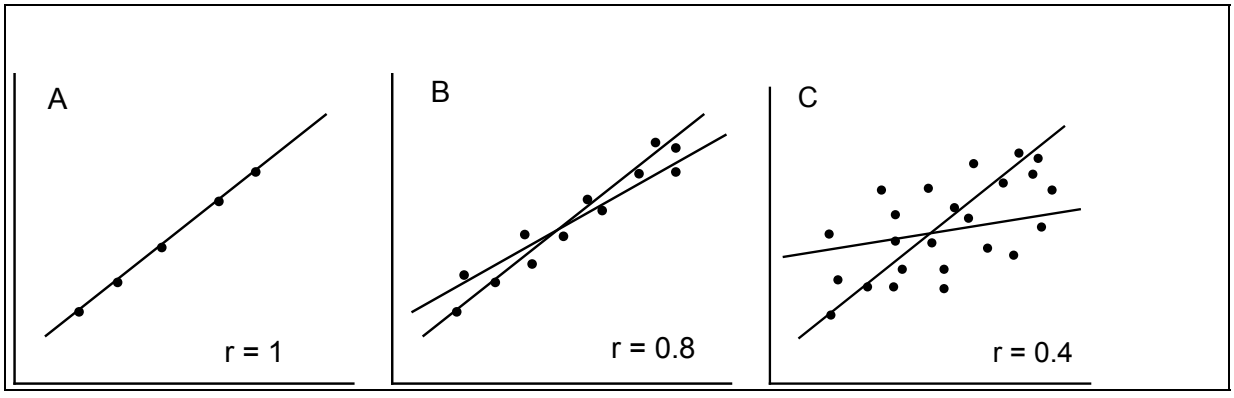
$$\mathbf{X = a+bY}$$

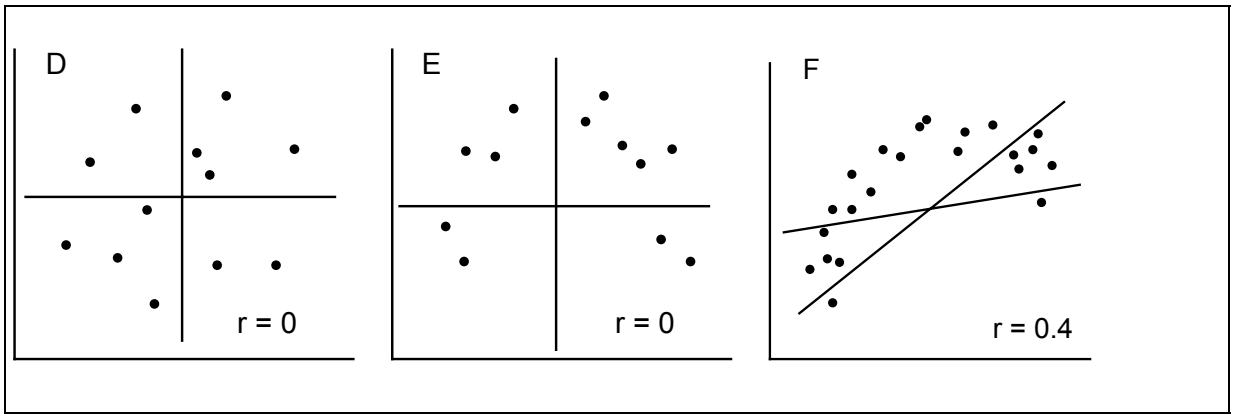
si ottiene la retta indicata dai punti in nero.

Entrambe passano da **baricentro della distribuzione**, individuato dall'**incontro delle due medie** ($\overline{X}$ e $\overline{Y}$), ma ognuna di esse è più vicina, in modo simmetrico, alla media della variabile indicata come effetto (la prima a Y e la seconda a X).

Il valore di correlazione lineare **r** può essere ricavato dai due coefficienti angolari **b**.

Le due rette coincidono solamente quando i punti sono disposti esattamente lungo una retta.





Le due rette, riportate in ognuna dei 6 figure precedenti, sono calcolate sulle stesse coppie di osservazioni, scambiando appunto X_1 e X_2. Esse

- i intersecano nel **baricentro della distribuzione**, il punto che rappresenta il valore medio di $\mathbf{X_1}$ e di $\mathbf{X_2}$;
- ma **non sono identiche o coincidenti** (eccetto nella figura A, in cui **r = 1**), poiché entrambe tendono ad avvicinarsi alla media della variabile assunta come dipendente.

Quando le due rette sono tra loro perpendicolari (figura D e figura E) con angoli di 90° e coincidono con le due medie, le due variabili sono indipendenti e tra loro non esiste alcuna correlazione (**r = 0**); inversamente, quando le due rette tendono ad avvicinarsi con un angolo minore, il valore assoluto della correlazione tende ad aumentare (figura C e figura B). Il valore massimo (**r = 1**) viene raggiunto quando le due rette coincidono e l'angolo tra esse è nullo (figura A).

Il segno della correlazione dipende dal coefficiente angolare delle due rette: è positivo, se il loro coefficiente angolare è positivo, mentre è negativo quando il coefficiente angolare è negativo. Pertanto il valore di **r** può variare tra **+1** e **-1**.

(Tra le figure non sono stati riportati valori di **r** negativi: la distribuzione dei punti avrebbe evidenziato una diminuzione dei valori della ordinata al crescere di quelli in ascissa e quindi le due rette avrebbero avuto una inclinazione verso il basso all'aumentare dell'ascissa.)

E' importante ricordare che **un valore assoluto basso o nullo di correlazione non deve essere interpretato come assenza di una qualsiasi forma di relazione tra le due variabili:**

- **è assente solo una relazione di tipo lineare**,
- ma tra esse **possono esistere relazioni di tipo non lineare**, espresse da curve di ordine superiore, tra le quali la più semplice e frequente è quella di secondo grado.

L'informazione contenuta in r riguarda solamente la quota espressa da una relazione lineare.

Per derivare la formula di **r** da quanto già evidenziato sulla regressione lineare semplice, è utile ricordare che essa può essere vista come **la media geometrica dei due coefficienti angolari (b) di regressione lineare**.

Infatti, indicando con

- $\mathbf{b_{x1\,/\,x2}}$ il **coefficiente angolare** della **prima retta** di regressione,
- $\mathbf{b_{x2\,/\,x1}}$ il **coefficiente angolare** della **seconda retta** di regressione,

il **coefficiente di correlazione r** può essere stimato

come

$$r = \sqrt{b_{X_1/X_2} \cdot b_{X_2/X_1}}$$

Poiché

$$b_{(i\,/\,j)} = \frac{Cod.ij}{Dev.i}$$

e dato che **le due Codevianze sono identiche**,

$$r = \sqrt{\frac{Cod.ij}{Dev.i} \cdot \frac{Cod.ji}{Dev.j}}$$

dopo semplificazione,

nella formulazione estesa con la consueta simbologia

si ottiene

$$r = \frac{\sum \left(X_1 - \overline{X}\right) \cdot \left(X_2 - \overline{X}\right)}{\sqrt{\sum \left(X_1 - \overline{X}\right)^2 \cdot \sum \left(X_2 - \overline{X}\right)^2}}$$

Per calcolare il coefficiente di correlazione da una serie di rilevazioni, si possono presentare due casi distinti:

- il primo, con **poche osservazioni**, quando i dati sono forniti come **coppie distinte di valori**;
- il secondo, con **molte osservazioni**, quando i dati sono stati raggruppati in **classi di frequenza**.

La formula sopra riportata è applicabile nel caso di osservazioni singole.

ESEMPIO. In 18 laghi dell'Appennino Tosco-Emiliano sono state misurate la conducibilità e la concentrazione di **anioni + cationi**, ottenendo le coppie di valori riportati nella tabella

Laghi	Conducibilità (X_1)	Anioni + Cationi (X_2)
SILLARA INF.	20	0,328
SILLARA SUP.	22	0,375
SCURO CERR.	22	0,385
VERDAROLO	26	0,445
SQUINCIO	24	0,445
SCURO PARMENSE	28	0,502
PALO	27	0,503
ACUTO	26	0,520
SCURO	29	0,576
COMPIONE INF.	35	0,645
GEMIO INF.	33	0,650
PETRUSCHIA	37	0,675
GEMIO SUP.	34	0,680
SANTO PARMENSE	35	0,746
BICCHIERE	37	0,764
BALLANO	39	0,845
BACCIO	41	0,936
VERDE	45	0,954

Calcolare il coefficiente di correlazione tra queste due variabili

a) - in modo diretto e

b) - mediante i due coefficienti angolari,

per meglio comprendere l'equivalenza delle due formule.

Risposta.

A) In modo diretto, con la formula che utilizza le singole coppie di valori

$$r = \frac{\sum \left(X_{1i} - \overline{X}_1\right) \cdot \left(X_{2i} - \overline{X}_2\right)}{\sqrt{\sum \left(X_{1i} - \overline{X}_1\right)^2 \cdot \sum \left(X_{2i} - \overline{X}_2\right)^2}}$$

si ottiene un valore di **r**

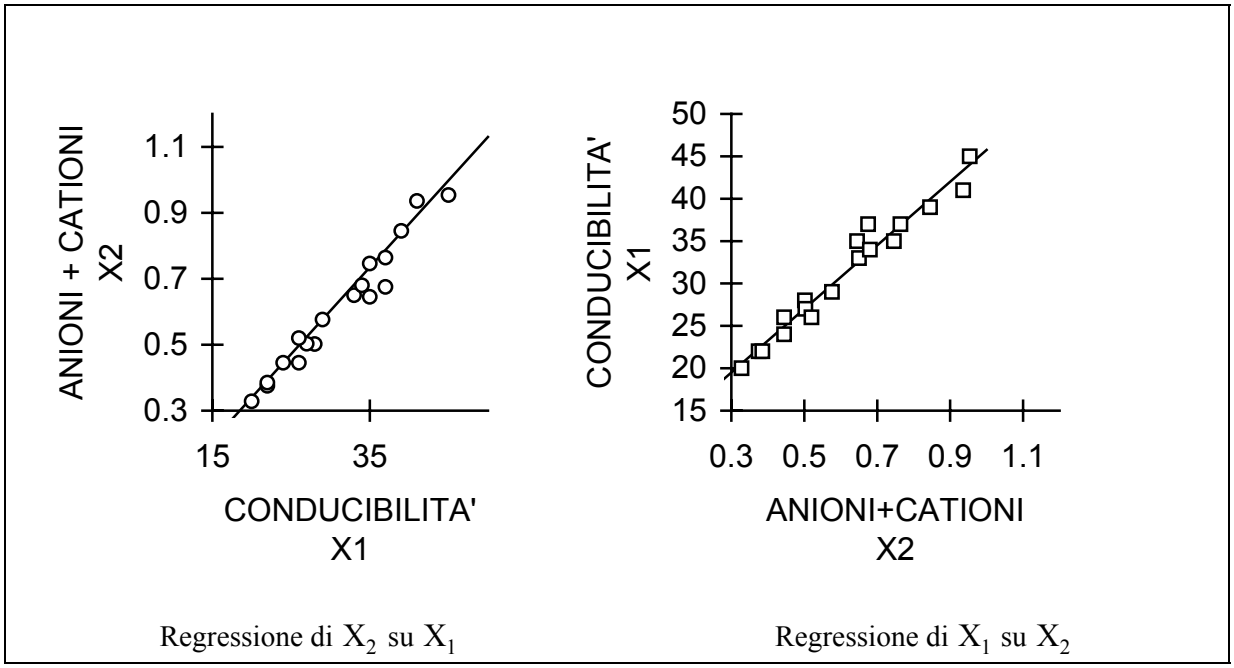
$$\mathbf{r} = \frac{22{,}893}{\sqrt{0{,}606 \cdot 887{,}778}} = \frac{22{,}893}{23{,}194} = \mathbf{0{,}987}$$

uguale a 0,987.

Utilizzando i coefficienti angolari delle due regressioni, che dai calcoli risultano

$$b_{X_2/X_1} = 0,026; \qquad b_{X_1/X_2} = 37,521$$

e che sono rappresentati nelle due figure seguenti



Regressione di X_2 su X_1 Regressione di X_1 su X_2

il coefficiente di correlazione

$$r = \sqrt{b_{X_2/X_1} \cdot b_{X_1/X_2}} = \sqrt{0,026 \cdot 37,521} = 0,9876$$

risulta uguale a 0,9876

con una differenza, dalla stima precedente, determinata dagli arrotondamenti.

Nel caso di **osservazioni raggruppate in classi**, il metodo per calcolare l'indice di correlazione resta sostanzialmente invariato, rispetto a quello presentato nel capitolo sulla statistica descrittiva.
Per ogni classe, come valore rappresentativo viene assunto il valore centrale; le differenze tra questi valori centrali di ogni classe ed il valore centrale di tutta la distribuzione devono essere moltiplicate per il numero di osservazioni.

Per semplificare i calcoli e per una esatta comprensione del fatto che **le variazioni di scala non incidono assolutamente sul valore di r** (**che è adimensionale**) è possibile utilizzare non i valori

osservati ma gli scarti delle grandezze da una qualsiasi origine arbitraria. Di norma è quella centrale, in quanto determina scarti minimi e simmetrici.
La classe di frequenza centrale o prossima al centro viene indicata con zero e le altre con il numero progressivo, positivo a destra e negativo a sinistra, di distanze unitarie da essa.

Per esempio, la distribuzione di $\mathbf{X_1}$ in 7 classi

50-69	70-89	90-109	110-129	130-149	150-169	170-189

che potrebbe utilizzare i valori centrali relativi (60, 80, 100, 120, 140, 160, 180) per il calcolo dell'indice r di correlazione può essere utilmente trasformata in una scala unitaria

-3	-2	-1	0	+1	+2	+3

mentre la distribuzione della variabile $\mathbf{X_2}$ in 6 classi

20-29	30-39	40-49	50-59	60-69	70-79

può essere trasformata in

-3	-2	-1	0	+1	+2

un'altra distribuzione arbitraria equivalente, seppure non simmetrica come la precedente.
E' intuitivo che, con questi nuovi dati, i prodotti e le somme necessarie alla stima del coefficiente di correlazione **r** risultano molto semplificati, per un calcolo manuale. Sono quindi tecniche del passato, superate dalle nuove possibilità offerte dall'informatica, con la quale non si pongono problemi di semplificazione dei calcoli. Restano però importanti i concetti: **l'indice di correlazione r tra due variabili è adimensionale**, fornisce lo stesso valore al variare delle scale di misura.

Ritornando al concetto dell'invarianza del valore di **r** rispetto al tipo di scala, nulla muterebbe nel suo valore se la prima o la seconda distribuzione fossero trasformate in una scala ancora differente, come la seguente

0	1	2	3	4	5	...	N

Con **dati raggruppati in distribuzioni di frequenze**, il coefficiente di correlazione **r** può essere ottenuto con la solita formula

$$r = \frac{Cod.X_1X_2}{\sqrt{Dev.X_1} \cdot \sqrt{Dev.X_2}}$$

in cui la **Codevianza** di $\mathbf{X_1}$ e $\mathbf{X_2}$ è data da

$$Cod.X_1X_2 = \sum f_{X1X2} \cdot d_{X1} \cdot d_{X2} - \frac{\left(\sum f_{X1} \cdot d_{X2}\right) \cdot \left(\sum f_{X2} \cdot d_{X2}\right)}{N}$$

e le **due devianze** da

$$Dev_{x1} = \sum f_{x1} \cdot d_{x1}^2 - \frac{\left(\sum f_{x1} \cdot d_{x1}\right)^2}{\sum f_{x1}}$$

la prima

e da

$$Dev_{x2} = \sum f_{x2} \cdot d_{x2}^2 - \frac{\left(\sum f_{x2} \cdot d_{x2}\right)^2}{\sum f_{x2}}$$

la seconda,

dove

- $\mathbf{d_{X1}}$ e $\mathbf{d_{X2}}$ sono gli scarti, misurati su una scala arbitraria, dei valori delle classi dall'origine scelta;
- $\mathbf{f_{X1}}$ e $\mathbf{f_{X2}}$ sono le frequenze dei valori di X_1 e di X_2 entro ciascuna classe;
- $\mathbf{f_{X1X2}}$ sono le frequenze delle coppie X_1-X_2 entro ciascuna coppia di classi.

ESEMPIO 1. Da una serie di rilevazioni effettuate su un campione d'acqua di 17 laghi (riportate nella tabella successiva)

A - costruire la relativa tabella a doppia entrata di distribuzione delle frequenze e

B - calcolare da essa il coefficiente di correlazione semplice r.

Laghi	Conducibilità X_1	Anioni + Cationi X_2
SCURO PR	28	0,502
COMPIONE INF.	35	0,645
SANTO PARMENSE	35	0,746
COMPIONE SUP.	45	0,815
BALLANO	39	0,845
BACCIO	41	0,936
PRADACCIO	53	1,218
OSIGLIA	54	1,259
SANTO MODENESE	79	1,382
NERO PIACENTINO	61	1,530
BUONO	71	1,771
NINFA	96	2,162
PRANDA	108	2,192
CERRETANO	99	2,272
SCAFFAIOLO	108	2,317
PADULE CERRETO	122	2,563
LAME	110	2,616

Risposte

A) Dai dati, è possibile ricavare la tabella a doppia entrata, come quella di seguito riportata

	X_1	21-36	37-52	53-68	69-84	85-100	101-116	117-133		
X_2	F_{X1X2}	-3	-2	-1	0	1	2	3	f_{X2}	Medie
0,5-0,8	-3	3 (27)							3	0,631
0,8-1,1	-2		3 (12)						3	0,865
1,1-1,4	-1			2 (2)	1				3	1,286
1,4-1,7	0			1					1	1,530
1,7-2,0	1				1				1	1,771
2,0-2,3	2					2 (4)	1 (4)		3	2,208
2,3-2,6	3						1 (4)	1 (9)	2	2,440
2,6-2,9	4						1 (8)		1	2,616
F_{X1}		3	3	3	2	2	3	1	17	
Medie		32,66	41,66	56,00	75,00	97,50	108,66	122,00		

Nel riquadro interno della tabella sono riportate le f_{x1x2} e (tra parentesi) i prodotti $fx \cdot d_{x1} \cdot d_{x2}$ che saranno utilizzati per il calcolo della codevianza. Non sono state riportate le frequenze nulle.
I vari passaggi necessari per stimare la Devianza di X_1 dai dati della distribuzione in classi sono riportati nella tabella successiva

X_1	f_x	d_x	$f_x \cdot d_x$	$f_x \cdot d_x^2$
21-37	3	-3	-9	27
37-53	3	-2	-6	12
53-69	3	-1	-3	3
69-85	2	0	0	0
85-101	2	+1	2	2
101-117	3	+2	6	12
117-133	1	+3	3	9
TOTALE	17		-7	65

Con la formula abbreviata

$$\sqrt{Dev_x} = \sqrt{\sum f_x \cdot d_x^2 - \frac{\left(\sum f_x \cdot d_x\right)^2}{\sum f_x}} = \sqrt{65 - \frac{(-7)^2}{17}} = 7{,}88$$

si ottiene la radice quadrata della devianza di X_1, utile ai calcoli successivi, che è uguale a 7,88.

Seguendo le stesse modalità, il calcolo della Devianza di X_2 e della sua radice quadrata (i cui passaggi sono riportati nella tabella successiva)

X_2	f_x	d_x	$f_x \cdot d_x$	$f_x \cdot d_x^2$
0,5-0,8	3	-3	-9	27
0,8-1,1	3	-2	-6	12
1,1-1,4	3	-1	-3	3
1,4-1,7	1	0	0	0
1,7-2,0	1	+1	1	1
2,0-2,3	3	+2	6	12
2,3-2,6	2	+3	6	18
2,6-2,9	1	+4	4	16
TOTALE	17		-1	89

$$\sqrt{89-\frac{(-1)^2}{17}}=9{,}43$$

fornisce un risultato di 9,43.

Dalle due tabelle è possibile ottenere i dati necessari alla stima della Codevianza

$$Cod_{X1X2}=\sum f_{X1X2}\cdot d_{X1}\cdot d_{X2}-\frac{\left(\sum f_{X1}\cdot d_{X1}\right)\cdot\left(\sum f_{X2}\cdot d_{X2}\right)}{N}=72-\frac{(-7)(-1)}{17}=71{,}588$$

dove $N=\sum f_{X1}=\sum f_{X2}$

che risulta uguale a 71,588

Il coefficiente di correlazione r

$$r=\frac{Cod_{X1X2}}{\sqrt{Dev_{X1}}\cdot\sqrt{Dev_{X2}}}=\frac{71{,}588}{9{,}431\cdot 7{,}88}=0{,}963$$

risulta uguale a 0,963.

E' semplice verificare empiricamente, come dimostrano i calcoli successivi, che anche cambiando i valori di d_{X1} e d_{X2} il coefficiente di correlazione non cambia.

	X_1	21-36	37-52	53-68	69-84	85-100	101-116	117-133		
X_2	F_{X1X2}	0	1	2	3	4	5	6	f_{X2}	Medie
0,5-0,8	0	3							3	0,631
0,8-1,1	1		3 (3)						3	0,865
1,1-1,4	2			2 (8)	1 (6)				3	1,286
1,4-1,7	3			1 (6)					1	1,530
1,7-2,0	4				1 (12)				1	1,771
2,0-2,3	5					2 (40)	1 (25)		3	2,208
2,3-2,6	6						1 (30)	1 (36)	2	2,440
2,6-2,9	7						1 (35)		1	2,616
F_{X1}		3	3	3	2	2	3	1	17	
Medie		32,66	41,66	56,00	75,00	97,50	108,66	122,00		

Si può infatti notare le f_{X1} ed f_{X2} sono rimaste inalterate, mentre sono cambiate le f_{X1X2}

La radice quadrata della Devianza di X_1

i cui passaggi sono riportati nella tabella successiva

X_1	f_x	d_x	$f_x \cdot d_x$	$f_x \cdot d_x^2$
21-37	3	0	0	0
37-53	3	1	3	3
53-69	3	2	6	12
69-85	2	3	6	18
85-101	2	4	8	32
101-117	3	5	15	75
117-133	1	6	6	36
TOTALE	17		44	176

$$\sqrt{\text{Dev}_x} = \sqrt{\sum f_x \cdot d_x^2 - \frac{\left(\sum f_x \cdot d_x\right)^2}{\sum f_x}} = \sqrt{176 - \frac{(44)^2}{17}} = \sqrt{176 - 113{,}882} = \sqrt{62{,}118} = 7{,}881$$

risulta uguale a 7,881

e la radice quadrata della devianza di X_2

X_2	f_x	d_x	$f_x \cdot d_x$	$f_x \cdot d_x^2$
0,5-0,8	3	0	0	0
0,8-1,1	3	1	3	3
1,1-1,4	3	2	6	12
1,4-1,7	1	3	3	9
1,7-2,0	1	4	4	16
2,0-2,3	3	5	15	75
2,3-2,6	2	6	12	72
2,6-2,9	1	7	7	49
TOTALE	17		50	236

$$\sqrt{Dev_{X2}} = \sqrt{\sum f_{X2} \cdot d_{X2}^2 - \frac{\left(\sum f_{X2} \cdot d_{X2}\right)^2}{\sum f_{X2}}} = \sqrt{236 - \frac{(50)^2}{17}} = \sqrt{236 - 147{,}059} = \sqrt{88{,}941} = 9{,}431$$

risulta uguale a 9,431.

La Codevianza di X_1 e X_2

$$Cod_{X1X2} = \sum f_{X1X2} \cdot d_{X1} \cdot d_{X2} - \frac{\left(\sum f_{X1} \cdot d_{X1}\right) \cdot \left(\sum f_{X2} \cdot d_{X2}\right)}{N} =$$

dove $N = \sum f_{X1} = \sum f_{X2}$

$$= 201 - \frac{44 \cdot 50}{17} = 201 - 129{,}412 = 71{,}588$$

risulta uguale a 71,588.

Essendo rimaste invariate sia la Codevianza che le Devianze, il coefficiente di correlazione semplice **r** non può che rimanere identico.

18.2. CONDIZIONI DI VALIDITA' E SIGNIFICATIVITA' DI r CON $\rho = 0$ E CON $\rho \neq 0$

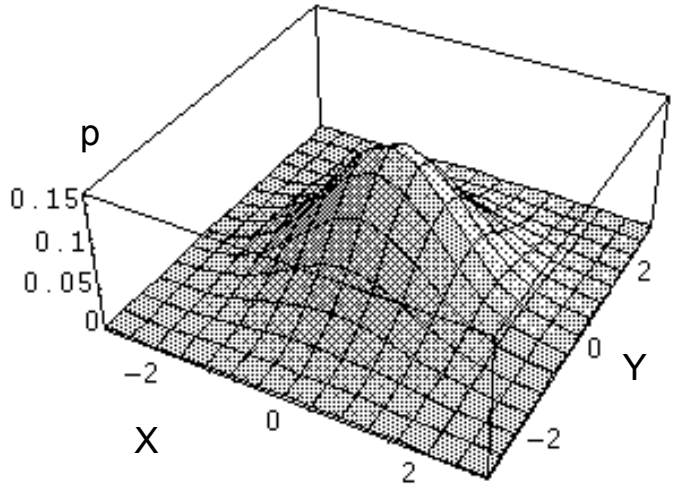
Le **condizioni di validità** della correlazione, il cui indice nel caso di una popolazione è indicato con ρ (rho), sono le stesse della regressione. Tuttavia, mentre nella regressione sono applicate solo alla variabile Y, nel caso della correlazione, che utilizza indistintamente entrambe le variabili, **richiede che sia X_1 che X_2 siano distribuite in modo approssimativamente normale.**

Con due variabili, l'ipotesi di normalità della distribuzione pretende la **distribuzione normale bivariata**, che è un'estensione a tre dimensioni della curva normale.

Mentre la superficie di una distribuzione univariata è determinata in modo compiuto da due parametri (media μ e deviazione standard σ), la superficie normale bivariata è determinata da cinque parametri:

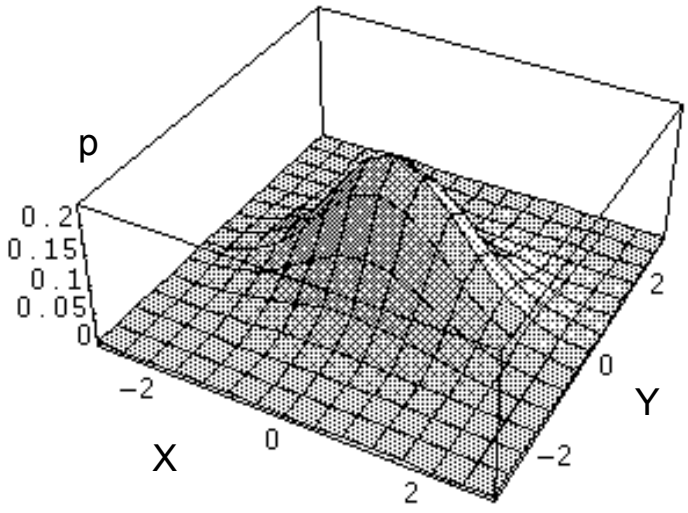
- media e deviazione standard della variabile X_1,
- media e deviazione standard della variabile X_2,
- coefficiente di correlazione (ρ) tra X_1 e X_2.

La sua rappresentazione grafica, nel caso in cui non esista correlazione ($\rho = 0$) tra le due variabili ed esse abbiano varianza uguale, determina una figura come quella riportata:



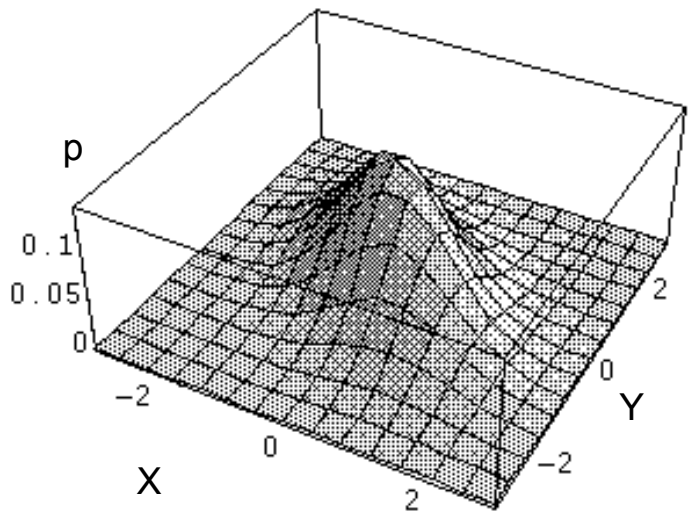
Distribuzione normale bivariata
X e Y sono due variabili **indipendenti ($\rho = 0$)** di **uguale varianza** ($\sigma_X^2 = \sigma_Y^2 = 1$)

La distribuzione normale bivariata assume la forma di una collina di forma circolare, che degrada nello stesso modo su tutti i versanti; la pendenza dipende dal valore della varianza.



Distribuzione normale bivariata
X e Y sono due variabili **indipendenti** ($\rho = 0$) con **varianze diverse** ($\sigma_X^2 = 1; \quad \sigma_Y^2 = 0,7$)

Quando le varianze sono diverse, sempre nel caso che non esista correlazione, la rappresentazione grafica assume la forma di una collina a pendenze diverse, con un declino più rapido dove la varianza è minore, con frequenze maggiori lungo la retta individuata da X medio e da Y medio.
Quando esiste correlazione, come nella figura successiva, la distribuzione bivariata tende ad assumere la forma di una cresta di montagna, distribuita in diagonale rispetto alle due medie. La cresta è tanto più sottile quanto più alto è il valore ρ della correlazione.



Distribuzione normale bivariata
X e Y sono due variabili **correlate** ($\rho = 0,6$) di **uguale varianza** ($\sigma_X^2 = \sigma_Y^2 = 1$)

Con **ρ = 1** la rappresentazione grafica diventa un piano perpendicolare alla base, posto in diagonale rispetto alle ascisse e alle ordinate. Il segno della correlazione determina solo la direzione di tale piano rispetto alla base.

Dopo il calcolo di un **coefficiente di correlazione r**, sempre valido come indice che misura la relazione tra due variabili in quanto solo descrittivo come il calcolo di una media o di una varianza, può porsi il **duplice problema** della sua **significatività**, cioè di verificare

a) l'ipotesi nulla **H_0: ρ = 0 (non significativamente diverso da zero)**

b) l'ipotesi nulla **H_0: ρ = ρ_0 (non significativamente diverso da un qualsiasi valore prefissato, ma diverso da zero)**

con ipotesi alternativa bilaterale oppure unilaterale in entrambi i casi.

A differenza dei test sulla media e sul coefficiente angolare **b** (oppure l'intercetta **a**), che possono assumere qualsiasi valore e quindi essere sempre distribuiti normalmente rispetto al valore della popolazione, un test di significatività pone problemi differenti di validità se intende verificare l'ipotesi nulla

a) $\rho = 0$

b) $\rho \neq 0$.

Nel primo caso ($\rho = 0$), i valori campionari **r** possono essere assunti come distribuiti in modo **approssimativamente normale e simmetrico** rispetto alla correlazione della popolazione (ρ).
Nel secondo caso ($\rho \neq 0$), i valori campionari **r** si distribuiscono in modo sicuramente asimmetrico intorno alla correlazione della popolazione (ρ) e in modo tanto più accentuato quanto più essa si allontana da zero e si avvicina a uno dei due estremi (**-1** o **+1**). E' intuitivo che, considerando ad esempio risultati positivi, con un valore reale di $\rho = 0{,}9$ il valore campionario **r** non potrà mai superare 1, mentre potrebbe essere 6 se non 5 oppure 4, in funzione del numero di dati

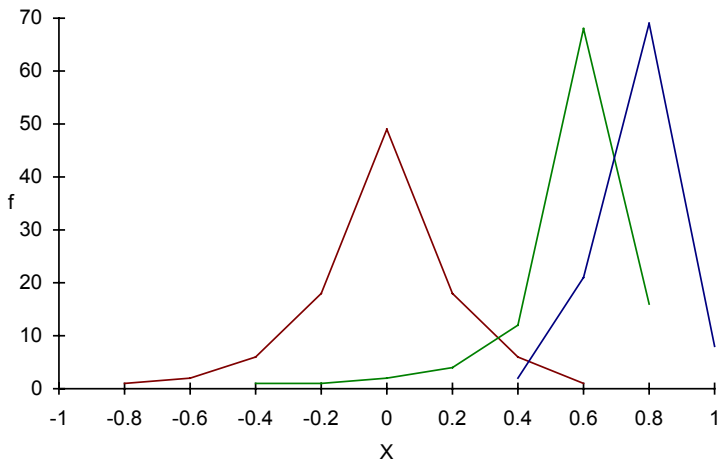


Grafico delle distribuzioni campionarie di 3 coefficienti di correlazione.
La distribuzione è simmetrica solo quando il suo valore atteso (ρ) è zero.

In questo secondo caso, occorre procedere ad una trasformazione di r, per rispettare la condizioni di validità.

