CAPITOLO XIII

TRASFORMAZIONI DEI DATI CON TEST PER NORMALITA' E PER OUTLIER

C A P I T O L O XIII

TRASFORMAZIONI DEI DATI;
TEST PER NORMALITA' E PER OUTLIER

13.1. MOTIVI DELLE TRASFORMAZIONE DEI DATI

Per essere applicati nel rispetto pieno delle condizioni di validità, i test di **statistica parametrica** già illustrati (il test **t** e il test **F** nelle loro svariate modalità) e quelli che saranno presentati nei capitoli prossimi (la regressione, la correlazione e la covarianza) richiedono che la distribuzione delle osservazioni sperimentali rispetti

- **condizioni di carattere formale,**
- **condizioni di carattere sostanziale**.

Senza la dimostrazione che i dati rilevati sono in accordo con questi presupposti, **qualunque conclusione raggiunta con un test parametrico può essere posta in discussione** e i risultati essere contestati.

Le condizioni di carattere formale sono fondamentalmente tre.

a) La completezza dei dati nei disegni sperimentali rigidi, come i blocchi randomizzati e i quadrati latini: **se mancano uno o più dati** occorre procedere alla loro integrazione.

b) La **presenza di dati uguali a zero o indefiniti**, che determinano l'impossibilità di ricavare la somma e quindi la media. Ad esempio, in caso di trasformazione in logaritmi un dato uguale a 0 (zero) determina un valore uguale a $-\infty$ (meno infinito); quando si misurano tempi di risposta a uno stimolo e la cavia non reagisce, il tempo diventa infinito o indeterminato. Questi problemi devono essere risolti con trasformazioni adeguate.

c) Una **diversa attendibilità dei risultati.** Percentuali e rapporti, calcolati alcuni su campioni di poche unità e altri su campioni grandi, non hanno la stessa attendibilità. Misure rilevate con precisione e altre indicate come **> X** oppure **≤ X** determinano una condizione insanabile per l'uso della statistica parametrica. Con questi dati è necessario ricorrere a test di statistica non parametrica, in quanto l'unica vera informazione è quella di rango.

Le condizioni di carattere sostanziale che un test parametrico deve rispettare sono fondamentalmente quattro e riguardano:

a) gli **effetti del trattamento,** che devono essere **additivi**;

b) gli **errori e le osservazioni,** che devono essere **indipendenti**;

c) la **distribuzione** degli errori e quella delle osservazioni, che devono essere **normali**;

d) se i dati sono suddivisibili in gruppi, le loro **varianza** devono essere **omogenee**.

A - Gli **effetti di due o più trattamenti possono combinarsi tra loro per addizione o per moltiplicazione**. L'argomento è stato presentato nell'analisi dell'interazione, con la rappresentazione grafica e il confronto tra tabelle di medie osservate e medie attese. In un trattamento a blocchi randomizzati senza repliche, occorre verificare se gli effetti sono indipendenti dalla media dei blocchi. Quando si hanno effetti moltiplicativi, si può ritornare al modello additivo mediante la trasformazione logaritmica dei dati, utilizzando le proprietà matematiche dei logaritmi.

B - L'indipendenza delle osservazioni è realizzata quando una rilevazione non è influenzata da quella precedente o comunque vicina. La dipendenza risulta più spesso da una correlazione nel tempo che nello spazio, per un trascinamento dell'informazione; può succedere quando lo strumento di misura viene alterato o semplicemente influenzato dall'osservazione precedente oppure un individuo può essere più simile a quelli vicini.
Si ha **indipendenza degli errori** quando i termini che definiscono la varianza d'errore sono distribuiti in modo casuale. Invece quando si evidenziano lunghe successioni di scarti positivi e di scarti negativi oppure scarti positivi e negativi tendono ad alternarsi con regolarità, sorgono forti sospetti sulla correttezza del campionamento. Le metodologie statistiche per evidenziare la presenza di questi fattori sono già presentate in alcuni test non parametrici.
La probabilità che una osservazione presenti un certo errore non deve dipendere né dal segno né dalla sua grandezza, ma essere assolutamente casuale.

Per ottenere l'indipendenza delle osservazioni e degli errori, è necessario che nella sperimentazione il ricercatore tenga in considerazione l'**effetto random**, l'**effetto ambiente** e l'**effetto trattamento,** attenendosi ai seguenti principi:
- la randomizzazione o scelta casuale del campione dalla popolazione dei dati possibili deve essere fondata su elementi obiettivi, come l'estrazione di numeri casuali da una tabella o da un sacchetto, per generazione casuale dal calcolatore; non deve mai essere lasciata all'arbitrio di un individuo (**effetto random**);
- ogni dato deve avere le stesse possibilità di essere influenzato da varie circostanze di tempo e di luogo (**effetto ambiente**);
- tutti gli individui del campione devono avere le stesse possibilità di essere sottoposti a un trattamento qualunque (**effetto trattamento**).

C - I test parametrici sono validi se **la distribuzione dei dati è normale e quindi quella degli errori è normale intorno alla media**. La verifica avviene con il controllo della simmetria e della curtosi. Le

conseguenze della **non normalità degli errori** spesso non sono gravi. Solamente una fortissima asimmetria ha effetti sul livello di significatività del **test F e** del **test t,** che **sono** ritenuti **robusti rispetto a questa condizione**; la correlazione e la regressione ne risentono maggiormente.

D - L'omogeneità delle varianze o **omoschedasticità** (omoschedalità in altri testi) viene verificata mediante i test già illustrati per due e per più campioni. Nella statistica parametrica, tutti i confronti tra le medie e la stima degli effetti aggiunti sono fondati sull'assunto che tutti i gruppi abbiano la stessa **varianza naturale o varianza vera (σ^2)**; se le varianze non sono omogenee, si determina una variazione del peso relativo dei gruppi sul valore della varianza d'errore.
Quando si rifiuta l'ipotesi di omoscedasticità, si può classificare l'**eteroscedasticità** come **regolare o irregolare**.
- La **eteroscedasticità** è detta **irregolare,** quando non si evidenzia alcun rapporto tra media e varianza. Può derivare da cause aberranti, come la presenza di un dato anomalo, oppure da una non corretta impostazione dell'esperimento. In questi casi, si deve verificare se si tratta di sbagli commessi dallo sperimentatore (come nella trascrizione dei dati) o di variazioni reali. Nel primo caso, si dovrebbe ripetere l'esperimento, se non è possibile individuare la causa e apportare la correzione. Nel secondo, si può procedere alla trasformazione dei dati, con uno dei metodi che verranno di seguito presentati.
- La **eteroschedasticità** è detta **regolare,** quando esiste una relazione di tipo noto, come nella distribuzione poissoniana, o comunque una relazione evidenziabile con i metodi della statistica descrittiva. In questo caso si opera la trasformazione dei dati, che spesso è specifica per ogni tipo di distribuzione; ad esempio per la poissoniana, quella ritenuta più adeguata è la trasformazione in radice quadrata.

Quando un ricercatore deve applicare un test a dati campionari, per i problemi derivanti dalla non-normalità, dalla eterogeneità delle varianze e dalla non additività, secondo il volume di Charles J. **Krebs** del 1999 (vedi ***Ecological Methodology***, 2nd ed. Addison Wesley Longman, Menlo Park, pp. 12 + 620) egli può scegliere tra quattro soluzioni:
1 - ricorrere a metodi non parametrici, anche se si determina una perdita nell'informazione della misura rilevata, poiché da una scala di rapporti o di intervalli si scende a una scala di rango o binaria;
2 - utilizzare una trasformazione dei dati, che elimina i tre problemi elencati in precedenza e offre il vantaggio di applicare ugualmente il test parametrico;
3 - utilizzare ugualmente il test parametrico senza trasformare i dati, contando sulla **robustezza del test**; è una soluzione accettata soprattutto quando il campione è grande ma, anche secondo Krebs, è una procedura da non raccomandare e che in questi ultimi anni è sempre più contestata;
4 - ricorrere ai nuovi metodi di ricampionamento (come il **jackknife** e il **bootstrap**), resi possibili dall'uso intensivo del computer.

Riprendendo in modo schematico i concetti illustrati, con la **trasformazione dei dati si effettua un tentativo**, che in varie situazioni raggiunge lo scopo, di ottenere

i **tre scopi principali**:

1 - **stabilizzare le varianze**,

2 - **linealizzare le relazioni tra variabili**,

3 - **normalizzare le distribuzioni**,

e **due scopi secondari**:

1 - **semplificare l'elaborazione di dati** che presentano caratteristiche non gradite,

2 - **rappresentare i dati in una scala ritenuta più adatta**.

13.2. ALCUNE TRASFORMAZIONI DEI DATI

Le trasformazioni riportate in letteratura e alle quali più frequentemente si ricorre sono:

- la lineare,
- la logaritmica,
- le potenze, che comprendono le radici e soprattutto la radice quadrata e cubica, la reciproca e la quadratica,
- le angolari,
- i probit, i logit, i normit.

La trasformazione lineare consiste nel **cambiamento di scala** o dell'origine delle misure, per facilitare la loro comprensione delle caratteristiche dei dati o i calcoli da effetuare. Può essere **moltiplicativa**, **additiva** e una **combinazione di queste due modalità**. E' il caso della trasformazione

- della temperatura da gradi Celsius a Fahrenheit (trasformazione additiva e moltiplicativa),
- di una lunghezza da pollici a centimetri m (trasformazione moltiplicativa),
- di una serie di dati che (ad es.) variano da 230 a 279 a valori da 1 a 49 (trasformazione additiva).

Questa ultima è la semplice sottrazione dello stesso valore a tutti i dati, che serve soprattutto per semplificare i calcoli.

In una **trasformazione moltiplicativa**, la **variabile trasformata** ($\mathbf{X_T}$) è ottenuta con una semplice moltiplicazione della **variabile originaria** ($\mathbf{X_0}$)

$$X_T = C \cdot X_0$$

dove **C** è la costante di conversione

(ad esempio, c = 2,54 per trasformare i pollici in cm e c = 0,394 nel capo opposto).

In questa trasformazione, seguono la stessa legge

- sia **la media** ($\overline{X}$)

$$\overline{X}_T = C \cdot \overline{X}_0$$

- sia **la deviazione standard e l'errore standard** (indicati genericamente con S)

$$S_T = C \cdot S_0$$

Non variando la forma della distribuzione né i rapporti tra media e varianza, la **trasformazione lineare** risulta **inutile quando si intende modificare le caratteristiche fondamentali della distribuzione**. A questo scopo, assumono importanza le trasformazioni non lineari, nelle quali a distanze uguali nella prima distribuzione non corrispondono distanze uguali anche in quella modificata.

Tutte le trasformazioni di seguito riportate, le più ricorrenti nelle applicazioni della statistica, non sono lineari.

La trasformazione in ranghi è una tecnica molto semplice, sempre più frequentemente raccomandata da autori di testi internazionali**.** Quando i dati sono abbastanza numerosi, utilizzare i ranghi al posto dei valori originari permette di ricostruire le condizioni di validità e di applicare tutti i test parametrici. Quando il campione è abbastanza numeroso (n > 30), i ranghi sono sempre distribuiti in modo normale; inoltre questa trasformazione elimina immediatamente l'effetto dei valori anomali.

E' utile soprattutto nel caso di disegni sperimentali complessi, a tre o più fattori con eventuale interazione o analisi gerarchica, per i quali nella statistica non parametrica non esistono alternative ai test di statistica parametrica. Questo accorgimento permette anche di superare **uno dei limiti fondamentali della statistica non parametrica, che offre test di significatività ma che raramente e con difficoltà è adattabile ai problemi, non meno importanti, di stima accurata dei parametri**.

La trasformazione logaritmica

$$Y = \log_a X$$

di solito avviene con **base 10** o con **base naturale (e)**, anche se non sono infrequenti quelli con **base 2**. Ha vari scopi.

Si applica quando la distribuzione ha simmetria positiva, per ottenere una distribuzione normale.

In variabili continue, è utile per rendere omogenee le varianze quando esse crescono all'aumentare della media.

Nel caso di effetti moltiplicativi tra variabili, come nell'interazione, per ritornare agli effetti additivi, richiesti dal modello statistico dell'ANOVA.

La tabella sottostante mostra come dati che possono variare da 2 a 20000 riducano il loro campo di variazione da 0,30 a 4,30 con logaritmi a base 10.

X	2	20	200	2000	20000
Y	0,30	1,30	2,30	3,30	4,30

La scelta della base è secondaria.

Qualunque trasformazione logaritmica (ad esempio a base ***e*** oppure a base **2** oppure **10**) determina effetti simili, anche se più o meno accentuati, poiché i dati trasformati differiscono solamente per una costante moltiplicativa.

Quando i coefficienti di variazione di gruppi a confronto sono approssimativamente costanti, le varianze aumentano in modo direttamente proporzionale alle medie; di conseguenza, confrontando i due gruppi A e B, tra i loro dati esiste la relazione

$$X_{Ai} = C \cdot X_{Bi}$$

dove **C** è la costante della proporzione.

La trasformazione dei dati con i logaritmi

$$\log X_{Ai} = \log C + \log X_{Bi}$$

rende le varianze omogenee, poiché i dati avranno una media differente ma la stessa forma di distribuzione.

La trasformazione logaritmica può essere applicata solamente a valori positivi, in quanto non esistono i logaritmi di valori negativi.

Quando si hanno **valori nulli**, poiché log 0 = - ∞ (meno infinito), la trasformazione richiede l'accorgimento di **aggiungere una costante** (con C = 1 oppure C = 0,5) a tutti i dati (non solo a quelli nulli)

$$Y = \log(X + C)$$

In varie situazioni, la trasformazione logaritmica ha effetti multipli: serve contemporaneamente a stabilizzare la varianza, a ridurre ad effetti additivi un effetto moltiplicativo, a normalizzare la distribuzione.

Si deve ricorrere alla trasformazione logaritmica quando si vuole ottenere una distribuzione normale da una distribuzione di dati caratterizzata da una forte asimmetria destra o positiva; si parla allora di **distribuzione log-normale.**

La **trasformazione in radice quadrata**

$$Y = \sqrt{X}$$

è uno dei casi più frequenti di **trasformazioni mediante potenze,** in cui c = 1/2.

$$Y = X^c$$

E' utile in particolare sia per normalizzare distribuzioni con asimmetria destra (ma meno accentuata rispetto alla trasformazione log) per omogeneizzare le varianze. Spesso è applicata a conteggi, quindi a **valori sempre positivi o nulli**, che seguono la distribuzione poissoniana.

In batteriologia, ematologia, fitosociologia è il caso di conteggi microbiologi oppure di animali o piante dispersi su una superficie. Poiché la varianza (npq) è proporzionale alla media (np), con i dati originali la condizione di omoschedasticità è spesso violata in partenza.

Quasi sempre i dati sono rappresentati da piccoli numeri, poiché all'aumentare della media la distribuzione poissoniana tende alla normale.

Quando si ha la **presenza di almeno uno zero** è consigliabile (per tutti i dati)
la trasformazione

$$Y = \sqrt{X + 0{,}5}$$

che risulta appropriata per valori piccoli, con medie inferiori a 1, in cui la semplice trasformazione in radice quadrata determinerebbe un ampliamento delle distanze tra i valori minori.

Anche delle trasformazione in radice quadrata sono state proposte alcune varianti.

Per ridurre **le relazioni tra varianza e media e quindi stabilizzare le varianze**,

- nel 1948 F. J. **Ascombe** (in ***The trasformation of Poisson, binomial and negative binomial data***, pubblicato su **Biometrika** vol. 35, pp. 246-254) ha proposto

$$Y = \sqrt{X + \frac{3}{8}}$$

- nel 1950 M. F. **Freeman** e J. W. **Tukey** (in ***Transformations related to the angular and square root***, pubblicato su **Annals of Mathematical Statistics** Vol. 21, pp. 607 - 611) hanno proposto

$$Y = \sqrt{X} + \sqrt{X + 1}$$

particolarmente adatta quando $X \leq 2$.