VALORI CRITICI IN TEST <u>BILATERALE</u>

DEL COEFFICIENTE DI CORRELAZIONE SEMPLICE r

(DF = N-2) **CON IPOTESI H_0: $\rho = 0$**

DF	α=0.05	α=0.01	α=0.001
1	0,9969	0,9999	1,0000
2	0,9500	0,9900	0,9990
3	0,8783	0,9587	0,9911
4	0,8114	0,9172	0,9741
5	0,7545	0,8745	0,9509
6	0,7067	0,8343	0,9249
7	0,6664	0,7977	0,8983
8	0,6319	0,7646	0,8721
9	0,6021	0,7348	0,8471
10	0,5760	0,7079	0,8233
11	0,5529	0,6835	0,8010
12	0,5324	0,6614	0,7800
13	0,5139	0,6411	0,7604
14	0,4973	0,6226	0,7419
15	0,4821	0,6055	0,7247
16	0,4683	0,5897	0,7084
17	0,4555	0,5751	0,6932
18	0,4438	0,5614	0,6788
19	0,4329	0,5487	0,6652
20	0,4227	0,5368	0,6524
21	0,4132	0,5256	0,6402
22	0,4044	0,5151	0,6287
23	0,3961	0,5052	0,6177
24	0,3882	0,4958	0,6073
25	0,3809	0,4869	0,5974
26	0,3739	0,4785	0,5880
27	0,3673	0,4705	0,5790
28	0,3610	0,4629	0,5703
29	0,3550	0,4556	0,5620
30	0,3494	0,4487	0,5541

DF	α=0.05	α=0.01	α=0.001
35	0,3246	0,4182	0,5189
40	0,3044	0,3932	0,4896
45	0,2875	0,3721	0,4647
50	0,2732	0,3541	0,4433
55	0,2609	0,3385	0,4245
60	0,2500	0,3248	0,4079
65	0,2405	0,3127	0,3911
70	0,2319	0,3017	0,3798
75	0,2242	0,2919	0,3678
80	0,2172	0,2830	0,3569
85	0,2108	0,2748	0,3468
90	0,2050	0,2673	0,3376
95	0,1996	0,2604	0,3291
100	0,1946	0,2540	0,3211
110	0,1857	0,2425	0,3069
120	0,1779	0,2324	0,2943
130	0,1710	0,2235	0,2832
140	0,1648	0,2155	0,2733
150	0,1593	0,2083	0,2643
160	0,1543	0,2019	0,2562
170	0,1497	0,1959	0,2488
180	0,1455	0,1905	0,2420
190	0,1417	0,1855	0,2357
200	0,1381	0,1809	0,2299
300	0,113	0,148	0,188
400	0,098	0,128	0,164
500	0,088	0,115	0,146
600	0,080	0,105	0,134
700	0,074	0,097	0,124
800	0,069	0,091	0,116
900	0,065	0,086	0,109
1000	0,062	0,081	0,104

VALORI CRITICI IN TEST UNILATERALE

DELCOEFFICIENTE DI CORRELAZIONE SEMPLICE r

(DF = N-2) **CON IPOTESI H_0: $\rho = 0$**

DF	α=0.05	α=0.01	α=0.001
1	0,988	1,000	1,000
2	0,900	0,980	0,998
3	0,805	0,934	0,986
4	0,729	0,882	0,963
5	0,669	0,833	0,935
6	0,621	0,789	0,905
7	0,582	0,750	0,875
8	0,549	0,715	0,847
9	0,521	0,685	0,820
10	0,497	0,658	0,795
11	0,476	0,634	0,772
12	0,457	0,612	0,750
13	0,441	0,592	0,730
14	0,426	0,574	0,711
15	0,412	0,558	0,694
16	0,400	0,542	0,678
17	0,389	0,529	0,662
18	0,378	0,515	0,648
19	0,369	0,503	0,635
20	0,360	0,492	0,622
21	0,352	0,482	0,610
22	0,344	0,472	0,599
23	0,337	0,462	0,588
24	0,330	0,453	0,578
25	0,323	0,445	0,568
26	0,317	0,437	0,559
27	0,311	0,430	0,550
28	0,306	0,423	0,541
29	0,301	0,416	0,533
30	0,296	0,409	0,526

DF	α=0.05	α=0.01	α=0.001
35	0,275	0,381	0,492
40	0,257	0,358	0,463
45	0,243	0,338	0,439
50	0,231	0,322	0,449
55	0,220	0,307	0,401
60	0,211	0,295	0,385
65	0,202	0,284	0,371
70	0,195	0,274	0,358
75	0,189	0,264	0,347
80	0,183	0,257	0,336
85	0,178	0,249	0,327
90	0,173	0,242	0,318
95	0,168	0,236	0,310
100	0,164	0,230	0,303
110	0,156	0,220	0,289
120	0,150	0,210	0,277
130	0,144	0,202	0,267
140	0,139	0,195	0,257
150	0,134	0,189	0,249
160	0,130	0,183	0,241
170	0,126	0,177	0,234
180	0,122	0,172	0,228
190	0,119	0,168	0,222
200	0,116	0,164	0,216
300	0,095	0,134	0,177
400	0,082	0,116	0,154
500	0,074	0,104	0,138
600	0,067	0,095	0,126
700	0,062	0,088	0,116
800	0,058	0,082	0,109
900	0,055	0,077	0,103
1000	0,052	0,073	0,098

Quando l'ipotesi nulla è

$$\mathbf{H_0:\ \rho = 0}$$

la significatività del coefficiente angolare **r** può essere verificata con 3 modalità, che ovviamente forniscono risultati identici:

1 – la tabella dei valori di **r**, in funzione di α e dei gdl (oppure del numero **n** di osservazioni),

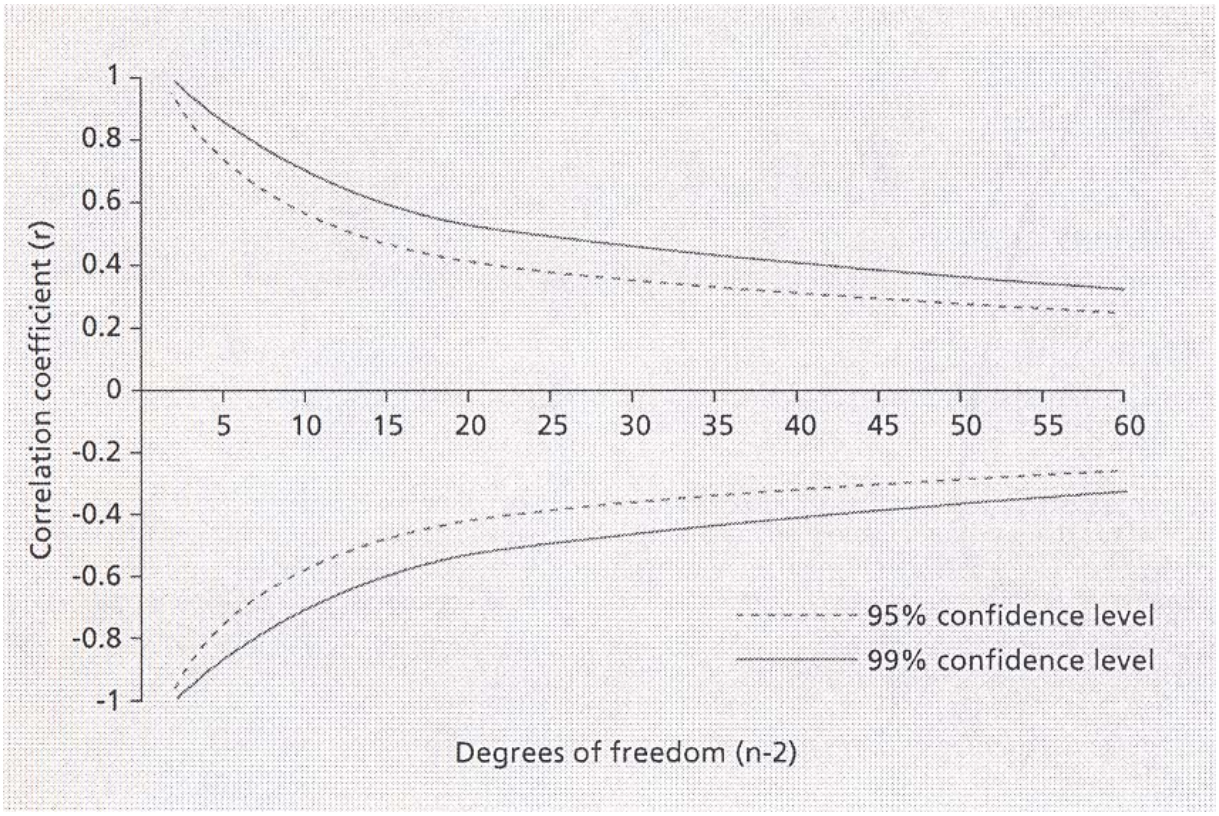
2 – il test **F** di Fisher-Snedecor,

3 – il test **t** di Student.

La prima modalità utilizza le tabelle sinottiche del valore di **r**, con gradi di libertà **n-2**, come sono stati riportati nelle pagine precedenti. Di conseguenza, è evidente che occorrono almeno 3 coppie d'osservazioni (DF = 1).

La semplice lettura dei valori critici nella tabella alle probabilità $\alpha = 0.05$, $\alpha = 0.01$ e $\alpha = 0.001$

DF	**$\alpha = 0.05$**	**$\alpha = 0.01$**	**$\alpha = 0.001$**
3	0,8783	0,9587	0,9911
200	0,1381	0,1809	0,2299
1000	0,062	0,081	0,104



e quella del grafico mostrano come sia errata l'affermazione semplicistica, riportata su alcuni testi, che un valore di correlazione **r = 0,3** sia indicativamente basso e un valore **r = 0,5** sia alto.
La significatività della correlazione è fortemente influenzata dai DF, in modo molto più marcato di quanto avviene nella distribuzione **t di Student** e nella distribuzione **F di Fisher-Snedecor**.
Dal semplice confronto delle due serie riportate nella tabellina precedente e dalla lettura del grafico grafico, risulta evidente che,

- con **pochi dati**, potrebbe non essere significativo alla probabilità α = **0.05** un valore di **r** apparentemente alto quale **0,85**;
- con **molti dati**, potrebbe essere altamente significativo, alla probabilità α **= 0.001,** anche un valore apparentemente basso, quale **0,25**.

Pochi testi riportano i valori critici di **r, validi per verificare l'ipotesi nulla H_0: ρ = 0**; quasi sempre si deve ricorre alla distribuzione **F** o a quella **t** che tutti i testi, anche elementari, riportano. Pure i programmi informatici, insieme con il valore di **r,** riportano la probabilità di **F** e/o di **t**.

Ricorrendo ai concetti spiegati nella regressione lineare semplice, anche nella verifica dell'ipotesi nulla relativa alla correlazione

$$\mathbf{H_0}: \rho = 0$$

il test **F**, con gdl **1** e **n-2**,

$$\mathbf{F_{1,n-2}} = \frac{\frac{r^2}{1}}{\frac{1-r^2}{n-2}}$$

è dato dal rapporto tra

- la **varianza dovuta alla regressione** (la devianza $\mathbf{r^2}$ / **1** df) e
- la **varianza d'errore** (la devianza d'errore **1 -** $\mathbf{r^2}$ / **n-2** df)

La formula semplificata diventa

$$\mathbf{F_{1,n-2}} = \frac{r^2 \cdot (n-2)}{1-r^2}$$

Con il test **t,** che ha df **n-2,**
ricordando nuovamente che

$$t_{n-2} = \sqrt{F_{1,n-2}}$$

la formula abitualmente utilizzata

è

$$t_{(n-2)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Con il test **F**, è possibile

- sia la verifica dell'ipotesi alternativa $\mathbf{H_1}$ bilaterale

$$\mathbf{H_1}: \rho \neq \mathbf{0}$$

- sia la verifica dell'ipotesi alternativa $\mathbf{H_1}$ unilaterale

$$H_1: \rho > \mathbf{0} \qquad \text{oppure} \qquad H_1: \rho < \mathbf{0}$$

assumendo sempre in una distribuzione bilaterale al posto delle probabilità 0.05, 0,01 e 0.001 rispettivamente le probabilità 0.10, 0.02, 0.002, come nelle tabelle precedenti sui valori critici di **r**. Ma è di più difficile comprensione, per chi non abbia ancora abbastanza familiarità con i test statistici, perché la distribuzione F con pochi gdl, come di solito nella pratica sperimentale, è asimmetrica.

La distribuzione **t**, in quanto simmetrica come la distribuzione **z**, permette di meglio comprendere la scelta delle probabilità in rapporto alla direzione dell'ipotesi alternativa. Per molti è quindi preferibile al test F, in particolare in test unilaterali, pure fornendo valori identici ai due metodi prima presentati.

ESEMPIO 1. La tavola sinottica di **r** per test bilaterali, con **df 15** alla probabilità α **= 0.05,** riporta il valore di **0,4821**.

Verificare la corrispondenza con il valori critici

a) della distribuzione **F** e

b) della **t** di Student,

che possono essere rintracciati nelle tabelle relative.

Risposta.

a) Con **r = 0,4821** e **n = 17**

la verifica dell'ipotesi nulla

$$\mathbf{H_0}: \rho = \mathbf{0}$$

con ipotesi alternativa bilaterale

$$\mathbf{H_1}: \rho \neq \mathbf{0}$$

mediante il test **F**

$$F_{1,\, n-2} = \frac{r^2 \cdot (n-2)}{1-r^2}$$

fornisce un risultato

$$F_{1,15} = \frac{0{,}4821^2 \cdot 15}{1 - 0{,}4821^2} = \frac{3{,}4486}{0{,}768} = 4{,}539$$

uguale a 4,539.

b) Mediante il test **t di Student**

$$t_{(n-2)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

fornisce

$$t_{(15)} = \frac{0{,}4821 \cdot \sqrt{15}}{\sqrt{1 - 0{,}4821^2}} = \frac{0{,}4821 \cdot 3{,}873}{\sqrt{0{,}7676}} = \frac{1{,}867}{0{,}876} = 2{,}13$$

un risultato uguale a 2,13.

E' semplice verificare, sulle tabelle dei valori critici di **F** e di **t**, che i due risultati corrispondono esattamente ai valori riportati per la probabilità α **= 0.05** in una distribuzione bilaterale e che

$$\mathbf{2{,}13^2 = 4{,}539}$$

a meno delle approssimazioni dei calcoli.

Per **<u>un test di significatività del coefficiente di correlazione r rispetto ad un qualsiasi valore di ρ_0 diverso da zero</u>**, quindi per verificare l'ipotesi nulla

$$\mathbf{H_0:\ \rho = \rho_0}$$

a causa dei motivi prima illustrati il valore di **r** deve essere trasformato.

Tra le diverse **proposte di trasformazione**, è ancora molto diffusa l'utilizzazione di quella di R. A. **Fisher** presentata

- nel 1915 nel dibattito sui **grandi campioni** (vedi l'articolo ***Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population,*** pubblicata su **Biometrika**, 10: 507-521)
- e nel 1921 per i **piccoli campioni** (vedi l'articolo ***On the "probable error" of a coefficient of correlation deduced a small sample***, pubblicato su **Metron** 1: 3-32).

Il valore di **<u>r è trasformato in un valore z</u>** (zeta minuscolo)
mediante

$$z = 0.5 \cdot \ln\left(\frac{1+r}{1-r}\right)$$

Con **questa trasformazione**,

- i valori positivi di **r**, che ovviamente variano da **0** a **+1**, cadono tra **0** e **+∞**
- i valori negativi di **r**, che ovviamente variano da **0** a **-1**, cadono tra **0** e **-∞**

in modo simmetrico. In realtà, nella pratica sperimentale dove i valori di **r** asintoticamente vicini a 1 sono rari, la variazione cade in un intervallo minore di poche unità, in modo simmetrico intorno alla zero.

Ad esempio

- **r = +0,88**

$$z = 0{,}5 \cdot \ln\left(\frac{1+0{,}88}{1-0{,}88}\right) = 0{,}5 \cdot \ln\frac{1{,}88}{0{,}12} = 0{,}5 \cdot \ln 15{,}66 = 0{,}5 \cdot 2{,}75 = +1{,}375$$

diventa **z = 1,375**

- **r = +0,98**

$$z = 0{,}5 \cdot \ln\left(\frac{1+0{,}98}{1-0{,}02}\right) = 0{,}5 \cdot \ln\frac{1{,}98}{0{,}02} = 0{,}5 \cdot \ln 99 = 0{,}5 \cdot 4{,}595 = +2{,}2975$$

diventa **z = +2,2975**

mentre

- **r = -0,88**

$$z = 0{,}5 \cdot \ln\left(\frac{1+(-0{,}88)}{1-(-0{,}88)}\right) = 0{,}5 \cdot \ln\frac{0{,}12}{1{,}88} = 0{,}5 \cdot \ln 0{,}0638 = 0{,}5 \cdot (-2{,}75) = -1{,}375$$

diventa **z = -1,375**

- **r = -0,98**

$$z = 0{,}5 \cdot \ln\left(\frac{1+(-0{,}98)}{1-(-0{,}98)}\right) = 0{,}5 \cdot \ln\frac{0{,}02}{1{,}98} = 0{,}5 \cdot \ln 0{,}0101 = 0{,}5 \cdot (-4{,}595) = -2{,}2975$$

diventa **z = -2,2975**

Anche il valore teorico od atteso di confronto (ρ_0) è trasformato nello stesso modo e viene indicato con ζ (zeta minuscolo dell'alfabeto greco).

La verifica di una differenza significativa tra un generico valore campionario **r** e il valore atteso ρ_0, con ipotesi nulla

$$\mathbf{H_0: \rho = \rho_0}$$

ed ipotesi alternativa bilaterale oppure unilaterale, è quindi effettuata con la distribuzione normale **Z (maiuscola)**

$$Z = \frac{z - \zeta}{\sigma_z}$$

dove

- **Z** (maiuscola) è il valore che serve per stimare la probabilità α nella distribuzione normale,
- **z** (minuscola) è il valore di **r** trasformato,
- ζ (zeta greca, minuscola) è il valore di ρ_0 trasformato,
- σ_z è l'**errore standard** di questa differenza (poiché **r** e ρ_0 sono valori medi),

dato approssimativamente da

$$\sigma_z = \sqrt{\frac{1}{n-3}}$$

ESEMPIO 2. Sulla base di numerosi campionamenti, su una rivista scientifica si afferma che la correlazione tra la presenza quantitativa della specie A e della specie B è positiva e pari a 0,85. Da una rilevazione campionaria con 30 osservazioni, il valore di **r** è risultato uguale a **+0,71**.
C'è motivo di ritenere che in questo caso si abbia un valore correlazione significativamente diversa?

Risposta. Per verificare l'ipotesi nulla

H_0: ρ = +0,85

con ipotesi alternativa bilaterale

H_1: $\rho \neq$ +0,85

per applicare la formula

$$Z = \frac{z - \zeta}{\sigma_z}$$

- dapprima si deve trasformare in **z** il valore **r = +0,71**

$$z = 0{,}5 \cdot \ln\left(\frac{1+0{,}71}{1-0{,}71}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}71}{0{,}29}\right) = 0{,}5 \cdot \ln 5{,}8965 = 0{,}5 \cdot 1{,}7744 = +0{,}887$$

ottenendo **z = +0,887**

- successivamente si deve trasformare in ζ il valore ρ_0 **= +0,85**

$$\zeta = 0{,}5 \cdot \ln\left(\frac{1+0{,}85}{1-0{,}85}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}85}{0{,}15}\right) = 0{,}5 \cdot \ln 12{,}3333 = 0{,}5 \cdot 2{,}5123 = +1{,}256$$

ottenendo **z = +1,256**

- e, con **n = 30**, si calcola l'errore standard σ_z

$$\sigma_z = \sqrt{\frac{1}{30-3}} = \sqrt{0{,}037} = 0{,}192$$

Per la significatività della differenza tra valore osservato (**r = +0,71**) e valore arreso (ρ_0 **= +0,85**), si ottiene

$$Z = \frac{0{,}887 - 1{,}256}{0{,}192} = -\frac{0{,}369}{0{,}192} = -1{,}92$$

un valore **Z = -1,92**.

In una distribuzione normale bilaterale è associato ad una probabilità $\alpha = 0.055$; di conseguenza, il test non risulta significativo, ma per una differenza trascurabile. Con **n >30** molto facilmente risulterebbe significativa.

Se il **test** fosse stato **unilaterale**, cioè se vi fosse stato motivo di chiedersi se il valore calcolato fosse significativamente minore di quello stimato,

con ipotesi alternativa unilaterale fosse stata

$$\mathbf{H_0: \rho < \rho_0}$$

il procedimento di calcolo sarebbe stato identico. Sarebbe variata solo la lettura della probabilità α, che in una distribuzione unilaterale sarebbe risultata uguale a 0.027 e quindi avrebbe determinato un test significativo.

18.3. SIGNIFICATIVITA' DELLA RETTA CON R^2 ?

Alla fine del capitolo sono riportati alcuni output di programmi informatici sulla regressione lineare semplice. Insieme con le risposte sulla significatività dei parametri

- **a** (intercetta),
- **b** (coefficiente angolare),
- è riportato il valore di $\mathbf{R^2}$ (***R-square***).

Vari ricercatori, per valutare la significatività della retta di regressione utilizzano non il relativo test **t** o il test **F,** il cui valore è sempre riportato come illustrato nel capitolo precedente, ma semplicemente riportano il valore di **r** come

$$r = \sqrt{R^2}$$

stimandone la significatività.

Il risultato numerico è identico a quello effettuato sulla retta, poiché il valore di **F**,

- sia nel test per la retta con coefficiente angolare **b**,
- sia in quello per la correlazione **r**

è dato dal rapporto tra la devianza della regressione e la devianza d'errore,

$$F = \frac{Varianza\ \ della\ \ regressione}{Varianza\ \ d'errore}$$

seppure il concetto sovente sia nascosto nelle formule abbreviate, di solito utilizzate.

Ad esempio, con le misure di peso ed altezza rilevati su 7 giovani donne

Peso (Y) in Kg.	52	68	75	71	63	59	57
Altezza (X) in cm.	160	178	183	180	166	175	162

è stata calcolata la retta di regressione

$$\hat{Y} = -73{,}354 + 0{,}796\,X$$

La significatività del coefficiente angolare **b** per verificare l'ipotesi nulla

$$\mathbf{H_0:\ \beta = 0}$$

con ipotesi alternativa bilaterale

$$\mathbf{H_1:\ \beta \neq 0}$$

può essere derivata dalla tabella riassuntiva (vedi tabulati nell'ultimo paragrafo, diversi dai calcoli manuali riportati nel capitolo precedente, a causa delle le approssimazioni),

Fonti di variazione	**Devianza**	**Gdl**	**Varianza**
Totale	**403,715**	**6**	**---**
Della Regressione	**323,208**	**1**	**323,208**
Errore	**80,506**	**5**	**16,101**

che fornisce tutti gli elementi utili al calcolo di **F**, ottenendo un valore che

$$F_{(1,5)} = \frac{323,208}{16,101} = 20,07$$

risulta uguale a **20,07** con df **1** e **5**.

Utilizzando gli stessi dati (come il precedente fornito dal tabulato del computer nell'ultimo paragrafo), il valore di $\mathbf{R^2}$ (R-square) risulta uguale a **0,8006** e $\mathbf{R^2_{adj}}$ (Adj R-sq) uguale a **0,7607**.
La significatività del test **F** per verificare l'ipotesi nulla

$$\mathbf{H_0: \rho = 0}$$

con ipotesi alternativa

$$\mathbf{H_1: \rho \neq 0}$$

mediante la formula

$$F_{1,\ n-2} = \frac{r^2 \cdot (n-2)}{1-r^2}$$

fornisce un **F** con df **1** e **5**

$$F_{1,5} = \frac{0,8006 \cdot 5}{1-0,8006} = \frac{4,003}{0,1994} = \mathbf{20,07}$$

uguale a **20,07.**
E' identico al precedente.

Ma, nonostante il risultato identico, il due metodi sottendono scopi differenti e hanno condizioni di validità differenti; di conseguenza, **usare la significatività di r al posto di b è errato**.
Negli ultimi anni, il coefficiente di correlazione ha assunto un ruolo nettamente più limitato rispetto al passato, quando sovente era preferito alla regressione lineare semplice: la sua genericità, cioè il non richiedere specificatamente una relazione di causa-effetto, veniva interpretata come maggiore possibilità di adattamento alla varietà delle condizioni ambientali. Più recentemente, si preferisce la regressione, in quanto dovrebbe indurre il ricercatore a ragionare con attenzione maggiore sui rapporti tra le due variabili, alla ricerca della relazione di causa effetto e alla sua direzione.

I **fattori principali che attualmente limitano l'uso della correlazione rispetto alla regressione lineare,** per cui anche i test di significatività non sono intercambiabili, sono almeno 5:
1 - le differenze nelle condizioni di validità tra correlazione e regressione: nella prima devono essere realizzate in entrambe le variabili X_1 e X_2, mentre nella seconda solo per la variabile Y;
2 - il diverso significato di relazione tra le due variabili, che nella correlazione è solo di co-variazione lineare e non di causa - effetto**;**

3 - **la quantità di informazione contenute nelle analisi e nei test di significatività:** nella correlazione è più ridotto, rispetto all'informazione data da **a, b, r^2** della regressione;

4 - **la maggiore complessità della verifica di differenze da valori teorici che non siano nulli e dei confronti tra risultati differenti nella correlazione,** a causa della sua asimmetria nella distribuzione per valori distanti da zero**;**

5 - **l'assenza di significato ai fini predittivi** della correlazione.

Attualmente, la correlazione viene preferita alla regressione solo quando non si vuole dichiarare, in quanto priva di significato, una relazione di causa - effetto tra le due variabili considerate.

18.4. INTERVALLO DI CONFIDENZA DI ρ

Pure nel caso della correlazione, la stima dell'intervallo di confidenza di un parametro richiede che i campioni siano distribuiti in modo simmetrico, rispetto al valore vero o della popolazione. Ma, come già evidenziato nel paragrafo precedente, a differenza di quanto avviene per la media campionaria $\bar{x}$ rispetto a **μ** e per il coefficiente angolare **b** rispetto a **β,** i valori campionari **r**

- sono distribuiti normalmente solo quando
- il valore di **ρ** è piccolo (teoricamente zero)
- e i campioni sono abbastanza grandi (teoricamente infiniti).

Quando il valore di **ρ** si allontana da zero, la distribuzione dei valori **r** campionari è sempre asimmetrica. Di conseguenza, pure conoscendo il valore di **r**

e la sua **varianza**

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

oppure il suo **errore standard** s_r

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

non è corretto calcolare l'errore fiduciale di **ρ**

attraverso

$$\rho = r \pm t_{\alpha/2,n-2} \cdot s_r$$

applicando alla correlazione

quanto è possibile per la retta

$$\beta = b \pm t_{\alpha/2,n-2} \cdot s_b$$

Infatti, a causa dell'asimmetria dei valori campionari attorno alla media **quando $\rho \neq 0$**

- **i limiti di confidenza**, che utilizzano il valore di **t** e l'errore standard, **non sono adatti a verificare la significatività** di un qualsiasi valore di correlazione rispetto ad un valore teorico atteso;
- **non è possibile utilizzare il test t per il confronto tra due indici di regressione r_1 e r_2**.

L'uso di questo test inferenziale è limitato al solo caso in cui si vuole verificare l'assenza di correlazione, espressa dall'ipotesi nulla **H_0: $\rho = 0$**.

Per l'uso generale, con qualsiasi valore di **r**, dell'intervallo di confidenza, sono stati proposti vari metodi. Ne possono essere citati cinque:

a) **la trasformazione di r in z** proposta da **Fisher,** valida soprattutto per grandi campioni**,**

b) il precedente **metodo di Fisher**, ma **con l'uso della distribuzione t di Student al posto della distribuzione normale**, più cautelativa per piccoli campioni,

c) **la procedura proposta da** M. V. **Muddapur** nel 1988, che utilizza la distribuzione **F** (vedi: ***A simple test for correlation coefficient in a bivariate normal population***. **Sankyd: Indian J. Statist**. Ser. B. 50: 60-68),

d) **la procedura proposta da** S **Jeyaratnam** nel 1992, analoga alla precedente, ma con l'uso della distribuzione **t** (vedi: ***Confidence intervals for the correlation coefficient***. **Statist. Prob. Lett**. 15: 389-393).

e) metodi grafici, come quelli riportati già nel 1938 da F. N. **David** in ***Tables of the Correlation Coefficient*** (ed. E. S. Pearson, London Biometrika Office).

Il terzo e il quarto metodo, citati anche da J. H. **Zar** nel suo test del 1999 (***Biostatistical Analysis***, 4th ed., Prentice Hall, New Jersey, a pagg. 383-384), offrono il vantaggio di stimare **un intervallo generalmente minore** di quello di Fisher, oltre all'aspetto pratico di **non richiedere trasformazioni** di **r** e quindi di essere **più rapidi**.

Questi test, come qualsiasi intervallo fiduciale, possono essere utilizzati anche per la verifica dell'ipotesi sulla differenza tra due medie, in un test bilaterale.

A) Il **<u>metodo di Fisher</u>** stima il limite inferiore **L_1** e il limite superiore **L_2** dell'intervallo di confidenza attraverso le relazioni

$$L_1 = z - Z_{\alpha/2} \cdot \sigma_z$$

$$L_2 = z + Z_{\alpha/2} \cdot \sigma_z$$

dove

- z è il valore campionario di **r** trasformato, attraverso

la relazione

$$z = 0{,}5 \cdot \ln\left(\frac{1+r}{1-r}\right)$$

- $Z_{\alpha/2}$ è il valore della **Z** nella distribuzione normale alla probabilità **α/2,** prescelta per definire l'intervallo,
- σ_z è l'**errore standard** (approssimato)

di **r trasformato in z**

$$\sigma_z = \sqrt{\frac{1}{n-3}}$$

che dipende da **n**.

Successivamente, i due valori stimati $\mathbf{L_1}$ e $\mathbf{L_2}$, calcolati ovviamente in una scala **z**, devono essere riportati sulla scala di **r**,

con la trasformazione

$$r = \frac{e^{2z}-1}{e^{2z}+1}$$

In questo ultimo passaggio, si perde la simmetria di $\mathbf{L_1}$ e $\mathbf{L_2}$ rispetto al valore centrale, quando **r≠0**.
L'asimmetria intorno ad **r** risulta tanto più marcata quanto più esso si avvicina a +1 oppure a -1.

B) Più recentemente, in vari testi è proposta la misura più cautelativa, che fornisce un intervallo maggiore in rapporto alla dimensione campionaria **n,** con la distribuzione **t di Student**.
Il limite inferiore $\mathbf{L_1}$ e il limite superiore $\mathbf{L_2}$ dell'intervallo di confidenza sono calcolati attraverso le relazioni

$$\mathbf{L_1} = z - t_{\alpha/2,\nu} \cdot \sigma_z$$

$$\mathbf{L_2} = z + t_{\alpha/2,\nu} \cdot \sigma_z$$

dove

- $t_{\alpha/2,\nu}$ è il valore alla probabilità **α/2** prescelta per definire l'intervallo, con **ν = n –2**,

mentre tutti gli altri parametri restano identici a quelli appena presentati.
Nulla cambia rispetto al metodo classico di Fisher, per quanto riguarda

- dapprima la trasformazione di **r** in **z**
- successivamente le trasformazioni dei valori $\mathbf{L_1}$ e $\mathbf{L_2}$ in scala **r**.

C) Il metodo proposto da **Muddapur,** meno noto del metodo classico di Fisher, più recente e più raro in letteratura, ma più rapido in quanto **non richiede alcuna trasformazione**,

con $\mathbf{L_1}$

$$\mathbf{L_1} = \frac{\left[\left(1+F\right)\cdot r\right]+\left(1-F\right)}{\left(1+F\right)+\left[\left(1-F\right)\cdot r\right]}$$

e $\mathbf{L_2}$

$$\mathbf{L_1} = \frac{\left[\left(1+F\right)\cdot r\right]-\left(1-F\right)}{\left(1+F\right)-\left[\left(1-F\right)\cdot r\right]}$$

dove

- **F** è il valore corrispondente nella distribuzione **F** di Fisher alla probabilità α per un test bilaterale e con df $\nu_1 = \nu_2 = \mathbf{n-2}$

fornisce una stima uguale a quella classica di Fisher.

D) Il metodo proposto da **Jeyaratnam,** che può essere letto come una variante del precedente, una sua formula abbreviata in quanto ancor più rapido

con $\mathbf{L_1}$

$$\mathbf{L_1} = \frac{r - \sqrt{\dfrac{t^2_{\alpha,\nu}}{t^2_{\alpha,\nu}+\nu}}}{1 - r\sqrt{\dfrac{t^2_{\alpha,\nu}}{t^2_{\alpha,\nu}+\nu}}}$$

e $\mathbf{L_2}$

$$\mathbf{L_2} = \frac{r + \sqrt{\dfrac{t^2_{\alpha,\nu}}{t^2_{\alpha,\nu}+\nu}}}{1 + r\sqrt{\dfrac{t^2_{\alpha,\nu}}{t^2_{\alpha,\nu}+\nu}}}$$

dove

- **t** è il valore corrispondente nella distribuzione **t di Student** alla probabilità α per un test bilaterale (come in tutti gli intervalli fiduciali) e con df $\nu = \mathbf{n-2}$
- $\nu = \mathbf{n-2.}$

ESEMPIO. Con 30 coppie di dati, è stato calcolato il coefficiente di correlazione lineare semplice **r = 0,71**.

Entro quale intervallo si colloca il valore reale o della popolazione (ρ), alla probabilità α **= 0.05?**

Calcolare i valori estremi $\mathbf{L_1}$ e $\mathbf{L_2}$ dell'intervallo fiduciale con i 4 diversi metodi.

Risposta

A) Con il metodo classico di **Fisher**

1 – dopo aver trasformato **r = 0,71** in **z**

con

$$z = 0{,}5 \cdot \ln\left(\frac{1+0{,}71}{1-0{,}71}\right) = 0{,}5 \cdot \ln 5{,}8965 = 0{,}5 \cdot 1{,}7744 = \mathbf{0{,}8872}$$

ottenendo z = 0,8872

2 – si stima il suo errore standard σ_z, ovviamente su scala **z**, con la formula approssimata

$$\sigma_z = \sqrt{\frac{1}{30-3}} = \sqrt{0{,}037} = \mathbf{0{,}1924}$$

ottenendo σ_z **= 0,1924**

3 – Successivamente, con $\mathbf{Z_{\alpha/2}}$ **= 1,96** (valore della distribuzione normale standardizzata alla probabilità α **= 0.05** bilaterale) si stimano i due limiti dell'intervallo $\mathbf{L_1}$ e $\mathbf{L_2}$:

con

$$L_1 = 0{,}8872 - 1{,}96 \cdot 0{,}1924 = 0{,}8872 - 0{,}3771 = \mathbf{0{,}5101}$$

si ottiene $\mathbf{L_1}$ **= 0,5101**

e con

$$L_2 = 0{,}8872 + 1{,}96 \cdot 0{,}1924 = 0{,}8872 + 0{,}3771 = \mathbf{1{,}2643}$$

si ottiene $\mathbf{L_1}$ **= 1,2643**

4 – Per il confronto di $\mathbf{L_1}$ e $\mathbf{L_2}$ con **r = 0,71**, è necessario ritrasformare i due valori **z** ottenuti nei corrispondenti valori in scala **r.** Ricordando che **e = 2,718**

per **z = 0,5101** con

$$r = \frac{2{,}718^{2 \cdot 0{,}5101} - 1}{2{,}718^{2 \cdot 0{,}5101} + 1} = \frac{2{,}718^{1{,}0202} - 1}{2{,}718^{1{,}0202} + 1} = \frac{2{,}7735 - 1}{2{,}7735 + 1} = \frac{1{,}7735}{3{,}7735} = \mathbf{0{,}470}$$

si ottiene $\mathbf{L_1}$ **= 0,470**

e per **z = 1,2643** con

$$r = \frac{2{,}718^{2 \cdot 1{,}2643} - 1}{2{,}718^{2 \cdot 1{,}2643} + 1} = \frac{2{,}718^{2{,}5286} - 1}{2{,}718^{2{,}5286} + 1} = \frac{12{,}5327 - 1}{12{,}5327 + 1} = \frac{11{,}5327}{13{,}5327} = \mathbf{0{,}852}$$

si ottiene $\mathbf{L_2 = 0{,}852}$

Con il metodo classico di Fisher, l'intervallo di confidenza di **r = 0,71** calcolato su **30** copie di dati è compreso tra i limiti **0,470** e **0,852** con probabilità $\alpha = \mathbf{0.05}$.

B) Utilizzando sempre il metodo di **Fisher**, ma con la **distribuzione t** al posto della distribuzione **z**,

1 – i primi due passaggi sono identici ai punti 1 e 2 precedenti, nei quali si era ottenuto

$$\mathbf{z = 0{,}8872} \qquad \text{e} \qquad \mathbf{\sigma_z = 0{,}1925}$$

2 – Dopo aver scelto il valore di **t** che, con $\alpha = \mathbf{0.05}$ **bilaterale** e df $\nu = \mathbf{28}$ **è**

$$\mathbf{t_{0.025,\ 28} = 2{,}048}$$

3 – si ottiene (sempre in scala **z**)

$$L_1 = 0{,}8872 - 2{,}048 \cdot 0{,}1925 = 0{,}8872 - 0{,}3942 = \mathbf{0{,}493}$$

un valore di $\mathbf{L_1 = 0{,}493}$

e

$$L_2 = 0{,}8872 + 2{,}048 \cdot 0{,}1925 = 0{,}8872 + 0{,}3942 = \mathbf{1{,}2814}$$

un valore di $\mathbf{L_2 = 1{,}2814}$

4 – Infine, come nella fase 4 precedente, si riportano questi due valori **z** in scala **r**:

per **z = 0,493** con

$$r = \frac{2{,}718^{2 \cdot 0{,}493} - 1}{2{,}718^{2 \cdot 0{,}493} + 1} = \frac{2{,}718^{0{,}986} - 1}{2{,}718^{0{,}986} + 1} = \frac{2{,}6802 - 1}{2{,}6802 + 1} = \frac{1{,}6802}{3{,}6802} = \mathbf{0{,}457}$$

si ottiene $\mathbf{L_1 = 0{,}457}$

e per **z = 1,2814** con

$$r = \frac{2{,}718^{2 \cdot 1{,}2814} - 1}{2{,}718^{2 \cdot 1{,}2814} + 1} = \frac{2{,}718^{2{,}5628} - 1}{2{,}718^{2{,}5628} + 1} = \frac{12{,}9686 - 1}{12{,}9686 + 1} = \frac{11{,}9686}{13{,}9686} = \mathbf{0{,}857}$$

si ottiene $\mathbf{L_2 = 0{,}857.}$

Con il metodo di Fisher nel quale sia utilizzata la distribuzione **t** con df **n-2**, l'intervallo di confidenza di **r = 0,71** calcolato su **30** copie di dati è compreso tra i limiti **0,457** e **0,857** con probabilità α **= 0.05**.

C) Con il metodo proposto da **Muddapur**,

1 - dapprima si trova il valore di **F** alla probabilità α **= 0.05** bilaterale con df $\nu_1 = \nu_2 = 28$; rilevato in una tabella molto più dettagliata di quella riportata nelle dispense esso risulta uguale a **2,13**;

2 – successivamente si calcolano

$\mathbf{L_1}$ con

$$L_1 = \frac{(1+2{,}13)\cdot 0{,}71+(1-2{,}13)}{(1+2{,}13)+(1-2{,}13)\cdot 0{,}71} = \frac{2{,}2223-1{,}13}{3{,}13-0{,}8023} = \frac{1{,}0923}{2{,}3277} = \mathbf{0{,}469}$$

ottenendo $\mathbf{L_1 = 0{,}469}$

e $\mathbf{L_2}$ con

$$L_1 = \frac{(1+2{,}13)\cdot 0{,}71-(1-2{,}13)}{(1+2{,}13)-(1-2{,}13)\cdot 0{,}71} = \frac{2{,}2223+1{,}13}{3{,}13+0{,}8023} = \frac{3{,}3523}{2{,}9323} = \mathbf{0{,}853}$$

ottenendo $\mathbf{L_1 = 0{,}853}$

D) Con il metodo proposto da **Jeyaratnam,**

1 - dapprima si trova il valore di **t** alla probabilità α **= 0.05** bilaterale con df ν **= 28**; esso risulta uguale a **2,13**;

2 – successivamente si calcolano

$\mathbf{L_1}$ con

$$L_1 = \frac{0{,}71-\sqrt{\dfrac{2{,}048^2}{2{,}048^2+28}}}{1-0{,}71\cdot\sqrt{\dfrac{2{,}048^2}{2{,}048^2+28}}} = \frac{0{,}71-\sqrt{\dfrac{4{,}1943}{32{,}1943}}}{1-0{,}71\cdot\sqrt{\dfrac{4{,}1943}{32{,}1943}}} =$$

$$L_1 = \frac{0{,}71-\sqrt{0{,}1303}}{1-0{,}71\cdot\sqrt{0{,}1303}} = \frac{0{,}71-0{,}361}{1-0{,}71\cdot 0{,}361} = \frac{0{,}349}{0{,}744} = 0{,}469$$

ottenendo $\mathbf{L_1 = 0{,}469}$

e $\mathbf{L_2}$ con

$$L_2 = \frac{0{,}71 + \sqrt{\frac{2{,}048^2}{2{,}048^2 + 28}}}{1 + 0{,}71 \cdot \sqrt{\frac{2{,}048^2}{2{,}048^2 + 28}}} = \frac{0{,}71 + \sqrt{\frac{4{,}1943}{32{,}1943}}}{1 + 0{,}71 \cdot \sqrt{\frac{4{,}1943}{32{,}1943}}} =$$

$$L_2 = \frac{0{,}71 + \sqrt{0{,}1303}}{1 + 0{,}71 \cdot \sqrt{0{,}1303}} = \frac{0{,}71 + 0{,}361}{1 + 0{,}71 \cdot 0{,}361} = \frac{1{,}071}{1{,}256} = 0{,}853$$

ottenendo $\mathbf{L_2 = 0{,}853}$

Questo calcolo diventa molto più rapido se dapprima, separatamente, si stima la parte sotto radice,

$$\sqrt{\frac{2{,}048^2}{2{,}048^2 + 28}} = \sqrt{\frac{4{,}1943}{32{,}1943}} = \sqrt{0{,}1303} = 0{,}361$$

che risulta uguale a **0,361**:

$$L_1 = \frac{0{,}71 - 0{,}361}{1 - 0{,}71 \cdot 0{,}361} = \frac{0{,}349}{0{,}744} = 0{,}469$$

$$L_2 = \frac{0{,}71 + 0{,}361}{1 + 0{,}71 \cdot 0{,}361} = \frac{1{,}071}{1{,}256} = 0{,}853$$

I risultati dei 4 metodi, con i dati dell'esempio, sono riportati nella tabella sottostante:

METODO	L_1	L_2
Fisher	0,470	0,852
Fisher, con distribuzione t	0,457	0,857
Muddapur	0,469	0,853
Jeyaratnam	0,469	0,853
r = 0,71 n = 30 α = 0.05		

E' sufficiente il semplice confronto, per verificare la loro corrispondenza. I calcoli sono stati fatti alla quarta cifra decimale per evitare arrotondamenti e meglio porre a confronto i risultati.

I motivi della trasformazione e suoi effetti sono illustrati da Fisher. Sempre in "***Metodi statistici ad uso dei ricercatori***, Torino 1948, Unione Tipografica Editrice Torinese (UTET), 326 p. traduzione di M Giorda, del testo **Statistical Methods for Research Workers** di R. A. **Fisher** 1945, nona ed., a pag. 184 e seguenti sono spiegati i motivi e gli effetti della trasformazione di **r** in **z**: "*Per piccoli valori di **r**, z è quasi uguale a **r**, ma quando r si avvicina all'unità , z cresce senza limiti. Per valori negativi di **r**, z è negativo. Il vantaggio di questa trasformazione di **r** in **z** sta nella distribuzione di coteste due quantità in campioni scelti a caso.*

Lo scostamento (errore standard*) di **r** dipende dal valore effettivo della correlazione ρ, come è rilevato dalla formula*

$$\sigma_r = \frac{1-\rho^2}{\sqrt{n-1}}$$

(dove n è il numero di coppie di dati)

*Poiché **ρ** è un'incognita, dobbiamo sostituirla con il valore osservato **r**, il quale, in piccoli campioni, non sarà però una stima molto accurata di **ρ**. L'errore tipo (***errore standard***) di **z** è di forma più semplice e, cioè, approssimativamente,*

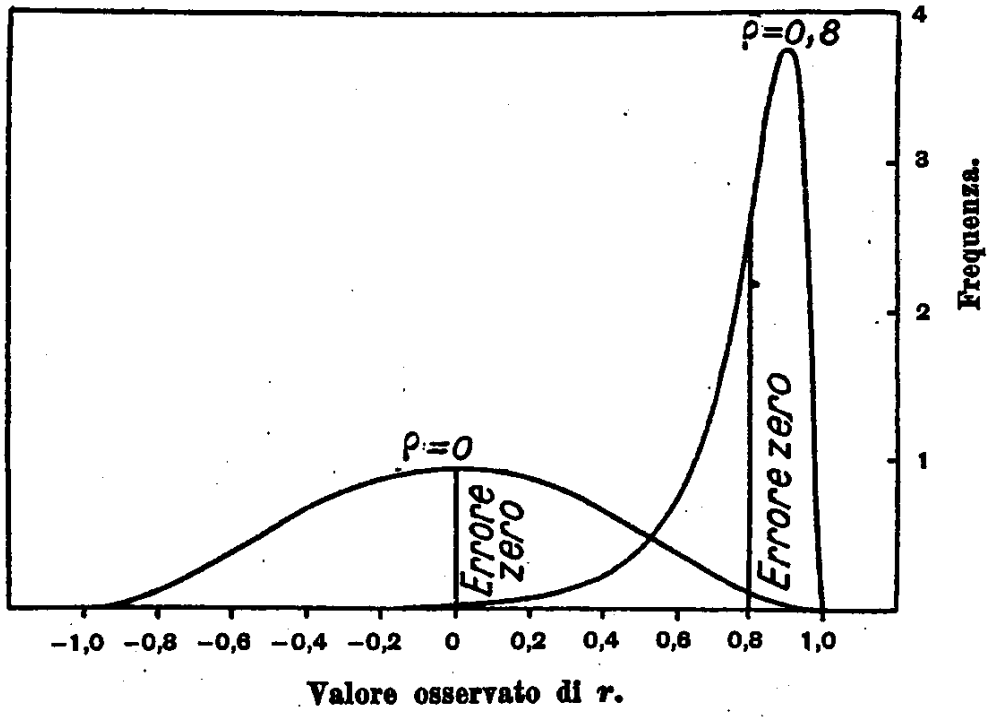
$$\sigma_z = \frac{1}{\sqrt{n-3}}$$

ed è praticamente indipendente dal valore della correlazione nella popolazione dalla quale si è tratto il campione.

*In secondo luogo, la distribuzione di **r** non è normale in piccoli campioni e, per correlazioni elevate, essa rimane lontana dalla normale anche nei grandi campioni. La distribuzione di **z** non è strettamente normale, ma tende rapidamente alla normalità quando il campione è accresciuto, qualunque possa essere il valore della correlazione.*

*Infine la distribuzione di **r** cambia rapidamente forma quando cambia **ρ**;conseguentemente non si può, con ragionevole speranza di successo, tentare di giustificare* (leggi: aggiustare) *l'asimmetria della distribuzione. Al contrario, la distribuzione di z essendo quasi costante nella forma, l'accuratezza delle prove* (leggi: la significatività del test) *può essere migliorata per mezzo di piccole correzioni dello scostamento dalla normalità. Tali correzioni sono, però, troppo piccole per assumere importanza pratica e noi non ce ne cureremo. La semplice assunzione che **z** è normalmente distribuita sarà in tutti i casi sufficientemente accurata.*

Questi tre vantaggi della trasformazione di ***r*** *in z possono notarsi comparando le* prossime due *figure.*



Nella prima *sono indicate le distribuzioni effettive di* ***r****, per 8 coppie di osservazioni, tratte da popolazioni aventi correlazioni 0 e 0,8:*

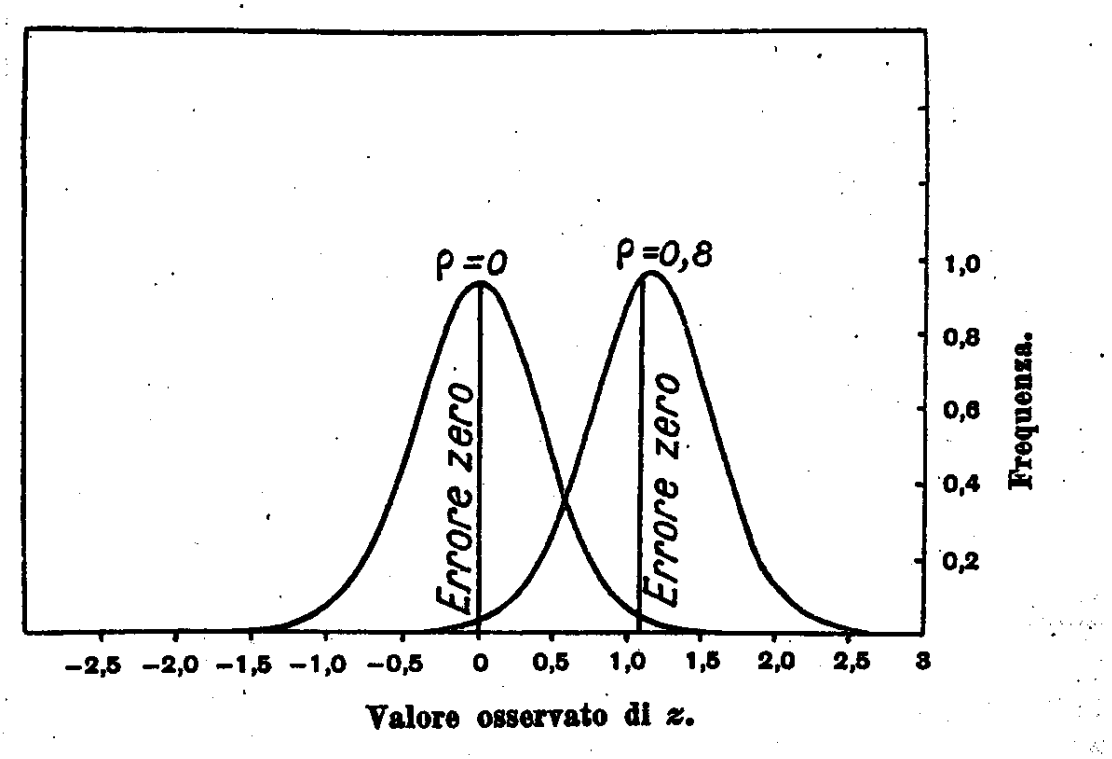
La seconda *presenta le corrispondenti curve di distribuzione di z.*

Le due curve della prima figura *sono molto differenti nelle loro altezze modali; entrambe sono nettamente curve non normali; anche nella forma esse sono in forte divario, una essendo simmetrica, l'altra molto asimmetrica.*

Al contrario, nella seconda figura *le due curve non differiscono sensibilmente in altezza; quantunque non esattamente normali nella forma, esse, anche per un piccolo campione di 8 coppie di osservazioni, vi si accostano talmente che l'occhio non può scoprire la differenza; questa normalità approssimativa, infine, eleva agli estremi limiti* $\rho = \pm 1$.

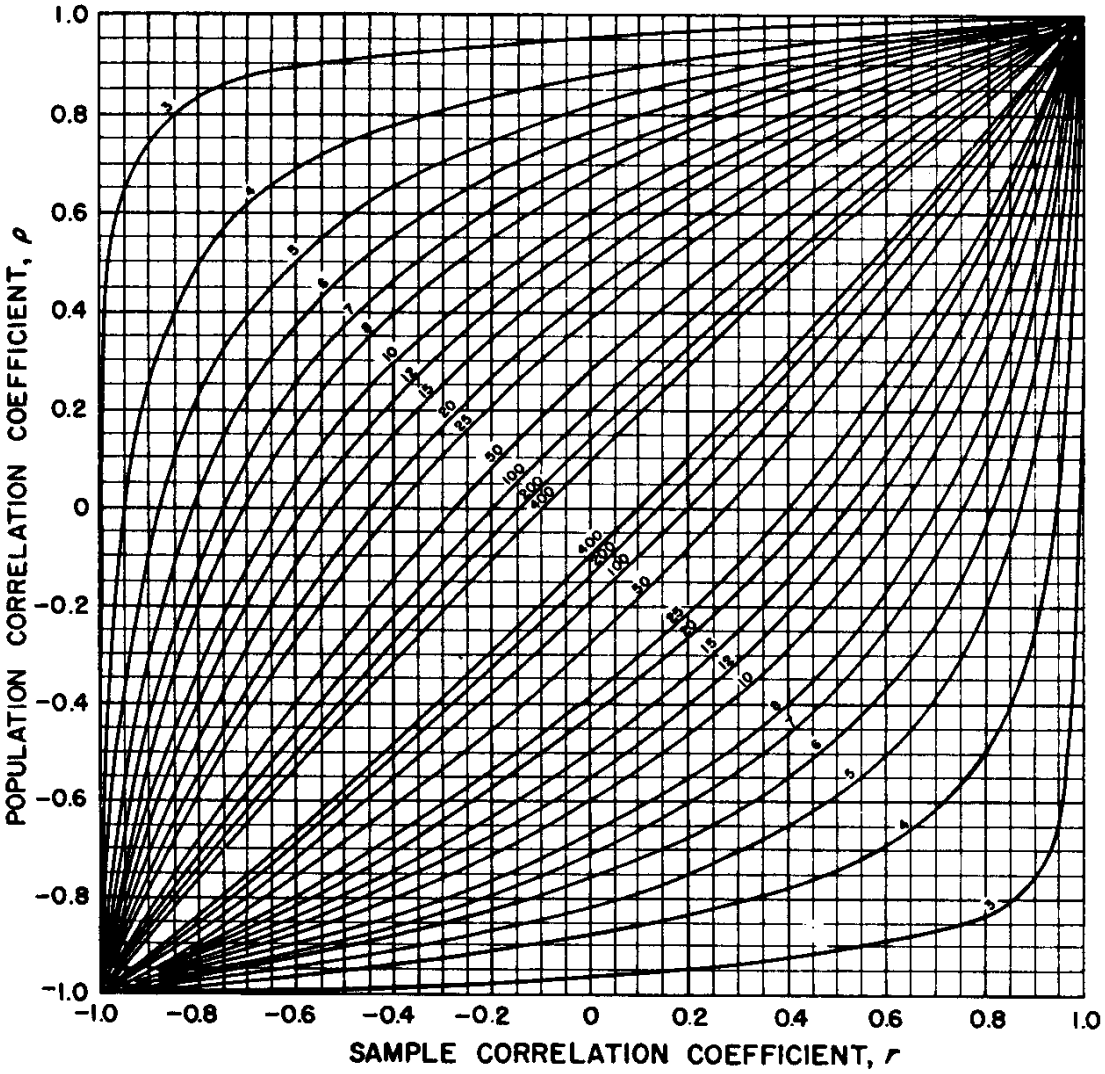
Una modalità addizionale è messa in evidenza dalla seconda figura *nella distribuzione per* $\rho = 0{,}8$. *Quantunque la curva stessa, a giudicare dall'apparenza, sia simmetrica, pure l'ordinata dell'errore*

zero non è posta al centro. Questa *figura, infatti, presenta il piccolo difetto che viene introdotto nella stima del coefficiente di correlazione quale è ordinariamente calcolato.*



Tra i metodi grafici possono essere ricordati quelli riportati già nel 1938 da F. N. **David** in ***Tables of the Correlation Coefficient*** (ed. E. S. Pearson, London Biometrika Office), almeno per l'importanza della rivista (**Biometrika**). Essi sono utilizzati anche nel manuale pubblicato dal Dipartimento di Ricerca della Marina Militare Americana nel 1960 (***Statistical Manual*** by Edwin L. **Crow**, Frances A. **Davis**, Margaret W. **Maxfield**, Research Department U. S: Naval Ordnance Test Station, Dover Publications, Inc., New York, XVII + 288 p.).
Come dimostrazione è qui riportato solamente quello per la probabilità $\alpha = 0.05$. Per le altre si rinvia alle indicazioni bibliografiche riportate.

Come in tutti i grafici l'**uso è semplice**, quasi intuitivo. I limiti consistono soprattutto nell'**approssimazione dei valori forniti**, determinati dalla lettura su curve molto vicine.



α = 0.05

Curve degli intervalli di confidenza

per il coefficiente di correlazione r

Ad esempio:

- letto, **sull'asse delle ascisse**, il valore di correlazione **r = +0,4** calcolato su un campione di **n = 6** dati,
- si sale verticalmente incontrando la curva con il numero 6 due volte:
- la prima in un punto che, proiettato sull'**asse delle ordinate**, indica ρ = -0,55
- la seconda in un punto che, proiettato sull'**asse delle ordinate**, indica ρ = +0,82.

Sono i due limiti dell'intervallo di confidenza, alla probabilità **α = 0.05**, per **r = 0,4** ottenuto in un **campione** di **n = 6** coppie di osservazioni.

E' importante osservare che essi sono **fortemente asimmetrici**, **intorno al valore campionario** calcolato.

La **stessa asimmetria** esiste se fossimo partiti dal **valore vero ρ o della popolazione**, per stimare la dispersione dei **valori campionari r,** sempre calcolati su gruppi di 6 coppie di dati e alla medesima probabilità di commettere un errore di Tipo I ($\alpha = 0.05$).

18.5. POTENZA A PRIORI E A POSTERIORI PER LA SIGNIFICATIVITA' DI r

Le due ipotesi nulle

- **H_0: $\rho = 0$** che indica che il valore reale di correlazione del campione è uguale a 0
- **H_0: $\rho = \rho_0$** che indica che il valore reale di correlazione del campione è uguale a un valore ρ qualsiasi, di norma diverso da 0

sono rifiutate correttamente quando nella realtà

$$\rho \neq 0 \quad \text{oppure} \quad \rho \neq \rho_0$$

il valore vero di ρ è diverso da 0 (zero).

Ma, in tali condizioni, la **distribuzione è asimmetrica**, come visto anche nella figura con l'ultimo esempio.

Pertanto, per il calcolo della potenza **1-β** in un test di significatività del coefficiente di correlazione **r,** è necessario ricorrere alla **trasformazione di r in z di Fisher**.

I **metodi di stima di 1-β presentano alcune differenze** se

A) l'ipotesi nulla è **H_0: $\rho = 0$**

B) l'ipotesi nulla è **H_0: $\rho = \rho_0$**

A) **Nel primo caso,** nella verifica dell'ipotesi **H_0: $\rho = 0$,** la stima del**la potenza a posteriori 1-β** è ottenuta dalla relazione:

$$Z_\beta = \left(z_{\alpha,\nu} - z\right) \cdot \sqrt{n-3}$$

dove

- Z_β è il valore che, nella **distribuzione normale unilaterale,** permette di ricavare direttamente (senza trasformazione) la probabilità β;
- $z_{\alpha,\nu}$ è il **valore critico di r**, da prendere in una delle due tabelle dei valori critici di **r**, per la probabilità α prefissata in una distribuzione **bilaterale o unilaterale** e con df **ν = n – 2**, **trasformato** in **z** con la formula di Fisher;
- z è il valore di **r sperimentale,** trasformato in **z** con la formula di Fisher;
- n è il numero di coppie di dati, con i quali è stato calcolato **r,**

- ricordando che sia il valore critico di **r** sia il valore sperimentale di **r** devono essere trasformati con la solita formula

$$z = 0{,}5 \cdot \ln\left(\frac{1+r}{1-r}\right)$$

Sempre per rifiutare l'ipotesi nulla **H_0: $\rho = 0$** quando il valore atteso è **$\rho_0 \neq 0$,**

la potenza a priori o numero minimo n di dati perché il valore campionario **r** sia significativo è

$$n = \left(\frac{Z_\beta + Z_\alpha}{\zeta_0}\right)^2 + 3$$

dove

- Z_β e Z_α sono i valori di **z** nella distribuzione normale per le probabilità α e β (senza trasformazione di Fisher),
- ζ_0 (zeta greca minuscola) è il valore ρ_0 atteso, trasformato in **z** secondo Fisher.

ESEMPIO 1. (PER L'IPOTESI **H_0: $\rho = 0$)** Con **20** dati è stato calcolato un valore di **r = 0,35** che non è risultato significativamente diverso da zero, alla probabilità **α = 0.05**. Infatti, nelle due tabelle relative è semplice osservare che il valore critico alla stessa probabilità **α = 0.05** in un test bilaterale con **df = 18** è **0,4438**. Si chiede:

a) quale è la potenza di questo test?

b) quante coppie di dati(**n**) occorre raccogliere, per rifiutare l'ipotesi nulla **H_0: $\rho = 0$** nel 90% dei casi, al livello di significatività del 5%?

Risposte.

Per il primo caso,

a) Applicando la formula

$$Z_\beta = \left(z - z_{\alpha,\nu}\right) \cdot \sqrt{n-3}$$

dove

- Z_β è il valore nella **distribuzione normale unilaterale** che permette di ricavare direttamente la probabilità **β;**
- z è il valore sperimentale di **r** (uguale a **0,35**), che trasformato in **z** con il metodo di Fisher

$$z = 0{,}5 \cdot \ln\left(\frac{1+0{,}35}{1-0{,}35}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}35}{0{,}65}\right) = 0{,}5 \cdot \ln 2{,}077 = 0{,}5 \cdot 0{,}73 = 0{,}365$$

risulta uguale a **0,365**

- $z_{\alpha,\nu}$ è il **valore critico** di **r**, alla probabilità **α = 0.05 bilaterale** con **ν = n - 2 = 18** (nella tabella relativa è **0,4438**); trasformato in **z** con il metodo di Fisher

$$z_{0.05,18} = 0{,}5 \cdot \ln\left(\frac{1+0{,}4438}{1-0{,}4438}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}4438}{0{,}5562}\right) = 0{,}5 \cdot \ln 2{,}593 = 0{,}5 \cdot 0{,}953 = 0{,}477$$

risulta uguale a **0,477.**

Con i dati dell'esempio,

$$Z_\beta = (0{,}365 - 0{,}477) \cdot \sqrt{20-3} = -0{,}112 \cdot 4{,}123 = -0{,}46$$

Z_β risulta uguale a **0,46**.

Nella distribuzione normale **unilaterale**, ad essa corrisponde una probabilità **β = 0,323**. Il segno indica solo la coda della distribuzione.
Di conseguenza, la **potenza 1-β** del test è 1-0,323 = **0,677** o **67,7%**

b) Per stimare **quanti dati n sono necessari** per un test con **β = 0.10** e **α = 0.05 bilaterale,** affinché un valore atteso di **ρ = 0,35** risulti significativamente **diverso da zero,**
usando la formula

$$n = \left(\frac{Z_\beta + Z_\alpha}{\zeta_0}\right)^2 + 3$$

dove con

- Z_β **= 1,28** (in una distribuzione normale **unilaterale** per una probabilità uguale a **0.10**),
- Z_α = **1,96** (in una distribuzione normale **bilaterale** per una probabilità uguale a **0.05**),
- ζ_0 = **0,365** ottenuto dalla trasformazione di **ρ = 0,35**

mediante

$$z = 0{,}5 \cdot \ln\left(\frac{1+0{,}35}{1-0{,}35}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}35}{0{,}65}\right) = 0{,}5 \cdot \ln 2{,}077 = 0{,}5 \cdot 0{,}73 = 0{,}365$$

si ottiene

$$n = \left(\frac{1{,}28 + 1{,}96}{0{,}365} \right)^2 + 3 = 8{,}877^2 + 3 = 78{,}8 + 3 = 81{,}8$$

un valore di **n = 81,8**. E' necessario rilevare almeno **82** coppie di dati.

Nel secondo caso, quando l'ipotesi nulla è **H_0: $\rho = \rho_0$** dove **$\rho_0 \neq 0$,**

la formula per calcolare **la potenza a posteriori 1-β** è leggermente più complessa di quella precedente, diventando

$$Z_\beta = \left(|z - \zeta_0| - z_{\alpha,\nu} \right) \cdot \sqrt{n-3}$$

dove, mantenendo uguali gli altri parametri,

- in $|z - \zeta_0|$
- **z** è il valore campionario di **r** trasformato in **z** con la formula di Fisher
- ζ_0 (zeta greca minuscola) è la stessa trasformazione del valore **ρ_0** atteso o ipotizzato,

ricordando che i tre valori entro parentesi, cioè (1) il valore sperimentale di **r**, (2) il valore atteso o teorico di confronto **ρ_0**, (3) il valore critico di **r** con **df n-2** devono essere trasformati con la formula di Fisher.

ESEMPIO 2 (PER L'IPOTESI **H_0: $\rho = \rho_0$** DOVE **$\rho_0 \neq 0$)** La correlazione tra le variabili **X_1** e **X_2** è stata valutata in **$\rho = 0{,}15$**. Con **100** dati, un ricercatore ha calcolato **r = 0,35** e ha motivo di credere che, nella nuova condizione sperimentale, la correlazione sia significativamente maggiore.
Calcolare la potenza del test, per una significatività **$\alpha = 0.05$**.

Risposta

1) Con **r = 0,35** il valore di **z**

$$z = 0{,}5 \cdot \ln\left(\frac{1 + 0{,}35}{1 - 0{,}35} \right) = 0{,}5 \cdot \ln\left(\frac{1{,}35}{0{,}65} \right) = 0{,}5 \cdot \ln 2{,}077 = 0{,}5 \cdot = 0{,}73 = 0{,}365$$

è uguale a **0,365**.

2) Con **ρ_0 = 0,15** il valore di ζ_0

$$z = 0{,}5 \cdot \ln\left(\frac{1+0{,}15}{1-0{,}15}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}15}{0{,}85}\right) = 0{,}5 \cdot \ln 1{,}353 = 0{,}5\cdot = 0{,}302 = 0{,}151$$

è uguale a **0,151,**

3) Il valore critico di **r** alla probabilità α **= 0.5** in una distribuzione unilaterale con **df = 98** non è riportato nelle 2 tabelle specifiche. Per interpolazione tra **0,168 con df 95** e **0,164 con df 100** può essere stimato uguale a **0,165.** Da esso si ricava

il valore di $z_{0.05,98}$

$$z_{0.05,98} = 0{,}5 \cdot \ln\left(\frac{1+0{,}165}{1-0{,}165}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}165}{0{,}835}\right) = 0{,}5 \cdot \ln 1{,}395 = 0{,}5 \cdot 0{,}333 = 0{,}166$$

che risulta uguale a **0,166**.

4) Da questi tre valori e con **n = 100**, si ottiene

$$Z_\beta = \left(\left|0{,}365 - 0{,}151\right| - 0{,}166\right) \cdot \sqrt{100-3} = \left(0{,}214 - 0{,}166\right) \cdot \sqrt{97} = 0{,}048 \cdot 9{,}849 = 0{,}47$$

un valore di Z_β **= 0,47**

In una distribuzione normale unilaterale, a **z = 0,47** corrisponde una probabilità di **0,316**.

Con questi dati, il valore di β è uguale a **0,316** e la potenza **1-**β del test richiesto è 1-0,316 = **0,684**.

18.6. DIFFERENZA TRA DUE COEFFICIENTI DI CORRELAZIONE, IN CAMPIONI INDIPENDENTI; CALCOLO DEL COEFFICIENTE COMUNE

Il confronto tra due coefficienti di correlazione $\mathbf{r_1}$ e $\mathbf{r_2}$, calcolati su **<u>due campioni indipendenti</u>**, per verificare

l'ipotesi nulla

$$\mathbf{H_0: \rho_1 = \rho_2}$$

con ipotesi alternativa sia bilaterale

$$\mathbf{H_1: \rho_1 \neq \rho_2}$$

che unilaterale

$$\mathbf{H_1: \rho_1 < \rho_2} \quad \text{oppure} \quad \mathbf{H_1: \rho_1 > \rho_2}$$

pone sempre il problema della forte asimmetria dei valori campionari alla quale si aggiunge quella della non omogeneità delle due varianze.

La **simmetria** e la **omoschedasticità** sono ricostruite mediante la trasformazione di Fisher, per cui il test diventa

$$Z = \frac{z_1 - z_2}{\sigma_{zi-z2}}$$

dove

- Z è il valore della distribuzione normale, unilaterale o bilaterale in rapporto all'ipotesi alternativa,
- z_1 e z_2 sono $\mathbf{r_1}$ e $\mathbf{r_2}$ trasformati con la formula di Fisher,
- σ_{z1-z2} è **l'errore standard della differenza** precedente, ottenuta

con

$$\sigma_{z1-z2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

come **formula generale** e

con la **formula semplificata**

$$\sigma_{z1-z2} = \sqrt{\frac{2}{n_1 - 3}}$$

quando i due campioni sono bilanciati($\mathbf{n_1 = n_2}$)

Con poche decine di dati, all'uso della distribuzione normale alcuni preferiscono l'uso della **tabella t** con **df = N-4**, in quanto più cautelativa. Permane il problema che solitamente i testi per la distribuzione t riportano solo le tavole sinottiche; quindi risulta impossibile stimare la probabilità in modo più preciso.

Nel caso di **due campioni dipendenti**, come possono essere i coefficienti di correlazione del padre e quello della madre con una caratteristica di un figlio, calcolate su varie coppie di genitori, la procedura è differente.

Se i due coefficienti di correlazione $\mathbf{r_1}$ e $\mathbf{r_2}$ non risultano significativamente differenti, anche sulla base della potenza del test e delle dimensioni del campione si può concludere che $\rho_1 = \rho_2$. In tali condizioni, spesso è utile calcolare un **coefficiente di correlazione comune o pesato** (***common or weighted correlation coefficient***); è la misura più corretta della correlazione tra le variabili $\mathbf{X_1}$ e $\mathbf{X_2}$, in quanto media ponderata dei risultati dei due esperimenti.

Sempre per i problemi di simmetria e omoschedasticità, per ottenere il coefficiente di regressione comune $\mathbf{r_w}$

- dopo la trasformazione di $\mathbf{r_1}$ e $\mathbf{r_2}$ rispettivamente in $\mathbf{z_1}$ e $\mathbf{z_2}$
- si calcola il valore medio $\mathbf{z_w}$

con

$$z_w = \frac{(n_1 - 3)\cdot z_1 + (n_2 - 3)\cdot z_2}{(n_1 - 3) + (n_2 - 3)}$$

che, nel caso di $\mathbf{n_1 = n_2}$, può essere semplificata in

$$z_w = \frac{z_1 + z_2}{2}$$

- Infine si trasforma $\mathbf{z_w}$ in $\mathbf{r_w}$

con

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Recentemente, sono stati proposti altri metodi, che dovrebbero dare stime più precise di $\mathbf{r_w}$.

ESEMPIO

a) Calcolare la significatività della differenza tra i due coefficienti di correlazione

- $\mathbf{r_1 = 0,22}$ con $\mathbf{n_1 = 30}$ e
- $\mathbf{r_2 = 0,31}$ con $\mathbf{n_2 = 50}$

b) Calcolare inoltre il coefficiente di correlazione ponderato $\mathbf{r_w}$.

Risposte

a) Dopo aver trasformato $\mathbf{r_1 = 0,22}$

$$z_1 = 0,5 \cdot \ln\left(\frac{1 + 0,22}{1 - 0,22}\right) = 0,5 \cdot \ln\left(\frac{1,22}{0,78}\right) = 0,5 \cdot \ln 1,564 = 0,5 \cdot 0,447 = 0,224$$

in $\mathbf{z_1 = 0,224}$

e $\mathbf{r_2 = 0,31}$

$$z_1 = 0{,}5 \cdot \ln\left(\frac{1+0{,}31}{1-0{,}31}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}31}{0{,}69}\right) = 0{,}5 \cdot \ln 1{,}899 = 0{,}5 \cdot 0{,}641 = 0{,}321$$

in $\mathbf{z_2 = 0{,}321}$

- si calcola l'errore standard della differenza $\mathbf{z_1 - z_2}$

$$\sigma_{z1-z2} = \sqrt{\frac{1}{30-3} + \frac{1}{50-3}} = \sqrt{0{,}037 + 0{,}021} = \sqrt{0{,}058} = 0{,}24$$

ottenendo σ_{z1-z2} **= 0,24**.

Da essi si ricava

$$Z = \frac{0{,}224 - 0{,}321}{0{,}24} = \frac{-0{,}097}{0{,}24} = -0{,}40$$

un valore di **Z = -0,40**.

Nella **distribuzione normale bilaterale**, poiché la domanda verte solo sulla significatività della differenza tra i due coefficienti di correlazione $\mathbf{r_1}$ e $\mathbf{r_2}$, a **Z = -0,40** corrisponde una probabilità α = **0,689**. Pure senza un'analisi della potenza del test, illustrata nel paragrafo successivo, la probabilità è così alta che si può concludere non esiste alcuna differenza tra i due coefficienti angolari.

Nel caso di un test unilaterale, rispetto alla procedura illustrata l'unica differenza consiste nella stima della probabilità α, non diversamente dal test t su due medie o due coefficienti angolari.

b) Per ottenere il coefficiente ponderato $\mathbf{r_w}$, con

- $\mathbf{z_1 = 0{,}224}$ e $\mathbf{n_1 = 30}$
- $\mathbf{z_2 = 0{,}321}$ e $\mathbf{n_2 = 50}$

- dapprima si stima $\mathbf{z_w}$

$$z_w = \frac{(30-3)\cdot 0{,}224 + (50-3)\cdot 0{,}321}{(30-3)+(50-3)} = \frac{6{,}048 + 15{,}087}{74} = \frac{21{,}135}{74} = 0{,}286$$

che risulta uguale a **0,286**

- infine con la trasformazione

$$r_w = \frac{2{,}718^{2\cdot 0{,}286} - 1}{2{,}718^{2\cdot 0{,}286} + 1} = \frac{2{,}718^{0{,}572} - 1}{2{,}718^{0{,}572} + 1} = \frac{1{,}77 - 1}{1{,}77 + 1} = \frac{0{,}77}{2{,}77} = 0{,}278$$

si ricava $\mathbf{r_w = 0{,}278}$.

Il coefficiente di correlazione comune o ponderato tra

- $\mathbf{r_1 = 0{,}22}$ con $\mathbf{n_1 = 30}$ **e**
- $\mathbf{r_2 = 0{,}31}$ con $\mathbf{n_2 = 50}$

è $\mathbf{r_w = 0{,}278.}$

18.7. POTENZA A PRIORI E A POSTERIORI DEL TEST PER LA SIGNIFICATIVITA' DELLA DIFFERENZA TRA DUE COEFFICIENTI DI CORRELAZIONE

La **potenza a posteriori 1-β** del test di significatività tra due coefficienti di correlazione in campioni indipendenti è valutata

mediante

$$Z_\beta = \frac{|z_1 - z_2|}{\sigma_{z1-z2}} - Z_\alpha$$

- dove, con la solita simbologia,
- z_1 e z_2 sono $\mathbf{r_1}$ e $\mathbf{r_2}$ trasformati con la formula di Fisher,
- σ_{z1-z2} è **l'errore standard della differenza** precedente, ottenuta

con

$$\sigma_{z1-z2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

- Z_β è il valore di **Z** che permette di ricavare la probabilità β in una **distribuzione normale unilaterale**,
- Z_α è il valore ricavato dalle tabella sulla base della probabilità α prefissata nella stessa distribuzione normale, ma **unilaterale o bilaterale in rapporto alla direzione dell'ipotesi $\mathbf{H_1}$**.

La potenza a priori o numero minimo di coppie di osservazioni (n) può essere valutata

con

$$n = 2\left(\frac{Z_\alpha + Z_\beta}{z_1 - z_2}\right)^2 + 3$$

utilizzando la simbologia consueta.

Come nel test **t**, la quantità **n** indica il numero minimo di dati per **ognuno dei due coefficienti di correlazione** a confronto, affinché l'ipotesi nulla $\rho_1 = \rho_2$ possa essere rifiutata alla probabilità α prefissata e con la potenza **1-β** desiderata.

Il bilanciamento di due campioni indipendenti permette di raggiungere la potenza massima del test, utilizzando il numero minimo di osservazioni per ogni gruppo. Ma nella pratica sperimentale, non sempre le rilevazioni nei due gruppi hanno lo stesso costo morale od economico: somministrare sostanze tossiche aumenta la mortalità delle cavie rispetto al placebo; le analisi del controllo e del trattato possono richiedere procedure differenti che esigono tempi di durata differente; tra due aree a confronto nelle quali prelevare i campioni, una può essere sul posto e l'altra molto distante. In vari settori della ricerca applicata, per valutare le trasformazioni intervenute nel periodo, è prassi richiedere il confronto di dati ancora da raccogliere rispetto a dati storici, il cui campione ovviamente non può essere ampliato.

Spesso, può essere utile diminuire al minimo le osservazioni di un gruppo oppure utilizzare il piccolo campione già raccolto, aumentando ovviamente quelle dell'altro gruppo, affinché il test non perda la potenza desiderata.

La **media armonica** permette di stimare

- quante osservazioni deve avere il campione **2** ($\mathbf{n_2}$),
- una volta che sia stato stimato il numero minimo di dati (**n**) e
- prefissato il numero di dati che si intende raccogliere o già raccolti per il campione **1** ($\mathbf{n_1}$)

mediante la relazione

$$n_2 = \frac{n_1(n+3) - 6n}{2n_1 - n - 3}$$

ESEMPIO 1

Il precedente test per la significatività della differenza tra i due coefficienti di correlazione

- $\mathbf{r_1 = 0{,}22}$ con $\mathbf{n_1 = 30}$
- $\mathbf{r_2 = 0{,}31}$ con $\mathbf{n_2 = 50}$

non ha permesso di rifiutare l'ipotesi nulla.

Si chiede

a) Quale era la potenza di questo test per una significatività con $\alpha = 0.05$**?**

b) Quanti dati per gruppo sarebbero necessari affinché nel **80%** dei casi il test risulti significativo alla probabilità $\alpha = 0.05$?