Ma, nonostante l'autorevolezza scientifica dei proponenti, nella pratica hanno avuto poca diffusione.

(Una presentazione più ampia di questa trasformazione è riportata nel capitolo sulle condizioni di validità della regressione e correlazione lineari).

La **trasformazione in radice cubica**

$$Y = \sqrt[3]{X}$$

viene utilizzata per popolazioni che vivono in uno spazio tridimensionale. Come in ecologia per la distribuzione di animali sul terreno si usa la radice quadrata, in idrobiologia per conteggi di plancton che non risentano della crescita esponenziale di tali popolazioni si ricorre abitualmente alla trasformazione in radice cubica.

Nell'analisi di popolazioni che vivono in una spazio tridimensionale, si usa la trasformazione logaritmica quando ha la prevalenza la differenza stagionale, in specie che hanno esplosioni demografiche, per cui si possono avere campioni con poche unità ed altri con varie migliaia di individui; si usa quella in radice cubica se i dati presentano differenze minori e la distribuzione è asimmetrica.

La **trasformazione reciproca**

$$Y = \frac{1}{X}$$

è particolarmente utile nell'**analisi di tempi**, come per confronti sulla sopravvivenza dopo somministrazione di un tossico ad elevata letalità o di reazione a stimoli. Di norma, la maggior parte delle reazioni cadono in un intervallo relativamente ristretto e con distribuzione approssimativamente normale; ma esistono anche individui che hanno tempi di reazione molto alti, con valori che alterano profondamente la distribuzione con una simmetria a destra.

Per tale asimmetria, sono quindi usate sia la trasformazione log sia quella in radice quadratica o cubica. La scelta tra esse dipende anche dalle caratteristiche della distribuzione dei dati.

La trasformazione reciproca serve per stabilizzare la varianza, quando essa aumenta in modo molto pronunciato rispetto alla media.

Quando uno o più individui non manifestano reazioni allo stimolo, il tempo diventa infinito: è impossibile fare la somma, calcolare la media e tutti le altre misure da essa derivate. La trasformazione reciproca, che attribuisce alla variabile $Y = \infty$ il valore zero, permette la stima di tutti parametri. Con essa, valori elevati di X corrispondono a valori di Y prossimi allo zero ed aumenti molto elevati in X producono effetti trascurabili in Y.

Per l'interpretazione sui risultati conviene ritornare alla scala originale, come per la media armonica.

Ad esempio, si supponga che i tempi di sopravvivenza in minuti di 5 insetti a una dose standard di DDT siano stati: 4, 5, 2, 10, ∞. Il quinto è sopravvissuto.

Quale è il tempo medio di sopravvivenza?

Dopo aver effettuato il reciproco ottenendo 0,25 0,20 0,50 0,10 0,0

si ricava la media 1,05/5 = 0,21.

Si ritorna alla scala originale in minuti, attraverso la relazione 1/0,21 = 4,76.

La **trasformazione quadratica**

$$Y = X^2$$

è utile in situazioni opposte a quelle fino ad ora presentate, cioè quando la **varianza tende a decrescere all'aumentare della media** e la distribuzione dei dati ha una forte **asimmetria negativa**. Sono fenomeni rari nella ricerca ambientale e biologica. Pertanto, il suo uso è poco diffuso.

La **trasformazione cubica**

$$Y = X^3$$

si utilizza quando la **asimmetria negativa è ancor più marcata**. Ma i casi sono rarissimi.

La **trasformazione angolare** o in **gradi** mediante **arcoseno**

$$Y = \text{arcsen}\sqrt{\frac{p}{100}}$$

quando p è la percentuale, altrimenti

$$Y = \text{arcsen}\sqrt{p}$$

quando p è la proporzione da 0 a 1

oppure la **trasformazione seno inverso**

$$Y = \text{sen}^{-1}\sqrt{\frac{X}{n}}$$

dove X è il numero di casi positivi su un campione di n dati.

Sono distribuzioni di tipo binomiale, che hanno un valore della varianza (pq) determinato da quella della media (p).

Per l'uso dei test parametrici, percentuali e frazioni presentano alcuni problemi, che richiedono analisi preliminari, poiché sono utilizzati per rendere le osservazioni indipendenti dalle dimensioni del campione. Per esempio, tra un primo esperimento che abbia fornito 3 risposte positive su 4 tentativi, un secondo che ne abbia dato 81 su 100 tentativi ed un terzo con 248 su 300 si può effettuare il confronto ricorrendo ai loro rapporti (rispettivamente 0,75 per il primo; 0,80 per il secondo; 0,83 per il terzo) oppure mediante percentuale (75%, 80% e 83%). Ma ognuno di questi dati ha una "attendibilità" diversa e un intervallo di confidenza differente; di conseguenza, non possono essere elaborati insieme.

Quando si dispone di percentuali e rapporti, occorre preliminarmente verificare su quali dimensioni del campione sono stati calcolati. L'analisi con test parametrici è accettabile solamente se le dimensioni sono relativamente simili: non è possibile elaborare insieme percentuali stimate su poche unità con altre stimate su un centinaio di individui od oltre. In statistica 3/4 non è

uguale a 15/20, se con il primo si intendono 3 risposte positive su 4 individui ed con il secondo 15 risposte positive su 20.
Con questi dati è possibile solamente un **test non parametrico,** poiché la **informazione reale** fornita da una serie di tali valori **è quella di rango**.

Una volta che sia stato chiarito questo aspetto, occorre passare alla trasformazione angolare. Una proporzione con media **p** ha una varianza uguale a **p(1-p)**: ha valori massimi per **p** prossimo a **0,5** e ha valori progressivamente minori per **p** che tende a **0** oppure a **1**. La trasformazione angolare ha la caratteristica opposta: determina variazioni maggiori agli estremi che al centro della scala, riconducendo i rapporti tra le varianze a valori di omoschedasticità.
La proporzione **p** che varia da 0 a 1 (o la percentuale da 0 a 100% tradotta in proporzione) è espressa in gradi φ che variano da 0 a 90, mediante la relazione già indicata. La tabella della pagina successiva permette di trasformare la proporzione direttamente in gradi.

Ad esempio,
- una proporzione p = 0,12 diventa φ = 20,3
- una proporzione p = 0,75 diventa φ = 60,0.

Questa trasformazione è poco precisa per valori di p prossimi a 0 oppure a 1. Per tali proporzioni sono proposte altre tabelle, anche se i computer e le semplici calcolatrici tascabili (utilizzando il simbolo $\boldsymbol{sin^{-1}}$) spesso contengano questa trasformazione.
Nella tabella per valori estremi prossimi a 0, ad esempio,
- una proporzione p = 0,0012 (o 1,2 per mille) diventa φ = 1,99
- una proporzione p = 0,9958 (o 99,58%) diventa φ = 9,10.

Nella tabella per valori estremi prossimi a 1, ad esempio,
- una proporzione p = 0,95 (o 95%) diventa φ = 77,08
- una proporzione p = 0,025 (o 2,5%) diventa φ = 86,28.

TABELLA DI TRASFORMAZIONE DI PROPORZIONI (con p da 0,01 a 0,99)

IN GRADI φ (da 5,7 a 84,3)

φ = arco seno $\sqrt{p}$

P	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	---	5,7	8,1	10,0	11,5	12,9	14,2	15,3	16,4	17,5
0,10	18,4	19,4	20,3	21,1	22,0	22,8	23,6	24,4	25,1	25,8
0,20	26,6	27,3	28,0	28,7	29,3	30,0	30,7	31,3	31,9	32,6
0,30	33,2	33,8	34,4	35,1	35,7	36,3	36,9	37,5	38,1	38,6
0,40	39,2	39,8	40,4	41,0	41,6	42,1	42,7	43,3	43,9	44,4
0,50	45,0	45,6	46,1	46,7	47,3	47,9	48,4	49,0	49,6	50,2
0,60	50,8	51,4	51,9	52,5	53,1	53,7	54,3	54,9	55,6	56,2
0,70	56,8	57,4	58,1	58,7	59,3	60,0	60,7	61,3	62,0	62,7
0,80	63,4	64,2	64,9	65,6	66,4	67,2	68,0	68,9	69,7	70,6
0,90	71,6	72,5	73,6	74,7	75,8	77,1	78,5	80,0	81,9	84,3

(Note)

Per ottenere valori più precisi, che considerino proporzioni alla terza cifra decimale, è sufficiente effettuare la stima mediante l'interpolazione lineare, ma solo per misure abbastanza grandi.

Per proporzioni piccole, utilizzare la relazione

$$\varphi = 57{,}3 \cdot \sqrt{p}$$

Quando i valori di p sono simmetrici rispetto a 0,50 si ha che i valori in gradi sono simmetrici rispetto a 45,0

Se la proporzione è grande, fare la trasformazione di p' dopo il calcolo di p' = 1-p.

TABELLA DI TRASFORMAZIONE DELLE PROPORZIONI PER VALORI ESTREMI:

DA **P = 0,0000** A **P = 0,0099** (PARTE SUPERIORE)

DA **P = 0,010** A **P = 0,100** (PARTE INFERIORE)

P	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,000	0,00	0,57	0,81	0,99	1,15	1,28	1,40	1,52	1,62	1,72
0,001	1,81	1,90	1,99	2,07	2,14	2,22	2,29	2,36	2,43	2,50
0,002	2,56	2,63	2,69	2,75	2,81	2,87	2,92	2,98	3,03	3,09
0,003	3,14	3,19	3,24	3,29	3,34	3,39	3,44	3,49	3,53	3,58
0,004	3,63	3,67	3,72	3,76	3,80	3,85	3,89	3,93	3,97	4,01
0,005	4,05	4,10	4,14	4,17	4,21	4,25	4,29	4,33	4,37	4,41
0,006	4,44	4,48	4,52	4,55	4,59	4,62	4,66	4,70	4,73	4,76
0,007	4,80	4,83	4,87	4,90	4,93	4,97	5,00	5,03	5,07	5,10
0,008	5,13	5,16	5,20	5,23	5,26	5,29	5,32	5,35	5,38	5,41
0,009	5,44	5,47	5,50	5,53	5,56	5,59	5,62	5,65	5,68	5,71

P	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,01	5,74	6,02	6,29	6,55	6,80	7,03	7,27	7,49	7,71	7,92
0,02	8,13	8,33	8,53	8,72	8,91	9,10	9,28	9,46	9,63	9,80
0,03	9,97	10,14	10,30	10,47	10,63	10,78	10,94	11,09	11,24	11,39
0,04	11,54	11,68	11,83	11,97	12,11	12,25	12,38	12,52	12,66	12,79
0,05	12,92	13,05	13,18	13,31	13,44	13,56	13,69	13,81	13,94	14,06
0,06	14,18	14,30	14,42	14,54	14,65	14,77	14,89	15,00	15,12	15,23
0,07	15,34	15,45	15,56	15,68	15,79	15,89	16,00	16,11	16,22	16,32
0,08	16,43	16,54	16,64	16,74	16,85	16,95	17,05	17,15	17,26	17,36
0,09	17,46	17,56	17,66	17,76	17,85	17,95	18,05	18,15	18,24	18,34
0,10	18,43									

TABELLA DI TRASFORMAZIONE DELLE PROPORZIONI PER VALORI ESTREMI:

DA **P = 0,900** A **P = 0,989** (PARTE SUPERIORE)

DA **P = 0,9900** A **P = 1,0000** (PARTE INFERIORE)

P	0	1	2	3	4	5	6	7	8	9
0,90	71,57	71,66	71,76	71,85	71,95	72,05	72,15	72,24	72,34	72,44
0,91	72,54	72,64	72,74	72,85	72,95	73,05	73,15	73,26	73,36	73,46
0,92	73,57	73,68	73,78	73,89	74,00	74,11	74,21	74,32	74,44	74,55
0,93	74,66	74,77	74,88	75,00	75,11	75,23	75,35	75,46	75,58	75,70
0,94	75,82	75,94	76,06	76,19	76,31	76,44	76,56	76,69	76,82	76,95
0,95	77,08	77,21	77,34	77,48	77,62	77,75	77,89	78,03	78,17	78,32
0,96	78,46	78,61	78,76	78,91	79,06	79,22	79,37	79,53	79,70	79,86
0,97	80,03	80,20	80,37	80,54	80,72	80,90	81,09	81,28	81,47	81,67
0,98	81,87	82,08	82,29	82,51	82,73	82,97	83,20	83,45	83,71	83,98

P	0	1	2	3	4	5	6	7	8	9
0,990	84,26	84,29	84,32	84,35	84,38	84,41	84,44	84,47	84,50	84,53
0,991	84,56	84,59	84,62	84,65	84,68	84,71	84,74	84,77	84,80	84,84
0,992	84,87	84,90	84,93	84,97	85,00	85,03	85,07	85,10	85,13	85,17
0,993	85,20	85,24	85,27	85,30	85,34	85,38	85,41	85,45	85,48	85,52
0,994	85,56	85,59	85,63	85,67	85,71	85,75	85,79	85,83	85,86	85,90
0,995	85,95	85,99	86,03	86,07	86,11	86,15	86,20	86,24	86,28	86,33
0,996	86,37	86,42	86,47	86,51	86,56	86,61	86,66	86,71	86,76	86,81
0,997	86,86	86,91	86,97	87,02	87,08	87,13	87,19	87,25	87,31	87,37
0,998	87,44	87,50	87,57	87,64	87,71	87,78	87,86	87,93	88,01	88,10
0,999	88,19	88,28	88,38	88,48	88,60	88,72	88,85	89,01	89,19	89,43
1,000	90,00									

La trasformazione delle proporzioni in arcoseno era già stata discussa da M. S. **Bartlett** nel 1947 (vedi l'articolo ***The use of transformations***, su **Biometrics**, Vol. 3, pp.39-52) in cui per proporzioni **p** calcolate su campioni di **n** dati con **X** casi favorevoli, quindi **p = X/n**

aveva proposto

$$Y = 2arcsin\sqrt{p + \frac{1}{2n}}$$

per proporzioni basse, vicine a 0,

che diventa

$$Y = 2arcsin\sqrt{0 + \frac{1}{4n}}$$

per il caso estremo di p = 0

e dall'altra parte

$$Y = 2arcsin\sqrt{p - \frac{1}{2n}}$$

per proporzioni alte, vicine a 1,

che diventa

$$Y = 2arcsin\sqrt{1 - \frac{1}{4n}}$$

per il caso estremo di p = 1

Nel 1948, F. J. **Anscombe**, (nell'articolo ***The transformation of Poisson, binomial, and negative binomial data***, pubblicato su **Biometrika** Vol. 35, pp. 246-254) come trasformazione migliore ha proposto la trasformazione in

$$Y = arcsin\sqrt{\frac{X + \frac{3}{8}}{n + \frac{3}{4}}}$$

dove, con la simbologia precedente,

X è il numero di casi favorevoli in un campione di **n** dati.