Risposte

A) Dapprima si trasformano **r_1= 0,22**

$$z_1 = 0{,}5 \cdot \ln\left(\frac{1+0{,}22}{1-0{,}22}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}22}{0{,}78}\right) = 0{,}5 \cdot \ln 1{,}564 = 0{,}5 \cdot 0{,}447 = 0{,}224$$

in **z_1 = 0,224**

e **r_2 = 0,31**

$$z_1 = 0{,}5 \cdot \ln\left(\frac{1+0{,}31}{1-0{,}31}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}31}{0{,}69}\right) = 0{,}5 \cdot \ln 1{,}899 = 0{,}5 \cdot 0{,}641 = 0{,}321$$

in **z_2 = 0,321**

Successivamente si calcola l'errore standard della differenza **z_1-z_2**

$$\sigma_{z1-z2} = \sqrt{\frac{1}{30-3} + \frac{1}{50-3}} = \sqrt{0{,}037 + 0{,}021} = \sqrt{0{,}058} = 0{,}24$$

ottenendo σ_{z1-z2} **= 0,24**

e nella tabella della distribuzione normale bilaterale per **α = 0.05** si rileva **Z = 1,96**.

Infine con

$$Z_\beta = \frac{|0{,}224 - 0{,}321|}{0{,}24} - 1{,}96 = 0{,}40 - 1{,}96 = -1{,}56$$

si ottiene **Z_β = - 1,56**

Ad un valore di **Z = 1,56** in una distribuzione normale unilaterale corrisponde una probabilità uguale 0,059; ma il valore negativo indica che essa si trova a sinistra della distribuzione, molto distante dal valore dell'ipotesi nulla; di conseguenza, la potenza del test è particolarmente bassa, pari appunto al 5,9%.

B) Una conferma di questa **potenza a posteriori** molto bassa può venire anche dalla stima della **potenza a priori** o numero minimo di dati necessari affinché il test sulla differenza tra i due coefficienti di correlazione risulti significativo. Con una potenza così bassa, ovviamente il numero (**n**) richiesto risulterà molto alto.

Poiché,

- per una probabilità α **= 0.05 bilaterale,** il valore di $\mathbf{Z_\alpha}$ è uguale a **1,96**
- mentre, per una probabilità β **= 0.20 unilaterale**, il valore di $\mathbf{Z_\beta}$ è uguale a **0,84**

e con $\mathbf{z_1 = 0{,}224}$ e $\mathbf{z_2 = 0{,}321}$

il numero minimo di dati per ogni per ogni gruppo

$$n = 2\left(\frac{1{,}96 + 0{,}84}{0{,}224 - 0{,}321}\right)^2 + 3 = \left(\frac{2{,}8}{0{,}097}\right)^2 + 3 = 833{,}2 + 3 = 837{,}2$$

è **n = 838** (sempre arrotondato all'unità superiore).

Il numero stimato è molto maggiore di quello dei dati raccolti nei due campioni (30 e 50). In vari settori della ricerca applicata, nei quali ogni misura campionaria ha costi non trascurabili, un numero così elevato indica l'impossibilità pratica di dimostrare la significatività della differenza tra i due coefficienti di correlazione. Pertanto, si può ritenere che tra essi non esista una differenza significativa, come d'altronde indicava il valore di probabilità α molto alto.

ESEMPIO 2. Un ricercatore dispone di un campione di **40** osservazioni, raccolte in una rilevazione di alcuni anni prima, nel quale il coefficiente di correlazione lineare semplice sulle quantità di due componenti chimici di un alimento è risultato uguale a **0,19**. Egli si aspetta che, per le trasformazioni intervenute nell'ultimo periodo nella coltivazione e nella conservazione dei cibi, tale correlazione sia aumentata. Con un campione di **30** misure ha infatti trovato un valore di **r = 0,48**.

Calcolare:

a) Quanti dati servono, per un esperimento con 2 campioni bilanciati, affinché il test risulti significativo alla probabilità α **= 0.05** con un rischio di commettere un errore di II Tipo pari a una probabilità β **= 0.20**?

b) Poiché il campione storico è di **40** dati, quanti ne deve raccogliere con il secondo campione per rispettare le probabilità α e β precedenti?

Risposte

Dopo aver effettuato il test di significatività, poiché se esso risultasse positivo il problema sarebbe già risolto, si stima il numero di dati minimo per ognuno dei due gruppi a confronto.

- Dapprima si trasforma $\mathbf{r_1 = 0{,}19}$ in $\mathbf{z_1}$

$$z_1 = 0{,}5 \cdot \ln\left(\frac{1 + 0{,}19}{1 - 0{,}19}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}19}{0{,}81}\right) = 0{,}5 \cdot \ln 1{,}469 = 0{,}5 \cdot 0{,}385 = 0{,}192$$

ottenendo $\mathbf{z_1 = 0{,}192}$

e $\mathbf{r_2 = 0{,}48}$ in $\mathbf{z_2}$

$$z_1 = 0{,}5 \cdot \ln\left(\frac{1+0{,}48}{1-0{,}48}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}48}{0{,}52}\right) = 0{,}5 \cdot \ln 2{,}846 = 0{,}5 \cdot 1{,}046 = 0{,}523$$

ottenendo $\mathbf{z_2 = 0{,}523.}$

- Successivamente, dalla distribuzione normale si ricava

il valore di **Z** per α **= 0.05 unilaterale** (perché il test chiede se vi è stato un aumento significativo) ottenendo $\mathbf{Z_\alpha = 1{,}645}$

e il valore di **Z** per β **= 0.20 unilaterale** ottenendo $\mathbf{Z_\beta = 0{,}84.}$

- Infine di ricava **n**

$$n = 2\left(\frac{1{,}645 + 0{,}84}{0{,}192 - 0{,}523}\right)^2 + 3 = \left(\frac{2{,}485}{0{,}331}\right)^2 + 3 = 7{,}508^2 + 3 = 56{,}4 + 3 = 59{,}4$$

che risulta uguale a 60.

Poiché è risultato un campione (**n = 60**) non molto più grande di $\mathbf{n_1 = 40}$, è possibile stimare il numero di dati necessario nel secondo ($\mathbf{n_2}$) per mantenere costante le probabilità richieste.

Applicando

$$n_2 = \frac{n_1(n+3) - 6n}{2n_1 - n - 3}$$

risulta

$$n_2 = \frac{40 \cdot (60+3) - 6 \cdot 60}{2 \cdot 40 - 60 - 3} = \frac{40 \cdot 63 - 360}{80 - 57} = \frac{2160}{23} = 93{,}9$$

che il secondo campione deve contenere almeno $\mathbf{n_2 = 94}$

E' una risposta che, per il numero non eccessivamente elevato di dati richiesti rispetto al campione già raccolto, rende possibile l'esperimento. Ad esempio, se il numero (**n**) di dati richiesto per ogni gruppo fosse stato oltre il doppio degli $\mathbf{n_1 = 40}$ già raccolti con il primo campione, il numero di dati da raccogliere con il secondo ($\mathbf{n_2}$) sarebbe stato molto grande, tale da rendere molto costoso, se non impossibile, l'esperimento. Inoltre con due campioni così sbilanciati, le condizioni di potenza e significatività del test sarebbero state profonde alterate.

Per altri concetti sul bilanciamento di due campioni indipendenti, si rinvia al capito sul test **t**, in quanto la procedura è simile.

18.8. TEST PER LA DIFFERENZA TRA PIU' COEFFICIENTI DI CORRELAZIONE COEFFICIENTE DI CORRELAZIONE COMUNE r_w E SUA SIGNIFICATIVITA'

Per il **confronto simultaneo tra più coefficienti di correlazione indipendenti** r_1, r_2, …, r_k, cioè per verificare l'ipotesi nulla

$$H_0: \rho_1 = \rho_2 = \quad = \rho_k$$

(seguendo le indicazioni di Jerrold H. **Zar** nel suo test del 1999 ***Biostatistical Analysis***, fourth ed. Prentice Hall, New Jersey, pp. 390-394)

si stima un valore χ^2_{k-1} con gdl **k-1**

$$\chi^2_{k-1} = \sum_{i=1}^{k}(n_i - 3)\cdot z_i^2 - \frac{\left[\sum_{i=1}^{k}(n_i - 3)\cdot z_i\right]^2}{\sum_{i=1}^{k}(n_i - 3)}$$

dove

- **k** è il numero di coefficienti di correlazione campionari **r** a confronto simultaneo
- n_i è il numero di dati campionari di ogni r_i
- z_i è il valore di ogni r_i trasformato nel corrispondente valore z_i con la formula

$$z = 0{,}5 \cdot \ln\left(\frac{1+r}{1-r}\right)$$

Se il **test non risulta significativo**, si può concludere che i diversi valori r_i sono stime campionarie dello stesso valore reale ρ; di conseguenza, come sua stima migliore, è utile calcolare il **coefficiente di correlazione comune** o **medio ponderato** r_w (***common r*** or ***weighted mean of r***)

con

$$z_w = \frac{\sum_{i=1}^{k}(n_i - 3)\cdot z_i}{\sum_{i=1}^{k}(n_i - 3)}$$

Successivamente, per un confronto con gli r_1, r_2, …, r_k originali, z_w può essere ritrasformato nella scala **r** (ricordando ancora che, in valore assoluto, **r** varia da **0** a **1**, mentre **z** varia da **0** a ∞)

attraverso

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

dove

- **e** è la costante neperiana (approssimata con 2,718).

A questo valore comune $\mathbf{r_w}$ **,** con un test molto più potente rispetto a quelli di ogni singolo $\mathbf{r_i}$ **,** poiché ha il vantaggio di essere calcolato su tutte le **N** coppie di dati,

in quanto

$$N = \sum_{i=1}^{k} n_i$$

può essere applicato

- sia il test per la verifica dell'ipotesi nulla

$$\mathbf{H_0: \rho = 0}$$

che $\mathbf{r_w}$ non sia **significativamente differente da zero**

mediante la formula proposta da J. **Neyman** nel 1959 (vedi capitolo: ***Optimal asymptotic test of composite hypothesis***, pp. 213-234. Nel libro di U. **Grenander** (ed.) ***Probability and Statistics: The Harald Cramér Volume,*** John Wiley, New York) e ripresa da S. R. **Paul** nel 1988

$$Z = \frac{\sum_{i=1}^{k} (n_i \cdot r_i)}{\sqrt{N}}$$

- sia il test per la verifica dell'ipotesi nulla

$$\mathbf{H_0: \rho = \rho_0}$$

che $\mathbf{r_w}$ non sia **significativamente differente da un valore prefissato ρ_0**

mediante la formula proposta da S. R. **Paul** nel 1988 (vedi, anche per la citazione precedente, l'articolo: ***Estimation of and testing significance for a common correlation coefficient***. sulla rivista **Communic. Statist. - Theor. Meth**. 17: 39-53, nel quale fa una presentazione e una disamina generale di questi metodi)

$$Z = (z_w - \zeta_0) \cdot \sqrt{\sum_{i=1}^{k} (n_i - 3)}$$

dove

- z_w è il valore $\mathbf{r_w}$ dopo trasformazione di Fisher,
- ζ_0 è il valore atteso o di confronto $\boldsymbol{\rho_0}$ dopo trasformazione di Fisher.

In questi due test sul coefficiente comune $\mathbf{r_w}$, le ipotesi alternative $\mathbf{H_1}$ possono essere bilaterali oppure unilaterali. L'unica differenza consiste nella stima della probabilità α: se in una distribuzione normale bilaterale oppure unilaterale.

Se i **k** campioni a confronto sono **dipendenti,** quali ad esempio i valori $\mathbf{r_i}$ di correlazione della variabile X_A con le variabili X_B, X_C, X_D calcolati sugli stessi prelievi, **questa tecnica non è corretta**. Si dovrà ricorrere alla statistica multivariata, con la correlazione multipla.

ESEMPIO 1. In una ricerca sulla correlazione lineare semplice tra le quantità di due conservanti X_A e X_B contenuti in campioni di alimenti, con tre verifiche indipendenti sono stati ottenuti i seguenti risultati:

1 - $\mathbf{r_1 = 0,48}$ con $\mathbf{n_1 = 120}$

2 - $\mathbf{r_2 = 0,31}$ con $\mathbf{n_2 = 100}$

3 - $\mathbf{r_3 = 0,48}$ con $\mathbf{n_3 = 150}$

Con questi dati,

a) verificare se tra questi $\mathbf{r_i}$ esiste una differenza significativa;

b) in caso di non rifiuto della ipotesi nulla, calcolare il coefficiente comune $\mathbf{r_w}$;

c) verificare se il coefficiente di correlazione comune $\mathbf{r_w}$ è significativamente maggiore di zero;

d) verificare se $\mathbf{r_w}$ si discosta significativamente dal valore $\boldsymbol{\rho_0}$ **= 0,32** indicato come valore reale in una pubblicazione presa a riferimento.

Risposte.

A) Dopo aver trasformato i valori di $\mathbf{r_i}$ nei corrispondenti valori $\mathbf{z_i}$

$$z_1 = 0{,}5 \cdot \ln\left(\frac{1+0{,}48}{1-0{,}48}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}48}{0{,}52}\right) = 0{,}5 \cdot \ln 2{,}846 = 0{,}5 \cdot 1{,}046 = 0{,}523$$

$$z_2 = 0{,}5 \cdot \ln\left(\frac{1+0{,}31}{1-0{,}31}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}31}{0{,}69}\right) = 0{,}5 \cdot \ln 1{,}899 = 0{,}5 \cdot 0{,}641 = 0{,}320$$

$$z_3 = 0{,}5 \cdot \ln\left(\frac{1+0{,}37}{1-0{,}37}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}37}{0{,}63}\right) = 0{,}5 \cdot \ln 2{,}175 = 0{,}5 \cdot 0{,}777 = 0{,}389$$

ottenendo: $\mathbf{z_1 = 0{,}523}$; $\mathbf{z_2 = 0{,}320;}$ $\mathbf{z_3 = 0{,}389}$

si calcola un valore del χ^2 con **gdl= 2**

applicando

$$\chi^2_{k-1} = \sum_{i=1}^{k}(n_i - 3)\cdot z_i^2 - \frac{\left[\sum_{i=1}^{k}(n_i - 3)\cdot z_i\right]^2}{\sum_{i=1}^{k}(n_i - 3)}$$

Scindendo per comodità di calcolo

$$\chi^2 = A - B$$

dove

$$A = \sum_{i=1}^{k}(n_i - 3)\cdot z_i^2$$

e

$$B = \frac{\left[\sum_{i=1}^{k}(n_i - 3)\cdot z_i\right]^2}{\sum_{i=1}^{k}(n_i - 3)}$$

con i dati dell'esempio

$$A = (120-3)\cdot 0{,}523^2 + (100-3)\cdot 0{,}320^2 + (150-3)\cdot 0{,}389^2$$

$$A = 117\cdot 0{,}274 + 97\cdot 0{,}102 + 147\cdot 0{,}151 = 32{,}058 + 9{,}894 + 22{,}197 = 64{,}149$$

risultano **A = 64,149**

e

$$B = \frac{\left[(120-3)\cdot 0{,}523 + (100-3)\cdot 0{,}320 + (150-3)\cdot 0{,}389\right]^2}{(120-3)+(100-3)+(150-3)}$$

$$B = \frac{(61{,}191 + 31{,}04 + 57{,}183)^2}{117+97+147} = \frac{149{,}417^2}{361} = \frac{22325{,}44}{361} = 61{,}843$$

B = 61,843

determinando

$$\chi^2 = 64{,}149 - 61{,}843 = 2{,}306$$

un valore di χ^2_2 **= 2,306**.

Poiché il valore critico alla probabilità α = **0.05** con **df = 2** è uguale a **5,991** non si può rifiutare l'ipotesi nulla. Una lettura più attenta della tabella dei valori critici evidenzia che il valore calcolato (**2,306**) è inferiore anche a quello riportato per la probabilità α = **0.25** che risulta uguale a **2,773**.
Si può quindi concludere che i tre coefficienti di correlazione $\mathbf{r_1}$, $\mathbf{r_2}$ e $\mathbf{r_3}$ sono statisticamente uguali.

B) Con questi risultati dell'analisi sulla significatività della differenza tra i **k** valori campionari, è utile calcolare il coefficiente di correlazione comune $\mathbf{r_w}$, come stima migliore del valore ρ della popolazione.
Con $\mathbf{z_1 = 0{,}523}$; $\mathbf{z_2 = 0{,}320;}$ $\mathbf{z_3 = 0{,}389}$ e $\mathbf{n_1 = 120,}$ $\mathbf{n_2 = 100,}$ $\mathbf{n_3 = 150}$
il valore di $\mathbf{z_w}$ comune

$$z_w = \frac{(120-3)\cdot 0{,}523 + (100-3)\cdot 0{,}320 + (150-3)\cdot 0{,}389}{(120-3)+(100-3)+(150-3)}$$

$$z_w \frac{61{,}191 + 31{,}04 + 57{,}183}{117+97+147} = \frac{149{,}417}{361} = 0{,}414$$

risulta uguale a **0,414**.
Ritrasformato in $\mathbf{r_w}$

$$r_w = \frac{2{,}718^{2\cdot 0{,}414}-1}{2{,}718^{2\cdot 0{,}414}+1} = \frac{2{,}718^{0{,}828}-1}{2{,}718^{0{,}828}+1} = \frac{2{,}289-1}{2{,}289+1} = \frac{1{,}289}{3{,}289} = 0{,}392$$

risulta uguale a **0,392**.

C) Per la verifica dell'ipotesi nulla $\mathbf{H_0: \rho = 0}$
con ipotesi alternativa unilaterale $\mathbf{H_1: \rho > 0}$
si utilizzano le varie osservazioni campionarie, per cui
da $\mathbf{r_1 = 0{,}48}$; $\mathbf{r_2 = 0{,}31;}$ $\mathbf{r_3 = 0{,}37}$ e $\mathbf{n_1 = 120,}$ $\mathbf{n_2 = 100,}$ $\mathbf{n_3 = 150}$
si ottiene

$$Z = \frac{(120\cdot 0{,}48)+(100\cdot 0{,}31)+(150\cdot 0{,}37)}{\sqrt{120+100+150}} = \frac{57{,}6+31{,}0+55{,}5}{\sqrt{370}} = \frac{144{,}1}{19{,}24} = 7{,}49$$

un valore di **Z** uguale a **7,49**.
Poiché nella **distribuzione normale unilaterale** il valore ottenuto è così grande da non essere nemmeno riportato nella tabella, ad esso è associata una probabilità α estremamente piccola. Di conseguenza, si può concludere che il valore medio $\mathbf{r_w}$ è significativamente maggiore di zero.

D) Per la verifica dell'ipotesi nulla $\mathbf{H_0: \rho = \rho_0}$
con ipotesi alternativa bilaterale $\mathbf{H_1: \rho \neq \rho_0}$

cioè che $\mathbf{r_w = 0,392}$ **sia significativamente differente da** $\boldsymbol{\rho_0 = 0,32}$

attraverso

$$Z = (z_w - \zeta_0) \cdot \sqrt{\sum_{i=1}^{k} (n_i - 3)}$$

dopo aver calcolato il coefficiente di correlazione comune in scala **z**, cioè $\mathbf{z_w = 0,414}$

si trasforma nella stessa scala il valore $\boldsymbol{\rho_0 = 0,32}$

ottenendo

$$z_{\rho_0} = \zeta_0 = 0,5 \cdot \ln\left(\frac{1+0,32}{1-0,32}\right) = 0,5 \cdot \ln\frac{1,32}{0,68} = 0,5 \cdot \ln 1,941 = 0,5 \cdot 0,663 = 0,331$$

Infine si calcola **Z**

$$Z = (0,414 - 0,331) \cdot \sqrt{117 + 97 + 147} = 0,083 \cdot 19 = 1,58$$

ottenendo un valore uguale a 1,58.

In una **distribuzione normale bilaterale**, a $\mathbf{Z = 1,58}$ corrisponde una probabilità $\boldsymbol{\alpha = 0,114}$. Non è possibile rifiutare l'ipotesi nulla: $\mathbf{r_w = 0,392}$ non è significativamente differente da $\boldsymbol{\rho_0 = 0,32}$.

<u>Correzioni per il bias della trasformazione di Fisher</u>

La trasformazione di **r** in **z** con la formula di Fisher determina un errore piccolo, ma sistematico e in eccesso, nel valore di **z**. E' un risultato noto da tempo,

- già evidenziato da H. **Hotelling** nel 1953 (vedi articolo: ***New light on the correlation coefficient and its trasformation***. **Journal Royal Statistical Society** B 15, pp. 193-232)
- e discusso dallo stesso R. A. **Fisher** nel 1958 (vedi testo: ***Statistical Methods for Research Workers***, 13th ed., Hafner , New York, 146 pp.).

Ne deriva, in particolare quando si ricorre a test per la significatività della differenza tra **k** campioni, utilizzando ad esempio la formula già presentata

$$\chi^2_{k-1} = \sum_{i=1}^{k} (n_i - 3) \cdot z_i^2 - \frac{\left[\sum_{i=1}^{k} (n_i - 3) \cdot z_i\right]^2}{\sum_{i=1}^{k} (n_i - 3)}$$

che gli errori si sommano e la differenza tra valore calcolato e valore reale non è più trascurabile. L'errore è tanto maggiore quanto più alto è **k.**

Per la correzione del ***bias*** nel valore di **z,** ottenuto con la trasformazione di **r**, le proposte sono numerose. Tra le più frequenti per aggiustare **z** possono esserne citate due:

A) **<u>La formula di Hotelling</u>** che, nell'articolo già citato del 1953,

propone di sottrarre **c**

$$c = \frac{3z + r}{4(n-1)}$$

al valore **z** stimato.

Ad esempio, con **r = 0,48** calcolato su **n = 120**

$$z_1 = 0{,}5 \cdot \ln\left(\frac{1+0{,}48}{1-0{,}48}\right) = 0{,}5 \cdot \ln\left(\frac{1{,}48}{0{,}52}\right) = 0{,}5 \cdot \ln 2{,}846 = 0{,}5 \cdot 1{,}046 = 0{,}523$$

si ottiene un valore di **z = 0,523**

Il valore corretto, ottenuto con la sottrazione della correzione

$$c = \frac{3 \cdot 0{,}523 + 0{,}48}{4 \cdot (120-1)} = \frac{2{,}009}{476} = 0{,}004$$

c = 0,004 diventa **z'** = z – c = 0,523 – 0,004 = **0,519**.

B) **<u>La formula di Fisher</u>**, proposta nel testo del 1955 per la correzione,

è

$$c = \frac{r}{2(n-1)}$$

Con lo stesso esempio, la correzione è

$$c = \frac{0{,}48}{2 \cdot (120-1)} = \frac{0{,}48}{236} = 0{,}002$$

c = 0,002 e il valore corretto diventa **z'** = z – c = 0,523 – 0,002 = **0,521**.

Per il calcolo della significatività della differenza tra più coefficienti, invece di correggere ogni valore $\mathbf{z_i}$ e usare la formula generale già presentata, è possibile utilizzare direttamente

<u>la formula corretta di Paul</u>,

$$\chi^2_{Paul} = \sum_{i=1}^{k} \frac{n_i \cdot (r_i - r_w)^2}{(1 - r_i \cdot r_w)^2}$$

che, come la precedente, utilizza la distribuzione $\boldsymbol{\chi^2}$ con **df = k-1**.

Ad esempio, applicata agli stessi dati della formula generale che ha stimato un valore di χ^2_2 **= 2,306,** con

r_1 = 0,48; **r_2 = 0,31;** **r_3 = 0,37** e **n_1 = 120,** **n_2 = 100,** **n_3 = 150** e **r_w = 0,392**

si ottiene

$$\chi^2_{Paul} = \frac{120 \cdot (0{,}48 - 0{,}392)^2}{[1 - (0{,}48 \cdot 0{,}392)]^2} + \frac{100 \cdot (0{,}31 - 0{,}392)^2}{[1 - (0{,}31 \cdot 0{,}392)]^2} + \frac{150 \cdot (0{,}37 - 0{,}392)^2}{[1 - (0{,}37 \cdot 0{,}392)]^2}$$

$$\chi^2_{Paul} = \frac{120 \cdot 0{,}088^2}{(1 - 0{,}188)^2} + \frac{100 \cdot 0{,}082^2}{(1 - 0{,}122)^2} + \frac{150 \cdot 0{,}022^2}{(1 - 0{,}180{,}145)^2}$$

$$\chi^2_{Paul} = \frac{0{,}924}{0{,}659} + \frac{0{,}67}{0{,}771} + \frac{0{,}075}{0{,}731} = 1{,}402 + 0{,}869 + 0{,}103 = 2{,}374$$

un valore pari a 2,374.

Sempre S. R. **Paul** nella sua ampia presentazione del 1988 (***Estimation of and testing significance for a common correlation coefficient***, pubblicato su **Communic. Statist. - Theor. Meth**. Vol. 17, pp. 39-53) suggerisce che <u>quando ρ è minore di **0,5**</u> (indicazione approssimata)

- per il calcolo del coefficiente medio z_w

al posto di

$$z_w = \frac{\sum_{i=1}^{k} (n_i - 3) \cdot z_i}{\sum_{i=1}^{k} (n_i - 3)}$$

sia usato

$$z_w = \frac{\sum_{i=1}^{k} (n_i - 1) \cdot z_i}{\sum_{i=1}^{k} (n_i - 1)}$$

- e nell'inferenza con ipotesi nulla H_0**:** $\rho = \rho_0$ per calcolare **Z**

al posto di

$$Z = (z_w - \zeta_0) \cdot \sqrt{\sum_{i=1}^{k} (n_i - 3)}$$

sia usata

$$Z = (z_w - \zeta_0) \cdot \sqrt{\sum_{i=1}^{k} (n_i - 1)}$$

18.9. CENNI SUI CONFRONTI MULTIPLI TRA PIU' r

Come nel caso delle medie, effettuato il test per la verifica dell'uguaglianza tra più coefficienti di correlazione, se si arriva alla conclusione che non tutti i coefficienti di correlazione sono tra loro uguali, si pone il problema di sapere tra quali la differenza sia significativa.

E' possibile pervenire alla soluzione sia

- con **confronti a priori**, attraverso la tecnica dei contrasti ortogonali,
- con **confronti a posteriori** o multipli, in modo analogo al test SNK o alla procedura di Tukey.

Senza riprendere dettagliatamente i concetti generali, già ampiamente discussi nei confronti tra **k** medie, è utile ricordare che i confronti a priori godono del vantaggio rilevante di utilizzare per ogni contrasto la stessa probabilità α del test generale, mentre i confronti a posteriori, sulla base del principio del Bonferroni, tendono ad utilizzare per ogni confronto una probabilità α/k, dove **k** è il numero di confronti possibili.

Per i **confronti a priori**, con la tecnica dei contrasti o confronti ortogonali,

- dapprima si stima un valore critico S_α per la probabilità α prefissata

con

$$S_\alpha = \sqrt{\chi^2_{\alpha,k-1}}$$

- successivamente si calcola **S**

$$S = \frac{\left|\sum_i c_i z_i\right|}{SE}$$

dove

$$SE = \sqrt{\sum_i c_i^2 \sigma_{zi}^2}$$

- c_i è il coefficiente del contrasto,
- z_i è la trasformazione di **r** in **z.**

Sono significativi i contrasti il cui valore **S** supera il valore critico S_α.

Per **i confronti a posteriori**, i metodi proposti sono numerosi.

Una procedura semplice prevede che, sempre dopo l'analisi complessiva con il χ^2 che deve risultare significativa, tra

- **due coefficienti di correlazione r_A e r_B calcolati su campioni di dimensioni n_A e n_B**
- **esista una differenza significativa** se il valore **q**

$$q = \frac{z_A - z_B}{\sqrt{\frac{1}{2}\left(\frac{1}{n_A - 3} + \frac{1}{n_B - 3}\right)}}$$

- è superiore a quello riportato nella tabella dei valori critici del q studentizzato con indici α, ∞, k

dove

- α è la probabilità complessiva,
- ∞ è il numero totale di dati utilizzati per il confronto dei k gruppi,
- p è il numero di passi tra i due valori r a confronto, dopo che sono stai tutti ordinati per rango.

Anche nel caso dei coefficienti di correlazione, il confronto tra k gruppi può essere tra un controllo e k-1 valori sperimentali. Il metodo è analogo a quello spiegato per le medie, poiché il numero di confronti possibili non è più **k(k-1)/2** ma scende a **k-1**.

18.10. LA CORRELAZIONE PARZIALE O NETTA DI PRIMO ORDINE E DI ORDINE SUPERIORE; LA CORRELAZIONE SEMIPARZIALE

Quando si analizzano le relazioni tra più variabili, la correlazione tra due di esse risente anche delle relazioni esistenti con le altre. Sovente, nella ricerca è richiesto di valutare l'associazione tra due variabili, eliminando l'influsso delle altre:

- è la **correlazione parziale o netta** (***partial correlation***),
- mentre quella discussa nei paragrafi precedenti è la **correlazione semplice o totale**.

Per esempio, nel caso in cui si intenda valutare le correlazioni tra 3 variabili come altezza, peso e circonferenza toracica, le relazioni esistenti tra circonferenza toracica ed altezza sono influenzate in modo rilevante da quelle esistenti tra ognuna di queste due con il peso. Nello stesso modo, la correlazione tra altezza e diametro del tronco di un albero risente della correlazione di entrambi con la sua età.

La correlazione parziale o netta è la stima della correlazione tra due variabili, dopo l'eliminazione degli effetti dovuti all'eventuale associazione con la terza (o il restante gruppo di k variabili).

Un metodo teoricamente possibile per valutare la correlazione netta tra due variabili sarebbe la misura della correlazione semplice o totale, mantenendo costante la terza variabile. Ma questa procedura presenta vari inconvenienti, facilmente identificabili:

- **la necessità di ripetere più volte i calcoli**;

- **l'impossibilità di estendere e generalizzare le conclusioni**, poiché per ogni valore della terza variabile si avrebbe una correlazione con un valore campionario differente;
- un **forte aumento della numerosità del campione** e quindi sia dei costi che dei tempi richiesti dalla ricerca.

Nel suo testo già citato, (nella traduzione italiana) Fisher scriveva: "*Grande parte dell'utilità dell'idea della correlazione risiede nella sua applicazione a gruppi con più di due variabili. In tali casi, in cui è nota la correlazione tra ciascuna coppia di tre variabili, se ne può eliminare una qualunque e trovar che in una popolazione scelta la correlazione delle altre due sarebbe da considerare come se la terza variabile fosse costante.*
Quando le stime delle tre correlazioni sono ***ottenibili dalla stessa massa di dati****, il processo di eliminazione fornirà una stima della correlazione parziale paragonabile in tutto e per tutto a una stima diretta".*

Nel linguaggio statistico, per misurare la correlazione parziale o netta tra due variabili si distinguono **correlazioni di vari gradi od ordini**, in rapporto al numero di variabili complessivamente utilizzate, ricordando che il concetto di correlazione riguarda la relazione esistente tra due.
- Quando si dispone solamente delle osservazioni relative a due variabili (come in precedenza),
la **correlazione è detta di grado zero o di ordine zero**;
- quando le variabili osservate sono tre, la correlazione tra due senza l'influenza della terza è detta
correlazione di 1° grado o di 1° ordine;
con quattro variabili, eliminata l'influenza di due,
- la correlazione è di 2° grado o di 2° ordine;
con **N** variabili, eliminata l'influenza delle altre **N-2**,
- **la correlazione tra due variabili è di grado od ordine (N-2)esimo**.

Nel caso di tre variabili, quando sono stati calcolati i coefficienti di correlazione semplice o totale, **il coefficiente di correlazione parziale o netta (**scritta come $\mathbf{r_{12,3}}$ e detta **correlazione tra le variabili $\mathbf{X_1}$ e $\mathbf{X_2}$ al netto degli effetti della variabile $\mathbf{X_3}$**) è data da

$$r_{12,3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1 - r_{13}^2) \cdot (1 - r_{23}^2)}}$$

con **gdl N-3**, dove
- $r_{12,3}$ è la correlazione parziale tra le variabili 1 e 2, a meno (o al netto) degli effetti della 3;
- r_{12}, r_{13}, r_{23} sono le correlazioni semplici tra le rispettive coppie di variabili.

Per la stima della correlazione netta, **le condizioni di validità** sono essenzialmente due:

- le correlazioni di ordine zero devono essere lineari;
- il numero **N** di osservazioni di ogni correlazione di ordine zero deve essere sempre superiore di alcune unità al numero delle variabili, poiché

il numero di gdl della correlazione parziale con k variabili è uguale a N-k.

ESEMPIO 1. Si considerino i 18 laghi dell'Appennino tosco-emiliano, già utilizzati precedentemente, per ognuno dei quali è stata misurata

- la conducibilità (X_1),
- la concentrazione di Ione Calcio + Ione Magnesio (X_2),
- la concentrazione di Solfati + Carbonati (X_3).

I valori sono riportati nella tabella successiva

Laghi	Conducibilità (X_1)	Ione Calcio + Ione Magnesio (X_2)	Solfati + Carbonati (X_3)
SILLARA INF.	20	0,063	0,137
SILLARA SUP.	22	0,077	0,149
SCURO CERRETO	22	0,078	0,095
VERDAROLO	26	0,125	0,156
SQUINCIO	24	0,120	0,107
SCURO PARMENSE	28	0,144	0,191
PALO	27	0,143	0,228
ACUTO OVEST	26	0,115	0,212
SCURO RIGOSO	29	0,185	0,244
COMPIONE INF.	35	0,194	0,322
GEMIO INF.	33	0,193	0,301
PETRUSCHIA	37	0,218	0,304
GEMIO SUP.	34	0,207	0,312
SANTO PARMENSE	35	0,254	0,311
BICCHIERE	37	0,250	0,352
BALLANO	39	0,315	0,354
BACCIO	41	0,364	0,415
VERDE	45	0,338	0,459

Calcolare i tre coefficienti di correlazione semplice o totale. Successivamente, al fine di valutare con maggiore precisione la correlazione esistente tra queste variabili, stimare i tre coefficienti di correlazione parziale.

Risposta. I tre coefficienti di correlazione semplice o totale (calcolati con un programma informatico) sono risultati:

$$r_{12} = 0.9628$$

$$r_{13} = 0{,}9704$$

$$r_{23} = 0{,}9388$$

Poiché la tabella di valori di correlazione semplice

con **16 df** (N-2 = 18-2) e alla probabilità α **= 0.001**, riporta il valore 0,7484

si deve concludere che i tre coefficienti di correlazione semplice sono tutti altamente significativi.

L'analisi con la correlazione netta permette di valutare se questi valori di correlazione tra coppie di variabili sono rafforzati dalla comune correlazione con la terza, per cui le correlazioni reali tra coppie di variabili sono minori.

Applicando la precedente formula

$$r_{12,3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1 - r_{13}^2) \cdot (1 - r_{23}^2)}}$$

le stime dei 3 coefficienti di correlazione parziale danno i risultati seguenti:

- $r_{12,3}$

$$r_{12,3} = \frac{0{,}9628 - 0{,}9704 \cdot 0{,}9388}{\sqrt{\left(1 - 0{,}9707^2\right) \cdot \left(1 - 0{,}9388^2\right)}} == \frac{0{,}9628 - 0{,}9110}{\sqrt{0{,}0583 \cdot 0{,}1187}} = \frac{0{,}0518}{0{,}0832} = 0{,}623$$

è uguale a 0,623;

- $r_{13,2}$

$$r_{13,2} = \frac{0{,}9704 - 0{,}9628 \cdot 0{,}9388}{\sqrt{\left(1 - 0{,}9628^2\right) \cdot \left(1 - 0{,}9388^2\right)}} = \frac{0{,}9704 - 0{,}9039}{\sqrt{0{,}073 \cdot 0{,}1187}} = \frac{0{,}0665}{0{,}0931} = 0{,}714$$

è uguale a 0,714;

- $r_{23,1}$

$$r_{23,1} = \frac{0{,}9388 - 0{,}9628 \cdot 0{,}9704}{\sqrt{\left(1 - 0{,}9628^2\right) \cdot \left(1 - 0{,}9704^2\right)}} = \frac{0{,}9388 - 0{,}9343}{\sqrt{0{,}0730 \cdot 0{,}0583}} = \frac{0{,}0045}{0{,}0652} = 0{,}069$$

è uguale a 0,069.

I valori critici sono riportati nella solita tabella dei coefficienti di correlazione;

questi **<u>coefficienti di correlazione con df 15</u> (N-k = 18-3)**

- alla probabilità α **= 0.05** danno un valore di **r** uguale a **0,4821**

- mentre alla probabilità α = **0.01** è uguale a **0,6055**
- e alla probabilità α = **0.001** è **0,7247**.

E' semplice osservare che i 3 coefficienti netti calcolati risultano minori di quelli semplici; quindi è ovvio dedurre che la correlazione totale tra due variabili era aumentata dalla comune correlazione con la terza variabile.

In merito alla loro significatività, in modo più dettagliato,
- il valore di $\mathbf{r_{12,3}}$ risulta significativo con probabilità inferiore a 0.01
- il valore di $\mathbf{r_{13,2}}$ è significativo con probabilità inferiore a 0.001
- il valore di $\mathbf{r_{23,1}}$ non è significativo; anzi è molto distante dalla significatività e prossimo alla totale assenza di correlazione.

ESEMPIO 2. Come secondo caso e approfondimento delle potenzialità del metodo, è utile riportare l'esempio sviluppato da Fisher. ***Eliminazione dell'età in correlazioni organiche con fanciulli in fase di sviluppo.*** (In "***Metodi statistici ad uso dei ricercatori***, Torino 1948, Unione Tipografica Editrice Torinese (UTET), 326 p. traduzione di M Giorda, del testo **Statistical Methods for Research Workers** di R. A. **Fisher** 1945, nona edizione (la prima è del 1925) a pag. 174.)

Con i *"dati di Munford e di Young, in un gruppo di fanciulli di differente età, si trovò che la* ***correlazione*** *fra* ***statura in pied****i e* ***perimetro toracico*** *era* ***+0,836****. Ci si potrebbe attendere che parte di questa associazione sia dovuta allo sviluppo generale in rapporto all'età crescente. Sarebbe quindi molto desiderabile, sotto vari aspetti, conoscere la correlazione tra le variabili in fanciulli d'una determinata età; ma nella fattispecie soltanto pochi fanciulli saranno esattamente della stessa età ed anche se compiliamo gruppi di età limitati a un anno, avremo in ciascun gruppo un numero molto inferiore al numero totale misurato. Al fine di utilizzare l'intero materiale, dobbiamo limitarci alla conoscenza delle* ***correlazion****i tra* ***statura in piedi*** *e* ***perimetro toracico*** *ed* ***età."***

Spiegando tutti i passaggi, dai valori di
- **statura in piedi (1)** e **perimetro toracico (2)** si ottiene $\mathbf{r_{12}}$ **= +0,836**
- **statura in piedi (1)** e **età (3)** si ottiene $\mathbf{r_{13}}$ **= +0,714**
- **perimetro toracico (2)** e **età (3)** si ottiene $\mathbf{r_{23}}$ **= +0,836**

Da essi ricava $\mathbf{r_{12,3}}$

$$r_{12,3} = \frac{0,836 - (0,714 \cdot 0,708)}{\sqrt{(1 - 0,714^2) \cdot (1 - 0,708^2)}} = \frac{0,836 - 0,505}{\sqrt{0,490 \cdot 0,499}} = \frac{0,331}{0,495} = 0,668$$

"Inserendo i termini numerici nella formula data, otteniamo $r_{12,3} = 0,668$, indicante che, quando l'età è eliminata, la correlazione, quantunque ancora considerevole, è stata notevolmente ridotta. Il valore medio stabilito dagli autori summenzionati per le correlazioni trovate raggruppando i fanciulli per anni è 0,653, un valore, cioè, non molto differente."

Nell'analisi dei coefficiente di correlazione parziale o netta si possono realizzare due situazioni:

- se **il valore parziale è maggiore di quello semplice** o addirittura diventa significativo, si deve dedurre che l'altro fattore nasconde la correlazione che effettivamente esiste;
- se **il coefficiente parziale è minore di quello semplice** o addirittura perde la significatività, si può dedurre che la terza variabile considerata è fortemente correlata con entrambe e le fa variare congiuntamente, **senza che tra esse esista una relazione diretta**.

La significatività della regressione risente in modo marcato di due fattori:

- il numero di osservazioni
- il campo di variazione di X_1 e X_2.

L'importanza del primo fattore è evidenziata dalla semplice lettura della tabella dei valori critici, che diminuiscono in modo rilevante all'aumentare del numero di gradi di libertà, come già sottolineato per la regressione lineare semplice.

Per il secondo fattore, si può osservare che

- quando i valori delle due variabili hanno un intervallo molto limitato, il coefficiente di correlazione ha un valore assoluto molto basso, difficilmente significativo;
- al contrario, quando anche solo una variabile ha un intervallo di variazione molto ampio, il coefficiente di correlazione è molto alto.

Per la corretta programmazione di un esperimento, è quindi conveniente

- **raccogliere in precedenza informazioni sulla loro variabilità e**
- **impostare la raccolta dei dati in modo che essa sia grande**.

Inoltre, poiché l'interpretazione della correlazione tra due variabili è fortemente influenzata

- sia dal numero di dati,
- sia dal campo di variazione,

è difficile confrontare la significatività di due coefficienti di correlazione con dati raccolti in condizioni diverse.

Con metodi molto simili alla correlazione parziale di primo ordine**, gli stessi principi possono essere estesi a 4 o più variabili, con la** correlazione parziale di secondo ordine e a quella di ordine superiore.

Prendendo in considerazione 4 variabili (X_1, X_2, X_3, X_4),

ognuna con lo **stesso numero N di osservazioni**,

si devono calcolare i 6 coefficienti i correlazione semplice o totale (r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34}).

La correlazione parziale di secondo ordine tra due di queste variabili X_1 e X_2 , mantenendo costanti le altre due X_3 e X_4 (scritta $r_{12,34}$),

può essere calcolata con la formula

$$r_{12,34} = \frac{r_{12,4} - r_{13,4} \cdot r_{23,4}}{\sqrt{\left(1 - r_{13,4}^2\right) \cdot \left(1 - r_{23,4}^2\right)}}$$

che utilizza tre correlazioni parziali di primo ordine ($r_{12,3}; r_{14,3}; r_{24,3}$).

Con più di 4 variabili (X_1, X_2, X_3, X_4, …, X_n), la formula generale, per calcolare la correlazione parziale tra la variabile X_1 e X_2 con le variabili X_3 e X_G (il gruppo di tutte le altre) mantenute costanti, diventa

$$r_{12,3G} = \frac{r_{12,G} - r_{13,G} \cdot r_{23,G}}{\sqrt{\left(1 - r_{13,G}^2\right) \cdot \left(1 - r_{23,G}^2\right)}}$$

Per i calcoli è necessario utilizzare un programma informatico, dato l'alto numero di operazioni da effettuare e quindi l'elevata probabilità di commettere errori.

ESEMPIO 3. (Continua l'esempio e la trattazione di Fisher)

"*In modo simile, possono successivamente essere eliminate due o più variabili; così con quattro variabili, possiamo prima eliminare la variabile 4, applicando tre volte la formula per trovare* $\boldsymbol{r_{12,4}}$, $\boldsymbol{r_{13,4}}$ *e* $\boldsymbol{r_{23,4}}$. *Quindi tornando ad applicare la medesima formula a questi tre nuovi valori, si ottiene*

$$r_{12,34} = \frac{r_{12,4} - r_{13,4} \cdot r_{23,4}}{\sqrt{\left(1 - r_{13,4}^2\right) \cdot \left(1 - r_{23,4}^2\right)}}$$

Il lavoro aumenta rapidamente col numero delle variabili da eliminare. Per eliminare **s** *variabili, il numero delle operazioni necessario, ciascuna variabile importando l'applicazione della stessa formula, è*

$$\frac{s \cdot (s+1) \cdot (s+2)}{6}$$

Per valori di **s** *da* ***1*** *a* ***6****, occorrono perciò* ***1***, ***4***, ***10***, ***20***, ***35***, ***56*** *operazioni.*

Gran parte di questa fatica può essere risparmiata usando le tavole di $\sqrt{1-r^2}$ *quali quelle pubblicate da J. R. Miner.*

Come le variabili indipendenti nella regressione, le variabili eliminate nell'analisi della correlazione non comparvero distribuite, anche approssimativamente, in distribuzioni normali. Del pari e questo è assai di frequente trascurato, errori casuali in esse introducono errori sistematici nei risultati. Per esempio, se la correlazione parziale delle variabili (1) e (2) fosse realmente zero, dimodochè $\boldsymbol{r_{12}}$ *fosse uguale a* $\boldsymbol{r_{13} \cdot r_{23}}$*, errori casuali nella misura o nella valutazione della variabile (3) tenderebbero a ridurre numericamente* $\boldsymbol{r_{12}}$ *e* $\boldsymbol{r_{23}}$ *in modo da rendere il loro prodotto numericamente minore di* $\boldsymbol{r_{12}}$*. Un'apparente correlazione parziale fra le prime due variabili sarà, perciò, prodotta da errori casuali nella terza.*

Mentre le correlazione parziale $r_{12,3}$ valuta gli effetti tra due variabili (X_1 e X_2) dopo che entrambe sono state aggiustate per la regressione con la terza variabile (X_3), per cui essa è la correlazione tra due variabili aggiustate ($X_{1,3}$ e $X_{2,3}$),

la **correlazione semiparziale** (***semipartial or part correlation***) $r_{1(2,3)}$

- valuta **la correlazione tra la variabile X_1 e la variabile X_2,**
- **dopo che solo la X_2 è stata aggiustata per la variabile X_3** (indicata con $X_{2,3}$).

La correlazione semi-parziale è quindi la correlazione tra una variabile non aggiustata (X_1) ed una seconda variabile aggiustata per la terza ($X_{(2,3)}$)

ed è calcolata con

$$r_{1(2,3)} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{23}^2}}$$

Un esempio di correlazione parziale riportato nei testi è la correlazione tra la perdita di peso in un gruppo N pazienti e il tempo che ognuno di essi dedica alla ginnastica, aggiustato per il consumo di calorie che l'esercizio, più o meno faticoso, comporta. Nella ricerca ambientale può essere il caso di salinità, ossigeno e temperatura

Con 3 variabili (X_1, X_2, X_3) sono teoricamente possibili 6 correlazioni parziali: oltre alla precedente $r_{1(2,3)}$, si hanno

$$r_{1(3,2)} = \frac{r_{13} - r_{12} \cdot r_{32}}{\sqrt{1 - r_{32}^2}}$$

$$r_{2(1,3)} = \frac{r_{21} - r_{23} \cdot r_{13}}{\sqrt{1 - r_{13}^2}}$$

ed in modo analogo le altre tre $r_{2(3,1)}$, $r_{3(1,2)}$, $r_{3(2,1)}$

anche se spesso hanno reale significato ecologico od ambientale solo una o due di esse. Non è quindi necessario calcolarle tutte, ma è bene limitare l'analisi a quelle che possono essere interpretate.
Anche nella correlazione semiparziale, una variabile può essere aggiustata per un numero più alto di variabili, non solo per una terza.

La correlazione semiparziale di secondo ordine $r_{1(2,34)}$ può essere stimata

- sia partendo dalle correlazioni parziali di primo ordine con

$$r_{1(2,34)} = \frac{\left(r_{12,3} - r_{14,3} \cdot r_{24,3}\right) \cdot \sqrt{1 - r_{13}^2}}{\sqrt{1 - r_{24,3}^2}}$$

- sia partendo da quella di secondo ordine con

$$r_{1(2,34)} = r_{12,34} \frac{\sqrt{1 - r_{13}^2}}{\sqrt{1 - r_{14,3}^2}}$$

Le correlazioni parziali e semi-parziali hanno la loro applicazione principale nello studio delle **inter-relazioni lineari** che esistono fra tre o più variabili. E' un concetto di estrema importanza nella teoria della **regressione multipla**, che è alla base della statistica multivariata.
Come dimostrato nella esposizione, Fisher attribuiva molta importanza a questi metodi di correlazione parziale, pure avvisando degli effetti determinati dagli errori casuali. Nella statistica moderna, come suggerito da vari autori, si preferisce ricorre alla regressione multipla.

18.11. ANALISI DELLA COVARIANZA PER DUE GRUPPI, CON TEST t DI STUDENT PER RETTE PARALLELE E PER RETTE NON PARALLELE

Per confrontare due medie, si ricorre al test t di Student. Per due o più, all'ANOVA.
L'uso del test t e dell'ANOVA richiedono che i due o più gruppi abbiano **medie uguali prima** dell'esperimento. Solamente se questa condizione è vera, gli effetti dei trattamenti possono essere misurati dalle **differenze riscontrate** tra le medie dei due o più gruppi **dopo** l'esperimento.
Per esempio, se si vogliono confrontare gli effetti di due o più farmaci sulla capacità respiratoria di cavie, i gruppi devono essere costruiti in modo tale da essere simili e quindi dare la stessa risposta media, prima della somministrazione del principio attivo. Pertanto, assegnare gli individui in modo casuale ai vari trattamenti, come richiede l'analisi totalmente randomizzata, ha lo scopo specifico di

rendere uguali gli effetti di tutti gli altri fattori che possono determinare differenze nei valori della capacità respiratoria oltre al farmaco, quali il sesso, l'età e/o la malattia.

In varie condizioni sperimentali, questo metodo non è possibile. Soprattutto per dati raccolti sul campo o per confronti tra popolazioni reali.
Per far risaltare il concetto con un esempio, si assuma di voler verificare se esiste una differenza significativa nella capacità respiratoria (Y espressa in litri) tra un gruppo di persone affette da malattie polmonari e uno di persone sane, che vivano nelle stesse condizioni. E' ovvio attendersi che i primi abbiano una media significativamente minore. Ma la capacità respiratoria è fortemente influenzata anche dall'età (X espressa in anni) e in persone adulte diminuisce con l'invecchiamento. Può quindi succedere che il gruppo di persone sane, scelte per il confronto in un gruppo ristretto, abbia un'età sensibilmente più avanzata di quella del gruppo degli ammalati. Ne deriverebbe, per l'effetto dell'età, che questi ultimi potrebbero avere una media di Y maggiore di quella dei primi; cioè che la capacità respiratoria media dei malati risulti significativamente maggiore di quella dei sani, contrariamente all'atteso e alla logica medica. Se invece il gruppo di sani fosse più giovane, la differenza tra la loro media e quella degli ammalati risulterebbe maggiore di quella realmente dovuta alla malattia.

Il confronto tra le medie dei volumi respiratori dei due o più gruppi deve quindi tenere in considerazione le età degli individui che li compongono: è **l'analisi della covarianza** (ANCOVA).

In quasi tutti i testi di statistica, questo argomento è trattato subito dopo il confronto tra due o più rette, poiché utilizza in buona parte gli stessi concetti e le stesse formule. In questo corso, è stato posto alla fine della parte dedicata alla statistica parametrica, come sua logica conclusione concettuale e metodologica. Infatti utilizza in modo congiunto
- sia il **test t** per il confronto tra due medie o il **test F** per in confronto tra k medie,
- sia la **regressione lineare**, per eliminare l'effetto del fattore di perturbazione o, con un concetto del tutto simile, per aggiustare i dati in funzione dell'altro effetto.

L'analisi della covarianza applicata a due gruppi può utilizzare il test t di Student o il test F. Il loro risultato è del tutto identico, come ripetutamente evidenziato nei capitoli precedenti, per la nota relazione

$$t_\nu = \sqrt{F_{1,\nu}}$$

Il test t offre il vantaggio di permettere confronti unilaterali, oltre a quelli bilaterali.

Il test può essere utilizzato sia quando **le due rette sono parallele**, sia quando esse **non lo sono**. Nell'applicazione successiva, che illustra contemporaneamente i concetti e i metodi applicati ad un esempio riportato da un testo internazionale (**Armitage** e **Berry**), è trattato prima il caso di due rette parallele poi quello di due rette non parallele.

Per illustrare, in modo semplice e nei minimi dettagli, tutti i passaggi logici e metodologici dell'**analisi della covarianza nel caso di due gruppi**, si supponga di voler confrontare la capacità respiratoria **media**

- di un gruppo di individui affetti da asma (campione 1) formato da 40 persone,
- con quella di un gruppo di controllo (campione 2), formato da 44 persone sane che vivono nelle stesse condizioni.

DUE RETTE PARALLELE

1) Dalle due distribuzioni campionarie di Y (capacità respiratoria espressa in litri) e di X (età in anni), devono essere ricavati i valori indicati nella prima colonna; con i dati raccolti, sono quelli riportati nelle altre tre colonne (Campione 1, Campione 2, Totale):

Calcoli preliminari	Campione 1	Campione 2	Totale
$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$	4.397	6.197	10594
$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$	26,58	20,61	47,19
$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\cdot\left(Y_i-\overline{Y}\right)$	-236,39	-189,71	426,10
$\overline{X}$	41,38	39,80	---
$\overline{Y}$	3,92	4,46	---
n_i	40	44	---

2) Da essi si ricavano i due coefficienti angolari, con le loro intercetta:

Rette	Campione 1	Campione 2
b	$\frac{-236,39}{4.397} = -0,0538$	$\frac{-189,71}{6.197} = -0,0306$
a	$3,92 - (-0,0538 \cdot 41,38) = 6,15$	$4,46 - (-0,0306 \cdot 39,80) = 5,68$
$\hat{Y}_i = a + b \cdot X_i$	$\hat{Y}_i = 6,15 + (-0,0538 \cdot X_i)$	$\hat{Y}_i = 5,68 + (-0,0306 \cdot X_i)$

3) Se, con il test di confronto già illustrato nel capitolo dedicato alla regressione, i due coefficienti angolari risultano uguali (cioè non significativamente differenti), si può stimare il **coefficiente angolare comune**

$$b = \frac{-426,10}{10.594} = -0,0402$$

ottenendo b = -0,0402.

4) Da esso è possibile

- ricavare la differenza (d) tra le due medie di Y ($\overline{Y}_1 - \overline{Y}_2$),
- considerando l'effetto della differente età media dei due gruppi ($\overline{X}_1 - \overline{X}_2$), come evidenziato nel grafico sottostante.

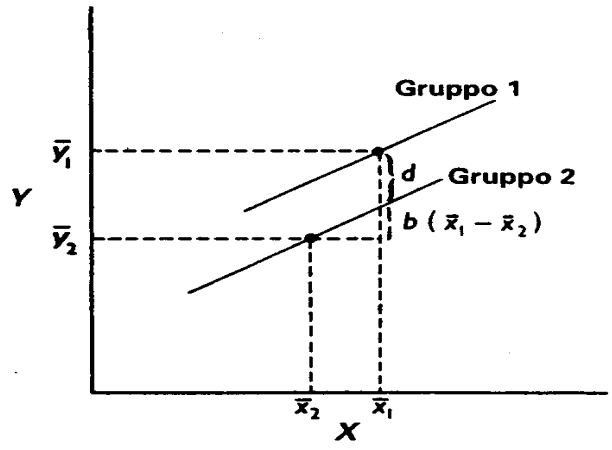
La distanza d tra le medie delle due Y (vedi sull'asse delle ordinate $\overline{Y}_1 - \overline{Y}_2$)

- è **aggiustata o ridotta** della quantità dovuta alla **differenza tra le due medie di X** (vedi sull'asse delle ascisse $\overline{X}_2 - \overline{X}_1$),
- **corretta per il coefficiente angolare comune** b ,

cioè della quantità

$$b \cdot \left(\overline{X}_1 - \overline{X}_2\right)$$

come evidenziato nella parte centrale del grafico successivo:



5) Di conseguenza questa distanza d determinata con formula generale che tiene i considerazione tutti i fattori enunciati mediante

$$d = \left(\overline{Y}_1 - \overline{Y}_2\right) - b \cdot \left(\overline{X}_1 - \overline{X}_2\right) =$$

$$d = \left(3{,}92 - 4{,}46\right) - \left(-0{,}0402\right) \cdot \left(41{,}38 - 39{,}80\right) = -0{,}540 + 0{,}064 = -0{,}476$$

risulta uguale a - 0,476

Per comprendere in modo chiaro l'operazione svolta, è utile **confrontare semplicemente le medie dei due gruppi riportate** nella tabella introduttiva e **ragionare su di esse**. Si evidenzia che

- la capacità respiratoria del gruppo 1, gli ammalati, è minore di quella del gruppo 2, cioè dei sani, di 0,54 litri (da 3,92 – 4,46);
- ma il gruppo di ammalati ha un'età media maggiore di 1,58 anni (da 41,38 –39,80);
- l'effetto sulla capacità respiratoria di questa maggiore età media è la perdita di litri 0,064 (come sono stimati da –0,0402 x 1,58 e dove –0,0402 è la perdita annuale);
- di conseguenza, la differenza reale nella capacità respiratoria del gruppo degli ammalati è 0,476 litri (derivando da 0,540 – 0,064).

6) Per le stime successive, è necessario calcolare

$$S_C^2 = \frac{\sum_{i=1}^{n1}\left(Y_{1i} - \overline{Y}_1\right)^2 + \sum_{i=1}^{n2}\left(Y_{2i} - \overline{Y}_2\right)^2 - \frac{\left(\sum_{i=1}^{n1}\left(X_{1i} - \overline{X}_1\right)\cdot\left(Y_{1i} - \overline{Y}_1\right) + \sum_{i=1}^{n2}\left(X_{2i} - \overline{X}_2\right)\cdot\left(Y_{2i} - \overline{Y}_2\right)\right)^2}{\sum_{i=1}^{n1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_{i=1}^{n2}\left(X_{2i} - \overline{X}_2\right)^2}}{n_1 + n_2 - 3}$$

Con i dati preliminari riportati nella tabella iniziale, la formula appare più semplice:

$$S_C^2 = \frac{26{,}58 + 20{,}61 - \frac{\left[\left(-236{,}39\right) + \left(-189{,}71\right)\right]^2}{4.397 + 6.197}}{40 + 44 - 3}$$

soprattutto con i totali

$$S_C^2 = \frac{47{,}19 - \frac{\left(-426{,}10\right)^2}{10.594}}{40 - 44 - 3} = 0{,}371$$

e si perviene al risultato $S_C^2 = 0{,}371$

7) Infine, per verificare la significatività della differenza nella capacità respiratoria media tra i due gruppi, ridotta della differenza media tra le età, si calcola il valore del t di Student

$$t_{(N-3)} = \frac{d}{\sqrt{S_C^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{\left(\overline{X}_1 - \overline{X}_2\right)^2}{\sum_{i=1}^{n1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_{i=1}^{n2}\left(X_{2i} - \overline{X}_2\right)^2}\right)}}$$

che con i dati dell'esempio

$$t_{(84-3)} = \frac{-0{,}476}{\sqrt{0{,}371 \cdot \left(\frac{1}{40} + \frac{1}{44} + \frac{\left(41{,}38 - 39{,}86\right)^2}{4.397 + 6.197}\right)}}$$

$$t_{(81)} = \frac{-0{,}476}{\sqrt{0{,}371 \cdot 0{,}04793}} = \frac{-0{,}476}{0{,}133} = -3{,}579$$

risulta **t = - 3,579** con gdl = 81.

Le tabelle del t di Student difficilmente riporta il valore critici del t con gdl = 81.

Si sceglie il valore inferiore riportato, come gdl = 80 per il quale in un test bilaterale il valore critico riportato con $\alpha = 0.005$ è 3,416.

Il valore calcolato è superiore e quindi si rifiuta l'ipotesi nulla, alla probabilità stimata.

8) Nella scelta del valore critico e nella formulazione dell'ipotesi, anche in questo tipo di test t occorre porre attenzione alla **direzione dell'ipotesi** alternativa. Tutto il problema ha voluto evidenziare non genericamente se le due medie sono differenti; ma se effettivamente, come logica medica vuole, la capacità respiratoria degli ammalati, a parità di età, è affettivamente minore di quella dei sani. E' quindi un test unilaterale: **più logico e più potente,** in quando dimezza la probabilità calcolata rispetto al test a due code. Si rifiuta l'ipotesi nulla, con probabilità P minore.

DUE RETTE NON PARALLELE

Il metodo precedente è valido quando le due rette sono parallele. Tradotto in termini di fisiologia, cioè con una lettura disciplinare come sempre occorrerebbe fare nell'analisi statistica, questa assunzione significa che la differenza nella capacità respiratoria dei sani e di quella degli ammalati si mantiene costante al variare dell'età. In realtà l'esperienza, in questo caso confermata da una lettura attenta dei dati, evidenzia che

- il coefficiente angolare degli ammalati (Campione 1) è b = - 0,0538
- il coefficiente angolare dei sani (Campione 2) è b = - 0,0306.

Con l'avanzare dell'età, la perdita di capacità respiratoria annuale degli ammalati (0,0538) è maggiore di quella dei sani (0,0306).

Sotto l'aspetto medico può apparire un risultato più logico dell'assunzione di parallelismo, anche se l'analisi statistica può non risultare significativa, a causa dei vari fattori aleatori che possono incidere su di essa, come un campione troppo piccolo, una grande variabilità tra gli individui che formano i due campioni, uno squilibrio per fattori non presi in considerazione in questa analisi quale il sesso, ecc. …

Può quindi essere utile effettuare il confronto non più tra le età medie ma tra età differenti, più giovani oppure più anziane. E' utile spesso fare **il confronto per valori specifici della covariata**. Per esempio si assuma di voler eseguire il confronto tra persone di 60 anni.

1) Definita l'età del confronto $X_k = 60$

la differenza d_k nella capacità respiratoria a quell'età

è

$$d_k = \left(\overline{Y}_1 - \overline{Y}_2\right) + b_1\left(X_k - \overline{X}_1\right) - b_2\left(X_k - \overline{X}_2\right)$$

Con i dati dell'esempio

$\overline{Y}_1 = 3{,}92$ $\overline{Y}_2 = 4{,}46$ $\overline{X}_1 = 41{,}38$ $\overline{X}_2 = 39{,}80$ $b_1 = -0{,}0538$ $b_2 = -0{,}0306$

si ottiene

$$d_{60} = \left(3{,}92 - 4{,}46\right) + (-0{,}0538) \cdot \left(60 - 41{,}38\right) - (-0{,}0306) \cdot \left(60 - 39{,}80\right)$$

$$d_{60} = \left(-0{,}54\right) + \left(-1{,}00\right) - \left(-0{,}62\right) = -0{,}92$$

che la differenza da 0,54 è salita a litri 0,92 sempre a svantaggio degli ammalati: $d_{60} = -0{,}92$

2) Si modifica anche **l'errore standard di questa differenza**.

Invece della formula con la quale si impiegano le due età medie

$$\sqrt{S_C^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{\left(\overline{X}_1 - \overline{X}_2\right)^2}{\sum_{i=1}^{n1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_{i=1}^{n2}\left(X_{2i} - \overline{X}_2\right)^2} \right)}$$

in questo caso **si deve utilizzare**

$$\sqrt{S_C^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{\left(X_k - \overline{X}_1\right)^2}{\sum_{i=1}^{n1}\left(X_{1i} - \overline{X}_1\right)^2} + \frac{\left(X_k - \overline{X}_2\right)^2}{\sum_{i=1}^{n2}\left(X_{2i} - \overline{X}_2\right)^2} \right)}$$

Con i dati dell'esempio precedente ai quali vanno aggiunti quelli delle due devianze di X

Calcoli preliminari	Campione 1	Campione 2
$\sum_{i=1}^{n}(X_i - \overline{X})^2$	4.397	6.197

si ottiene

$$\sqrt{0{,}371 \cdot \left(\frac{1}{40} + \frac{1}{44} + \frac{(60-41{,}38)^2}{4.397} + \frac{(60-39{,}80)^2}{6.197} \right)}$$

$$\sqrt{0{,}371 \cdot (0{,}025 + 0{,}02273 + 0{,}07885 + 0{,}06584)} = \sqrt{0{,}371 \cdot 0{,}1924} = 0{,}267$$

un **errore standard** uguale a 0,267

3) Infine si stima **il valore t**

$$t_{(81)} = \frac{-0{,}92}{0{,}267} = -3{,}445$$

che risulta **t = - 3,445** con gdl = 81.

Si può ugualmente rifiutare l'ipotesi nulla alla probabilità P < 0.005 soprattutto se il confronto è unilaterale.

4) Per meglio comprendere il metodo, quindi ai fini di un suo uso corretto ed utile, è importante sottolineare che:

- la scelta dell'età di confronto non deve uscire dal campo di variazione sperimentale delle X, poiché entrerebbe in discussione la validità della retta stimata;
- come era logico attendersi, all'aumentare dell'età **è aumenta la differenza** tra le due capacità respiratorie medie;
- ma, allontanandosi dalla media, è aumentato anche l'errore standard della differenza, come bene evidenzia la formula ($X_k - \overline{X}$);
- ne consegue che la differenza è risultata meno significativa del semplice confronto tra le due medie; considerato in valore assoluto il valore di t con le due medie è stato t_{81} = 3,579 mentre confrontando la capacità media stimata per l'età di 60 anni è stato t_{81} = 3,445

Non sempre si ottiene questo peggioramento. Oltre che dalla distanza dalla media, dipende dal segno del coefficiente angolare, dal suo valore in modulo, dalla distanza dell'età scelta per il confronto rispetto alla media dei due gruppi.

18.12. ANALISI DELLA COVARIANZA PER K GRUPPI (ANCOVA) E RIDUZIONE PROPORZIONALE DELLA VARIANZA D'ERRORE

Con k gruppi, il caso più semplice è quello di un disegno sperimentale completamente randomizzato. Ricorrendo alla simbologia consueta, mentre nell'analisi della varianza a un criterio di classificazione il modello è

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

nell'analisi della covarianza diviene

$$Y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij}$$

Nell'analisi della covarianza è possibile utilizzare una **variabile concomitante X**, chiamata **covariata**, che spiega parte della variabilità della variabile dipendente Y. Consente un calcolo più preciso degli effetti dei trattamenti, riducendo la variabilità casuale, cioè non controllata.
Per esempio, si supponga di voler stimare l'effetto di tre tossici (α_i) sul peso (Y_i) delle cavie. L'analisi della varianza ad un solo criterio di classificazione richiede che si formino tre gruppi, inizialmente identici per caratteristiche degli individui; dopo un periodo di somministrazione del principio attivo, le differenze tra i pesi medi dei tre gruppi permettono di valutare l'effetto dei tossici sulla crescita ponderale delle cavie. Ma avere tre gruppi di cavie con le stesse dimensioni iniziali spesso è difficile, soprattutto se l'esperimento non avviene in laboratorio ma in natura, con animali catturati.

Un altro caso sovente citato nei testi di statistica applicata è il confronto tra il peso (Y_i) di organi di animali sottoposti a trattamenti (α_i) diversi, in cui il peso dell'organo dipende sia dall'effetto del trattamento che dalle dimensioni (X_i) dell'animale. E' possibile eliminare dal peso dell'organo quella quantità che dipende delle dimensioni dell'animale [$\beta(x_{ij} - \bar{x})$], per valutare in modo più preciso la quota che può essere attribuita al trattamento.
Si parla di analisi ***after the fact***, che permette di valutare quale sarebbe stato il risultato, se la variabile concomitante X fosse stata costante per tutti i gruppi.

Attualmente, poiché i calcoli vengono eseguiti dai programmi informatici, **è diventato importante capire più che eseguire**. Il test richiede alcuni passaggi logici, che possono essere riassunti in **5 punti**.

1 - Si applica l'**analisi della varianza alle Y**, per verificare l'ipotesi nulla

$$H_0: \ \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

se le medie delle popolazioni dalle quali sono estratti i k campioni a confronto sono uguali, con ipotesi alternativa

H_1 che non tutte le μ sono uguali.

Con le formule abbreviate abituali si calcolano:

- la **devianza totale delle Y**

$$\sum Y_{ij}^2 - \frac{\left(\sum Y_{ij}\right)^2}{n} \qquad \text{con df } \mathbf{n - 1}$$

- la **devianza tra trattamenti delle Y**

$$\sum \frac{\left(\sum Y_i\right)^2}{n_i} - \frac{\left(\sum Y_{ij}\right)^2}{n} \qquad \text{con df } \mathbf{k - 1}$$

- e, per differenza, la **devianza d'errore delle Y**

Dev. Y totale - Dev. Y tra tratt. con df **n - k**

dove

- n = numero totale di dati,
- n_i = numero di dati di ogni gruppo,
- k = numero di gruppi a confronto.

Stimata la varianza tra trattamenti e la varianza d'errore, si perviene al test F

$$F_{(k-1,\ n-k)} = \frac{s^2 tratt.}{s^2 errore}$$

che **raramente risulta significativo, se l'effetto della regressione di Y su X è elevato**.

2 - Per correggere i dati calcolati, si valuta **l'effetto della regressione** ricordando che è la sua devianza è stimata dal rapporto

$$\frac{Cod._{XY}^2}{Dev._X}$$

A questo scopo si devono calcolare:

- la **codevianza XY totale**

$$\sum\left(X_{ij}\cdot Y_{ij}\right)-\frac{\left(\sum X_{ij}\right)\cdot\left(\sum Y_{ij}\right)}{n}$$

- la **codevianza XY tra trattamenti**

$$\sum\frac{\left(\sum X_{ij}\cdot\sum Y_{ij}\right)}{n_i}-\frac{\left(\sum X_{ij}\right)\cdot\left(\sum Y_{ij}\right)}{n}$$

- e, per differenza, la **codevianza XY d'errore**

Cod. XY totale - Cod. XY tra trattamenti

(A questo proposito, è importante ricordare che le **Codevianze possono risultare sia positive che negative**. Per esempio, con una **Codevianza XY totale** positiva si può ottenere anche una **Codevianza XY tra trattamenti** che risulti negativa; di conseguenza, la **Codevianza XY d'errore** può risultare maggiore di quella totale.)

3 - Sempre per stimare l'effetto della regressione, è preliminare effettuare il calcolo delle devianze delle X, con le stesse modalità seguite per la Y:

- la **devianza totale delle X**

$$\sum X_{ij}^2-\frac{\left(\sum X_{ij}\right)^2}{n}$$

- la **devianza tra trattamenti delle X**

$$\sum\frac{\left(\sum X_i\right)^2}{n_i}-\frac{\left(\sum X_{ij}\right)^2}{n}$$

- e, per differenza, la **devianza d'errore delle X**

Dev. X totale - Dev. X tra tratt.

4 - Con i valori calcolati al punto 2 e al punto 3, si stimano le **Devianze dovute alla regressione** del coefficiente **b** comune; servono solo la devianza totale e quella d'errore, ottenute da:

- devianza totale dovuta alla regressione

$$\frac{(Cod.XY.totale)^2}{Dev.X.totale}$$

- e la **devianza d'errore dovuta alla regressione**

$$\frac{(Cod.XY.d'errore)^2}{Dev.X.d'errore}$$

- la devianza tra trattamenti dovuta alla regressione non serve, per il motivo che spiegato nel passaggio successivo

5 - E' così possibile ottenere le **devianze delle Y ridotte** o devianze dovute alle deviazioni della regressione, sottraendo alla devianza totale e alla devianza d'errore delle Y, calcolate al punto 1, quelle rispettive calcolate al punto 4.

Si stimano:

- la devianza totale delle Y ridotte

Dev. Y totale - Dev. Y totale della regressione con df **(n - 1) - 1**

(In questa operazione, che trasferisce il confronto dei singoli valori della Y dalla media generale alla retta di regressione comune, si **perde un altro df**)

- e **la devianza d'errore delle Y ridotte**

Dev. Y d'errore - Dev. Y d'errore della regressione con df **(n - k) - 1**

(Come in precedenza, rispetto alla devianza d'errore delle Y, calcolata al punto 1, ha 1 df in meno)

La **devianza tra trattamenti delle Y ridotte** è ottenuta per differenza tra queste due immediatamente precedenti:

Dev. delle Y ridotte totale - Dev. delle Y ridotte d'errore con df **k - 1**

che mantiene i suoi **df = k - 1**.

Questa stima della devianza tra trattamenti è ottenuta per differenza e non più per semplice sottrazione della devianza tra trattamenti che poteva essere calcolata al punto 4, in modo analogo, da quella tra trattamenti calcolata al punto 1 perché ne avrebbe stimato solamente una parte: infatti essa deve comprendere

- sia gli scostamenti delle medie di gruppo intorno a una retta di regressione calcolata per la variabilità tra i gruppi, cioè interpolate per le medie dei gruppi,
- sia la differenza tra le pendenze delle rette di regressione parallele, calcolate entro gruppi, con la pendenza della retta di regressione interpolata tra le medie di gruppo.

6 - Calcolate la varianza tra trattamenti e la varianza d'errore sulle Y ridotte, **il test F, che considera l'effetto della regressione sui valori delle Y,** è dato dal loro rapporto

$$F_{(k-1,\ n-k-1)} = \frac{s^2 tra.trattamenti.delle.Y.ridotte}{s^2 d'errore.delle.Y.ridotte}$$ con df **k - 1** e **(n - k) - 1**.

ESEMPIO 1. A tre gruppi di cavie sono state somministrate tre sostanze tossiche (A, B, C) che, alterando il metabolismo, determinano un forte aumento ponderale.
Poiché sono stati utilizzati animali di dimensioni diverse, per valutare correttamente gli effetti sul peso (Y) deve essere considerata anche la lunghezza (X) delle cavie.

TRATTAMENTI

A		B		C	
X	Y	X	Y	X	Y
25	18	15	18	19	16
23	16	12	15	21	19
19	13	17	20	18	18
24	16	11	12	17	15
21	14	19	22	19	17
---	---	16	18	---	---

(Il metodo non richiede che i tre gruppi abbiano lo stesso numero di osservazioni, essendo del tutto analogo all'analisi della varianza a un criterio di classificazione. Per facilitare il calcolo manuale, pesi ed altezze sono stati riportati in valori trasformati, che non modificano i risultati; inoltre, sempre per facilitare i calcoli, sono state scelti campioni molto piccoli, spesso insufficienti per un esperimento reale).

Con questi dati, effettuare l'analisi della varianza e della covarianza, per valutare compiutamente l'effetto delle tre sostanze sul peso finale delle cavie.

Risposta. Prima di procedere alle analisi, è sempre di elevata utilità una rappresentazione grafica dei dati e delle medie a confronto.