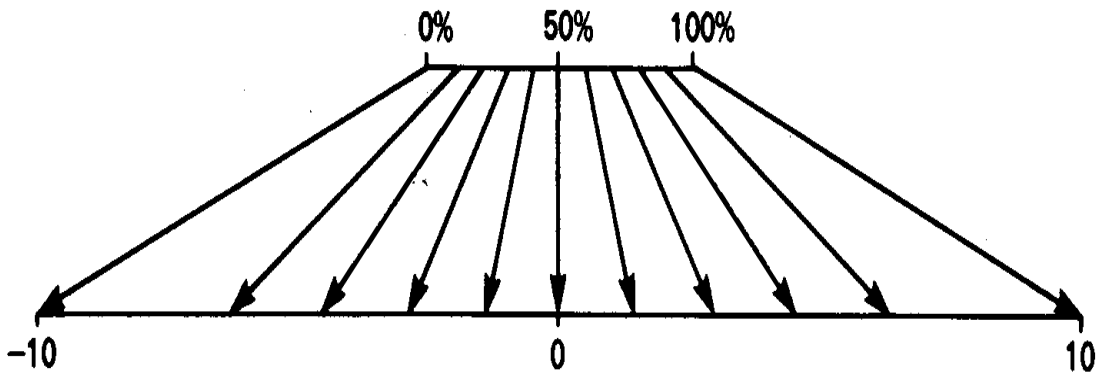
Nel 1950, M. F. **Freeman** e J. W. **Tukey** (con l'articolo ***Transformations related to the angular and the square root***, pubblicato su **Annals of Mathematical Statistics** Vol. 21, pp. 607 - 611) hanno proposto un ulteriore raffinamento, quindi secondo alcuni preferibile,

con

$$Y = \frac{1}{2}\left(arcsin\sqrt{\frac{X}{n+1}} + arsin\sqrt{\frac{X+1}{n+1}} \right)$$

che fornisce risultati molti vicini a quelli di Anscombe, eccetto per valori di p che siano estremi, molto vicini a 0 oppure a 1.

La trasformazione di percentuali o proporzioni con i differenti metodi presentati ha sempre l'effetto di ampliare le differenze verso gli estremi, come illustrato nella figura successiva



In essa a quella per omogeneizzare le varianze è stato aggiunta una trasformazione lineare, che non le modifica, ma fa in modo che il 50% diventi 0 e la distribuzione sia simmetrica intorno a esso.
E' vantaggiosa a livello interpretativo quando il fenomeno atteso ha una frequenza del 50%.

Le percentuali richiedono poi particolare attenzione nella interpretazione dei risultati.
Ad esempio, se un farmaco nuovo determina una riduzione del numero di decessi dal 3% a 1% e in una malattia diversa un altro farmaco determina una riduzione dei decessi dal 12% al 7% chi ha avuto il risultato migliore?
E' vero che il primo ha abbassato la mortalità del 66% (2/3) e il secondo solo del 42% (5/12). Ma in termini di sopravvivenza, su 100 pazienti il primo ha determinato la sopravvivenza di due e l'altro di cinque persone. Ne consegue che è sempre importante presentare chiaramente il problema e lo scopo del confronto.

La **trasformazione seno inverso iperbolico**

$$Y = \sqrt{k}\ \text{sen h}^{-1}\sqrt{\frac{X}{k}}$$

occupa una posizione intermedia tra la trasformazione logaritmica da applicare in variabili poissoniane altamente disturbate e la trasformazione in radice quadrata per variabili poissoniane standard.

La **trasformazione tangente iperbolica inversa**

$$Y = 1/2 \log_e \frac{1+r}{1-r} = \tan h_r^{-1}$$

è analoga alla trasformazione logaritmica ed è applicata a variabili che variano da -1 a +1.
E' utile per normalizzare la distribuzione dei coefficienti di correlazione (r). Come vedremo, essi sono distribuiti normalmente solo per valori intorno allo zero, mentre diventano sempre più asimmetrici avvicinandosi ai valori estremi di +1 e –1.

La **trasformazione log-log**

$$Y = \log_e (-\log_e p)$$

e la **trasformazione log-log complementare**

$$Y = \log_e (-\log_e (1-p))$$

si applicano a percentuali di sopravvivenza, nello studio dei tempi di eliminazione di un gruppo di cavie in dosaggi biologici.

La **trasformazione probit**

$$P = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{z-5} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right) dx$$

(**probits** da ***probability units***) è definita come la deviata normale equivalente, aumentata di 5.
Nello studio della relazione dose-risposta, la percentuale di individui che rispondono all'effetto causato della dose viene di solito rappresentato con una curva cumulata. Essa ha forma sigmoide, se la curva della distribuzione originaria è normale, con la conseguenza che a parità di errore nella dose l'errore nella risposta non è costante, ma varia secondo il punto in cui incontra perpendicolarmente la sigmoide. Per un errore costante nella risposta, occorre trasformarla in una retta.

La curva percentuale cumulata può essere linearizzata in vari modi. Uno dei più diffusi consiste appunto nei probits, ottenuti con due passaggi logici:
1- Sostituire ai valori di p dell'ordinata quelli corrispondenti all'ascissa della distribuzione normale standardizzata

$$Y' = \frac{X-\mu}{\sigma}$$

A causa della simmetria della distribuzione normale, il 50% dei valori Y' è negativo e l'altro 50% è positivo. E' noto e può anche essere osservato sulla tabella della distribuzione normale che meno di 2 valori su 10.000 hanno un valore di Y' inferiore a -3,5.

2 - Successivamente a tutti i valori trasformati in Y' aggiungere la quantità 5: si eliminano tutti i valori negativi.

Questi valori trasformati mediante la relazione

$$Y = 5 + Y' = 5 + \frac{X - \mu}{\sigma}$$

sono i **probits**.

Nei suoi effetti, questa trasformazione è analoga a quella angolare, in quanto i valori verso gli estremi della distribuzione sono più dilatati di quelli collocati nella parte centrale. Il campo di variazione della scala probit tende all'infinito; la scala dei probit si distingue da quella angolare soprattutto nei valori prossimi a 0 e a 1. La trasformazione in probits, rendendo lineare la sigmoide di una cumulata tratta dalla distribuzione normale, permette di trattare la stima dei parametri della distribuzione normale (μ e σ) come quello dei parametri di una regressione lineare (intercetta α e coefficiente angolare β). Ma la stima corretta dei parametri della retta richiede che i punti sperimentali abbiano la stessa varianza; di conseguenza i valori dei probits dovrebbero essere ponderati.

L'effetto linearizzante della trasformazione probit è stato ampiamente utilizzato nelle **carte di probabilità,** usate per verificare in modo semplice e con un grafico se una distribuzione era normale. La diffusione dei calcolatori, che permettono stime rapide ed esatte dei valori di asimmetria e curtosi di una serie di dati campionari, ha reso superflui questi metodi grafici.

La **trasformazione normit**

$$P = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{\mu}{2}\right) du$$

è un'altra trasformazione di percentuali cumulative basate sull'integrale di probabilità della curva normale. Fornisce valori diversi dai probits.

La **trasformazione logit** viene anche essa applicata a osservazioni percentuali ed è ottenuta con

$$Y = \log_e \frac{p}{1-p}$$

L'effetto di questa trasformazione **logistica** o logit è simile a quella probit e può determinare analisi del tutto uguali, in particolare nello studio del dosaggio con risposte quantali.

L'attuale diffusione dell'informatica, che ha superato le difficoltà derivanti dalla complessità dei calcoli e dal tempo richiesto nei calcoli manuali, ha annullato la necessità di linearizzare le distribuzioni. Di conseguenza, le trasformazioni probit e logit sono sempre meno usate.

13.3 ALTRI EFFETTI DELLE TRASFORMAZIONI

Quando si analizza un fenomeno biologico o ambientale, sovente le modalità per misurarlo sono numerose. E' sempre importante che la misura prescelta

- abbia la caratteristica di descrivere meglio delle altre il fattore analizzato e come tale sia facilmente interpretabile,
- determini una distribuzione dei dati sperimentali, che sia in accordo le caratteristiche richieste dalla corretta applicazione del test prescelto.

Ad esempio per valutare la resistenza veloce di atleti su un percorso di 4 Km, come misura della loro capacità atletica è possibile utilizzare

- il tempo impiegato complessivamente (esempio, 12 minuti da un atleta),
- il tempo impiegato per unità di distanza (3 minuti per Km),
- la distanza percorsa per ogni minuto (4 Km / 12 minuti = 0,333 Km/minuto), che è il suo reciproco,
- la distanza teorica percorsa in un'ora alla stessa velocità (20 km/ora).

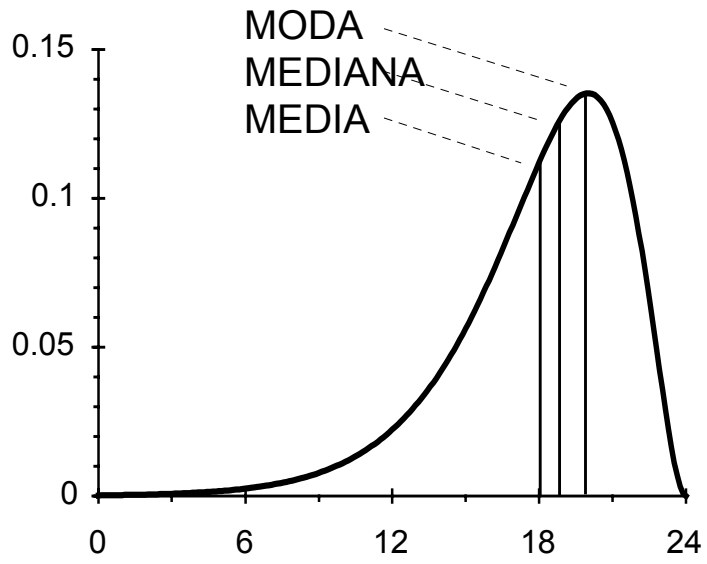
Come più volte accennato, la trasformazione di una serie di dati è giustificata dagli effetti che essa determina sulla forma della distribuzione.

Entro il campo di variazione di una serie osservazioni sperimentali, l'effetto di una trasformazione non è costante. Alcune hanno un effetto più forte sui valori minori, altre su quelli maggiori. Dipende dall'azione congiunta del tipo di trasformazione e della dimensione dei valori.

Nel capitolo scritto da John D. **Emerson** e intitolato ***Introduction to Transformation*** nel testo del 1991 di David Caster **Hoaglin**, Frederick **Mosteller** e John W. **Tukey** dal titolo ***Fundamentals of Exploratory Analysis of Variance*** (A Wiley-Interscience Publication, John Wiley $ Sons, Inc. New York, XVII + 430 p.), le trasformazione sono classificate

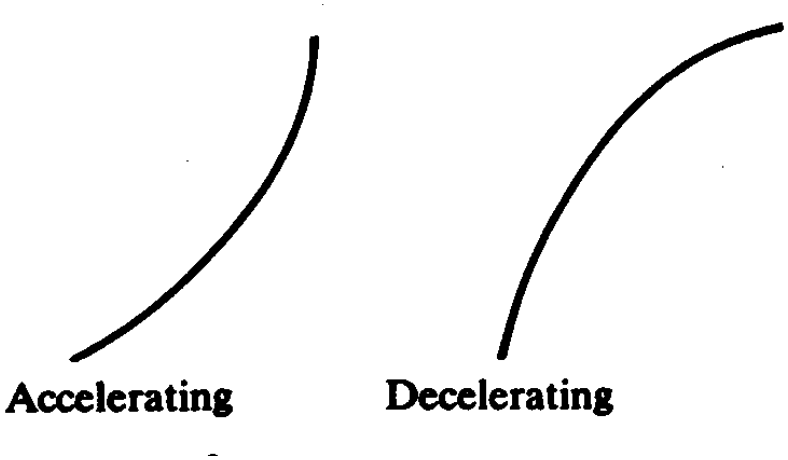
- sulla base degli effetti che esse hanno sulla forma della distribuzione.

Quando i dati hanno **asimmetria sinistra o negativa**, come nella figura sottostante

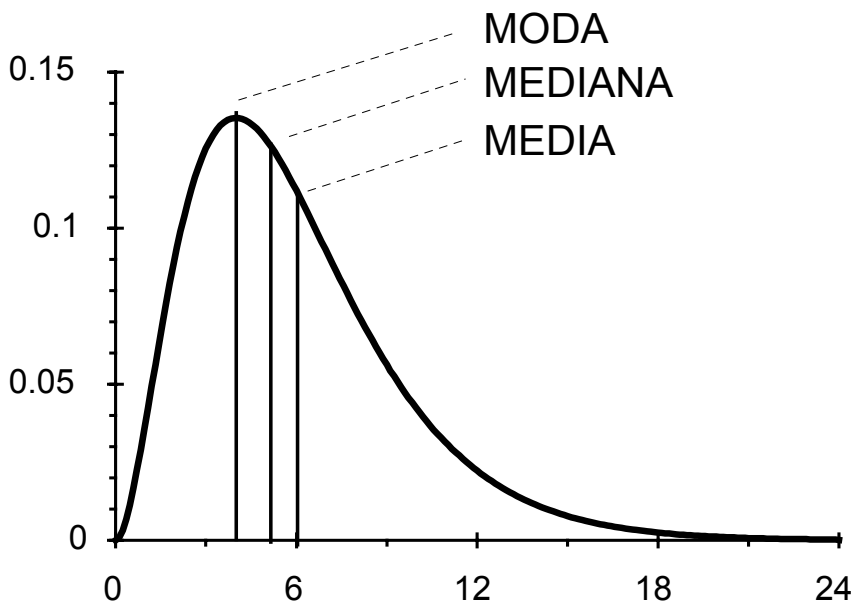


per ottenere una distribuzione normale è necessario distanziare maggiormente i valori alti. Si deve quindi usare una trasformazione che è definita ***accelerating*****,** oppure ***curved up*** o ***cupped*** come sono, ad esempio

- X^2 oppure X^3 quando $X > 1$ (come nel grafico),
- $\sqrt[2]{X}$ oppure $\sqrt[3]{X}$ quando $X < 1$.

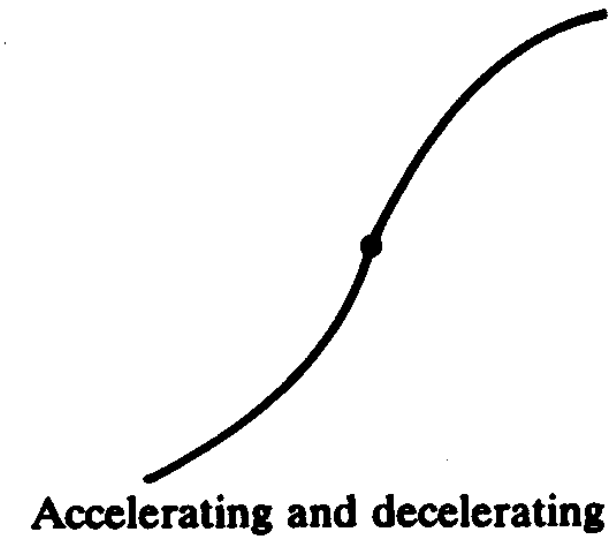


Quando la distribuzione dei dati ha **asimmetria destra o positiva**, come nella figura successiva



per ottenere una distribuzione normale è necessario distanziare maggiormente i valori bassi. Si deve quindi usare una trasformazione ***decelerating*****,** oppure ***curved down*** o ***capped*** come sono quasi tutte, da quelle logaritmica, alla radice quadrata e alla reciproca che inoltre inverte il rango dei valori.