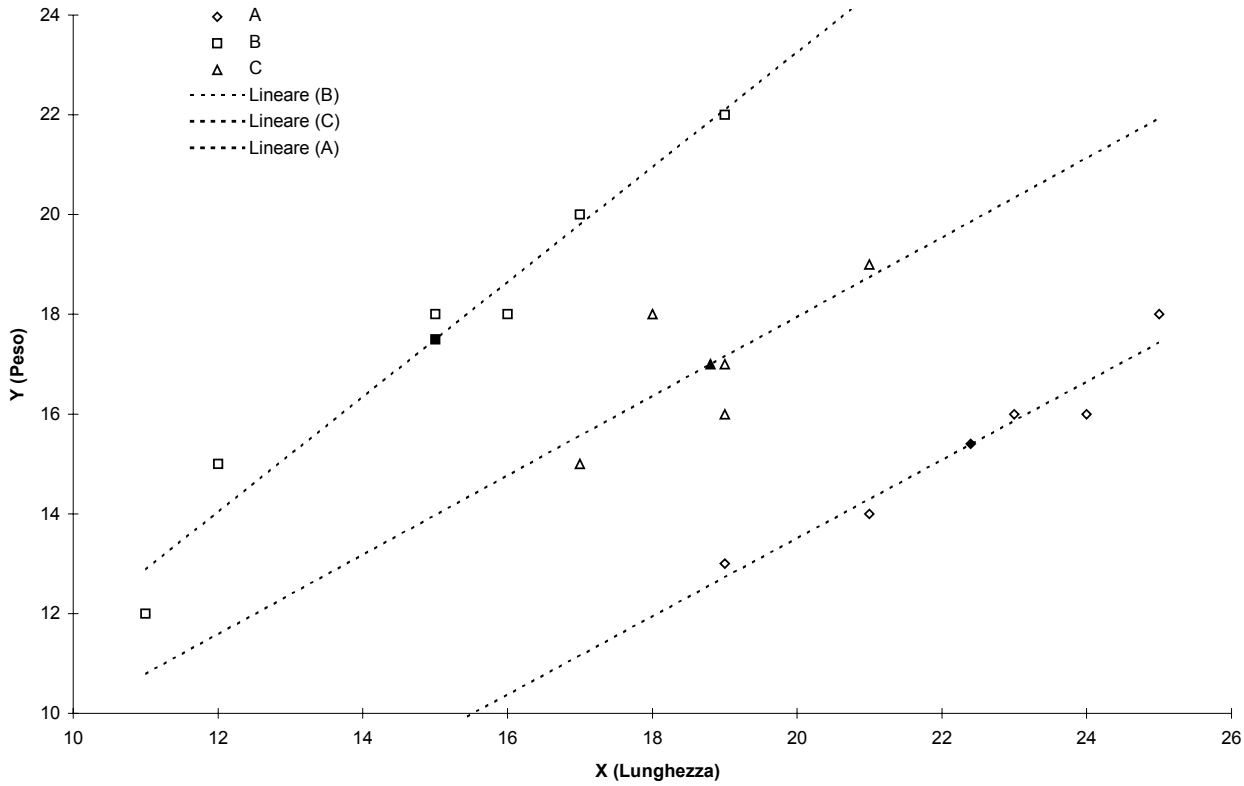
Il **diagramma di dispersione dei 3 gruppi** mostra che le differenze tra le tre medie dei valori campionari di Y sono ridotte e che la regressione lineare tra lunghezza X e peso Y per ogni gruppo è evidente, con coefficienti angolari simili.
Per l'interpretazione dei risultati e per i calcoli successivi con le formule abbreviate, è utile determinare preliminarmente le seguenti serie di somme e medie:

Somma $X_A = 112$	Somma $X_B = 90$	Somma $X_C = 94$	Somma $X = 296$
Somma $Y_A = 77$	Somma $Y_B = 105$	Somma $Y_C = 85$	Somma $Y = 267$
$n_A = 5$	$n_B = 6$	$n_C = 5$	$n = 16$

media $X_A = 22{,}40$	media $X_B = 15{,}00$	media $X_C = 18{,}80$	media $X = 18{,}50$
media $Y_A = 15{,}40$	media $Y_B = 17{,}50$	media $Y_C = 17{,}00$	media $Y = 16{,}6875$

Somma $X^2_A = 2532$	Somma $X^2_B = 1396$	Somma $X^2_C = 1776$	Somma $X^2 = 5704$
Somma $Y^2_A = 1201$	Somma $X^2_B = 1901$	Somma $X^2_C = 1445$	Somma $X^2 = 4557$

Somma $XY_A = 1743$	Somma $XY_B = 1628$	Somma $XY_C = 1605$	Somma $XY = 4976$



Seguendo lo stesso schema precedentemente descritto, i calcoli da effettuare possono essere raggruppati in 5 fasi.

1 - Per l'**analisi della varianza ad 1 criterio di classificazione sui valori di Y** (peso), si devono stimare i valori di

- la **devianza Y totale**

$$4557 - \frac{267^2}{16} = 4557 - 4455{,}56 = 101{,}56$$

che risulta uguale a 101,56 ed ha 15 df,

- la **devianza Y tra trattamenti**

$$\frac{77^2}{5} + \frac{105^2}{6} + \frac{85^2}{5} - \frac{267^2}{16} = 1185{,}8 + 1837{,}5 + 1445 - 4455{,}56 = 12{,}74$$

che è uguale a 12,74 ed ha 2 df

- la **devianza Y d'errore**

$$101{,}44 - 12{,}74 = 88{,}7$$

che è uguale a 88,7 ed ha 13 df (15-2).

I risultati possono essere presentati in una tabella

	Devianze Y	DF	Varianze	F	Prob.
Totale	101,44	15	---	---	---
Tra Tratt.	12,74	2	6,37	<1	Non stimata
Errore	88,70	13	6,823	---	---

dalla quale risulta evidente che il valore di F è inferiore a 1; pertanto, le differenze tra le medie campionarie non sono assolutamente significative.

2 - Per tenere in considerazione l'effetto della regressione, occorre calcolare le codevianze tra X e Y; quindi:

- la **codevianza XY totale**

$$4976 - \frac{296 \cdot 267}{16} = 4976 - 4939{,}5 = 36{,}5$$

che risulta uguale a **36,5**

- la **codevianza XY tra trattamenti**

$$\frac{112 \cdot 77}{5} + \frac{90 \cdot 105}{6} + \frac{94 \cdot 85}{5} - \frac{296 \cdot 267}{16} = 1724{,}8 + 1575 + 1598 - 4939{,}5 = -41{,}7$$

che ha un valore negativo (**- 41,7**)

- la **codevianza XY d'errore**

$$36{,}5 - (-41{,}7) = 78{,}2$$

che risulta maggiore di quella totale (**78,2**).

3 - Per procedere alle stime richieste è necessario calcolare anche le devianze di X:

- la **devianza X totale**

$$5704 - \frac{296^2}{16} = 5704 - 5476 = 228$$

che è uguale a **228**

- la **devianza X tra trattamenti**

$$\frac{112^2}{5} + \frac{90^2}{6} + \frac{94^2}{5} - \frac{296^2}{16} = 2508{,}8 + 1350 + 1767{,}2 - 5476 = 150$$

risulta uguale a **150,**

- la **devianza X d'errore**

$$228 - 150 = 78$$

uguale a **78**.

4 - Le devianze dovute alla **regressione b comune**, necessarie alla stima della Y ridotte, sono:

- la **devianza totale della regressione**

$$\frac{36{,}5^2}{228} = 5{,}84$$

che risulta uguale a **5,84**,

- **la devianza d'errore della regressione**

$$\frac{78{,}2^2}{78} = 78{,}4$$

che risulta uguale a **78,4**.

5 - In conclusione, **le devianze dovute alle deviazioni dalla regressione o devianze delle Y corrette** sono:

- la **devianza totale delle Y corrette**

$$101,44 - 5,84 = 95,6$$

uguale a **95,6** con **14** df a causa della **perdita di un altro df** dovuto alla correzione per la regressione (16 - 1 - 1),

- la **devianza d'errore delle Y corrette**

$$88,7 - 78,4 = 10,3$$

uguale a **10,3** con df **12** (anch'esso **perde un altro df,** poiché diventa **l'errore intorno alla retta),**

e, per differenza,

- la **devianza tra trattamenti delle Y corrette**

$$95,6 - 10,3 = 85,3$$

che risulta uguale a 85,3 con 2 df.

6 - Con questi ultimi dati, è possibile applicare l'analisi della varianza dei valori di Y che considerano l'effetto di regressione sulla X

	Devianze Y ridotte	DF	Varianze	F	Prob.
Totale	95,6	14	---	---	---
Tra tratt.	85,3	2	42,65	49,69	<0.0001
Errore	10,3	12	0,85833	---	---

e permettono un test F

$$F_{(2,12)} = \frac{42,65}{0,85833} = 49,69$$

che risulta altamente significativo.

ESEMPIO 2. Riprendendo i dati presentati da William L. **Hays** nel suo testo del 1994 (**Statistics**, 5th ed. Holt, Rinehart and Winston, Fort Worth, Texas), si assuma di confrontare cinque differenti modalità (A, B, C, D, E) di trasformazione farmacologica di un prodotto naturale di base (X), dal quale viene derivato il prodotto industriale finito (Y).

La tabella riporta la concentrazione della sostanza nel prodotto da trasformare (X) e in quello trasformato (Y)

A		B		C		D		D	
X	Y	X	Y	X	Y	X	Y	X	Y
10	18	22	40	30	38	35	25	11	15
20	17	31	22	31	40	37	45	16	17
15	23	16	28	18	41	41	50	19	20
12	66	17	31	22	40	30	51	25	23
57	77	86	121	101	159	143	171	71	75

Si vuole valutare se la produzione media del prodotto finito ($\overline{Y}_i$) nelle 5 aziende è significativamente differente, tenendo in considerazione le differenze presenti nella materia prima ($\overline{X}_i$).

Risposta. Presentando solo i calcoli da effettuare, i passaggi possono essere schematizzati in 4 punti.

1) Come nell'analisi della varianza a un criterio di classificazione o completamente randomizzata, si calcolano le devianze e i loro gdl, ovviamente per le **Y**.
Oltre ai totali di gruppo riportati nella tabella, dapprima si stimano

$$\sum_{i=1}^{n} Y_i^2 = 20.851 \qquad \sum_{i=1}^{n} Y_i = 603 \qquad n = 20$$

Con le solite formule delle devianze si ricavano:

- Dev. Y Totale $20.851 - \frac{603^2}{20}$ = 2.670,55 con gdl = 19

- Dev. Y Tra $\frac{\left(77^2 + 121^2 + 159^2 + 171^2 + 75^2\right)}{4} - \frac{603^2}{20}$ = 1998,80 con gdl = 4

- Dev. Y errore 2.670,55 – 1.988,80 = 671,75 con gdl = 15

2) Con le stesse modalità si calcolano le devianze delle X, poiché servono successivamente per correggere le devianze delle Y appena stimate.

Oltre ai totali di gruppo riportati nella tabella, si eseguono le somme

$$\sum_{i=1}^{n} X_i^2 = 12.066 \qquad \sum_{i=1}^{n} X_i = 458 \qquad n = 20$$

e da essi si ricavano:

- Dev. X Totale $12.066 - \frac{458^2}{20} =$ 1.577,8

- Dev. X Tra $\frac{\left(57^2 + 86^2 + 101^2 + 143^2 + 71^2\right)}{4} - \frac{458^2}{20} =$ 1095,8

- Dev. X errore 1.577,8 – 1.095,8 = 482,0

3) Per lo stesso scopo si stimano le codevianze; oltre ai totali di gruppo riportati nella tabella, si stimano

$$\sum_{i=1}^{n} X \cdot Y = 15.140 \qquad \sum_{i=1}^{n} X_i = 458 \qquad \sum_{i=1}^{n} Y_i = 603 \qquad n = 20$$

e da essi si ricavano:

- Codev. XY Totale $15.140 - \frac{458 \cdot 603}{20} =$ 1.331,30

- Codev. XY Tra $\frac{\left(57 \cdot 77 + 86 \cdot 121 + 101 \cdot 159 + 143 \cdot 171 + 71 \cdot 75\right)}{4} - \frac{458 \cdot 603}{20} =$ 1349,3

- Codev. XY errore 1.3321,30 – 1.349,30 = -18,0

(Osservare che con le codevianze, che possono essere negative, si mantiene la proprietà additiva; di conseguenza, quella tra trattamenti può essere maggiore di quella totale e quindi la codevianza d'errore risultare negativa)

4) Infine si ricavano le **devianze delle Y aggiustate**

$$\text{Dev. Y aggiustate} = Dev.Y - \frac{(Codev.XY)^2}{Dev.X}$$

ottenendo

- Dev. Y aggiustate Totale $2.670{,}55 - \frac{1.331{,}3^2}{1.577{,}8} =$ 1.547,24 con gdl = 18

(perde un altro gdl)

- Dev. Y aggiustate errore $671{,}75 - \frac{-18^2}{482} =$ 671,08 con gdl = 14

(perde anch'esso un altro gdl)

Da queste due, per differenza, si stima

- Dev. Y aggiustate Tra 1.547,24 – 671, 08 = 876, 16 con gdl = 4

E' utile riportare questi dati conclusivi delle devianze aggiustate nella solita tabella dell'ANOVA

Fonte di variazione	Dev. Agg.	DF	S^2	F	P
Totale	1547,24	18	---	---	---
Tra	876,16	4	219,04	4,57	< 0,025
Errore	671,08	14	47,93	---	---

In essa sono stati aggiunti il valore di F

$$F_{(4,14)} = \frac{219{,}04}{47{,}93} = 4{,}57$$

e la probabilità P < 0.025 che permette di rifiutare l'ipotesi nulla.

Infatti con DF 4 e 14 i valori critici riportati nelle tabelle sono

- F = 3,89 alla probabilità $\alpha = 0.025$
- F = 5,04 alla probabilità $\alpha = 0.01$

Allo scopo di valutare il vantaggio apportato dall'analisi della covarianza alla significatività delle differenze tra le medie della Y è utile fornire la stima della **riduzione proporzionale della varianza d'errore** r_e^2, dovuta alla correzione per la regressione comune:

$$r_e^2 = \left(\frac{Codev.errore.XY}{\sqrt{(Dev.errore.X)\cdot(Dev.errore.Y)}} \right)^2$$

Nell'ultimo esempio, con i dati

Codev. Errore di XY = -18 Dev. Errore della X = 482,0 Dev. Errore della Y = 671,75

si ottiene

$$r_e^2 = \left(\frac{-18}{\sqrt{482,0 \cdot 671,75}} \right)^2 = \left(\frac{-18}{569} \right)^2 = (-0,0316)^2 = 0,00099$$

che la correzione relativa è stata minore di 0.001, quindi totalmente trascurabile.

Nell'esempio 1, con

Codev. Errore di XY = 78,2 Dev. Errore della X = 78 Dev. Errore della Y = 88,7

si ottiene

$$r_e^2 = \left(\frac{78,2}{\sqrt{78 \cdot 88,7}} \right)^2 = \left(\frac{78,2}{83,18} \right)^2 = (0,94)^2 = 0,884$$

una **correzione relativa** che è superiore all'88 per cento. E' molto importante, tale da rendere significativo il test F sull'uguaglianza delle medie di Y, mentre prima non lo era.

A conclusione della dimostrazione sperimentale del metodo, è utile **rivedere i concetti di base della regressione e le sue condizioni di validità**.

Il caso presentato è **il modello dell'analisi della Covarianza per un solo fattore** (***Single-Factor Covariance Model***) del tutto simile all'analisi della varianza ad effetti fissi.

Con una simbologia leggermente modificata da quella sempre utilizzata, al solo scopo di rendere più facilmente leggibile la figura successiva, questo modello può essere scritto come

$$Y_{ij} = \mu + \tau_i + \gamma(X_{ij} - \overline{X}) + \varepsilon_{ij}$$

dove

- μ = media generale di tutte le Y,
- τ_i = effetto fisso del trattamento i: è la differenza tra la media del trattamento (μ_i) e la media generale (μ); pertanto deve esistere la relazione

$$\sum \tau_i = \sum (\mu_i - \mu) = 0$$

- γ = coefficiente angolare della retta di regressione, per la relazione generale esistente tra Y e X

Il valore di X_{ij}, detta **variabile concomitante**, è assunta come costante.

Gli errori (ε_{ij}) sono indipendenti e devono determinare una varianza costante lungo tutta la retta.

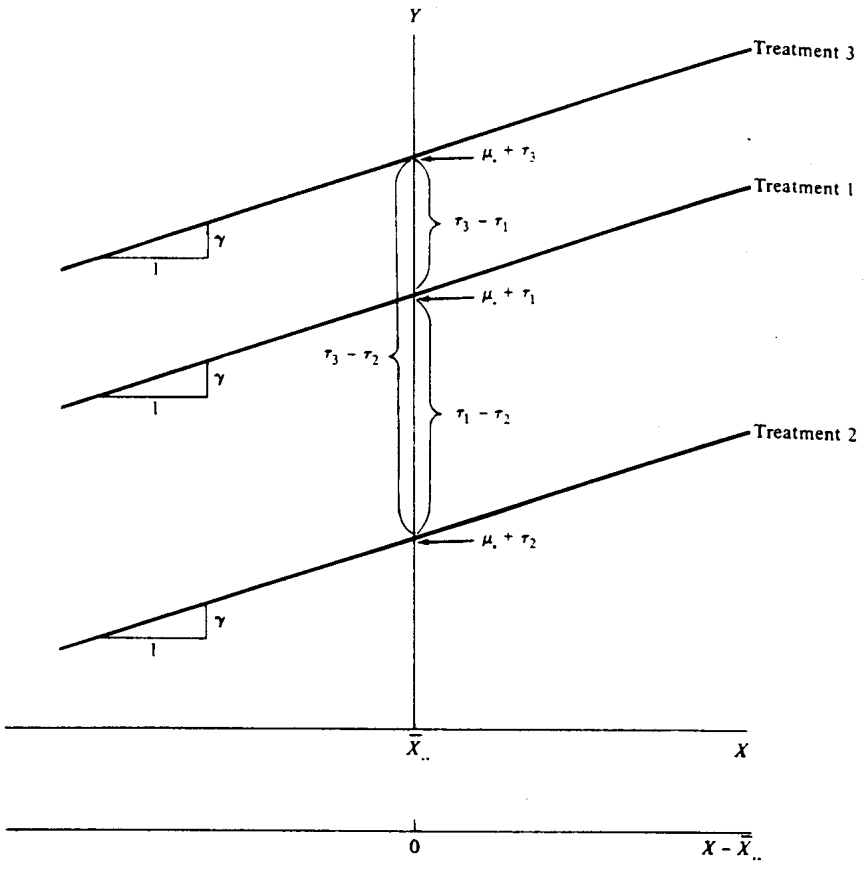
L'analisi della covarianza ha come ipotesi nulla

$$H_0: \ \tau_1 = \tau_2 = ... = \tau_k = 0$$

contro l'ipotesi alternativa

H_1: non tutte le τ sono uguali a 0

In una analisi della covarianza con 3 trattamenti, i vari concetti possono essere rappresentati graficamente come nella figura prossima:



In essa sono evidenziati gli aspetti più importanti dell'analisi della covarianza.

A) I valori delle Y nei tre trattamenti sono rappresentate lungo altrettante rette, tra loro parallele, che esprimono la relazione con la variabile concomitante X. Vale a dire, riprendendo l'esempio sulla relazione tra capacità respiratoria espressa in litri (Y_{ij}) e età (X_{ij}) di ogni individuo, la perdita in funzione dell'età (γ) deve essere uguale in ogni gruppo.

B) I vari gruppi a confronto possono avere medie delle X differenti ($\overline{X}_i$); il confronto viene attuato rispetto alla media generale ($\overline{\overline{X}}$), in modo tale che $\overline{X}_i - \overline{\overline{X}} = 0$;

Le **condizioni di validità** sono cinque:
1) la **normalità degli errori**,
2) l'**uguaglianza della varianza** per i vari trattamenti,
3) l'**uguaglianza del coefficiente angolare** delle rette di regressione dei vari trattamenti,
4) la **relazione di regressione** tra la variabile X e la variabile Y **deve essere lineare**,
5) **gli errori non devono essere correlati**.

La **terza assunzione**, quella che tutti i trattamenti devono avere lo stesso coefficiente angolare (nel testo sempre indicato con β, per segnalare che si tratta di quello della popolazione e quindi non considera le piccole differenze campionarie tra i b) **è cruciale**.
Se i coefficienti angolari sono differenti, nell'analisi statistica occorre separare i trattamenti. Per ottenere questo risultato, il metodo più semplice è quello illustrato per il caso di due campioni, nel paragrafo precedente.

Con un approccio differente dai metodi esposti, l'analisi della covarianza può essere affrontata anche con il calcolo matriciale. E' un argomento che rientra nella statistica multivariata, per la quale è necessaria una impostazione differente e più complessa di quella qui esposta. Ad essa si rimanda per approfondimenti ulteriori.

L'analisi della covarianza ha avuto ampi sviluppi, in **due direzioni** diverse.
- Da una parte, in analogia con l'analisi della varianza, permette di considerare contemporaneamente **vari fattori e le loro interazioni**, in riferimento a **una sola covariata**. Se l'analisi è limitata a due sole variabili, con la consueta simbologia il modello additivo è

$$Y_{ijk} = \mu + \alpha_i \gamma_j + \alpha\gamma_{ij} + \beta(X_{ijk} - \overline{X}) + \varepsilon_{ijk}$$

- Dall'altra, con la covarianza multipla si possono seguire **più covariate** (X_1 , X_2 ,…, X_n). Il modello più semplice, due covariate in una analisi della varianza ad 1 solo criterio, è

$$Y_{ijk} = \mu + \alpha_i + \beta_1(X_{1ij} - \overline{X}_1) + \beta_2(X_{2ij} - \overline{X}_2) + \varepsilon_{ijk}$$

Dalla loro combinazione, **più variabili e più covariate**, si ottengono modelli additivi che appaiono molto complessi. Superati con l'informatica i problemi di calcolo, attualmente i limiti alla complessità del disegno sperimentale sono posti solamente dall'interpretazione dei risultati, particolarmente difficile nel caso di più interazioni.

Per la trattazione dell'analisi della covarianza a disegni complessi, più variabili e più covariate con eventuale interazione, si invia a testi specifici.

18.13. GLI OUTLIER NELL'ANALISI DI REGRESSIONE E CORRELAZIONE

Nel caso della statistica bivariata, la **ricerca degli outlier assume una importanza** ancora maggiore di quella che ricopre nella statistica univariata:

- l'individuazione è **più complessa**, meno evidente alla semplice lettura del dato perché bidimensionale;
- soprattutto **gli effetti possono essere molto grandi** sui risultati della regressione e della correlazione, fino a invertire il segno della relazione;
- **i metodi diventano più sofisticati**, meno immediati nell'applicazione e meno intuitivi nei concetti.

Nella statistica univariata, gli outlier aumentano sempre la varianza e quindi riducono la significatività di un test. Nella regressione e nella correlazione, possono avere anche l**'effetto opposto di rendere molto significativi i test sulla linearità e sulla correlazione**, quando in realtà sulla base di tutti gli altri dati non si sarebbe rifiutata l'ipotesi nulla. Dipende dalla **collocazione dell'outlier**, rispetto alle altre coppie di valori.

In termini tecnici, oltre al ***masking effect*** (descritto nei paragrafi dedicati agli outlier nella statistica univariata), nella statistica bivariata si può avere un **importante *swamping effect***, cioè la capacità di sommergere l'informazione fornita complessivamente da tutte le altre coppie di dati. Un esempio è riportato nell'ultimo paragrafo dedicato alla discussione, nella statistica univariata, se gli outlier debbano essere compresi o esclusi nell'analisi statistica.

Per l'importanza che ricoprono, nella multivariata gli outlier hanno una letteratura molto ampia. Questa presentazione è limitata alle situazioni più semplici, con due variabili continue.

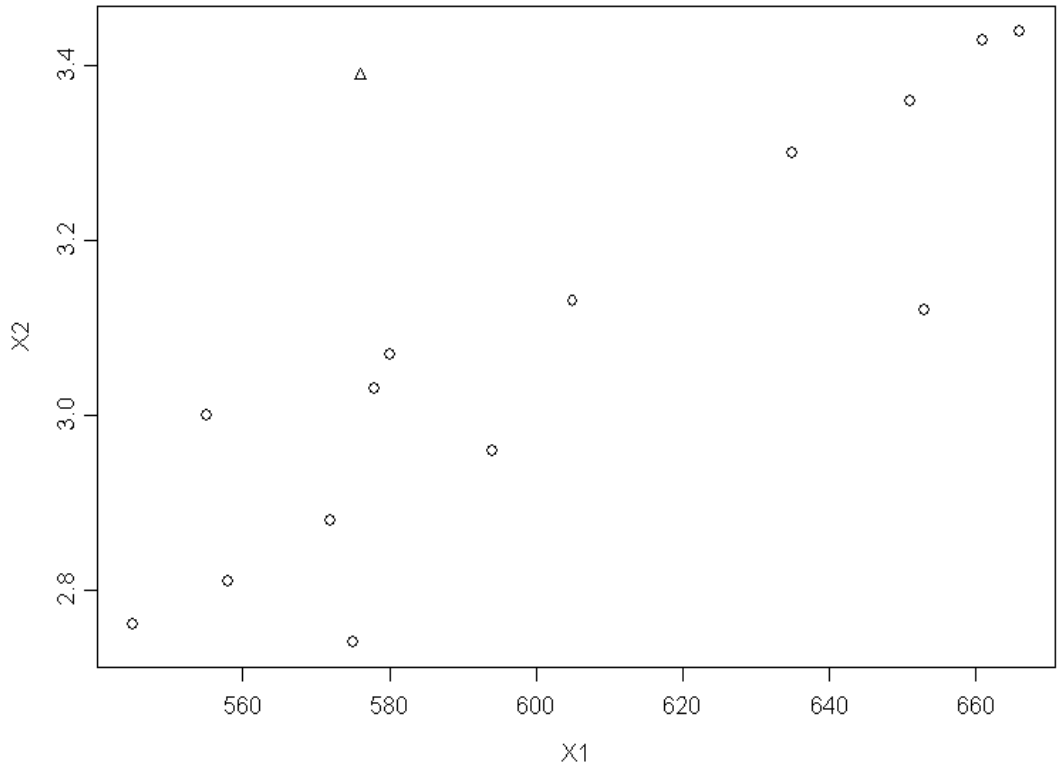
Come nella univariata, quando si sospetta la presenza di un outlier, è preliminare a qualsiasi analisi statistica l'**accertamento che non si tratti di un errore**, commesso in una delle tante fasi di elaborazione dell'informazione, dalla raccolta al trasferimento dei dati. Molto spesso **l'analisi combinata di due parametri ne facilita l'individuazione**. Ad esempio, non è possibile una gravidanza o indici fisiologici ad essa correlati, in donne troppo giovani o troppo anziane; richiede chiaramente una verifica, se un peso di 40 Kg è associato a una persona con altezza di 190 cm. Almeno una delle due variabili ha un valore errato, da correggere prima dell'analisi statistica.
Se è impossibile, è necessario eliminare la coppia di valori.

Quando la coppia di dati **sono le misure effettivamente ottenute nell'esperimento, per la ricerca dell'outlier si pone il problema di una loro valutazione in rapporto all'**ambiente statistico**, vale a dire all'**informazione fornita dagli altri dati**.**
Il **primo approccio,** semplice ma importante, alla verifica della eventuale **presenza di uno o più outlier** è sempre la rappresentazione grafica. In questo caso, al posto dell'istogramma, si utilizzano i punti in un **diagramma cartesiano,** detto **diagramma di dispersione**. L'anomalia di una coppia di dati risulta molto più evidente di quanto appaia alla semplice lettura delle due variabili, poiché
- separatamente sia il valore di X sia quello di Y possono rientrare nella distribuzione degli altri valori,
- ma congiuntamente possono individuare un punto che è nettamente separato.

Ad esempio, in questa serie di 15 coppie di dati

X_1	576	635	558	578	666	580	555	661	651	605	653	575	545	572	594
X_2	3,39	3,30	2,81	3,03	3,44	3,07	3,00	3,43	3,36	3,13	3,12	2,74	2,76	2,88	2,96

che rappresentano la quantità di ammoniaca (X_1) e di solfati (X_2) presenti in altrettanti campioni di acqua inquinata, con una semplice lettura, per quanto attenta, è impossibile individuare il potenziale valore anomalo. La rappresentazione grafica mostra con sufficiente evidenza che il punto che si differenzia maggiormente dagli altri é individuato dalla prima coppia di valori (X_1 = 576 e X_2 = 3,39), rappresentato con un triangolo (in alto a sinistra) nel grafico.



La lettura di una sola dimensione avrebbe condotto a rilevare unicamente che

- 576 è tra i valori minori della variabile X_1 , ma che ne ha cinque minori,
- 3,39 è tra i valori maggiori della variabile X_2 , ma ne ha due maggiori.

Ma la identificazione di quel punto come **outlier,** seppure **visivamente evidente**, sotto **l'aspetto statistico** non è ovvia. Inoltre è sempre importante, quando si rifiuta l'ipotesi nulla

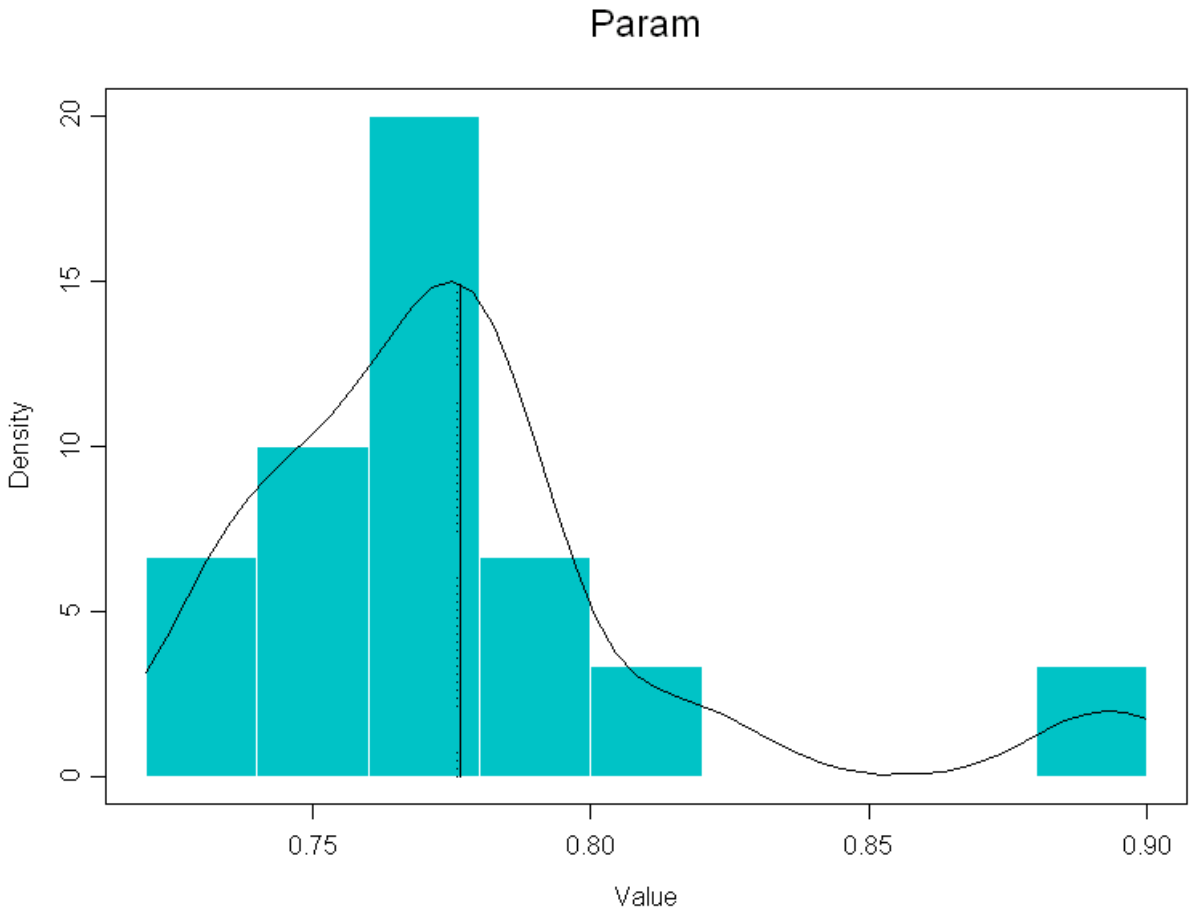
H_0: il punto X_i, Y_i non è un outlier contro **H_1: il punto X_i, Y_i è un outlier**

e quindi si decide che si tratta di un valore anomalo,

- **conoscere la probabilità α di commettere un errore di I tipo**, almeno in modo approssimato.

I metodi statistici proposti sono numerosi. Non tutti portano alle stesse conclusioni, quando la situazione non è del tutto palese.

In alcuni casi, possono essere impiegati anche metodi che in realtà hanno altre finalità.



Distribuzione dei 15 pseudo-valori di **r** calcolati dalle 15 coppie di dati campionari.
(Le ordinate sono state moltiplicate per 3 per motivi grafici.

Ad esempio, una delle tecniche più recenti ed efficaci, quando si devono utilizzare dati che hanno una forma della distribuzione ignota o comunque non normale, è il **jackknife.**

Come descritto ampiamente nel capitolo in cui è riportato, quando è applicato alla correlazione, calcola tanti valori r di correlazione quante sono le coppie di dati, escludendone ogni volta una.
In questo caso, con la sua applicazione ai dati della tabella e come riportato nell'ultimo grafico,
- sono stati calcolati i 15 valori di r riportati in ascissa,
- mentre sull'asse delle ordinate è riportata la loro frequenza.

Il valore di correlazione di circa 0,90 è il risultato del **jackknife** quando nelle $n-1$ coppie di dati non è compreso il punto anomalo già sospettato. Il confronto con l'istogramma, collocato a sinistra della figura e formato da tutti gli altri valori r che comprendono quel punto, evidenzia gli effetti del punto

outlier sul valore di r. Questo r cade lontano dalla media degli altri r ed è fuori dal loro intervallo di confidenza.

Ma se

- quel valore r di correlazione è un outlier, rispetto agli altri r,
- anche il punto X_i, Y_i che lo determina è un outlier, rispetto agli altri punti.

Per una analisi statistica **ampia e ragionata** degli outlier, **è utile conoscere l**'impostazione classica,

- **sia per una** maggiore disponibilità di metodologie **da applicare alle varie situazioni,**
- **sia per** giustificare scelte differenti**, come il metodo jackknife nell'esempio precedente.**

Il punto di partenza della metodologia classica **o tradizionale distingue se sui dati raccolti si utilizza**

- **la regressione lineare semplice oppure la correlazione semplice.**

Infatti

- **nel modello di** regressione lineare**, gli outlier sono analizzati esclusivamente per la** variabile dipendente**, spesso indicata con Y,**
- mentre nel modello di **correlazione lineare**, gli outlier sono individuati analizzando congiuntamente le **due variabili X_1 e X_2**.

In realtà, i due gruppi di metodi sono applicati agli stessi dati, trattandosi sempre di statistica bivariata. Inoltre esistono relazioni strette tra il coefficiente angolare b e il coefficiente di regressione lineare r.

Nella **regressione**, dove la variabile X serve per stimare la variabile Y, la tecnica per evidenziare gli outlier è **sempre** basata sull'**analisi dei residui** (***residuals*** o ***raw residuals***)

con r_i,

$$r_i = Y_i - \hat{Y}_i$$

vale a dire sulle differenze tra ogni Y_i osservata e la corrispondente $\hat{Y}_i$ calcolata per la stessa X_i.

18.14. L'ANALISI DEI RESIDUI PER L'IDENTIFICAZIONE DEGLI OUTLIER; RESIDUALS, STUDENTIZED RESIDUALS, STANDARDIZED RESIDUALS

Nella regressione, un **outlier** può essere definito come l'**osservazione che produce un residuo molto grande**. Alcune tecniche semplici sono riportate da

- James E. De **Muth** nel suo testo del 1999 ***Basic Statistics and Pharmaceutical Statistical Applications*** (edito da Marcel Dekker, Inc. New York, XXI + 596 p. a pag. 538 - 543).

Ad esso si rimanda per approfondimenti. I metodi sono presentati con lo sviluppo di un esempio, qui riportato con maggiori dettagli nei passaggi logici.

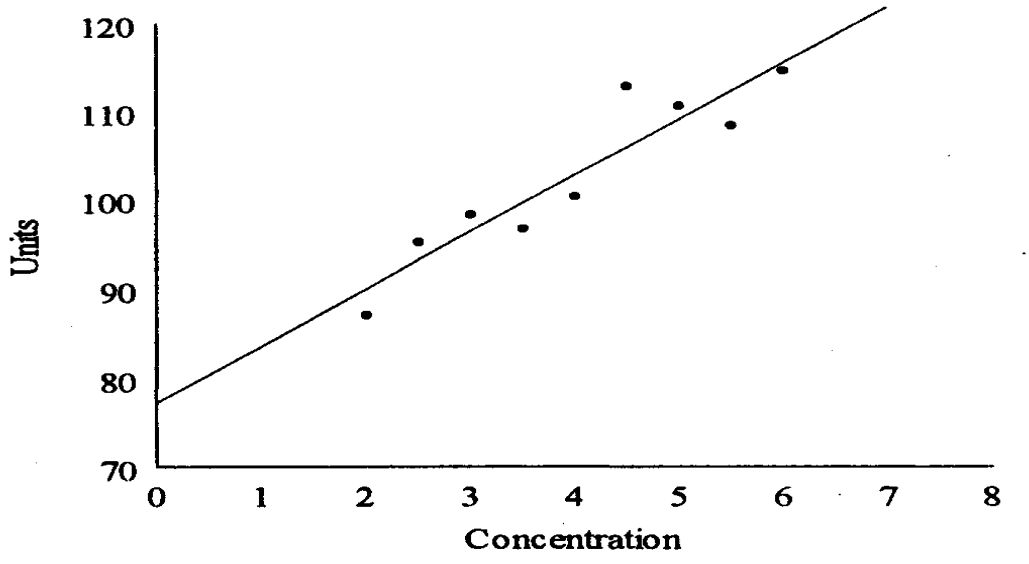
Nello studio di un fenomeno di fermentazione, in cui

- X_i rappresenta la concentrazione della sostanza, (prima riga),
- Y_i è la quantità fermentata nell'unità di tempo (seconda riga),
- $\hat{Y}_i$ è la quantità stimata sulla base della retta di regressione calcolata: $\hat{Y}_i = a + bX_i$ (terza riga),
- r_i indica lo scarto tra i due valori della variabile Y: $r_i = Y_i - \hat{Y}_i$ (quarta riga)

come i dati della tabella seguente

X_i	2,0	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0	$\sum$
Y_i	87,1	95,2	98,3	96,7	100,4	112,9	110,7	108,5	114,7	
$\hat{Y}_i$	89,980	93,165	96,350	99,535	102,720	105,905	109,090	112,275	115,460	
r_i	**-2,840**	**+2,035**	**+1,950**	**-2,835**	**-2,230**	**+6,895**	**+1,610**	**-3,735**	**-0,760**	**0,000**

l'analisi degli **outlier** richiede dapprima una lettura del **diagramma di dispersione dei punti osservati** X_i e Y_i (vedi pagina successiva), rispetto alla loro retta di regressione.



Ad occhio,

- il punto di coordinate X_i = 4,5 e Y_i = 112,9 che determina lo scarto maggiore (r_i = +6,895)
- non appare così distante dagli altri da poter essere giudicato un outlier.

Ma

- la distanza del punto Y_i = 112,9 dalla sua proiezione sulla retta $\hat{Y}_i$ = 105,905 non semplice da valutare. Soprattutto, per decidere, si richiedono test che permettano di **stimare la probabilità** α.

Come prima analisi, con i dati X_i e Y_i, è importante verificare la significatività della retta di regressione

$$\mathbf{H_0:}\ \beta = 0 \qquad \text{contro} \qquad \mathbf{H_1:}\ \beta \neq 0$$

Con il calcolo di F si ottiene:

Fonte	$S.Q.$	df	S^2	F	P
Totale	705,94	8	---	---	---
Regressione	608,65	1	608,65	43,78	< 0.001
Errore	97,29	7	13,90	---	---

La tabella dei risultati (F = 43,78 per df 1 e 7, con P < 0.001) dimostra che **linearità è altamente significativa**. Attraverso la **varianza d'errore** (S_e^2 = 13,90), è poi ricavabile **un valore** importante **per l'analisi dei residui**,

- **la deviazione standard degli errori (**S_e**),**

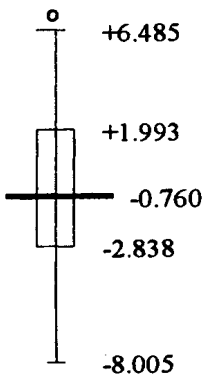
che in questo caso risulta

$$S_e = \sqrt{13,9} = 3,728$$

Nella figura precedente, a una valutazione ***occhiometrica***, il valore di Y per X = 4,5 non appariva molto distante dagli altri. Ma è un outlier oppure solo un valore estremo in una distribuzione normale? Utilizzando la retta di regressione (non sono riportati i suoi parametri) si calcolano

- i valori attesi ($\hat{Y}_i$ nella terza riga della tabella)
- e per differenza i residui o errori (r_i nella quarta riga della tabella; in molti testi indicati con e_i).

Con la serie dei residui r_i, si è ritornati a dati univariati. Quindi con essi sono possibili tutte le analisi già presentate per la statistica univariata, a partire dagli istogrammi e dal **Box and Wiskers** di **Tukey**.



Secondo i calcoli di Tukey e come appare in questa figura si ottiene una prima risposta:

- il cerchio vuoto che identifica **il residuo maggiore (+6,895) è un outlier,** in quanto è superiore al valore **VAS (+6,485)** o **cinta interna** o ***inner fence***. (Rivedere i paragrafi della univariata).

Ne consegue che **il punto corrispondente**, di coordinate (X = 4,5 e Y = 112,9), **è giudicato statisticamente un outlier**.

La sua presenza rimetterebbe in discussione la validità della regressione calcolata in precedenza e quindi la significatività dell'analisi, che richiedono la normalità della distribuzione degli errori.

Per facilitare la **lettura statistica** del **grafico dei residui,** è prassi utilizzare una loro **rappresentazione standard** che rimedia alle difficoltà precedenti, poiché è indipendente dalla collocazione (intercetta a) e dalla pendenza della retta (coefficiente angolare b).

In questo grafico (nella pagina successiva),

- la retta è sempre orizzontale, parallela all'asse delle ascisse sulla quale sono riportati i valori X_i ,
- mentre i valori dei residui sono letti sull'asse delle ordinate.

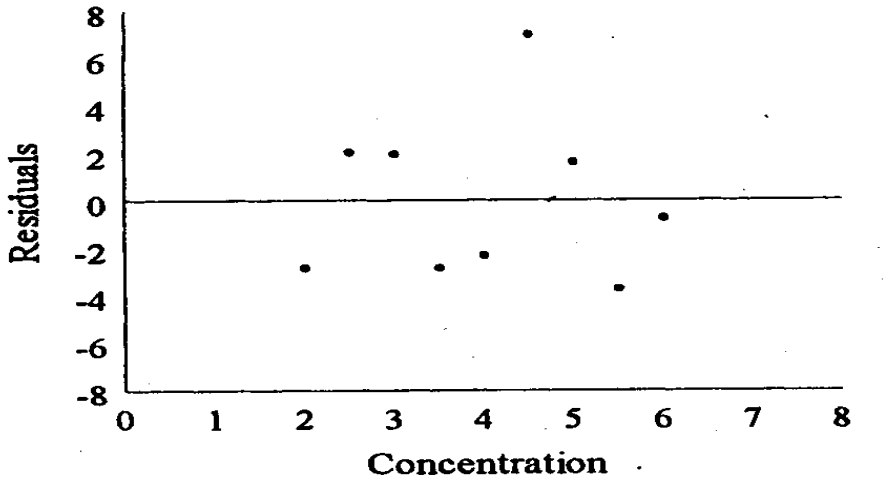
Diventa più semplice osservare che

- la **distanza del punto outlier** dalla retta orizzontale appare con evidenza maggiore,

rispetto al precedente diagramma di dispersione, costruito con i dati originali distribuiti intorno alla retta di regressione.

Con chiarezza ugualmente maggiore, risulta una **proprietà importante dei residui** (già rimarcata nella tabella):

- **la loro somma è uguale a zero**.



Un'altra **convenzione diffusa nell'analisi degli outlier**, in quanto facilita il confronto tra variabili diverse e casi differenti uniformando le dimensioni, è la trasformazione dei residui in **residui studentizzati** (***studentized residuals***).

Essa **rende uguale la scala di valutazione**,

attraverso la relazione

$$t = \frac{r_i}{\sqrt{S_e^2}}$$

Ad esempio,

riprendendo la tabella dei dati, il **primo residuo** (r_i = -2,840)

diventa

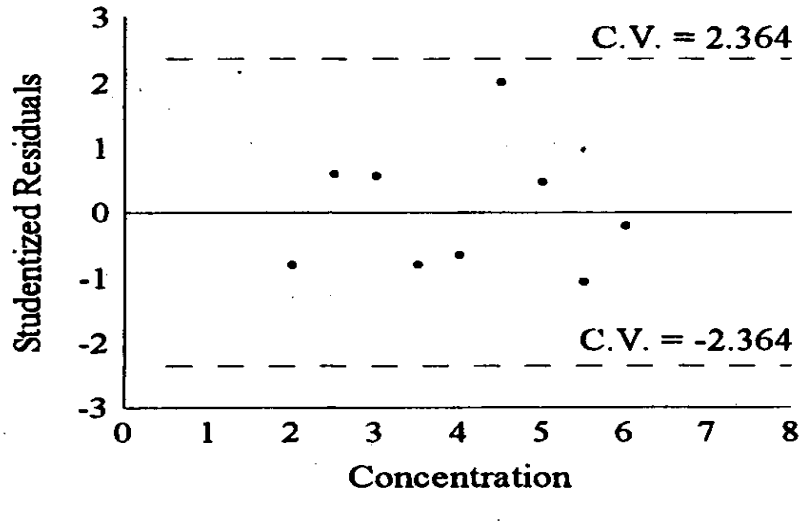
$$t = \frac{-2{,}840}{\sqrt{13{,}90}} = \frac{-2{,}840}{3{,}728} = -0{,}762$$

un **residuo studentizzato t** = -0,762.

I precedenti r_i risultano trasformati in r_i **studentizzati** come nella tabella successiva:

X_i	2,0	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0
r_i	**-2,840**	**+2,035**	**+1,950**	**-2,835**	**-2,230**	**+6,985**	**+1,610**	**-3,735**	**-0,760**
r_i studentizzati	**-0,762**	**+0,546**	**+0,523**	**-0,760**	**-0,622**	**+1,874**	**0,432**	**-1,002**	**-0,204**

Anche di questi **residui studentizzati** (***studentized residuals***) è bene fare la rappresentazione grafica (utilizzando i valori della prima e della terza riga della tabella precedente). Nel grafico, senza essere espressamente dichiarato, con la presenza delle due linee tratteggiate è riportato anche il risultato di un **altro test su gli outlier**, che è bene esplicitare.



Il valore C.V. = 2,364 **(con segno positivo e negativo sopra le linee tratteggiate, parallele alla media)**

- è il **valore critico** del **t di Student** con **7 gradi di libertà**,
- per la probabilità $\alpha = 0.05$ in una **distribuzione bilaterale** (nelle tabelle allegate in realtà è 2,365).

In questa rappresentazione grafica dei **residui studentizzati,** è reso visibile un concetto:

i residui studentizzati, ottenuti con

$$t = \frac{r_i}{\sqrt{S_e^2}}$$

sono **altrettanti test t**

- in cui, con i dati dell'esempio, **nessun residuo supera il valore critico (2,364)**,

Hanno **gradi di libertà** $n-2$, poiché sono residui intorno alla retta di regressione, per tracciare la quale servono due punti.

A differenza del precedente metodo di Box and Wiskers, in questo **secondo test**

- **nessun residuo** risulta essere **un outlier**,

- se il valore critico è scelto alla probabilità α = **0.05** in una **distribuzione bilaterale.**

Quindi non è significativo quel punto che, con il precedente test di Tukey, risultava un outlier (+6,985). In questo caso, il metodo dei residui studentizzati fornisce un valore (+1,874) nettamente inferiore a quello critico (2,364).

Ma l'analisi **t di Student** con **k residui** solleva il problema del **principio del Bonferroni**, che spesso su questi problemi viene trascurato. Per ogni **confronto t** di Student,

- la **probabilità α'** da utilizzare dovrebbe essere la probabilità totale α_T = 0.05 divisa per k.

I **residui studentizzati,** anche se solamente su alcuni testi, in modo non appropriato sono chiamati anche **residui standardizzati** (***standardized residuals***). Per questi ultimi, al posto della **devianza standard campionaria** S_e, è utilizzata

- la **deviazione standard della popolazione** σ_e.

$$Z = \frac{r_i}{\sqrt{\sigma_e^2}}$$

Quando il campione è molto grande, i residui studentizzati e i residui standardizzati tendono a coincidere, come il valore t di Student **tende a convergere verso il valore della** Z**.**

Rimane la difficoltà di definire quando un campione è sufficientemente grande. Nella pratica sperimentale, spesso questa tecnica è utilizzata anche con campioni piccoli.

Con i **residui standardizzati**, al posto del valore critico **t di Student** che ha gradi di libertà ($n - 2$), si utilizza la **distribuzione Z**. I suoi valori critici sono sempre minori del t.

Ad esempio, alla probabilità α = 0.05

- per i **residui studentizzati** è stato utilizzato come valore critico **t = 2,364** (con gdl = 7)
- mentre per i **residui standardizzati** il valore critico corrispondente è **Z = 1,96**.

Tuttavia, nell'analisi degli outlier spesso vengono utilizzate stime approssimate. Quindi

- per la **probabilità α = 0.05** con i residui standardizzati viene assunto **il valore 2, non 1,96.**.

Ma la probabilità (5%) corrispondente è alta: verrebbero indicati come outlier valori che frequentemente non li sono.

Ne consegue che, per decidere che **un dato è un outlier**, è prassi diffusa **utilizzare 3 come valore critico** e non 2 (vedi outlier nella statistica univariata). La probabilità P è nettamente minore di 0.05.

18.15. HAT VALUE O LEVERAGE, STUDENTIZED DELETED RESIDUALS.

La ricerca degli outlier nella regressione lineare è strettamente associata al problema più ampio della validità stessa della regressione**, che è fondata su tre assunzioni:**

1 – la media della popolazione della variabile dipendente**, per l'intervallo di valori della variabile indipendente che è stato campionato,** deve cambiare in modo lineare **in rapporto ai valori della variabile indipendente;**

2 – per ogni valore della variabile indipendente, **i valori possibili della variabile dipendente devono essere distribuiti normalmente**;

3 – **la deviazione standard della variabile dipendente** intorno alla sua media (la retta), per un dato intervallo di valori della variabile indipendente, **deve essere uguale per tutti i valori della variabile indipendente**.

La presenza anche di un solo outlier nelle Y, per un certo valore della X, modificando

- la media, che in questo caso è la retta (assunto 1),
- la forma della distribuzione (assunto 2),
- la deviazione standard (assunto 3),

rende irrealizzate queste condizioni di validità.

Anche quando non si intende analizzare se è presente almeno un outlier, la distribuzione effettiva dei dati può determinare una o più di queste condizioni. Ne consegue che, per affermare statisticamente la validità di una retta, sarebbe importante applicare sempre alcune delle **tecniche** diagnostiche **della regressione** (***regression diagnostics***), per valutare i suoi residui.

Nel testo

- di R. L. **Mason**, R.F: **Gunst** e J. L. **Hess** del 1989 ***Statistical Design and Analysis of Experiments*** (edito da John Wiley and Sons, New York, pp. 510-257)
- e in quello più recente di Stanton A. **Glanz** e Bryan K. **Slinker** del 2001 ***Primer of applied regression and analysis of variance*** (2^{nd} ed. Mc Graw-Hill, Inc., New York, 27 + 949),

per citarne solamente due, tra quelli che affrontano questi argomenti, sono riportati vari metodi per una impostazione più generale e approfondita dell'analisi dei residui, in grado di assicurare la validità dell'analisi della regressione e della correlazione.

Glanz e **Slinker,** con esempi spesso divertenti, sviluppano una serie di applicazioni dei **vari test di diagnostica della regressione**, utilizzando dati totalmente inventati. Sono le misure di alcuni marziani, dei quali si vuole conoscere le caratteristiche fondamentali, attraverso l'analisi statistica. La pratica di avvalersi di dati non reali, in esercizi e in dimostrazioni di statistica applicata, è criticata da molti studiosi, in quanto può condurre a problemi e situazioni irreali. E' l'opposto degli scopi specifici della disciplina. Ma in questo caso, data la competenza degli autori e nel contesto di tanti dati sperimentali, servono per illustrare con semplicità e rapidità una casistica numerosa e complessa di situazioni reali.

Di 11 marziani è stata misura la lunghezza del piede (in **cm**) e il quoziente d'intelligenza (in ***zorp***)

Marziano	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Piede (cm)	10	8	13	9	11	14	6	4	12	7	5
Intellig. (zorp)	8,04	6,95	7,58	8,81	8,33	9,96	7,24	4,26	10,84	4,82	5,68

E' indicato come **ESEMPIO A**.

Non è noto se esista una relazione di causa-effetto tra le due variabili.

Tuttavia, come metodo esplorativo, è applicata l'analisi della regressione lineare semplice e viene calcolato il coefficiente di correlazione r (la cui significatività, ovviamente, è identica a quella del coefficiente angolare b).

Assumendo

- come variabile indipendente (X) la lunghezza del piede
- e come variabile dipendente (Y) il quoziente d'intelligenza,

il programma informatico fornisce i seguenti risultati:

CAMPIONE A

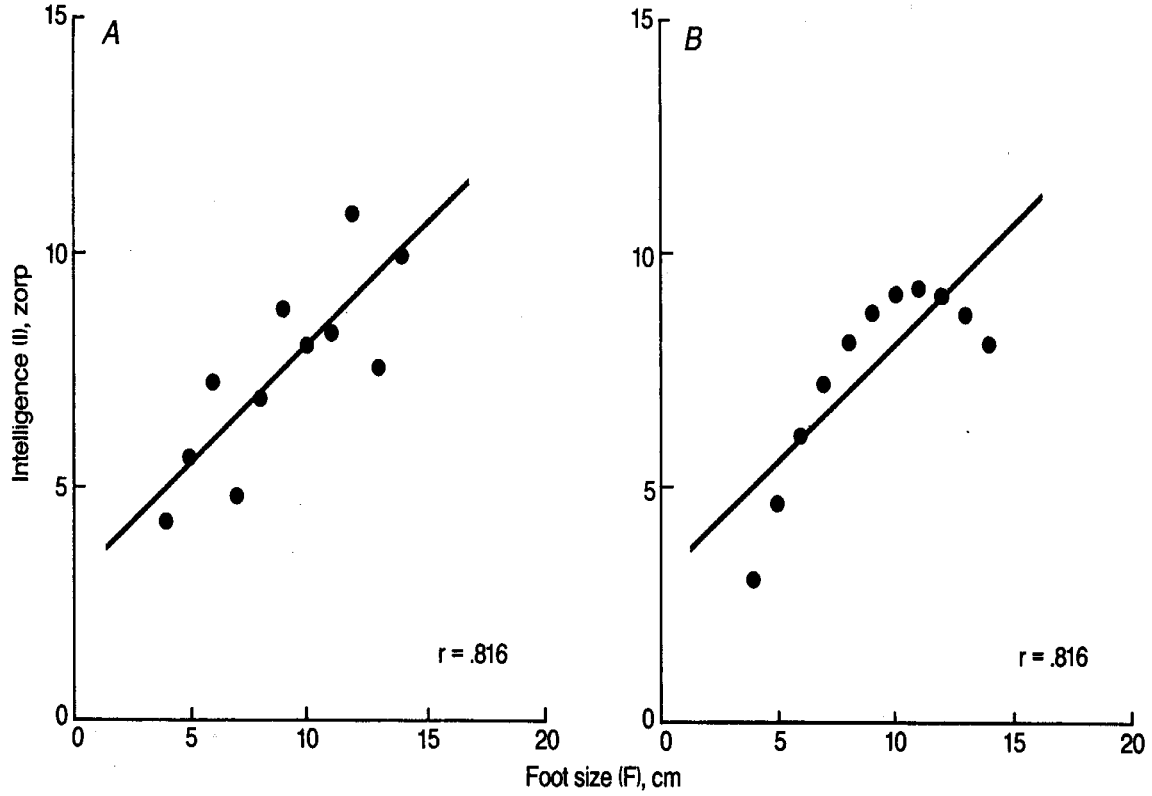
The regression equation is: a **= 3,00** b **= 0,500**

Predictor	Coeff.	St.dev.	t-ratio	P
a	3,000	1,125	2,67	0,026
b	0,5001	0,1179	4,24	0,002

s = 1,237 Rsq= 0,667 Rsq(adj)=0,629 **r = 0,816**

Analysis of variance

SOURCE	DF	SS	MS	F	P
Regression	1	27,510	27,510	17,99	0,002
Error	9	13,763	1,529		
Total	10	41,273			


A
B
C
D
Intelligence (I), zorp
Foot size (F), cm
r = .816
0
5
10
15
20

Come evidenzia anche il diagramma di dispersione riportato nella pagina precedente (grafico A, in alto a sinistra),

- la **retta di regressione** è $\hat{Y}_i = 3{,}0 + 0{,}5 \cdot X_i$
- il **coefficiente di correlazione** è $r = 0{,}816$
- la linearità e la correlazione **sono significative** (con t = 4,24 o F = 17,99 e P = 0,002);
- è significativamente **differente da zero anche l'intercetta** a (t = 2,67 e P = 0.026).

Le **altre tre figure (B, C, D)** sono state costruite da **Glanz** e **Slinker** in modo tale che i dati, non riportati, forniscono le seguenti **tre analisi della regressione**:

CAMPIONE B

The regression equation is: a **= 3,00** b **= 0,500**

Predictor	Coeff.	St.dev.	t-ratio	P
a	3,001	1,125	2,67	0,026
b	0,5000	0,1180	4,24	0,002

s=1,237 Rsq=0,666 Rsq(adj)=0,629 **r=0,816**

Analysis of variance

SOURCE	DF	SS	MS	F	P
Regression	1	27,500	27,500	17,97	0,002
Error	9	13,776	1,531		
Total	10	41,276			

CAMPIONE C

The regression equation is: a **= 3,00** b **= 0,500**

Predictor	Coeff.	St.dev.	t-ratio	P
a	3,002	1,124	2,67	0,026
b	0,4997	0,1179	4,24	0,002

s=1,236 Rsq=0,666 Rsq(adj)=0,629 **r=0,816**

Analysis of variance

SOURCE	DF	SS	MS	F	P
Regression	1	27,470	27,470	17,97	0,002
Error	9	13,756	1,528		
Total	10	41,226			

CAMPIONE D

The regression equation is: a **= 3,00** b **= 0,50**

Predictor	Coeff.	St.dev.	t-ratio	P
a	3,002	1,124	2,67	0,026
b	0,4999	0,1178	4,24	0,002

s=1,236 Rsq=0,667 Rsq(adj)=0,630 **r=0,816**

Analysis of variance

SOURCE	DF	SS	MS	F	P
Regression	1	27,490	27,490	18,00	0,002
Error	9	13,742	1,527		
Total	10	41,232			

Dalla lettura di queste tre tabelle, risulta con evidenza che i dati con i quali sono stati costruiti i tre grafici (B, C, D) hanno in comune con il grafico A

- la **stessa retta di regressione**: $\hat{Y}_i = 3{,}0 + 0{,}5 \cdot X_i$
- lo **stesso coefficiente di correlazione:** $r = 0{,}816$
- lo **stesso errore standard**: $s = 1{,}237$

Le piccole differenze nei test di significatività sono trascurabili.

Ma i **quattro diagrammi di dispersione** risultano visivamente **molto differenti**. Effettivamente hanno caratteristiche diverse, che è successivamente saranno quantificate in indici.

- La **figura A** (in alto, a sinistra) rappresenta una **situazione corretta**, in cui sono rispettate le tre condizioni di validità e nella quale pertanto **non sono presenti outlier**.

- La **figura B** (in alto, a destra) riproduce una **situazione non corretta**, in cui non sono rispettate tutte le condizioni di validità, ma nella quale **non sono presenti outlier**. Infatti la collocazione dei punti lungo la retta indica che **la regressione esiste, ma che essa non è lineare**. E' un esempio classico di ***model misspecification***, di **scelta errata del modello di regressione**.

- La **figura C** (in basso, a sinistra) mostra una **situazione non corretta**, in cui non sono rispettate tutte le condizioni di validità e nella quale **è presente un outlier**, con un leggero ***swamping effect***. Poiché la retta è fondata sul principio dei minimi quadrati, il valore anomalo ha un peso determinante

sul coefficiente di regressione b, attirandolo verso se. Questa capacità di attrazione di un punto è tanto maggiore, quanto più grande è la distanza del dato dal baricentro della distribuzione.

- La **figura D** (in basso, a destra) rappresenta un'altra **situazione non corretta**, nella quale **non sono rispettate tutte le condizioni di validità**; soprattutto è **presente un outlier**, molto distante dagli altri e quindi con un peso sproporzionato sui **coefficienti di regressione** b e di correlazione r.

In termini tecnici, si dice che

- è un ***leverage point*** o ***hat value***
- che ha un importante ***swamping effect***.

Vale a dire che, come visibile nel diagramma di dispersione, è collocato in una posizione dove ha una forte capacità di **sommergere l'informazione** data da tutte le altre coppie di dati.

La retta e la correlazione di questa figura D **non sarebbero significativi, senza la presenza di quel dato anomalo**. Se il dato anomalo è un errore, è doveroso eliminarlo. Ma anche se è corretto, occorre molta cautela per poterlo utilizzare nel calcolo della regressione e della correlazione. Secondo **Glanz** e **Slinker:** ***Even if the point is valid,*** *you should be* ***extremely cautions when using the information*** *in this figure to draw conclusions….* ***Such conclusions are essentially based on the value of a single point****.* ***It is essentials to collect more data … before drawing any conclusions.***

Il problema statistico è come arrivare a conclusioni sulla validità delle analisi non sulla base di **descrizioni qualitative**, ma attraverso **metodologie statistiche** condivise **che quantificano le diverse caratteristiche**. In modo più dettagliato, a pag. 118 sempre **Glanz** e **Slinker** scrivono: *These graphical differences are also quantitavely reflected in the value of* ***regression diagnostics*** *associated with the individual data points.* ***These differences are the key to indentifying problems*** *with the* ***regression model*** *or* ***errors*** *in the data. The fact we can fit a linear regression equation to a set of data - even if it yields a reasonably high and* ***statistically significant correlation coefficient and small error of the estimate – does not ensure that the fit is appropriate or reasonable****.*

Come nel paragrafo precedente,

- le informazioni fondamentali sulla validità della regressione e della correlazione
- sono basati sui residui $r_i = Y_i - \hat{Y}_i$, detti anche ***raw residuals***, per distinguerli più nettamente dagli altri ***residuals***, diversamente aggettivati, che derivano da questi per elaborazioni successive.

L'**analisi della normalità** della distribuzione dei **residui grezzi** (***raw residuals***), dei **residui studentizzat**i o di quelli **standardizzati** può essere effettuata con le tecniche illustrate per la statistica univariata.

Quindi, si rimanda ad esse. Anche su questi dati è utile

- costruire il grafico dei residui,
- applicare a essi il test di Tukey con il metodo Box-and-Wiskers,
- calcolare e rappresentare graficamente i residui studentizzati, alla ricerca degli outlier.

Ma sono possibili e vantaggiose anche altri analisi, sebbene non esauriscano l'elenco:

- stimare il ***leverage*** o ***hat value*** di ogni punto, che valutata l'**influenza potenziale** sulla regressione;
- calcolare gli ***Studentized deleted residuals*** o ***externally Studentized residuals***;
- calcolare la **distanza di Cook** (***Cook's distance***), che valuta l'**influenza effettiva o reale (*actual influence*)** di ogni punto sui risultati della regressione; è **chiamata distanza** ma è una **misura d'influenza** del dato sul risultato complessivo della regressione.

Il ***leverage*** o ***hat value*** è un termine usato nell'analisi della regressione multipla, per definire il peso che le singole osservazioni hanno sul valore della regressione. Sono di particolare interesse i dati con un valore estremo, in una o più variabili indipendenti. Per il principio dei minimi quadrati, la retta è forzata a passare vicino a quei punti, che pertanto hanno una grande capacità di attrarre verso di loro la retta e quindi di determinare residui piccoli.

Nel caso della regressione lineare semplice, quindi con una sola variabile indipendente, il **leverage** h_i del punto X_i , Y_i è stimato con

$$h_i = \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

Questo numero, che deve essere **calcolato per ogni punto**,

- **varia da 0 a 1**,
- è determinato dalla distanza del **valore della variabile X dalla sua media**
- **rapportato alla devianza totale della X**.

Nell'**esempio A**, dove $\overline{X} = 9{,}0$ e $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = 110$

- per il punto del marziano I con X = 10 e Y = 8,04

il **leverage**

$$h_i = \frac{1}{11} + \frac{(10 - 9{,}0)^2}{110} = 0{,}0909 + 0{,}0091 = 0{,}1000$$

è piccolo (uguale a 0,1000) poiché il suo valore di **X è vicino alla media**;

- per il punto del marziano VIII con X = 4 e Y = 4,26

il **leverage**

$$h_i = \frac{1}{11} + \frac{(4-9{,}0)^2}{110} = \frac{1}{11} + \frac{25}{110} = 0{,}0909 + 0{,}2273 = 0{,}3182$$

è **maggiore** (uguale a 0,3182) poiché il suo valore di **X è più lontano dalla media**.

Il **leverage** è definito come

- una **influenza potenziale** del punto sulla regressione e correlazione, determinato dalla distanza del valore X_i dalla sua media $\overline{X}$.

Con i dati dell'**esempio A**, si osserva appunto che

- il valore minimo di leverage è quello del marziano IV, poiché il suo valore della variabile X coincide con la media,
- mentre è massimo per i marziani VI e VIII, che sono agli estremi per la variabile X

Marziano	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Piede X	10	8	13	9	11	14	6	4	12	7	5
Intel. Y	8,04	6,95	7,58	8,81	8,33	9,96	7,24	4,26	10,84	4,82	5,68
Raw Res e_i	+0,039	-0,051	-1,921	+1,309	-0,171	-0,041	+1,239	-0,740	+1,839	-1,681	+0,179
Leverage h_i	0,1000	0,1000	0,2364	0,0909	0,1273	0,3182	0,1727	0,3182	0,1727	0,1273	0,2364
Stud. Res. r_i	+0,033	-0,043	-1,778	+1,110	-0,148	-0,040	+1,102	-0,724	+1,634	-1,454	+0,166
Stud.Del.Res.	+0,031	-0,041	-2,081	+1,127	-0,140	-0,038	+1,117	-0,705	+1,838	-1,568	+0,157

Idealmente, per una buona retta di regressione**,**

- tutti i punti dovrebbero avere la stessa influenza sui parametri della retta di regressione;

- pertanto i valori di *leverage* **dei punti campionati dovrebbero essere uguali e piccoli.**

Nella regressione multipla il **valore medio** di ***leverage*** è $h_i = \frac{k+1}{n}$

dove k è il numero di variabili indipendenti

Nella **regressione lineare semplice**, dove k = 1

- il **valore medio** del ***leverage*** è $\frac{1+1}{n} = \frac{2}{n}$

Ne consegue anche che nella regressione lineare semplice

- la somma dei ***leverage*** di n dati è uguale a 2:

$$\sum_{i=1}^{n} h_i = 2$$

I valori possibili di ***leverage*** variano da un minimo di $1/n$ a un massimo di 1.

Nella prassi statistica, sono giudicati **alti i valori maggiori di 0,4**; altri statistici suggeriscono di controllare quelli che sono **oltre il doppio del valore medio**.

Con i dati dell'esempio precedente, con $k = 1$ e $n = 11$

si ha che il **valore medio** di ***leverage*** è

$$\frac{k+1}{n} = \frac{1+1}{11} = 0{,}1818$$

Sempre nella lettura dei valori di ***leverage***, si evidenzia che essi sono massimi (0,3182) per i marziani VI e VIII, benché non siano molto maggiori del valore medio. Se ne può dedurre che

- **la retta e/o la correlazione sono calcolate**, per questo aspetto, **in condizioni ottimali**,
- poiché **tutti i punti forniscono un contributo analogo al valore totale**.

Il **leverage è una potenzialità**, non un peso effettivo sulla determinazione della retta di regressione e sul valore della correlazione.

Stime del **peso effettivo** sono fornite da

- **l'*internally Studentized residual*** spesso chiamato semplicemente ***Studentized residual***, generando confusione con quelli definiti prima nello stesso modo ma con formula differente;
- **l'*externally Studentized residual*** chiamato anche ***Studentized deleted residual***;
- la **distanza di Cook (*Cook's distance*).**

A differenza della simbologia utilizzata nel paragrafo precedente,

il **residuo grezzo** o ***raw residual*** (e_i) del punto i

come spesso avviene può essere indicato con

$$e_i = Y_i - \hat{Y}_i$$

Da esso è ricavato il **residuo Studentizzato** o ***Studentized residual*** r_i

con

$$r_i = \frac{e_i}{S_e \cdot \sqrt{1 - h_i}}$$

dove

- S_e è la deviazione standard dei residui; con i dati dell'esempio A, è $S_e = 1{,}237$
- h_i è il valore di leverage del valore X_i relativo al residuo.

Nell'**esempio A**, dove S_e = 1,237

- per il punto del marziano I con e_i = +0,039 e h_i = 0,1000

lo **studentized residual**

$$r_i = \frac{+0{,}039}{1{,}237 \cdot \sqrt{1-0{,}1000}} = \frac{+0{,}039}{1{,}237 \cdot 0{,}949} = \frac{+0{,}039}{1{,}174} = +0{,}033$$

è r_i = +0,033;

- per il punto del marziano VIII con e_i = -0,740 e h_i = 0,3182

lo **studentized residual**

$$r_i = \frac{-0{,}740}{1{,}237 \cdot \sqrt{1-0{,}3182}} = \frac{-0{,}740}{1{,}237 \cdot 0{,}826} = \frac{-0{,}740}{1{,}022} = -0{,}724$$

è r_i = -0,724.

(I valori dei residui studentizzati per tutti gli 11 marziani sono riportati nella tabella precedente.)

Il **valore dei residui studentizzati risulta grande,** quando contemporaneamente **sono grandi**

- sia il **valore del residuo** e_i **,**
- sia il **valore di leverage** h_i **.**

In questo caso dello ***Studentized residual***, **la deviazione standard dei residui** S_e **è calcolata usando tutti gli** n **del campione**; per questo motivo, con una dizione più completa e precisa, l'indice

$$r_i = \frac{e_i}{S_e \cdot \sqrt{1-h_i}}$$

è noto anche come ***internally Studentized residual***.

Ma per analizzare l'effetto degli outlier, è utilizzato spesso un **altro indice studentizzato dei residui**.

Per ogni residuo,

- la deviazione standard S_e **è calcolata senza il punto** i , cioè dopo aver tolto dal calcolo della retta e da quelli successivi per arrivare all'errore **il punto** i .

La **simbologia** della **deviazione standard** diventa $S_{e(-i)}$.

Il residuo $e_i = Y_i - \hat{Y}_i$ con il nuovo denominatore

- è **indicato con** r_{-i}

$$r_{-i} = \frac{e_i}{S_{e(-1)} \cdot \sqrt{h_i}}$$

- ed è chiamato ***externally deleted residual*** oppure ***Studentized deletd residual***.

Il motivo fondamentale di questa metodologia deriva dal fatto che, se il punto i è un outlier, con la sua presenza determina un valore alto della deviazione standard S_e**.**

Per costruire un test più sensibile alla scoperta dell'outlier i e eliminare il suo ***masking effect***, è quindi opportuno non considerare i valori del punto i nel calcolo di S_e, e successivamente utilizzare appunto la nuova **deviazione standard** $S_{e(-i)}$.

Per il calcolo dei ***Studentized deletd residual*****,** esiste un problema pratico rilevante. A pag. 137 del testo già citato **Glantz** e **Slinker** scrivono: *Although most regression programs report Studentized residuals,* ***they often do not clearly state which definition is used****; to be sure, you should check the program's documentation to see which equation is used to compute the Studentized residual.*

A questo scopo e come stima più rapida degli ***externally Studentized residual*** r_{-i} è opportuno utilizzare gli ***internally Studentized residual*** r_i (facilmente ricavabili dalla varianza d'errore dell'ANOVA, come mostrato nei paragrafi precedenti),

mediante la relazione

$$r_{-i} = r_i \cdot \sqrt{\frac{n-k-2}{n-k-1-r_i^2}}$$

dove, nella statistica bivariata, $k = 1$.

Spesso

- **i valori dei *Studenzized residuals*** r_i

- e quelli dei corrispondenti valori ***Studentized deleted residuals*** r_{-i} sono simili.

La tabella successiva mostra come per gli 11 marziani le differenze non siano molto importanti:

- solamente il valore del marziano III da –1,778 diventa –2,081 con un aumento del 17% in valore assoluto
- e quello del marziano IX aumenta del 12 % ma partendo da un valore minore.

Marziano	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
1) r_i	+0,033	-0,043	-1,778	+1,110	-0,148	-0,040	+1,102	-0,724	+1,634	-1,454	+0,166
2) r_{-i}	+0,031	-0,041	-2,081	+1,127	-0,140	-0,038	+1,117	-0,705	+1,838	-1,568	+0,157

Tuttavia **a volte sono molto differenti**. **In alcuni casi** r_{-i} (avendo ridotto la varianza d'errore) **può essere molto maggiore di** r_i (in valore assoluto)**.**

Quale dei due valori utilizzare nel test per gli outlier?

In questi casi,

- con r_{-i} **il test sull'outlier è più prudenziale**,
- con r_i **il test sull'outlier è più potente.**

E' ovvio che l'interesse dell'utente per uno dei due risultati può influire sulla scelta di quale residuo utilizzare.

Sempre **Glantz** e **Slinker** (a pag. 138) suggeriscono il test più potente, scrivendo: ***…the value of*** r_{-i} ***can greatly exceed*** r_i***, in some instances. Thus,*** r_{-i} ***is slightly preferred.***

18.16. LA DISTANZA EUCLIDEA TRA LE STATISTICHE DELLA RETTA E LA DISTANZA DI COOK; APPLICAZIONI DEL JACKKNIFE.

Le tecniche che utilizzano **i residui** sono giudicate **non consistenti** per **valutare la presenza di un outlier** quando

1 - i dati **non sono distribuiti in modo normale**,

2 - e/o il valore anomalo può **influire** in modo **potenzialmente** sproporzionato sia sul coefficiente angolare b e sull'intercetta a della regressione, sia sul valore dell'indice r di correlazione.