Alcune trasformazioni hanno entrambi gli effetti, ovviamente in aree diverse della distribuzione dei dati

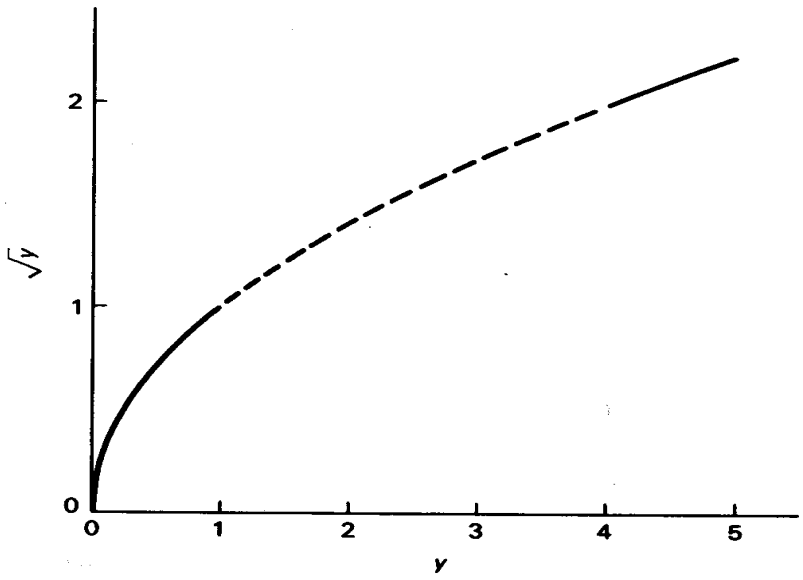


Accelerating and decelerating

Ad esempio, la trasformazione in X^2

- è ***accelerating*** quando $X > 1$
- è ***decelerating*** quando $X < 1$.

mentre la trasformazione in $\sqrt{X}$ ha il comportamento opposto



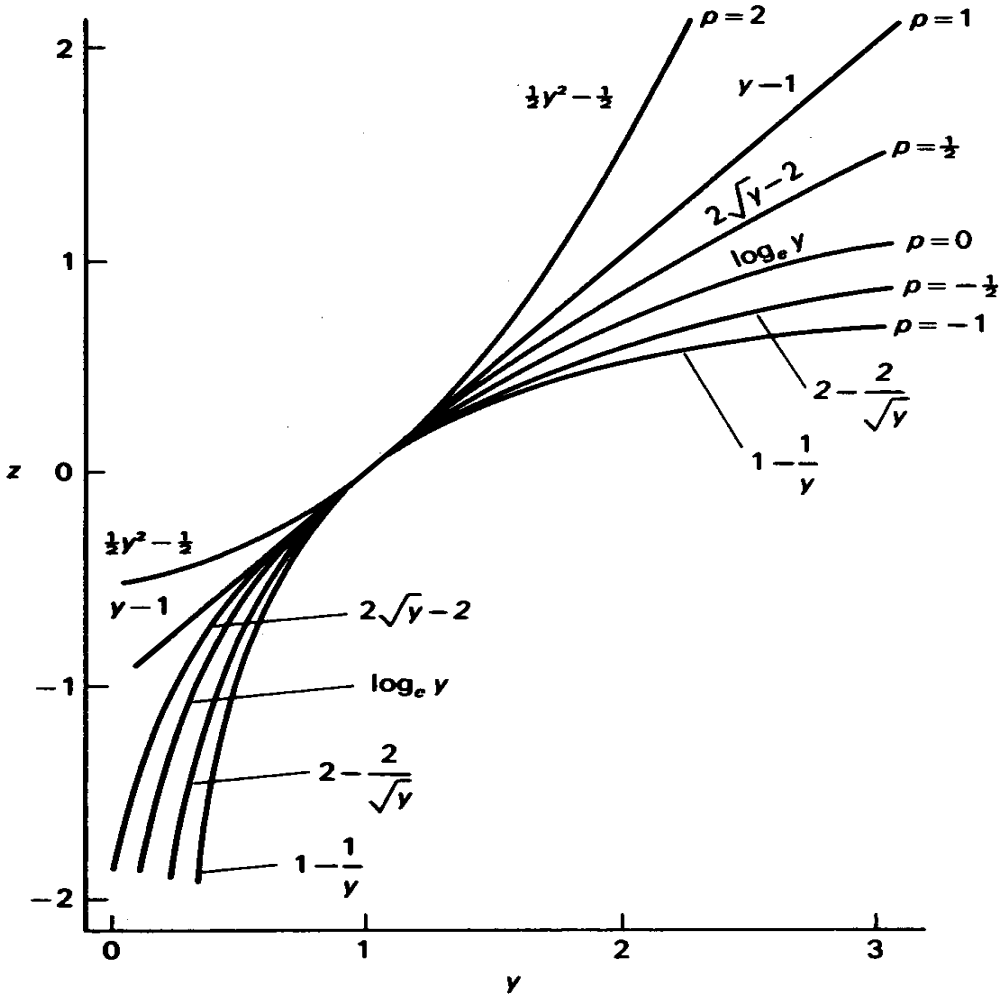
Questa proprietà di modificare i rapporti e le distanze tra i dati dipende dal fatto che la intensità con la quale le trasformazioni agiscono sui singoli valori cambia in funzione della dimensione, come evidenzia la figura precedente.

In essa sono illustrati gli effetti di una trasformazione in radice quadrata. Ad esempio l'azione tra i valori di X (Y nella figura) è molto più forte per i valori compresi tra 0 e 1, rispetto a quelli compresi tra 4 e 5. E' il concetto di ***strenght*** di una trasformazione, che non deve essere confuso con quello di ***accelerating*** (in questo caso tra 0 e 1) e quello di ***decelerating*** (oltre 1), illustrati in precedenza.

Tutte quelle presentate sono **trasformazioni monotoniche**: non modificano l'ordine dei valori.

Al massimo, come in quella reciproca, l'invertono. Questo concetto assume importanza nella statistica non parametrica: **i test fondati sui ranghi non modificano il loro risultato, con qualsiasi trasformazione monotonica**.

Le trasformazioni sono anche classificate in famiglie. Nel grafico successivo sono riportati alcuni membri di una famiglia di trasformazioni di potenza.



Il grafico, che comprende alcune trasformazioni non descritte nel paragrafo precedente, evidenzia con chiarezza sufficiente i loro effetti. Per ulteriori approfondimenti di questi aspetti, si rinvia al testo citato dal quale sono state riprese le figure di questo paragrafo.

13.4. LA SCELTA DELLA TRASFORMAZIONE IDONEA: IL METODO DI BOX-COX

La trasformazione da applicare ad una serie di dati campionari, per rispettare le condizioni di validità dei test parametrici, spesso è conosciuta a priori, sulla base di quanto noto sulle caratteristiche del fenomeno analizzato e del tipo di scala utilizzato per misurarlo. Prima di applicare un test parametrico, è sempre utile ricercare in letteratura **la trasformazione più adeguata, per normalizzare la distribuzione dei dati** raccolti.

Le trasformazioni possibili e le indicazioni per le diverse situazioni, riportate nel paragrafo precedente, sono derivate da queste esperienze. Ma quando si analizza un fattore nuovo, **è difficile individuare la trasformazione più appropriata.**

Il problema si pone soprattutto quando un fenomeno può essere misurato in modi diversi. Per esempio, la velocità di un gruppo di soggetti può essere valutata sia misurando il **tempo** impiegato per concludere un percorso, sia utilizzando il **rapporto tra la distanza e il tempo**. **Le due serie di dati non hanno la stessa forma di distribuzione e quindi l'analisi statistica potrebbe condurre a inferenze differenti**.

Per scegliere il tipo di misura più adeguato, esistono due criteri:

- il primo dipende dalla conoscenza scientifica dell'argomento: **è la misura che meglio valuta il fenomeno e lo rende più comprensibile**;
- il secondo è di tipo tecnico-statistico: **è la misura che ha una forma di distribuzione dei dati più rispettosa delle condizioni di validità del test**, cioè determina una distribuzione normale o approssimativamente tale.

Ma spesso i dati, raccolti sulla base della misura scelta, non sono distribuiti in modo normale nemmeno in modo approssimato. Si pone quindi il problema della loro trasformazione.

Nel **1964,** G. E. **P. Box e** D. R. **Cox** (con l'articolo ***An analysis of transformations (with Discussion)***, pubblicato su **Journal of the Royal Statistical Society**, Series B 26, pp. 211-252) hanno proposto un metodo iterativo e concettualmente complesso, divenuto operativamente semplice e di vasta applicazione con l'uso dei computer, per **individuare quale trasformazione dei dati può meglio normalizzare la loro distribuzione**.

Il metodo ricorre a una famiglia di trasformazioni di potenze. Si ottiene una **X** trasformata ($\mathbf{X_{TRAS}}$) mediante

- la formula

$$X_{TRAS} = \frac{X^{\lambda} - 1}{\lambda}$$

quando $\lambda \neq 0$

- oppure con

$$X_{TRAS} = \log(X)$$

quando $\lambda = 0$

dove il valore di $\boldsymbol{\lambda}$ viene fatto variare da –3 a +3.

Il **valore di λ che meglio normalizza la distribuzione** è quello che **rende massima la funzione L** (nota come **log-likelihood function)**,
con

$$L = -\frac{\nu}{2}\ln s^2_{TRAS} + (\lambda - 1)\frac{\nu}{n}\sum \ln X$$

in cui

- L = valore del log-likelihooh,
- ν = numero di gdl, corrispondente a **n-1**
- n = numero totale di dati,
- **s^2_{TRAS}** = varianza dei dati trasformati, utilizzando l'equazione precedente con λ
- λ = stima del parametro che indica la trasformazione,
- X = valore del dato originale.

Inoltre è **possibile calcolare l'intervallo fiduciale di λ,** entro il quale è conveniente scegliere la trasformazione più adeguata. Benché possa teoricamente assumere qualsiasi valore da –3 a +3 in una scala continua, in pratica λ ha significato pratico solo per alcuni valori. La individuazione di λ non avviene quindi solo sulla base di calcoli, ma attraverso la scelta ragionata entro i limiti fiduciali di S_λ. Questa risposta ricavata dai dati sperimentali e le indicazioni attinte dalle varie esperienze spesso coincidono.
Nel testo del 1978 di George E. P. **Box**, William G. **Hunter** e J. Stuart **Hunter,** "***Statistics for Experimenters. An introdution to Design, Data Analysis and Model Building***", pubblicato da John Wiley & Sons, New York, p. 653 (a pag. 239), per la probabilità α
si propone di stimare S con

$$S = S_\lambda\left(1 + \frac{t^2_{\nu,\alpha/2}}{\nu}\right)$$

- S definisce il limite massimo della deviazione standard minima.

Secondo questi autori, nella successiva analisi della varianza applicata ai dati trasformati, **la devianza e la varianza d'errore perderebbero 1 gdl**, appunto perché vincolate alla condizione di essere le minori possibili nei confronti del fattore considerato.
Il valore di λ individuato corrisponde all'esponente a cui elevare la variabile da trasformare,
cioè

$$\mathbf{X' = X^\lambda}$$

L'elenco dettagliato dei valori abituali di **λ** e delle trasformazioni corrispondenti riporta:

- **λ = 3** indica una trasformazione con elevamento al cubo, cioè $\mathbf{X^3}$ (poiché la distribuzione dei dati originali ha un g_1 molto negativo);

- **λ = 2** indica una trasformazione con elevamento al quadrato, cioè $\mathbf{X^2}$ (da applicare quando la distribuzione dei dati originali ha un indice di asimmetria g_1 meno negativo del precedente);

- **λ = 1** indica una trasformazione lineare, che non modifica la curva della distribuzione , cioè **X** (poiché la distribuzione dei dati ha già una forma simile alla normale);

- **λ = ½** indica una trasformazione con radice quadrata, cioè $\sqrt[2]{X}$ (g_1 è leggermente positivo);

- **λ = 1/3** indica una trasformazione con radice cubica, cioè $\sqrt[3]{X}$ (g_1 è positivo);

- **λ = 0** indica una trasformazione logaritmica, cioè **ln X** oppure **log X** (g_1 è fortemente positivo),

- **λ = -1/3** indica una trasformazione reciproca, con **X** sotto radice cubica, cioè $\frac{1}{\sqrt[3]{X}}$ (g_1 positivo);

- **λ = -1/2** indica una trasformazione reciproca, con **X** sotto radice quadrata, cioè $\frac{1}{\sqrt[2]{X}}$ (g_1 positivo);

- **λ = -1** indica una trasformazione reciproca di **X,** cioè $\frac{1}{X}$ (g_1 positivo);

- $\lambda = -2$ una trasformazione reciproca con **X** al quadrato, cioè $\frac{1}{X^2}$ (g_1 positivo);

- $\lambda = -3$ indica una trasformazione reciproca con **X** al cubo, cioè $\frac{1}{X^3}$ (g_1 positivo).

Quando tra i dati originali è compreso 0 (zero),
l'equazione

$$L = -\frac{\nu}{2}\ln s^2_{TRAS} + (\lambda - 1)\frac{\nu}{n}\sum \ln X$$

è senza soluzione poiché ln 0 = - ∞.
In questi casi, prima della trasformazione occorre aggiungere 0,5 oppure 1 a tutti i valori originari.

ESEMPIO. Una applicazione della trasformazione più adeguata ad una distribuzione di frequenza secondo il metodo di Box-Cox può essere rintracciata nell'ottimo volume di metodi applicati all'ecologia di Charles J. **Krebs** del 1999 (***Ecological Methodology***, 2nd ed., Addison Wesley Longman, Menlo Park, California, pp. XII + 620, nelle pagg. 552-554). Poiché la procedura richiede molti calcoli ed è utile alla comprensione del metodo che tutti i passaggi siano riportati in dettaglio, il campione utilizzato è molto piccolo.

Si assuma che siano state rilevate le seguenti 6 misure

55	23	276	73	41	97

Con estrema evidenza dalla semplice lettura dei dati, anche senza esperienza di analisi statistiche, non appare logico assumere che questi valori siano stati estratti da una popolazione distribuita in modo normale; se non altro è evidente l'asimmetria destra, per la presenza di un valore (276) molto più alto degli altri.
Si tratta di individuare la trasformazione più adeguata per questi dati, affinché la loro distribuzione possa assumere forma normale, almeno in modo approssimato.

Risposta. Per stimare **L (log-likelihood function)**,
con

$$L = -\frac{\nu}{2}\ln s^2_{TRAS} + (\lambda - 1)\frac{\nu}{n}\sum \ln X$$

- in cui
- $\nu = 5$
- n = 6

si devono prima ricavare sia i valori lnX sia le X_{TRAS} da cui ricavare la loro varianza (s^2_{TRAS}), per una serie di valori di λ, che normalmente variano da –3 a + 3.
Nell'esempio citato, il valore L è calcolato per i seguenti valori di λ: -3, -2, -1, -0,5, 0, +0,5, +1, +2.; ma potrebbe essere fatto per tutti i 60 decimali compresi nell'intervallo tra –3,0 e +3,0.
Ognuno degli 8 valori λ indicati richiede vari passaggi, per ottenere il valore di L corrispondente.

Poiché i calcoli sono simili, la illustrazione è limitata al solo caso di **λ = -2,**
cioè alla trasformazione

$$X' = \frac{1}{X^2}$$

Come primo passo, si trasformano i valori di X mediante la relazione

$$X_{TRAS} = \frac{X^{\lambda} - 1}{\lambda}$$

Per X = 55 e λ = -2 si ricava

$$X_{TRAS} = \frac{55^{-2} - 1}{-2} = \frac{\frac{1}{3025} - 1}{-2} = \frac{0{,}0003306 - 1}{-2} = \frac{-0{,}9996694}{-2} = 0{,}4998347$$

una X_{TRAS} = 0,4998347 (è utile riportare vari decimali)
e si eleva al quadrato questo risultato, ottenendo X^2_{TRAS} = 0,2498247

Effettuando questo calcolo per ognuno dei 6 valori si ottiene la serie seguente

X	X_{TRAS}	X^2_{TRAS}
55	0,4998347	0,2498247
23	0,4990548	0,2490557
276	0,4999934	0,2499934
73	0,4999062	0,2499062
41	0,4997026	0,2497027
97	0,4999469	0,2499469
Totale	2,9984386	1,4984396

Di essa si calcolano i totali $\sum X_{TRAS} = 2{,}9984386$ e $\sum X^2_{TRAS} = 1{,}4984396$

Successivamente, utilizzando la formula abbreviata per la varianza

$$s^2_{TRAS} = \frac{\sum X^2_{TRAS} - \frac{\left(\sum X_{TRAS}\right)^2}{n}}{n-1}$$

con i dati dell'esempio si ricava

$$s^2_{TRAS} = \frac{1{,}4984396 - \frac{2{,}9984386^2}{6}}{5} = \frac{1{,}4984396 - 1{,}498439006}{5} = 1{,}23^{-7}$$

$s^2_{TRAS} = 1{,}23^{-7}$

Infine, dopo aver calcolato anche $\sum \ln X$

X	55	23	276	73	41	97	Totale
$\ln X$	4,007333	3,135494	5,620401	4,290459	3,713572	4,574711	25,34197

che risulta uguale a 25,34197

si ricava L

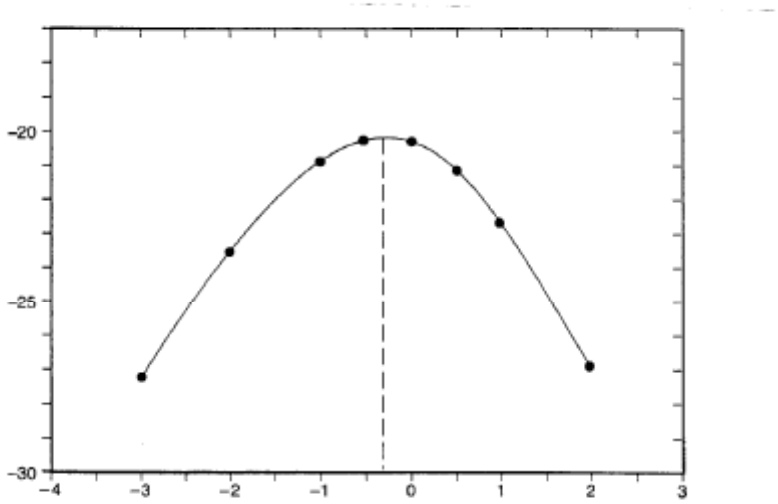
$$L = -\frac{5}{2} \cdot (\ln 1{,}23^{-7} + (-2-1) \cdot \frac{5}{6} \cdot 25{,}34197 = -23{,}6$$

che risulta L = -23,6.

Questa procedura illustrata per $\lambda = -2$ deve essere ripetuta per tutti i valori λ desiderati.
Per gli 8 valori di λ indicati,
si ottiene la seguente serie di valori L

λ	-3	-2	-1	-0,5	0	+0,5	+1	+2
L	-27,2	-23,6	-20,9	-20,2	-20,3	-21,1	-22,7	-26,9

La rappresentazione grafica può aiutare a comprendere come il valore massimo di L (con il segno) è collocato tra $\lambda = -0,5$ e $\lambda = 0$.



Rappresentazione grafica della funzione Log-likelihood (valori L riportati in ordinata)
per valori di λ (riportati in ascissa) uguali a –3, -2, -1, -0,5, 0, +0,5, +1, +2.

Poiché

- **$\lambda = -1/2$** indica una trasformazione reciproca, con **X** sotto radice quadrata, cioè $\frac{1}{\sqrt[2]{X}}$
- **$\lambda = -1/3$** indica una trasformazione reciproca, con **X** sotto radice cubica, cioè $\frac{1}{\sqrt[3]{X}}$
- **$\lambda = 0$** indica una trasformazione logaritmica, cioè **ln X** oppure **log X**

la scelta della tarsformazione da eseguire è limitata a questi tre.

Il tipo di misura effettuata e le caratteristiche di distribuzione del fenomeno studiato possono essere di aiuto, nella scelta definitiva e più corretta tra queste tre trasformazioni.

In modo acritico, è possibile utilizzare il valore esatto stimato dalla rappresentazione grafica, che corrisponde anche al valore che è

fornito dall'analisi con il computer:

$$\lambda = -0{,}29 \left(-\frac{1}{3{,}45}\right)$$

cioè

$$X' = \frac{1}{\sqrt[3{,}45]{X}}$$

Ma questa trasformazione è stimata sui dati campionari; un'altra rilevazione avrebbe senza dubbio indicato una trasformazionediversa da

$$\sqrt[3{,}45]{\ }$$

In conclusione, al posto della serie dei valori di X misurati, è conveniente scegliere una delle 4 trasformazioni indicate

	X	55	23	276	73	41	97
$\lambda = -1/2$	$X' = \frac{1}{\sqrt[2]{X}}$	0,135	0,209	0,060	0,117	0,156	0,102
$\lambda = -1/3$	$X' = \frac{1}{\sqrt[3]{X}}$	0,263	0,352	0,154	0,239	0,290	0,218
$\lambda = -0{,}29$	$X' = \frac{1}{\sqrt[3{,}45]{X}}$	0,313	0,403	0,196	0,288	0,341	0,266
$\lambda = 0$	X' = lnX	4,007	3,135	5,620	4,290	3,714	4,575

Dalla semplice lettura si evidenzia che tra valore minimo (23) e valore massimo (276) le distanze relative sono molto più ridotte. La trasformazione che le riduce maggiormente è quella logaritmica (ln). E' la trasformazione che avrebbe suggerito un esperto di ecologia, sapendo che si trattava della crescita esponenziale di una popolazione.

13.5. EFFETTI DELLE TRASFORMAZIONI SUI RISULTATI DELL'ANOVA

Nei test **t** ed **F**, la trasformazione dei dati per normalizzare la distribuzione ottiene l'effetto di rendere minima la varianza d'errore. E' quindi un criterio di scelta: **la trasformazione più adeguata è quella che rende minima la varianza d'errore e quindi rende i test più significativi** e con ciò più potenti.
La complessità dei problemi da risolvere per scegliere la trasformazione più adeguata e il dibattito che sempre si pone sulla reale validità dell'analisi attuata possono essere meglio illustrati con la discussione ampia di un esempio, tratto dal testo già citato di George E. P. **Box**, William G. **Hunter** e J. Stuart **Hunter** "***Statistics for Experimenters. An introdution to Design, Data Analysis and Model Building***", pp. 228-240).

ESEMPIO. Per verificare gli effetti di 3 sostanze tossiche (A, B, C) sulla sopravvivenza di cavie di età diversa (I, II, III, IV), ad ognuno dei 12 gruppi (3 trattamenti x 4 blocchi) sono stati assegnati 4 individui. Per ognuno di essi è stato misurato il tempo di sopravvivenza, tradotto in una grandezza unitaria equivalente a 10 ore.

I risultati sono riportati nella tabella sottostante

Età	Sostanze Tossiche TRATTAMENTI											
BLOCCHI	A				B				C			
I	0,31	0,45	0,46	0,43	0,36	0,29	0,40	0,23	0,22	0,21	0,18	0,23
II	0,8	1,10	0,88	0,72	0,92	0,61	0,49	1,24	0,30	0,37	0,38	0,29
III	0,43	0,45	0,63	0,76	0,44	0,35	0,31	0,40	0,23	0,25	0,24	0,22
IV	0,45	0,71	0,66	0,62	0,56	1,02	0,71	0,38	0,30	0,36	0,31	0,33

(In essa, il valore 0,31 della prima cavia appartenente alla classe d'età I e al tossico A indica che essa è sopravvissuta 3,1 giorni).
E' un disegno fattoriale a due fattori con repliche (3 trattamenti x 4 blocchi con 4 repliche per ogni esperimento; quindi 48 dati), che permette di verificare l'eventuale significatività sia di ognuno dei due fattori, sia della loro interazione.

L'analisi della varianza (ovviamente ottenuta con un programma informatico) fornisce i seguenti risultati

Fonte di variazione	Devianza	Gdl	Varianza	F	P
Totale	**3,005**	**47**	---	---	---
Tra gruppi	**2,204**	**11**	**0,200**	**9,010**	**.000**
Tra tossici	1,033	2	0,517	23,222	.000
Tra età	0,921	3	0,307	13,806	.000
Interazione	0,250	6	0,0417	1,874	.112
Entro gruppi (errore)	**0,801**	**36**	0,0222	---	---

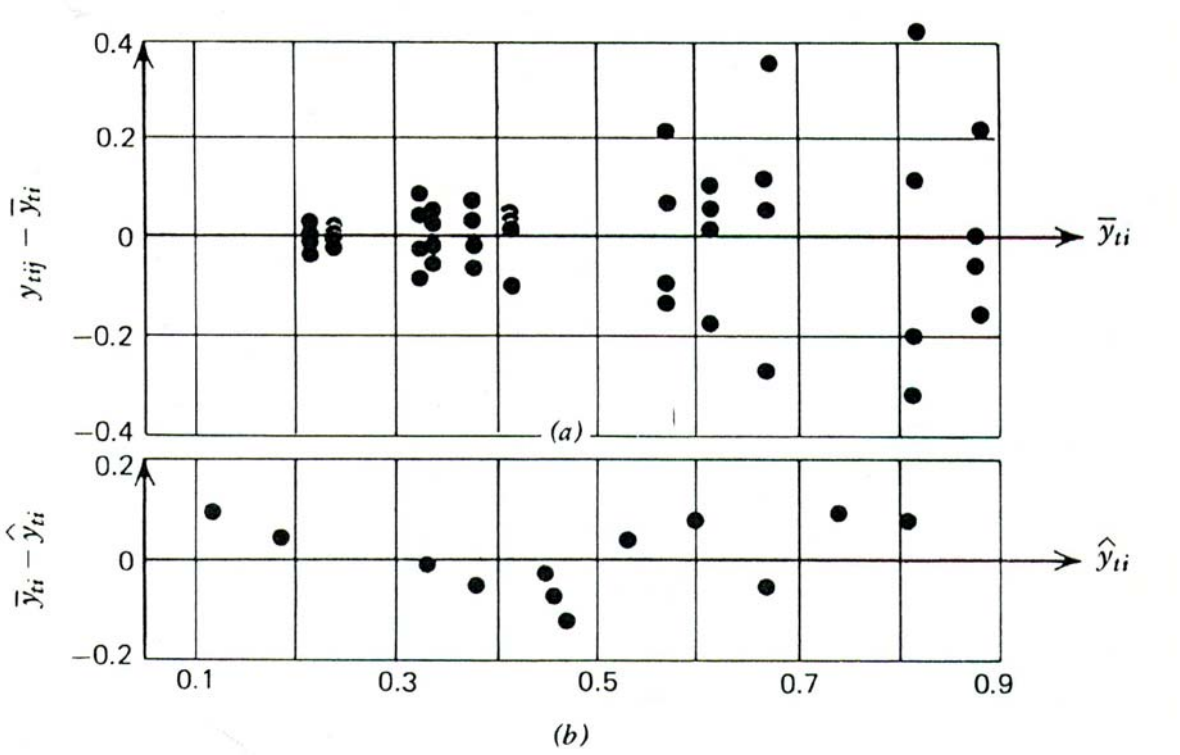
Essi permettono di rifiutare l'ipotesi nulla, relativamente al confronto tra tossici e tra età; per l'interazione si può sostenere una significatività tendenziale (P = .112), che potrebbe forse essere dimostrata con un aumento delle dimensioni del campione.

Ma l'analisi effettuata è valida?

Per valutare se sono state rispettate le condizioni di validità, è utile analizzare i residui. Il modo più semplice è quello della loro rappresentazione grafica, che può riguardare sia la variabilità entro gruppi che l'interazione. A questo scopo è utile **costruire due grafici**, che i programmi informatici più sofisticati permettono di stampare con facilità:

- il primo (vedi grafico **a**) può essere ottenuto riportando sull'asse delle ascisse (trasferito al centro) la media di ogni gruppo (in questo caso quello di casella $\overline{X}_{ij}$) e sull'asse delle ordinate gli scarti di ognuna delle **n** (4) repliche da essa ($X_{ijk} - \overline{X}_{ij}$);
- il secondo (vedi grafico b) è costruito riportando sull'asse delle ascisse le medie attese $\hat{X}_{ij}$ in ogni gruppo (con $\hat{X}_{ij} = \overline{X}_i + \overline{X}_j - \overline{\overline{X}}$) e sull'asse delle ordinate le differenze tra le medie osservate e queste medie attese ($\overline{X}_{ij} - \hat{X}_{ij}$).

(Nelle figure successive, tratte dal testo citato, la variabile è indicata con Y)



Dall'analisi dei due grafici appare con evidenza che

1 - gli scarti di ognuna delle 4 repliche dalla media del loro gruppo aumentano al crescere del valore della media;

2 - gli scarti tra le medie osservate e quelle attese tendono ad una relazione di tipo curvilineo, all'aumentare del valore delle medie.

Per conclusioni condivise sul primo punto, occorrerebbe effettuare i confronti tra varianze. Ma le analisi inferenziali sulla omogeneità delle varianze (test di Hartley, Cochran, Bartlett, Levene) sono molto tolleranti: non rifiutare l'ipotesi nulla non significa che essa sia vera, in particolare quando i dati sono pochi.

Di conseguenza, è lecito il sospetto che l'analisi della varianza applicata in precedenza non sia valida, in quanto potrebbe non essere rispettata la condizione di omoschedasticità. Si impone quindi una trasformazione dei dati.

Ma quale è la trasformazione più adeguata? Il fatto che la varianza entro casella o errore cresca all'aumentare della media suggerisce di utilizzare una trasformazione per $\mathbf{g_1}$ positivo (forte asimmetria destra); ma esse sono tante, da quella in radice quadrata a quella logaritmica, oppure il reciproco.

Per meglio comprendere gli effetti delle trasformazioni, un primo tentativo può essere effettuato con la radice quadrata. I valori diventano quelli riportati nella tabella successiva

Trasformazione in radice quadrata ($\sqrt[2]{X}$ arrotondata alla seconda cifra decimale)

Età	Sostanze Tossiche TRATTAMENTI		
BLOCCHI	A	B	C
I	0,56 0,67 0,68 0,66	0,60 0,54 0,63 0,48	0,47 0,46 0,42 0,48
II	0,91 1,05 0,94 0,85	0,96 0,78 0,70 1,11	0,55 0,61 0,62 0,54
III	0,66 0,67 0,79 0,87	0,66 0,59 0,56 0,63	0,48 0,50 0,49 0,47
IV	0,67 0,84 0,81 0,79	0,75 1,01 0,84 0,62	0,55 0,60 0,56 0,57

e l'analisi della varianza applicata ad essi fornisce i seguenti risultati

Fonte di variazione	Devianza	Gdl	Varianza	F	P
Totale	**1,365**	**47**	---	---	---
Tra gruppi	**1,071**	**11**	**0,0974**	**11,938**	**.000**
Tra tossici	0,561	2	0,280	34,389	.000
Tra età	0,431	3	0,144	17,601	.000
Interazione	0,079	6	0,013	1,62	.169
Entro gruppi (errore)	**0,294**	**36**	**0,00815**	---	---

Da essi emerge che:

1 - il test F tra gruppi (F = 11,9 con 11 gdl), quello tra tossici che interessa maggiormente (F = 34,4 con 2 gdl) e quello tra età (F = 17,6 con 3 gdl) sono tutti più significativi di quanto risultassero in precedenza, con i dati originari;

2 - il test F per l'interazione (F = 1,62 con 6 gdl) è meno significativo di quanto suggerito dall'analisi p- recedente.

I risultati sono migliori; ma questa è la trasformazione più adeguata oppure ne esistono altre preferibili?