Esistono altri metodi, fondati su principi differenti dai precedenti, che anche nelle due condizioni precedenti permettono di misurare **l'influenza reale** di un punto sul risultato complessivo della statistica calcolata. Essi si fondono essenzialmente sul concetto di

- valutare di quanto cambiano i risultati,
- quando un punto specifico viene eliminato.

Tra le metodologie più diffuse, sono da ricordare:

A - il **metodo grafico** che descrive le variazione delle statistiche a (intercetta) e b (coefficiente angolare) della regressione, ottenute eliminando un punto ogni volta, rispetto all'analisi con tutti i dati;
B - la **distanza di Cook** (**Cook's distance**), che nonostante il nome in realtà è una **misura di influenza**.
C - varie applicazioni del metodo **jackknife**, anche se spesso chiamate con nomi diversi, quali DFBETA e SDBETA, DFFIT e SDFIT, che ora sono possibili con i programmi informatici, poiché richiedono lunghi calcoli ripetuti.

A - Per il **metodo grafico**, con i dati dell'**esempio A**, utilizzando i dati di tutti gli 11 marziani, è stata calcolata
- la **retta** di **regressione**

$$\hat{Y}_i = 3{,}00 + 0{,}50 \cdot X_i$$

Dati dell'esempio A

Marz.	Tutti	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Piede X	---	10	8	13	9	11	14	6	4	12	7	5
Intel. Y	---	8,04	6,95	7,58	8,81	8,33	9,96	7,24	4,26	10,84	4,82	5,68
a	3,00	3,00	3,01	2,41	2,89	2,99	2,98	2,50	3,54	3,34	3,49	2,90
b	0,50	0,50	0,50	0,59	0,50	0,50	0,50	0,54	0,45	0,44	0,47	0,51
h_i	---	0,1000	0,1000	0,2364	0,0909	0,1273	0,3182	0,1727	0,3182	0,1727	0,1273	0,2364
St. r_i	---	+0,033	-0,043	-1,778	+1,110	-0,148	-0,040	+1,102	-0,724	+1,634	-1,454	+0,166
Cook	---	0,000	0,000	0,489	0,062	0,002	0,000	0,127	0,123	0,279	0,154	0,004

Per valutare l'effetto dei singoli punti (da I a XI) sulle statistiche della regressione, con le 11 coppie di dati originali X_i e Y_i si calcolano altrettante rette, togliendo ogni volta uno degli 11 punti.
I valori sono quelli riportati nella tabella precedente. Ad esempio,
- togliendo il valori del marziano I,
la retta è uguale

$$\hat{Y}_i = 3{,}00 + 0{,}50 \cdot X_i$$

- mentre togliendo i dati del marziano III,

la retta diventa

$$\hat{Y}_i = 2{,}41 + 0{,}59 \cdot X_i$$

E' evidente il concetto che quanto più **un punto** X_i, Y_i

1) è lontano dal baricentro $\overline{X}, \overline{Y}$ del diagramma di dispersione

2) **non è allineato** con gli altri,

- tanto più **le statistiche** a_{-i} **e** b_{-i} **(calcolate senza quel punto) sono lontane** dai valori a e b calcolati con tutti i dati.

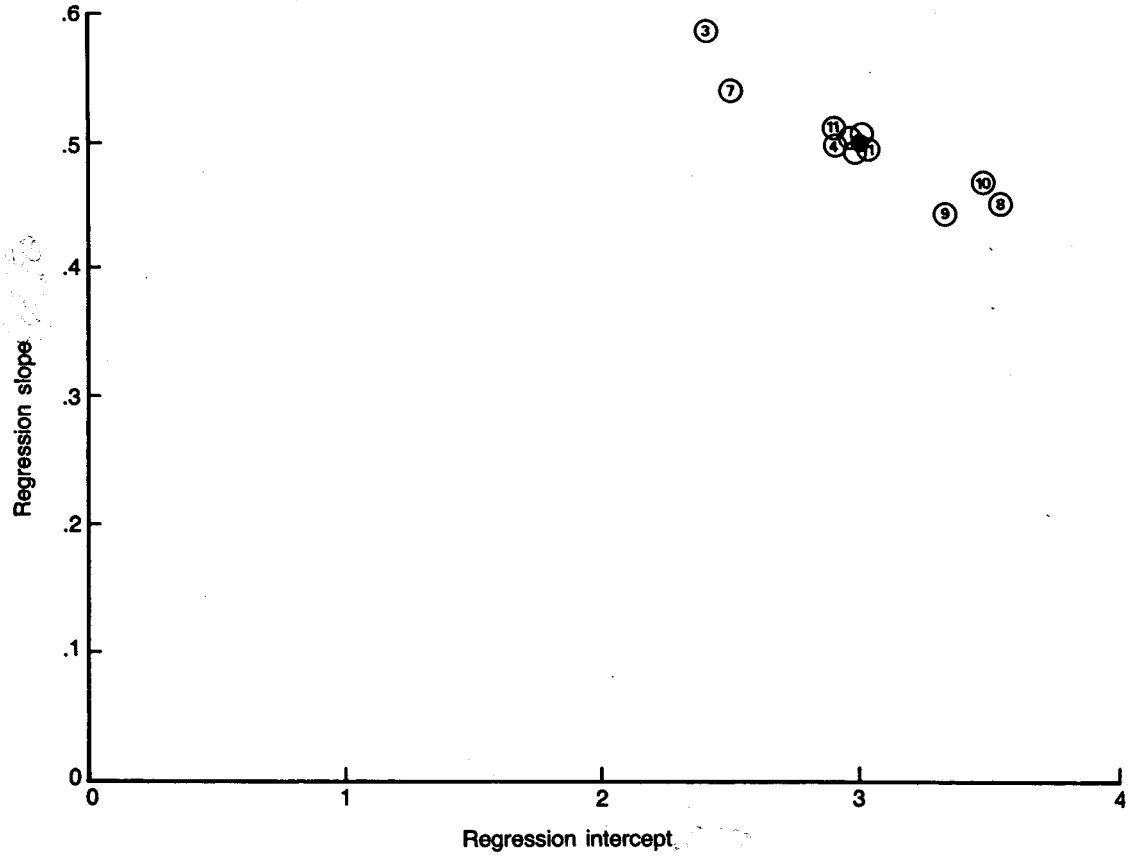
Nella figura successiva, in un diagramma cartesiano dove

- i valori delle **intercette** sono riportati sull'**asse delle ascisse** (in bianco le 11 a_{-i} e in nero a);

- i valori dei **coefficienti angolari** sono riportati in **ordinata**, (in bianco le 11 b_{-i} e in nero la b)

si evidenzia l'effetto dei punti che si discostano maggiormente dalle due medie.

Diagramma di a_{-i} e b_{-i} dell'**esempio A**



Le informazioni bidimensionali contenute nelle diverse statistiche a_{-i} e b_{-i} delle 11 regressioni rispetto a quella calcolata con tutti i dati di coordinate a e b possono essere ridotte a una sola dimensione, sommandole in altrettante distanze cartesiane d_i

con

$$d_i = \sqrt{(a_{-i} - a)^2 + (b_{-i} - b)^2}$$

Le n distanze d_i misurano la **lontananza in uno spazio bidimensionale**,

- tra il punto (in nero) di coordinate a , b
- e ognuno degli altri 11 punti (in bianco e numerati) di coordinate a_{-i} , b_{-i} .

Dalla lettura del grafico, ad esempio, emerge con evidenza che è maggiore

- la distanza d_3 relativa ai dati del marziano III

$$d_3 = \sqrt{(2{,}41 - 3{,}00)^2 + (0{,}59 - 0{,}50)^2} = \sqrt{0{,}2601 + 0{,}0081} = 0{,}5179$$

- rispetto alla distanza d_4 relativa ai dati del marziano IV

$$d_4 = \sqrt{(2{,}89 - 3{,}00)^2 + (0{,}50 - 0{,}50)^2} = \sqrt{0{,}0121 + 0{,}0} = 0{,}11$$

Interessante sotto l'**aspetto descrittivo** e fondato sullo stesso **concetto del jackknife**, questo metodo presenta **due gravi inconvenienti**:

1 - ogni distanza d_i è **la somma di due unità non omogenee**, quali i valori dell'intercetta a e del coefficiente angolare b , che misurano caratteristiche differenti e pertanto **non sommabili** della retta;

2 - come già in precedenza i valori dei ***raw residuals*** *(*e_i*)* sono influenzati dall'unità di misura della variabile Y, **queste distanze** d_i **risentono dell'unità di misura con le quali sono state rilevate la variabile X e la variabile Y**.

Se, nell'esempio utilizzato, l'altezza fosse stata misurata in piedi oppure in metri invece che in centimetri, le statistiche a e b sarebbero state differenti e quindi anche le n distanze d_i .

Nonostante questi gravi limiti,

- le d_i permettono ugualmente alcune analisi entro il singolo esperimento,

- utilizzando gli stessi metodi già illustrati per i residui e_i (**raw residuals**).

Ancora una volta, con la serie delle distanze d_i, dai dati bivariati si è ritornati a dati univariati.

Quindi diventano possibili

- tutte le **analisi già presentate per la statistica univariata**,
- a partire dagli **istogrammi** e dal **Box and Wiskers** di Tukey.

Inoltre, **le distanze d_i possono essere standardizzate e/o studentizzate**, utilizzando la varianza d'errore dell'analisi della regressione. Essa può essere stimata sia con tutti i dati, sia eliminando ogni volta i dati del punto di cui si calcola la distanza.
In queste analisi, è indispensabile l'uso di programmi informatici, che spesso usano metodi diversi. Occorre porre attenzione alle istruzioni (quando fornite).

B – Un'altra **misura molto diffusa e adimensionale della distanza**, cioè indipendente dalle unità di misura con le quali sono state rilevate la variabile X e la variabile Y,
è la **distanza D_i di Cook** (***Cook's distance***),

$$D_i = \frac{r_i^2}{k+1} \cdot \frac{h_i}{1-h_i}$$

dove

- r_i è la misura del **residuo studentizzato** (***Studentized residual*** o ***internally Studentized residual***) del punto i e misura la discrepanza (***discrepancy***),
- h_i è il **leverage** o ***hat value*** del punto i, cioè la sua **influenza potenziale**.

Il valore D_i **è grande,** quando il **punto i ha un effetto importante** sul valore del coefficiente di regressione. Il valore della distanza D_i di Cook con k **abbastanza grande** e n nettamente maggiore di 10, tende a essere distribuito

- come la **distribuzione F di Fisher**,
- con gradi di libertà $k+1$ al numeratore e $n-k-1$ al denominatore.

Pertanto, nella **statistica multivariata** e in **campioni abbastanza grandi**, permette l'inferenza per la verifica dell'ipotesi

H_0: il punto i **non è** un outlier contro H_1: il punto i **è** un outlier

Per la **retta di regressione lineare semplice**, utilizzando i dati riportati nell'ultima tabella,
- per il marziano III che ha h_i =0,2364 e r_i = -1,778

- la distanza di Cook

$$D_3 = \frac{(-1{,}778)^2}{1+1} \cdot \frac{0{,}2364}{1-0{,}2364} = 1{,}5806 \cdot 0{,}3096 = 0{,}4893$$

è D_3 = 0,4893 (nella tabella, arrotondato in 0,489).

Una ulteriore dimostrazione dei concetti che sono implicati nella misura della **distanza di Cook** è fornita dall'analisi statistica dei dati dell**'esempio C**.

Nel diagramma di dispersione dell'**esempio C** (vedere la figura precedente con i 4 diagrammi), si evidenzia che un punto è chiaramente lontano dalla sequenza degli altri, collocati lungo una curva.
Quindi la retta di regressione non è adatta, in quanto il modello è diverso dalla linearità.

La lettura coordinata
- del **grafico che riporta le distanze** a_{-i} e b_{-i}
- e della **tabella dei valori che conducono al calcolo delle distanze di Cook**

evidenzia che nell'**esempio C** il marziano 3 ha caratteristiche che lo distinguono più nettamente dal gruppo degli altri dieci, rispetto all'**esempio A**.
In questo caso, il **residuo studentizzato** r_i = 2,999 è un valore che merita attenzione, anche se come outlier non è statisticamente significativo.
In un **test t** bilaterale (per gradi di libertà $n-2$ = 9 e alla probabilità α = 0,025) il valore critico è uguale a 2,685. Se fosse analizzato da solo, con una scelta a priori, sarebbe significativo con probabilità P < 0.025.

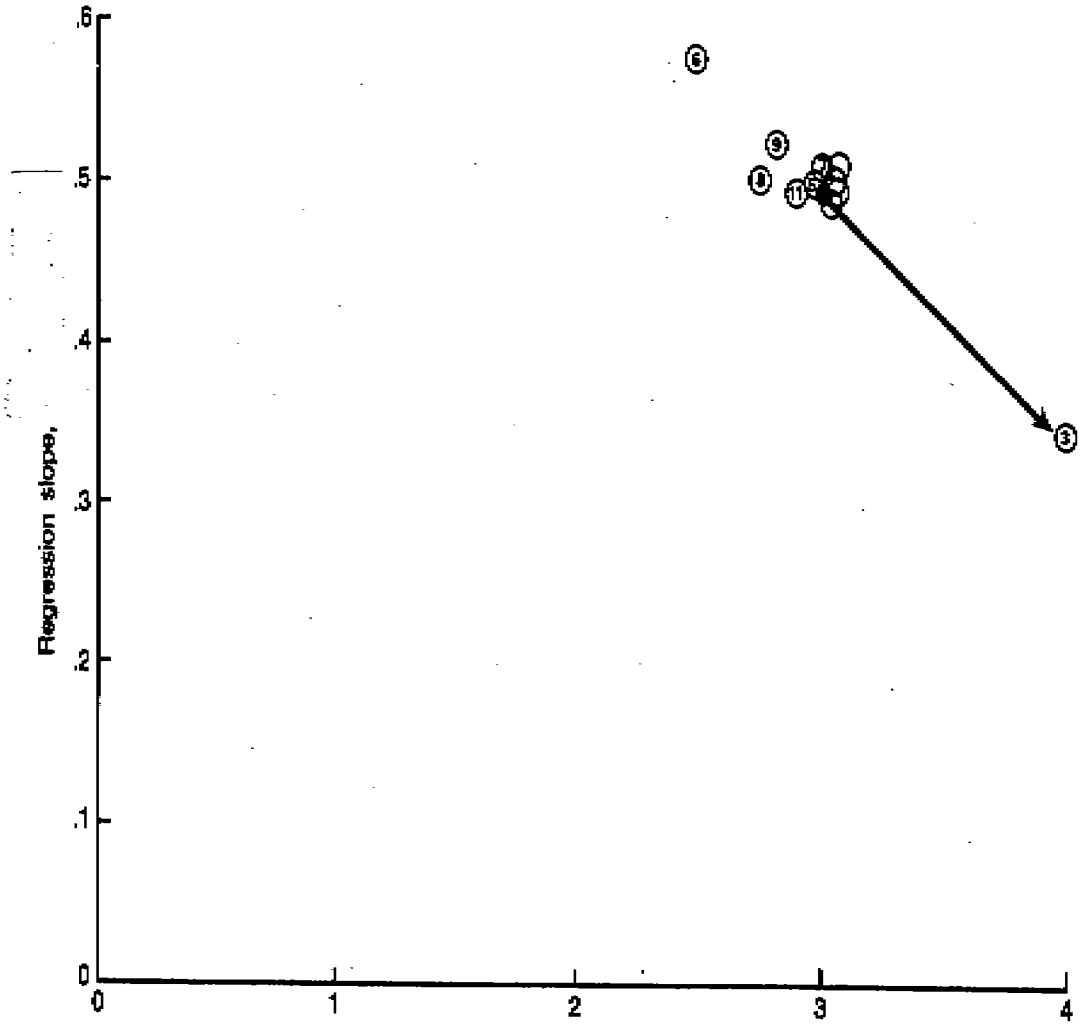
Ma è un punto su 11 complessivi.
Pertanto, secondo vari autori di testi di statistica è necessario **applicare il principio del Bonferroni**:
- per essere significativo alla **probabilità complessiva** (***experiment-wise***) α = 0,05
- il valore del **test t per un singolo punto** deve essere maggiore di quello critico per la probabilità specifica (***comparison-wise***) α = 0.05/11 = 0,00454.

Anche la **distanza di Cook** (ultima riga della tabella) risulta alta, abbinando
- un **residuo** (chiamato anche **discrepanza**) con un valore **alto**
- a un **leverage** di livello **medio**.

Dati dell'esempio C

Marz.	Tutti	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
a	3,00	3,01	3,05	4,01	3,04	2,95	2,46	2,97	2,72	2,84	3,03	2,88
b	0,50	0,51	0,50	0,35	0,50	0,51	0,58	0,50	0,53	0,53	0,50	0,51
h_i	---	0,1000	0,1000	0,2364	0,0909	0,1273	0,3182	0,1727	0,3182	0,1727	0,1273	0,2364
St. r_i	---	-0,460	-0,196	+2,999	-0,331	-0,597	-1,135	+0,070	+0,381	-0,755	-0,070	+0,212
Cook	---	0,012	0,002	1,393	0,006	0,026	0,301	0,001	0,034	0,060	0,000	0,007

Diagramma di a_{-i} e b_{-i} dell'**esempio C**



C - Recentemente, vari programmi informatici per l'**analisi degli outlier** presentano metodi che sono l'applicazione del **jackknife** a statistiche tra loro differenti, ma sempre con la stessa logica di base presentata in precedenza per la correlazione. Nel caso della regressione lineare semplice, si hanno

- **DFBETA,** una distanza D_{ij} uguale a

$$D_{ij} = b_j - b_{j(-i)}$$

che è ottenuta per ogni dato i, sottraendo al valore di b_j, calcolato utilizzando tutti i dati, il valore $b_{j(-i)}$ calcolato escludendo il valore i;

- **SDBETA,** una versione standardizzata dell'indice precedente, ottenuta dividendo D_{ij} per una stima ***deleted*** dell'errore standard del coefficiente b;

- **DFFIT** che è il valore di Y predetto ($\hat{Y}_i$) quando è escluso il caso i;

- **SDFIT,** la versione standardizzata del precedente DFFIT;

- il **grafico cartesiano** (***plot***) dei **valori SDFIT,** che sono riportati in ordinata mentre i valori della X sono riportati in ascissa.

Quando per la stessa analisi sono presentati **più metodi**, che quasi sempre si rifanno a **principi statistici differenti** e forniscono **risposte non coincidenti**, alla conclusione del dibattito tecnico è abituale la domanda pratica: "Ma **quale test usare?"**

Come risposta, è utile giovarsi delle parole di **Glantz** e **Slinker** (a pag. 144): *The diagnostics we have discussed so far – residuals, standardized residuals, leverage, Studentized residuals, and Cook's distance – are all designed to provide different approaches to indentifying points that do not fit with the assumed regression model.* ***No single one of these diagnostics tells the whole story****. They should de used together to identify possibly erroneous data points or problems with the regression model (equation) itself.*

In altri termini, è intelligente fornire una **risposta articolata,** che evidenzi le differenti risposte e nella quale **la scelta conclusiva della significatività o della non significatività dell'outlier è giustificata in modo scientificamente credibile**. E' lecito fornire una riposta sola, indicando solamente un test, quando l'evidenza del risultato è assoluta; vale a dire, quando tutti i test hanno fornito risposte uguali. Ma è la conclusione alla quale si perviene sempre, parlando di test dove sono possibili più metodi.

Le varie metodologie presentate in questi paragrafi utilizzano i residui, cioè gli scarti tra valori osservati e valori attesi delle Y. A questi residui sono applicare trasformazioni che permettono un loro uso più generale. In letteratura, non sempre i nomi di questi differenti residui sono indicati con precisione; si parla solo di residuals quando si tratta in realtà di standardized residuals oppure di standardized residuals quando invece si tratta di deleted standardized residuals.

E’ quindi utile ricordare i termini scientifici inglesi più diffusi, con la loro definizione:

- **Residuals, Raw Residuals, Unstandardized Residuals**: sono le differenze tra valori osservati e attesi. La loro somma è 0 e quindi anche la loro media è 0.
- **Standardized Residuals**: sono i residui divisi l'errore standard della popolazione; hanno media 0 e deviazione standard 1.
- **Studentized Residuals** o **Internally Studentized Residuals**: sono i residui divisi la deviazione standard del campione, che quindi varia da caso a caso; hanno media 0 e deviazione standard maggiore di 1. In vari testi, i termini **standardized residuals e Studentized residuals** sono usati come sinonimi.
- **Deleted Residuals**: sono i residui quando nel calcolo del coefficiente di regressione è escluso un dato campionario; sono le differenze tra i valori della dipendente e i corrispondenti valori predetti aggiustati, con l'eliminazione del dato campionario.
- **Studentized Deleted Residuals** o **Externally Studentized Residual**s: sono i residui precedenti (deleted residuals) studentizzati; l'effetto di un valore è eliminato dal calcolo dell'errore standard; questi residui possono essere ampi a causa della distanza del Y_i osservato dal valore $\hat{Y}_i$ stimato e del leverage; la media è 0 e la varianza è leggermente maggiore di 1.

18.17. LETTURA DI TRE TABULATI DI PROGRAMMI INFORMATICI SU REGRESSIONE E CORRELAZIONE LINEARE SEMPLICE.

Con le misure di peso (in Kg.) e di altezza (in cm.) di 7 giovani, come riportato nella tabella,

Individui	1	2	3	4	5	6	7
Peso (Y)	52	68	75	71	63	59	57
Altezza (X)	160	178	183	180	166	175	162

effettuare l'analisi statistica con un programma informatico.

Dopo aver espressamente indicato quale è la variabile dipendente (il peso) e quella indipendente (l'altezza), le risposte fornite dall'output in linea generale sono le seguenti.

Riquadro 1.

Nella parte inferiore, sono riportati i parametri della retta di regressione: l'intercetta ed il coefficiente angolare, con i relativi errori standard.

Nella quinta colonna sono indicati i valori del t di Student, per la verifica dell'ipotesi nulla H_0 che il parametro in oggetto sia significativamente diverso da 0.

La sesta ed ultima colonna, sempre nella parte inferiore a destra del riquadro 1), mostra il valore di probabilità, per un test bilaterale.

Nella parte superiore del riquadro è riportata l'analisi della varianza, con tutti i valori relativi ai parametri indicati. Il valore di F è il quadrato di quello del t di Student ed, ovviamente, le due probabilità coincidono.

Sotto l'analisi della varianza sono riportati altri indicatori utili ad eventuali confronti ed interpretazioni ulteriori dei risultati:

- **Root MSE** è la radice quadrata della varianza (Mean Square, sovente tradotto in italiano con quadrato medio);
- **Dep mean** è la media della variabile dipendente;
- **C. V.** è il coefficiente di variazione (sempre della variabile dipendente);
- **R-square** è il valore di R^2, o R oppure r^2 già trattato nella discussione sul valore predittivo della retta;
- **Adj. R-sq** (simboleggiato sovente con $\overline{R}^2$) è il valore di R Adjusted, che considera l'effetto dei gdl ed è calcolato come

$$\overline{R}^2 = 1 - \frac{(1 - R^2) \cdot (n - 1)}{dfe}$$

dove

- **n** è il numero di dati,
- **dfe** sono i gdl della varianza d'errore.

1)

Dependent Variable:PESO

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	323.20784	323.20784	20.073	0.0065
Error	5	80.50644	16.10129		
Total	6	403.71429			

Root MSE	4.01264	R-square	0.8006
Dep Mean	63.57143	Adj R-sq	0.7607
C.V.	6.31202		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-73.354062	30.59903924	-2.397	0.0618
ALTEZZA	1	0.796078	0.17768273	4.480	0.0065

Nel riquadro 2 è riportata l'analisi della correlazione. Sono stati utilizzati gli stessi dati dell'esempio precedente, relativi alla regressione lineare tra peso ed altezza in 7 giovani, per facilitare il confronto tra i due risultati. Sovente, i programmi della regressione forniscono analisi delle caratteristiche della distribuzione delle due serie di dati, presentati nel capitolo della statistica descrittiva ed utili alla verifica delle condizioni di validità della correlazione e della regressione, che sono molto più dettagliate di quelle riportate nel riquadro sottostante

2)
Correlation Analysis

1 'WITH' Variables: ALTEZZA
1 'VAR' Variables: PESO

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
ALTEZZA	7	172.00	9.2195	1204	160	183
PESO	7	63.57	8.2028	445	52	75

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 7

	PESO
ALTEZZA	0.89475 0.0065

I risultati indicati nella parte superiore del riquadro 2) non hanno bisogno di ulteriori spiegazioni.
Nel parte inferiore, è riportato il valore di correlazione r di Pearson è la probabilità relativa alla sua significatività, come verifica dell'ipotesi nulla H_0: $\rho = 0$

Nei riquadri 3) e 4) sono descritti i risultati dell'analisi della covarianza,

3)

General Linear Models Procedure
Class Level Information

Class	Levels	Values
GRUPPO	3	A B C

Number of observations in data set = 16

Dependent Variable: PESO

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	91.1380128	30.3793376	35.40	0.0001
Error	12	10.2994872	0.8582906		
Corrected Total	15	101.4375000			

R-Square	C.V.	Root MSE	PESO Mean
0.898465	5.551699	0.92644	16.6875

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GRUPPO	2	12.7375000	6.3687500	7.42	0.0080
LUNGHE	1	78.4005128	78.4005128	91.34	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GRUPPO	2	85.2948111	42.6474055	49.69	0.0001
LUNGHE	1	78.4005128	78.4005128	91.34	0.0001

con i dati dell'esempio sul peso.

Sono stati utilizzati 16 dati campionari, suddivisi in tre gruppi ed indicati con le lettere A, B e C.

Sempre nel riquadro 3) sono riportati i risultati di varie analisi della varianza.

La parte superiore fornisce la varianza d'errore e la parte inferiore le varianze relative ai confronti delle medie dei 3 gruppi (df = 2) con il metodo delle Y ridotte e la stima della significatività della regressione lineare (df = 1).

4)

General Linear Models Procedure

Tukey's Studentized Range (HSD) Test for variable: PESO

NOTE: This test controls the type I experimentwise error rate.

Alpha= 0.05 Confidence= 0.95 df= 12 MSE= 0.858291
Critical Value of Studentized Range= 3.773

Comparisons significant at the 0.05 level are indicated by '***'.

GRUPPO Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
B - C	-0.9966	0.5000	1.9966	
B - A	0.6034	2.1000	3.5966	***
C - B	-1.9966	-0.5000	0.9966	
C - A	0.0369	1.6000	3.1631	***
A - B	-3.5966	-2.1000	-0.6034	***
A - C	-3.1631	-1.6000	-0.0369	***

Nel riquadro 4) sono riportati i confronti multipli tra le tre medie, con il metodo di Tukey.
Per ogni coppia di medie è riportata la differenza, sia positiva che negativa, con i miti dell'intervallo di confidenza. La differenza risulta significativa alla probabilità prefissata (nel tabulato uguale a 0.05) quando l'intervallo fiduciale, che ovviamente comprende la differenza media, esclude lo 0.

18.18. CONFRONTO TRA QUATTRO OUTPUT INFORMATICI SULLA REGRESSIONE LINEARE SEMPLICE: SAS, MINITAB, SYSTAT, SPSS

Quando si passa dallo studio dalla teoria e dalle formule statistiche alla loro applicazione con dati elaborati mediante programmi informatici, un **problema pratico** non trascurabile è la capacità di **leggere e interpretare i risultati** degli output. Nel momento in cui si passa dalle aule ai laboratori, spesso si trovano situazioni differenti da quelle apprese dai libro. E' un passaggio professionale, ai quali un corso di statistica spesso non prepara, anche quando viene accompagnato da applicazioni al computer.

Al primo approccio, output nuovi presenta difficoltà pratiche, poiché in molti casi:

- sono impostati graficamente in modo dissimile;
- usano **termini tecnici differenti**, tra loro e da quelli del testo adottato;
- riportano **analisi statistiche, che sono diverse**, almeno in parte.

Sono situazioni che si presentano anche nel caso più semplice della regressione e della correlazione lineari semplici, dove l'output è limitato a una o al massimo a due pagine.

Per preparare a questa situazione, alcuni testi riportano e confrontano gli output di programmi informatici.

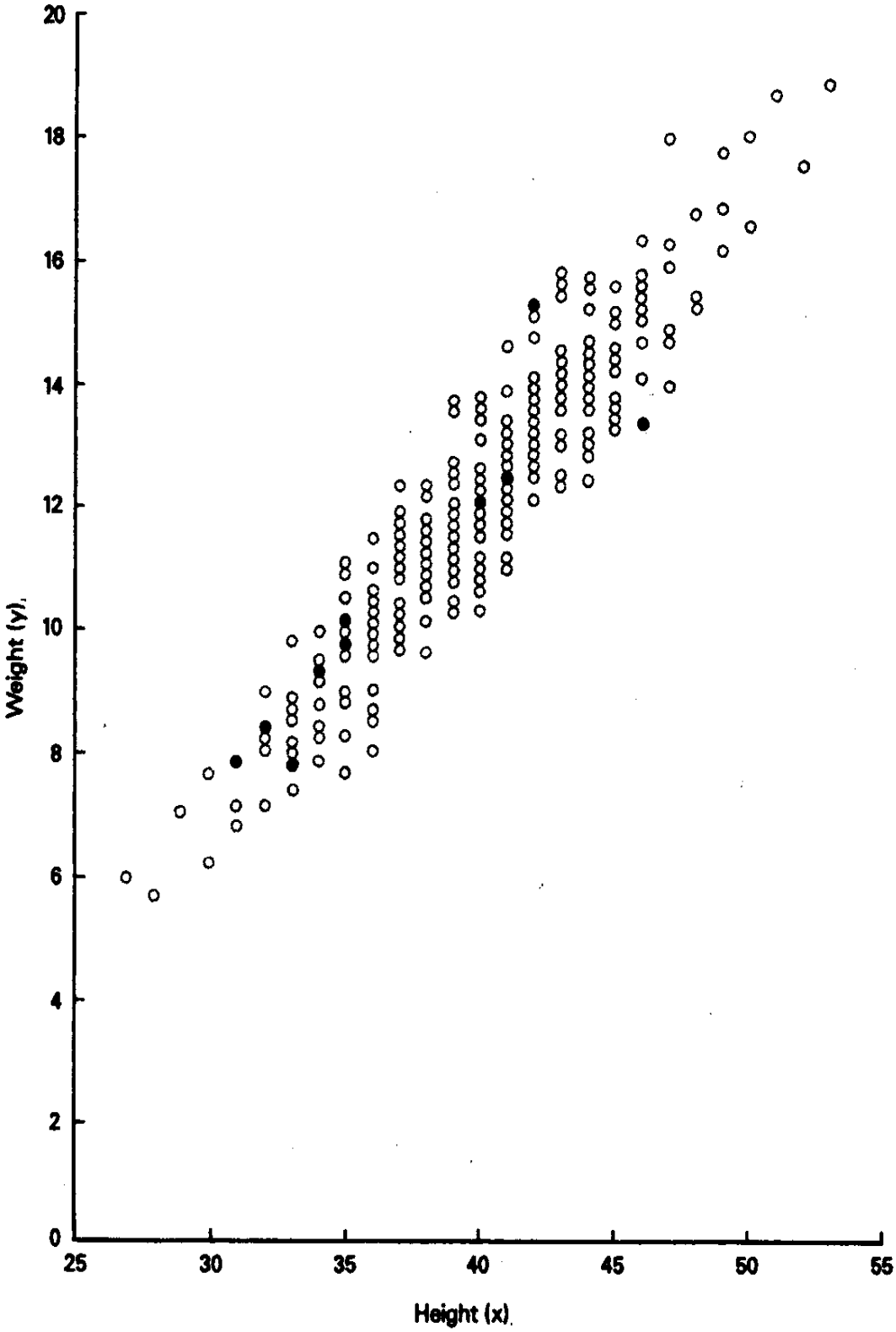
Per queste dispense, sono state riprese alcune pagine del volume di Stanton A. **Glanz** e Bryan K. **Slinker** del 2001 ***Primer of applied regression and analysis of variance*** (2nd ed. Mc Graw-Hill, Inc., New York, 27 + 949).

Nella diagramma di dispersione successivo, i cerchi rappresentano una popolazione di 200 marziani, dalla quale è stato estratto di un campione di 10 individui, indicati dai cerchi anneriti. Di queste 10 unità campionarie, sono stati misurati l'altezza (sull'asse delle ascisse) e il peso (sull'asse delle ordinate), allo scopo di studiare con le metodologie statistiche le caratteristiche di questi esseri misteriosi.

Nelle due pagine successive sono riportati gli output di quattro programmi statistici a grande diffusione internazionale: SAS, MINITAB, SYSTAT, SPSS, scelti tra i tanti sul mercato, secondo la versione in commercio nell'anno 2000.

E' evidente la diversa impostazione grafica, nella quale è necessario individuare le informazioni che forniscono i parametri della retta, della correzione e la loro significatività.

Un primo problema da risolvere è il **differente numero di cifre decimali per ogni parametro:** si va dalle otto del SAS, alle due o tre degli altri programmi. Il numero da riportare nell'articolo o nel rapporto scientifico dipende dalla precisione delle misure introdotte nell'input e dalle dimensioni del campione.


20
18
16
14
12
10
8
6
4
2
0
Weight (y)
25
30
35
40
45
50
55
Height (x)

SAS

DEP VAR: W Weight

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	1	44.35755625	44.35755625	47.706	0.0001
ERROR	8	7.43844375	0.92980547		
TOTAL	9	51.79600000			

ROOT MSE	0.9642642	R-SQUARE	0.8564
DEP MEAN	10.38	ADJ R-SQ	0.8384
C.V.	9.289636		

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB> $\|T\|$
INTERCEP	1	-6.0076	2.39213153	-2.511	0.0363
H	1	0.44410849	0.06429857	6.907	0.0001

MINITAB

The regression equation is
W = -6.008 + 0444 H

Predictor	Coef	Stdev	t-ratio	P
Constant	-6.008	2.392	-2.51	0.036
H	0.4441	0.06430	6.91	0.000

S = 0.9643 R-sq = 85,6% R-sq(adj) = 83,8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	44.358	44.358	47.71	0.000
Error	8	7.438	0.930		
Total	9	51.796			

SYSTAT

```
DEP VAR: W    N: 10   MULTIPLE R: 0.925   SQUARES MULTIPLE R: 0.856

ADJUSTED SQUARED MULTIPLE R: 0.838   STANDARD ERROR OD ESTIMATE: 0.964


EFFECT  COEFFICIENT  STD ERROR  STD COEF  TOLERANCE      T  P(2 TAIL)

CONSTANT     -6.008      2.392     0.000            -2.511     0.036
H             0.444      0.064     0.925      1.000  6.907     0.000


                          ANALYSIS OF VARIANCE

SOURCE      SUM-OF-SQUARES     DF    MEAN-SQUARE     F-RATIO         P
REGRESSION          44.358      1         44.358      47.706     0.000
RESIDUAL             7.438      8          0.930
```

SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.925[a]	.856	.838	.9643

[a] Predictor: (Constant), H

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	44.358	1	44.358	47.706	.000[a]
	Residual	7.438	8	0.930		
	Total	51.796	9			

[a] Predictor: (Constant), H

[b] Dependent Variable: W

Coefficients[a]

Model		Understandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	-6.008	2.392		-2.511	.036
	H	.444	.064	.925	6.907	.000

[a] Dependent Variable: W

Con i dati di questo esempio, basterebbe un decimale o al massimo due, data la precisione all'unità con la quale sono state misurate la variabile dipendente e quella indipendente.
Nella lettura dei tabulati è bene seguire un percorso logico, secondo la sequenza:
1 - individuare la variabile dipendente e quella indipendente, con indicazione dei valori della retta di regressione;
2 - valutare la significatività dell'intercetta e del coefficiente angolare, con il test t di Student bilaterale; se l'ipotesi era unilaterale, tale probabilità deve essere dimezzata;
3 - interpretare i risultati del test della linearità, con il test F;
4- e il valore di R-quadro, per un giudizio sulla predittività della retta.

Oltre a
- piccole differenze nelle indicazioni, quali ***R-SQUARE*** in SAS, ***R-sq*** in MINITAB, come ***SQUARED MULTIPLE R*** in SYSTAT e ***R Square*** in SPSS,
- e al fatto che **il valore della correlazione r = 0,925** sia riportato solamente in SYSTAT dove è indicato con ***MULTIPLE R*** e in SPSS dove è indicato con R,
tra i quattro output esistono alcune diversità nel linguaggio:
- l'intercetta è chiamata ***intercept*** nel SAS mentre è chiamata ***constant*** in MNITAB, SPSS e SYSTAT;
- i coefficienti dell'equazione della regressione sono indicati con ***coefficient*** in MINITAB e SYSTAT, mentre sono chiamati ***parameter estimate*** in SPSS e ***B*** in SPSS;
- la devianza (SS) e la varianza (S) della regressione sono chiamate ***model*** nel SAS e ***regression*** negli altri tre programmi;
- l'errore standard, vale a dire la radice quadrata della varianza d'errore, è indicata con ***std. error of the estimate*** in SPSS e SYSTAT, mentre è indicato con ***s*** nel programma MINITAB e ***root MSE*** nel SAS.