E' semplice dimostrare che, con la trasformazione reciproca, i dati diventano

Trasformazione in reciproco ($\frac{1}{X}$ arrotondata alla seconda cifra decimale)

Età	Sostanze Tossiche TRATTAMENTI		
BLOCCHI	A	B	C
I	3,23 2,22 2,17 2,33	2,78 3,45 2,50 4,35	4,55 4,76 5,56 4,35
II	1,22 0,91 1,14 1,39	1,09 1,64 2,04 0,81	3,33 2,70 2,63 3,45
III	2,33 2,22 1,59 1,32	2,27 2,86 3,23 2,50	4,35 4,00 4,17 4,55
IV	2,22 1,41 1,52 1,61	1,79 0,98 1,41 2,63	3,33 2,78 3,23 3,03

e l'analisi della varianza fornisce risultati

Fonte di variazione	Devianza	Gdl	Varianza	F	P
Totale	**65,505**	**47**	---	---	---
Tra gruppi	**56,862**	**11**	**5,169**	**21,531**	**.000**
Tra tossici	34,877	2	17,439	72,635	.000
Tra età	20,414	3	6,805	28,343	.000
Interazione	1,571	6	0,262	1,090	.387
Entro gruppi (errore)	**8,643**	**36 (35)**	**0,240**	---	

ancor più significativi per i due fattori, ma che escludono la significatività, anche solo tendenziale, della loro interazione:

- il test F tra gruppi fornisce un valore pari a 31,531 (contro 11,938 precedente e 9,010 del primo caso);

- il test F tra tossici fornisce un valore pari a 72,635 (contro 34,389 precedente e 23,222 del primo caso);
- il test F tra età fornisce un valore pari a 28,343 (contro 17,601 precedente e 13,806 del primo caso);
- il test F dell'interazione fornisce un valore pari a 1,090 (contro 1,623 precedente e 1,874 del primo caso).

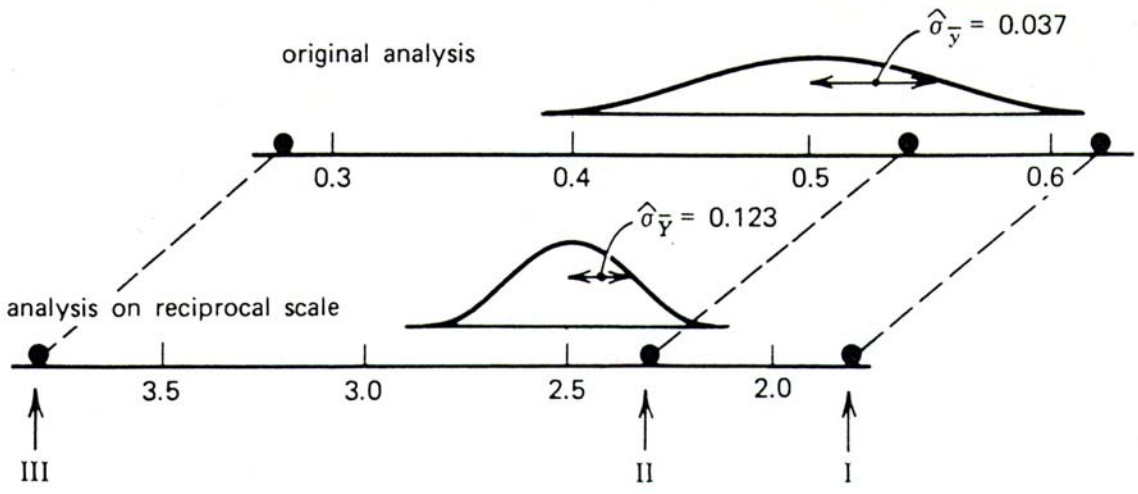
La figura successiva, che riporta
- nella parte superiore, **i dati originari** con la media dei tre tossici e
la loro **deviazione standard**

$$\sigma_{Xoriginari} = \sqrt{\frac{0{,}0222}{16}} = 0{,}037$$

- nella parte inferiore, **i dati trasformati in reciproco** con la media dei tre tossici e
la loro **deviazione standard**

$$\sigma_{Xtrasformati} = \sqrt{\frac{0{,}240}{16}} = 0{,}123$$

rapportati alla stessa scala

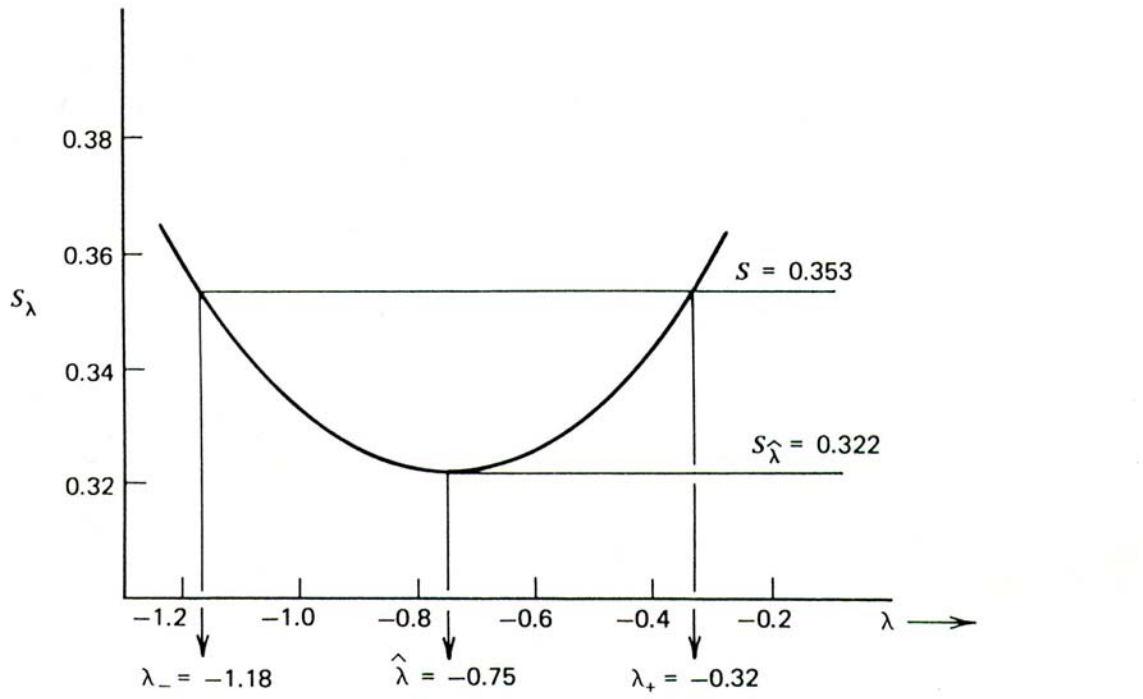


dimostra visivamente gli effetti della trasformazione sulla riduzione della varianza d'errore; nel caso specifico, sulla deviazione standard delle tre medie a confronto.

Per scegliere la trasformazione più adeguata all'esempio riportato, con una procedura sistematica che consideri tutte le possibilità migliori, **Box *e al.*** nel loro testo considerano solamente gli effetti di riga e di colonna, quindi una serie di valori $\mathbf{S_\lambda}$, derivati dalla somma dei quadrati dei residui con 42 gdl.
Nella tabella sottostante, per ogni valore di $\boldsymbol{\lambda}$ è riportato il corrispondente valore di $\mathbf{S_\lambda}$ da essi stimato:

$\boldsymbol{\lambda}$	-2,5	-2,0	-1,6	-1,4	-1,2	-1,0	-0,8	-0,6	-0,4	-0,2	0,0	0,5	1,0
$\mathbf{S_\lambda}$,	1,333	0,664	0,463	0,401	0,359	0,333	0,323	0,326	0,343	0,375	0,424	0,635	1,051

Dai valori di λ (in ascissa) e di S_λ in ordinata è stato ricavato il grafico



Da esso emerge che, con i criteri precedentemente definiti, la trasformazione più adeguata è

- $\boldsymbol{\lambda}$ **= -0,75** corrispondente al valore minimo di $\mathbf{S_\lambda}$ **= 0,322**

Questa risposta solleva 2 problemi:

- il valore di λ = **-0,75** è una risposta campionaria e non è accettabile impostare la trasformazione solo su un risultato sperimentale, poiché sarebbe differente nei vari casi affrontati;
- una trasformazione con elevamento alla potenza **–0,75** è insolita e priva di significato specifico, mentre l'esperienza ha dimostrato che in questi casi (tempi di risposta ad uno stimolo) quella adeguata è la trasformazione reciproca.

La stima dell'intervallo fiduciale permette di giungere ad una risposta generale.

Per $\alpha = 0.05$ e con $\nu = 42$, poiché $t_{42,\,0.025} = 2{,}021$

si ottiene un valore di S

$$S = 0{,}322 \cdot \left(1 + \frac{2{,}021^2}{42}\right) = 0{,}353$$

pari a 0,353.

Di conseguenza, è accettabile un valore S_λ fino al limite di 0,353. Sulla figura precedente, simmetrica rispetto al valore centrale, corrispondono valori di λ che sono compresi tra **–1,18** e **–0,32**.

Poiché λ = **-1** è compreso in questo intervallo fiduciale, la trasformazione reciproca è adatta ai dati sperimentali raccolti, in pieno accordo con la teoria sulle misure di tempo.

In questo esempio, che descrive una realtà complessa ma frequente nella ricerca ambientale, con la trasformazione che normalizza la distribuzione dei dati si possono risolvere contemporaneamente i problemi derivanti da più cause:

- la **non additività** dei due fattori considerati (per la presenza di una interazione tendenzialmente significativa),
- la **non omoschedasticità** dei gruppi a confronto,
- la **non normalità** della distribuzione dei dati.

A questi è da aggiungere il caso in cui i dati presentino una variabilità elevata,

cioè quando il rapporto

$$X_{massimo} / X_{minimo}$$

è grande,

- indicativamente **maggiore di tre** (come nel caso dell'esempio 1).

13.6. TEST PER LA VERIFICA DI NORMALITA', SIMMETRIA E CURTOSI, CON I METODI PROPOSTI DA SNEDECOR-COCHRAN

Prima e dopo la trasformazione dei dati, occorre misurare e verificare le caratteristiche fondamentali della loro distribuzione, per verificare se esiste **normalità, simmetria , curtosi**.

Il confronto di queste due serie di indici, quelli prima della trasformazione e quelli dopo, permette di valutarne l'effetto.

Inoltre, la scelta del test, soprattutto se parametrico o non parametrico, dipende in larga misura da queste risposte.

I metodi proposti in letteratura sono numerosi. Disponendo di una distribuzione di frequenza, è possibile ricorrere a tre test differenti per verificare:

- la **normalità** (***normality***),
- la **simmetria** (***skewness***),
- la **curtosi** (***kurtosis***).

Tra essi, per **campioni sufficientemente grandi,** possono essere ricordati quelli proposti da George W. **Snedecor** e William G. **Cochran** nel loro testo (***Statistical Methods***, 6th ed. The Iowa State University Press, Ames Iowa, U.S.A.). Nelle varie ristampe, dalla metà del Novecento e per oltre un trentennio è stato uno dei testi di riferimento più importanti per gli studiosi di statistica applicata.

I metodi da essi proposti e qui ripresi sono parte della impostazione classica, che è bene conoscere anche quando gli attuali programmi informatici ricorrono a procedure differenti, ritenute più potenti o più precise.

Per **valutare la normalità di una distribuzione** di dati sperimentali Snedecor e Cochran propongono di ricorre al test χ^2, chiamato appunto **test per la bontà dell'adattamento** (***goodness of fit test***), confrontando

- la distribuzione osservata
- con quella attesa, costruita mediante la media e la varianza del campione applicate alla normale.

L'ipotesi nulla è che non esistano differenze significative tra la distribuzione dei dati raccolti e quella normale corrispondente, con **stessa media** e **stessa varianza**.

L'ipotesi alternativa è che la distribuzione osservata se ne discosti in modo significativo, per un effetto combinato di asimmetria e curtosi.

Per presentare in modo dettagliato la procedura di verifica, si supponga di avere raccolto 500 misure di una scala continua e che la loro distribuzione di frequenza sia quella riportata nelle prime due colonne della tabella seguente.

Classe	Freq. Osservate	Freq. Attese	χ^2
< 130	9	20,30	6,29
130 – 139	35	30,80	0,57
140 – 149	68	55,70	2,72
150 – 159	94	80,65	2,21
160 – 169	90	93,55	0,13
170 – 179	76	87,00	1,39
180 – 189	62	64,80	0,12
190 – 199	28	38,70	2,96
200 – 209	27	15,85	3,85
210 – 219	4	7,10	1,35
220 – 229	5	2,20	6,04
230 – 239	1	0,50	
240 +	1	0,15	
Totale	**500**	**500,00**	**27,63**

Partendo dai dati campionari, è necessario:

- stimare le **frequenze attese** (riportate nella terza colonna),
- calcolare **il valore del χ^2**,

con i seguenti passaggi logici.

1 – Si individuano i valori centrali ($\overline{X}_i$) di ogni classe (ad esempio, per la classe 130-139 è 135); per le due classi estreme, occorre ipotizzare che esse abbiano la stessa ampiezza e quindi i due valori centrali siano rispettivamente 125 e 245 (come già evidenziato nel primo capitolo, dedicato alla statistica descrittiva, è sempre conveniente non fare classi aperte, appunto per favorire l'esatta individuazione del valore centrale).

2 - Si calcola la **media generale** ($\overline{\overline{X}}$) della distribuzione osservata,

con

$$\overline{\overline{X}} = \frac{\sum_{i=1}^{k} \left(\overline{X}_i \cdot n_i \right)}{n}$$

dove

- k è il numero di classi (nell'esempio k = 13),
- n_i è il numero di osservazioni della classe i,
- n è il numero totale di osservazioni (nell'esempio n = 500).

3 – Si calcola la deviazione standard (s) della distribuzione osservata,
con

$$S = \sqrt{\frac{\sum_{i=1}^{k} \left(\overline{X}_i - \overline{\overline{X}}\right)^2 \cdot n_i}{n}}$$

4 – Si stima il valore di Z per gli estremi di ogni classe (X_i); per le ultime due classi deve essere calcolato solo per il valore più vicino alla media generale; il valore di Z è calcolato mediante
la relazione

$$Z = \frac{X_i - \overline{\overline{X}}}{s}$$

5 – Dal valore Z di ogni estremo di classe si ricava, attraverso la tavola della distribuzione normale, la frequenza relativa corrispondente (già illustrato negli esempi del primo capitolo); per differenza, si stima la frequenza relativa di ogni classe.

6 – Rapportando a n (uguale a 500 nell'esempio) queste frequenze relative, si ottiene la frequenza attesa di ogni classe, come riportato nella terza colonna.

7 – Per ogni classe si stima il χ^2, mediante
la formula classica

$$\chi^2 = \frac{\left(Freq.Oss. - Freq.Att.\right)^2}{Freq.Att.}$$

8 – Poiché è condizione di validità di questo test che ogni frequenza attesa non sia inferiore a 5, nel caso dell'esempio le ultime tre classi devono essere raggruppate in una sola, riducendo così il numero totale di classi da 13 a 11.

9 – La somma degli 11 valori χ^2 fornisce il valore del χ^2 totale (uguale a 27,63), che ha 8 gdl.

10 – Infatti, benché esso sia stato ottenuto dalla somma di 11 valori, la distribuzione attesa è stata calcolata sulla base di tre quantità ricavate da quella osservata: la **media**, la **deviazione standard** e il **numero totale di osservazioni**. Di conseguenza, i gdl di questo χ^2 sono 11-3 = 8.

11 – Il valore critico per 8 gdl alla probabilità $\alpha = 0.005$ è $\chi^2 = 21{,}96$.

Poiché il valore calcolato (27,63) è superiore a quello critico, si rifiuta l'ipotesi nulla alla probabilità specificata: la distribuzione osservata è significativamente differente da una distribuzione normale, che abbia la stessa media e la stessa varianza.

Già il semplice confronto tabellare tra la distribuzione osservata e quella attesa evidenziava alcune differenze: ma il test permette di valutare tale scostamento in modo oggettivo. Il **chi quadrato per la normalità è un test generalista**: somma gli effetti di tutti gli scostamenti dalla normalità e non è diretto ad evidenziare gli effetti di una causa specifica. Nei dati della tabella precedente, appare evidente che la distribuzione osservata è asimmetrica; ma occorre essere in grado di fornirne un indice numerico e valutarne la significatività con test specifici.

Il **test per la *skewness*** (termine introdotto da Karl Pearson nel 1985, con la funzione β) di una popolazione di dati è fondato sul valore medio della quantità

$$(X-\mu)^3$$

dove **X** è ogni singolo valore e **μ** è la media della popolazione.
La **misura fondamentale della skewness**, in una popolazione di **n** dati,
è indicata con $\mathbf{m_3}$

$$m_3 = \frac{\sum_{i=1}^{n}(X_i-\mu)^3}{n}$$

e è chiamata **momento terzo intorno alla media** (***third moment about the mean***) o **momento di terzo ordine**.
Il suo valore
- è uguale a **0** (zero) quando la distribuzione dei dati è perfettamente **simmetrica**,
- è **positivo** quando la distribuzione è caratterizzata da una **asimmetria destra** (i valori oltre la media sono più frequenti),
- **è negativo** quando la distribuzione ha una **asimmetria sinistra** (i valori oltre la media sono meno frequenti).

Ma il valore assoluto di questo indice è fortemente dipendente dalla scala utilizzata (una distribuzione di lunghezze misurata in millimetri ha un valore di asimmetria maggiore della stessa distribuzione misurata in metri).

Per rendere questa misura **adimensionale**, cioè **indipendente dal tipo di scala e uguale per tutte le distribuzioni che hanno la stessa forma,** occorre dividere il momento di terzo ordine (m_3) per σ^3.

Da questo concetto sono derivati i due indici più diffusi, tra loro collegati da una relazione matematica semplice:

- l'indice β_1 di **Pearson**

$$\beta_1 = \left(\frac{m_3}{\sigma^3}\right)^2$$

- l'indice γ_1 di **Fisher**

$$\gamma_1 = \frac{m_3}{\sigma^3}$$

Quando calcolati su una distribuzione sperimentale, essi sono indicati rispettivamente con $\mathbf{b_1}$ e $\mathbf{g_1}$
Di conseguenza, il valore della **skewness di una distribuzione sperimentale** è

$$\sqrt{b_1} = g_1 = \frac{m_3}{m_2\sqrt{m_2}}$$

dove m_3 è

$$m_3 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{n}$$

e m_2 è

$$m_2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}$$

Nel calcolo del momento di secondo ordine, cioè della varianza, anche il testo di Snedecor-Cochran indica n, al posto del consueto -1, corrispondente ai gradi di libertà: offre il vantaggio pratico di semplificare i calcoli e per campioni grandi determina una differenza minima.

Riprendendo la stessa distribuzione di frequenza gia utilizzata in precendenza per la verifica della normalità mediante il test χ^2,

Classe	Limite Inf.	f	**U**	$f \cdot U$	$\mathbf{U^2}$	$f \cdot U^2$	$\mathbf{U^3}$	$f \cdot U^3$	$\mathbf{U^4}$	$f \cdot U^4$
< 130	120-	9	-4	-36	16	144	-64	-576	256	2.304
130 – 139	130-	35	-3	-105	9	315	-27	-945	81	2.835
140 – 149	140-	68	-2	-136	4	272	-8	-544	16	1.088
150 – 159	150-	94	-1	-94	1	94	-1	-94	1	94
160 – 169	160-	90	0	0	0	0	0	0	0	0
170 – 179	170-	76	1	76	1	76	1	76	1	76
180 – 189	180-	62	2	124	4	248	8	496	16	992
190 – 199	190-	28	3	84	9	252	27	756	81	2.268
200 – 209	200-	27	4	108	16	432	64	1.728	256	6.912
210 – 219	210-	4	5	20	25	100	125	500	625	2.500
220 – 229	220-	5	6	30	36	180	216	1.080	1.296	6.480
230 – 239	230-	1	7	7	49	49	343	343	2.401	2.401
240 +	240-	1	8	8	64	64	512	512	4.096	4.096
Totale	**---**	**500**	**---**	**+86**	**---**	**2.226**	**---**	**+3.332**	**---**	**32.046**

i calcoli possono essere semplificati, rispetto alla formula presentata con i momenti, indicando le classi con valori prestabiliti.

Poiché l'**indice è adimensionale** e quindi **le classi possono avere valori convenzionali**, diversi da quelli effettivamente rilevati, è conveniente modificare la scala delle classi: si indica con **0** (zero) la classe centrale (più frequente), con interi negativi quelle inferiori e con interi positivi quelle superiori (vedi quarta colonna, indicata con **U,** dove le classi sono state fatte variare da –4 a 8, ma potevano ugualemente variare da –6 a 6 o qualsiasi altra serie di valori convenzionali).
La metodologia abbreviata, utile per i calcoli manuali, richiede che questo valore sia elevato al quadrato (vedi colonna $\mathbf{U^2}$) e al cubo (vedi colonna $\mathbf{U^3}$).
L'elevamento alla quarta ($\mathbf{U^4}$) è richiesto nel test successivo, utile per la verifica della significatività del grado di curtosi.

Dopo aver ottenuto le somme

$$n = 500 \qquad \sum f \cdot U = +86 \qquad \sum f \cdot U^2 = +2.226 \qquad \sum f \cdot U^3 = +3.332$$

si ricavano $\mathbf{h_1}$, $\mathbf{h_2}$ e $\mathbf{h_3}$

con

$$h_1 = \frac{\sum f \cdot U}{n} = \frac{+86}{500} = +0{,}172$$

$$h_2 = \frac{\sum f \cdot U^2}{n} = \frac{2.226}{500} = 4{,}452$$

$$h_3 = \frac{\sum f \cdot U^3}{n} = \frac{+3.332}{500} = +6{,}664$$

Infine da essi **m_2** con

$$m_2 = h_2 - h_1^2 = 4{,}452 - (+0{,}172)^2 = 4{,}452 - 0{,}029584 = 4{,}4224$$

e **m_3** con

$$m_3 = h_3 - 3h_1h_2 + 2h_1^3 = +6{,}664 - (3 \cdot +0{,}172 \cdot +4{,}452) + (2 \cdot +0{,}172^3)$$

$$m_3 = 6{,}664 - 2{,}2972 + 0{,}010177 = 4{,}376977$$

I momenti di secondo ordine (**m_2**) e di terzo ordine (**m_3**) intorno alla media, per i dati sperimentali raccolti, sono

m_2 = 4,4224 e **m_3 = 4,376977**.

Infine con

$$\sqrt{b_1} = g_1 = \frac{m_3}{m_2\sqrt{m_2}} = \frac{4{,}376977}{4{,}4224 \cdot \sqrt{4{,}4224}} = \frac{4{,}376977}{9{,}300087} = 0{,}4706$$

si ottiene l'**indice di skewness**

$$\sqrt{b_1} = g_1 = 0{,}4706\,.$$

In **campioni grandi** (in alcuni testi **n > 150**; in altri, più rigorosi, **n ≥ 500**) estratti casualmente da una popolazione normale,

- questi indici **sono distribuiti in modo approssimativamente normale**,
- con media μ

$$\mu = 0$$

- e deviazione standard σ

$$\sigma = \sqrt{\frac{6}{n}}$$

TAVOLA DEI VALORI CRITICI
DELL'INDICE DI SKEWNESS $\sqrt{b_1} = g_1$
(in valore assoluto)

n	σ	Test bilaterale $\alpha = 0.10$	Test bilaterale $\alpha = 0.02$
		Test unilaterale $\alpha = 0.05$	Test unilaterale $\alpha = 0.01$
25	0,4354	0,711	1,061
30	0,4052	0,662	0,986
35	0,3804	0,621	0,923
40	0,3596	0,587	0,870
45	0,3418	0,558	0,825
50	0,3264	0,534	0,787
60	0,3009	0,492	0,723
70	0,2806	0,459	0,673
80	0,2638	0,432	0,631
90	0,2498	0,409	0,596
100	0,2377	0,389	0,567
125	0,2139	0,350	0,508
150	0,1961	0,321	0,464
175	0,1820	0,298	0,430
200	0,1706	0,280	0,403
250	0,1531	0,251	0,360
300	0,1400	0,230	0,329
350	0,1298	0,213	0,305
400	0,1216	0,200	0,285
450	0,1147	0,188	0,269
500	0,1089	0,179	0,255

Nell'esempio (**n = 500**),

la deviazione standard (σ) dell'indice di **skewness** è

$$\sigma = \sqrt{\frac{6}{500}} = 0{,}1095$$

E' quindi possibile valutare,

- con un **test bilaterale** se l'asimmetria è diversa da 0

$$H_0: \gamma_1 = 0 \quad \text{contro} \quad H_1: \gamma_1 \neq 0$$

- oppure con un **test unilaterale** se esiste asimmetria destra

$$H_0: \gamma_1 \leq 0 \quad \text{contro} \quad H_1: \gamma_1 > 0$$

o se essa è sinistra

$$H_0: \gamma_1 \geq 0 \quad \text{contro} \quad H_1: \gamma_1 < 0$$

Nel caso di **<u>campioni grandi</u> (**in alcuni testi **n > 150**; in altri, più rigorosi, **n ≥ 500)**, utilizzando l'indice di **skewness** calcolato ($\sqrt{b_1}$ o g_1), si valuta la sua significatività

ricavando Z con la formula

$$Z = \frac{g_1}{\sqrt{\frac{6}{n}}}$$

derivata dalla formula generale

$$Z = \frac{g_1 - 0}{\sigma}$$

Con i dati dell'esempio

$$Z = \frac{g_1}{\sqrt{\frac{6}{n}}} = \frac{0{,}4706}{\sqrt{\frac{6}{500}}} = \frac{0{,}4706}{0{,}1095} = 4{,}29$$

risulta un valore di Z (4,29) molto alto, al quale nella tavola della distribuzione normale corrisponde una probabilità $\alpha < 0.0001$ sia per un test a una coda sia per un test a due code. In conclusione, si può affermare che nella distribuzione osservata è presente una asimmetria destra altamente significativa.

Nel caso di **campioni piccoli (n ≤ 150)**, occorre utilizzare una distribuzione specifica che fornisce una approssimazione più accurata. E' possibile ricorrere alla tabella dei valori critici riportata in precedenza, tratta dal testo di Snedecor e Cochran e valida sia per $\sqrt{b_1}$ sia per $\mathbf{g_1}$.

Alla probabilità α prefissata sono significativi gli indici $\sqrt{b_1}$ o $\mathbf{g_1}$ che, in valore assoluto, sono maggiori di quelli riportati nella tabella.

Il **test per la kurtosis** (raramente chiamata anche ***peakedness*** o ***tailed-ness***) di una popolazione di dati è fondato sul valore medio della quantità

$$(X-\mu)^4$$

diviso per σ^4.

In una popolazione distribuita in modo normale risulta uguale a 3.

Gli indici di curtosi $\mathbf{b_2}$ di Pearson e $\mathbf{g_2}$ di Fisher sono ricavati da

$$b_2 - 3 = g_2 = \frac{m_4}{m_2^2} - 3$$

dove

$$m_4 = \frac{\sum (X-\overline{X})^4}{n}$$

Per stimare $\mathbf{b_2}$ da una distribuzione di frequenza (utilizzando la stessa impiegata per l'asimmetria) dopo aver calcolato oltre ai parametri precedenti

anche

$$\sum f \cdot U^4 = 32.046$$

e

$$h_4 = \frac{\sum f \cdot U^4}{n} = \frac{32.046}{500} = 64{,}092$$

con

$$m_4 = h_4 - 4h_1h_3 + 6h_1^2h_2 - 3h_1^4$$

si ricava

$$m_4 = 64{,}092 - 4 \cdot (+0{,}172) \cdot (+6{,}664) + 6 \cdot (0{,}172)^2 \cdot 4{,}452 - 3 \cdot (+0{,}172)^4$$

$$m_4 = 64{,}092 - 4{,}5848 + 0{,}7902 - 0{,}0026 = 60{,}2948$$

$m_4 = 60{,}2948$.

Infine si ottengono b_2

$$b_2 = \frac{m_4}{m_2^2} = \frac{60{,}2948}{4{,}4224^2} = \frac{60{,}2948}{19{,}5576} = 3{,}0829$$

e g_2

$$g_2 = b_2 - 3 = 3{,}0829 - 3 = 0{,}0829$$

In **campioni grandi** (in alcuni testi **n > 1.000**; in altri, più rigorosi, **n ≥ 2.000**), il valore $\mathbf{g_2}$ è distribuito in modo **approssimativamente normale**

con

$$\mu = 0 \quad \text{e} \quad \sigma = \sqrt{\frac{24}{n}}$$

Nell'esempio (**n = 500**),

la deviazione standard (σ) dell'indice di **kurtosis** è (molto) approssimativamente

$$\sigma = \sqrt{\frac{24}{500}} = 0{,}2191$$

E' quindi possibile valutare,

- con un test bilaterale se l'indice di curtosi $\mathbf{g_2}$ è diversa da 0

$$H_0: \gamma_2 = 0 \quad \text{contro} \quad H_1: \gamma_2 \neq 0$$

- oppure con un test unilaterale se la curva è platicurtica

$$H_0: \gamma_2 \leq 0 \quad \text{contro} \quad H_1: \gamma_2 > 0$$

o se essa è leptocurtica

$$H_0: \gamma_2 \geq 0 \quad \text{contro} \quad H_1: \gamma_2 < 0$$

Nel caso di **campioni grandi**, utilizzando l'indice di curtosi calcolato (b_2–3 oppure g_2), si valuta la sua significatività

ricavando Z con la formula

$$Z = \frac{g_2}{\sqrt{\frac{24}{n}}}$$

derivata dalla formula generale

$$Z = \frac{g_2 - 0}{\sigma}$$

VALORI CRITICI SUPERIORI E INFERIORI DI b_2

PER IL TEST DI KURTOSIS ALLE PROBABILITA' $\alpha = 0.05$ E $\alpha = 0.01$

$$b_2 = \frac{m_4}{m_2^2}$$

n	$\alpha = 0.05$		$\alpha = 0.01$	
	Superiore	Inferiore	Superiore	Inferiore
50	3,99	2,15	4,88	1,95
75	3,87	2,27	4,59	2,08
100	3,77	2,35	4,39	2,18
125	3,71	2,40	4,24	2,24
150	3,65	2,45	4,13	2,29
200	3,57	2,51	3,98	2,37
250	3,52	2,55	3,87	2,42
300	3,47	2,59	3,79	2,46
350	3,44	2,62	3,72	2,50
400	3,41	2,64	3,67	2,52
450	3,39	2,66	3,63	2,55
500	3,37	2,67	3,60	2,57
550	3,35	2,69	3,57	2,58
600	3,34	2,70	3,54	2,60
650	3,33	2,71	3,52	2,61
700	3,31	2,72	3,50	2,62
750	3,30	2,73	3,48	2,64
800	3,29	2,74	3,46	2,65
850	3,28	2,74	3,45	2,66
900	3,28	2,75	3,43	2,66
950	3,27	2,76	3,42	2,67
1.000	3,26	2,76	3,41	2,68
1.200	3,24	2,78	3,37	2,71
1.400	3,22	2,80	3,34	2,72
1.600	3,21	2,81	3,32	2,74
1.800	3,20	2,82	3,30	2,76
2.000	3,18	2,83	3,28	2,77

Con i dati dell'esempio, anche se in realtà il campione è troppo piccolo per questo test,

$$Z = \frac{g_2}{\sqrt{\frac{24}{n}}} = \frac{0{,}0829}{\sqrt{\frac{24}{500}}} = \frac{0{,}0829}{0{,}2191} = 0{,}38$$

risulta un valore di Z (0,38) piccolo.

Ad esso corrisponde una probabilità

- $\alpha = 0{,}3620$ in un test bilaterale,
- $\alpha = 0{,}1810$ in un test unilaterale.

Sono probabilità comunque molto alte, che non solo non rifiutano l'ipotesi nulla ma permettono di affermare cha la distribuzione è molto simile alla normale, per quanto riguarda la curtosi.

Per **campioni piccoli**, (**n < 2.000**) il testo di **Snedecor-Cochran** riporta i valori critici di $\mathbf{b_2}$ alla probabilità $\alpha = 0.05$ e $\alpha = 0.01$ stimati da Pearson.

Dalla loro lettura, è semplice osservare che **non sono distribuiti in modo simmetrico** intorno a 3.

Per **n = 500**, alla probabilità $\alpha = 0.05$ il limite superiore è +0,37 rispetto al valore centrale, mentre il limite inferiore è –0,33. I due scarti diventano molto simili, alla seconda cifra decimale, solo quando n = 2.000.

Alla probabilità $\boldsymbol{\alpha}$ prescelta, sono significativi tutti i valori di $\mathbf{b_2}$ esterni a questo intervallo.

Per la significatività di g_2 è sufficiente sottrarre 3 ai valori della tabella.

Durante i primi decenni del Novecento, sono state utilizzate le quantità g_1 e g_2 per stimare la distribuzione di dati in popolazioni non normali, caratterizzate dai parametri γ_1 e γ_2. E' stato dimostrato, come afferma il Teorema del Limite Centrale, che in una distribuzione di medie campionarie ($\overline{X}$) le misure di skewness e kurtosis tendono entrambe a zero con l'aumento delle dimensioni (n) del campione

$$\gamma_1(\overline{X}) = \frac{\gamma_1}{\sqrt{n}} : \gamma_2(\overline{X}) = \frac{\gamma_2}{\sqrt{n}}$$

Un altro aspetto interessante è che la curtosi(γ_2) aumenta la varianza di un campione(s^2) rispetto al valore reale della popolazione (σ^2),

attraverso la relazione

$$s^2 = \frac{2\sigma^4}{\nu}\left(1 + \frac{\nu}{\nu+1} \cdot \frac{\gamma_2}{2}\right)$$

dove ν sono i gdl del campione.

Se in una popolazione con varianza σ^2 l'indice di curtosi è $\gamma_2 = 1$, la varianza del campione (s^2) è 1,5 volte più ampia di quella risultante in una distribuzione normale (quindi con $\gamma_2 = 0$). Se la curtosi è $\gamma_2 = 2$ la varianza del campione è circa il doppio di quella corrispondente in una popolazione normale.

13.7. METODI GRAFICI E ALTRI TEST (LILLIEFORS, D'AGOSTINO-PEARSON) PER NORMALITA', SIMMETRIA E CURTOSI (CENNI DEI TEST DI GEARY E DI SHAPIRO-WILK)

Per analizzare la normalità di una distribuzione, oltre al test χ^2 i test proposti sono numerosi. Alcuni, ovviamente bilaterali, considerano gli effetti congiunti di asimmetria e curtosi; sono chiamati anche **test omnibus** (***Omnibus test for departure from normality***). Attualmente i più utilizzati sono:

- il test di **Lilliefors**, derivato dal test di Kolmogorov-Smirnov chiamato anche ***distance test***, essendo fondato sulla distanza massima tra la distribuzione cumulata osservata e quella cumulata attesa,
- il test proposto da R. B. **D'Agostino** e E. S. **Pearson**.

Altri test, come già visto nel paragrafo precedente, possono prendere in considerazione solo la simmetria e la curtosi, permettendo anche l'ipotesi unilaterale. Tra questi test possono essere brevemente ricordati

- il test di R. C. **Geary** (vedi gli articoli, entrambi del 1947 e sullo stesso volume, ***Frequency distribution of*** $\sqrt{b_1}$, pubblicato su **Biometrika**, Vol. 34, pp.: 68-97 e ***Testing for normality***, su **Biometrika,** Vol. 34, pp.:209-242),
- il test di S. S. **Shapiro &** M. B. **Wilks** (vedi del 1965 l'articolo ***An analysis of variance test for normality (complete sample)***, pubblicato su **Biometrika,** Vol. 52, pp.: 591-611 e del 1968 l'articolo ***Approximations for the null distribution of the W statistic***, pubblicato su **Technometrics**, Vol. 10, pp.: 861-866).

Essi non utilizzano i momenti di 3° e 4° ordine, ma un indicatore (**U** per **Geary** e **W** per **Shapiro & Wilk**) fondato sul rapporto tra le due misure della variabilità. Per Geary

U = deviazione media / deviazione standard

Cioè

$$U = \frac{\frac{\sum |X_i - \bar{X}|}{n}}{\sqrt{m_2}} = \frac{\sum |X_i - \bar{X}|}{n \cdot s}$$

Calcolato su una **popolazione normale**, **U = 0,7979**.

Per lo studio delle **curtosi**,

- una **curtosi positiva** (curva platicurtica) produce **valori bassi**, inferiori a 0,7979
- una **curtosi negativa** (curva leptocurtica) produce **valori alti**, superiori a 0,7979.

Il confronto tra i valori di g_2 e di U, ovviamente calcolati sugli stessi dati, dimostrano un buon accordo. Il valore U offre due vantaggi

- è tabulato anche per campioni di piccole dimensioni,
- è più facile e rapido da calcolare.

Come quello di Shapiro & Wilk è comunque un test poco diffuso e fondato su una base teorica meno solida.

L'esempio riportato nel paragrafo percedente ha dimostrato che

- **il test con il χ^2 ha poca potenza**,

per verificare la bontà dell'**adattamento alla normale di una distribuzione osservata.**

Questo problema è stato risolto con la richiesta di numero molto alto di osservazioni; ma nella ricerca ambientale e biologica, raramente si raccolgono alcune centinaia di dati.

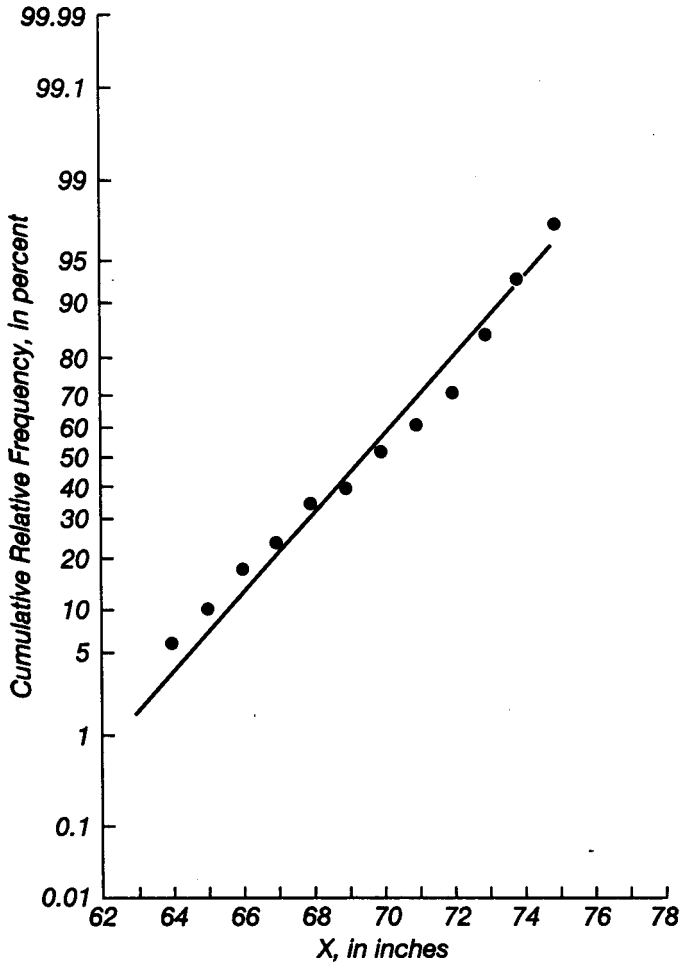
Il test di Kolmogorov-Smirnov, che può essere applicato alla verifica della normalità per un campione, offre il vantaggio di poter essere utilizzato anche con pochi dati. Inoltre, quando la scala è una variabile continua, gli intervalli di classe possono essere molto piccoli e tra loro differenti: ne deriva un'analisi più sensibile, in particolare quando sono importanti le frequenze verso gli estremi.

Per analizzare la normalità di una distribuzione, con la diffusione dei computer in questi anni sono stati rilanciati **i metodi grafici**. Tra essi, è diffuso quello che

- sull'asse delle ascisse riporta i valori della scala utilizzata,
- sull'asse delle ordinate riporta le frequenza relative cumulate di ogni classe, espresse in percentuale.

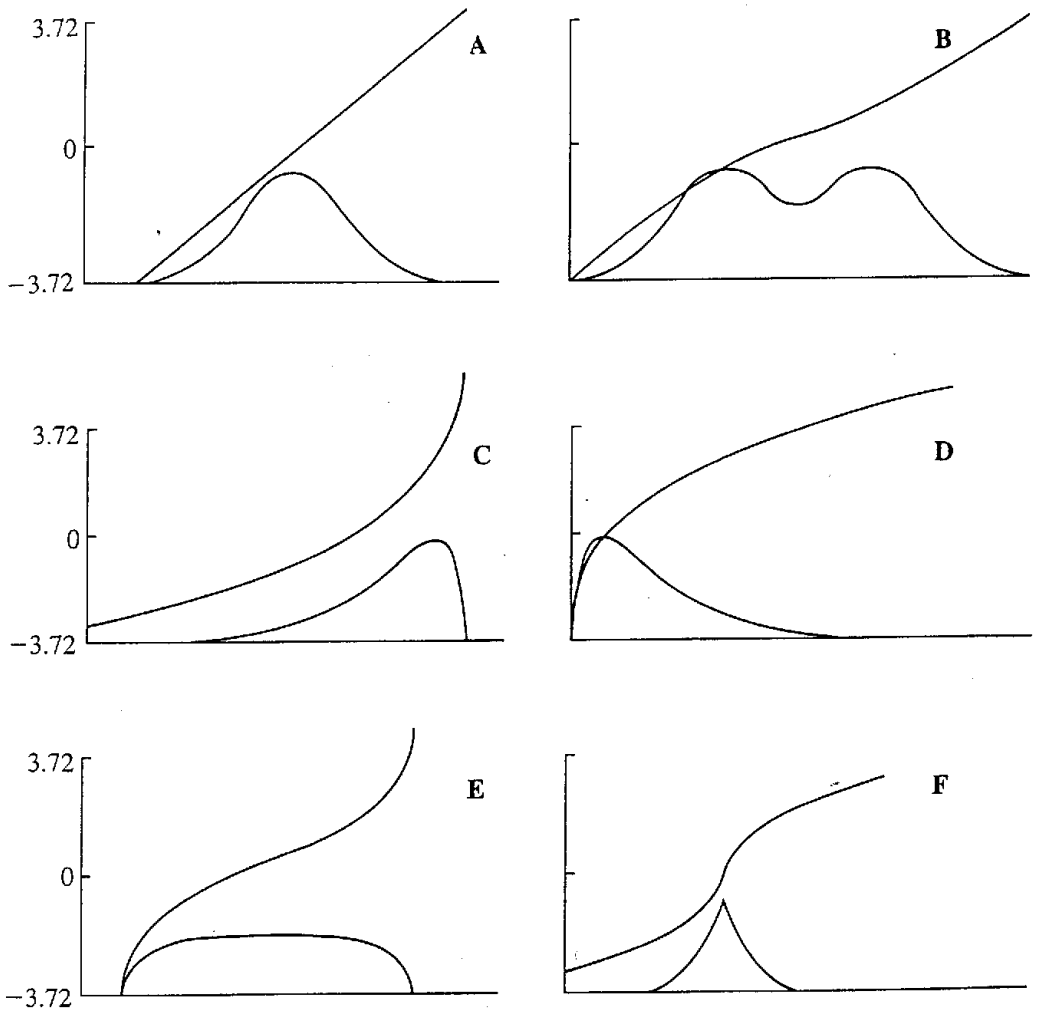
Per illustrare questa metodologia, viene riproposta la distribuzione dell'altezza di 70 studenti universitari, misurata in pollici, tratta dal testo di Jerrold **Zar** del 1999 (***Biostatistical Analysis***, 4th ed. Prentice Hall, Upper Saddle River, New Jersey):

Altezza (X)	Freq. Oss.	% su totale	Cumulata (Y)
63	2	2,9	2,9
64	2	2,9	5,8
65	3	4,2	10,0
66	5	7,1	17,1
67	4	5,7	22,8
68	6	8,6	31,4
69	5	7,1	38,5
70	8	11,5	50,0
71	7	10,0	60,0
72	7	10,0	70,0
73	10	14,3	84,3
74	6	8,6	92,9
75	3	4,2	97,1
76	2	2,9	100,0
Totale	70	100,0	----



I valori della X sono distribuiti in modo approssimativamente normale, i punti della percentuale cumulata (Y) sono distribuiti in modo lineare, utilizzando carte specifiche come quella riportata. La pendenza della retta è inifluente, dipendendo solo dalla scala dalla scala delle ascisse.

Altri grafici sono più complessi da costruire manualmente perché richiedono il calcolo di Z per tutti i valori campionari di X; ma sono altrettanto semplici da ottenere con il computer.
Essi sull'asse delle ordinate riportano il valore di Z, entro un campo di variazione estremamente ampio, che contiene oltre il 99% dei dati. Nelle figure sottostanti, sono riportati schemi grafici in cui il valore di Z varia da –3,72 a +3,72; corrripondono alla percentuali che in ogni coda della distribuzione escludono una frequenza pari a 0.0001 (o 0.01%)



Quando la distribuzione è perfettamente normale ($\mathbf{g_1}$ e $\mathbf{g_2}$ uguali a 0), la cumulata delle frequenze ha una forma esattamente lineare (Fig. A).

I vari tipi di scostamento dalla normalità forniscono curve di forma differente e tipica. Di conseguenza, dalla forma della cumulata è possibile dedurre la probabile forma della distribuzione di frequenza sottostante. Questo metodo risulta più semplice dell'uso della normale, in quanto lo scostamento da una retta risulta visivamente più evidente di quella da una normale, soprattutto quando i dati sono pochi.

Nelle altre cinque figure del grafico precedente, sono rappresentate rispettivamente:

B) una distribuzione bimodale,

C) una distribuzione con asimmetria sinistra ($\mathbf{g_1}$ negativo),

D) una distribuzione con asimmetria destra ($\mathbf{g_1}$ positivo),

E) una distribuzione platicurtica ($\mathbf{g_2}$ positivo),

F) una distribuzione leptocurtica ($\mathbf{g_2}$ negativo).

Il test proposto da H. **Lilliefors** nel 1967 (vedi l'articolo ***On the Kolmogorov-Smirnov test for normality with mean and variance unknown***, pubblicato su **Journal of the American Statistical Association** Vol. 62, pp.: 399-402) è particolarmente utile in **campioni di dimensioni minime**. I suoi valori critici (vedi tabella nella pagina successiva) iniziano da $n = 4$.

Come scrive Lilliefors, è una modificazione del test di Kolmogorov-Smirnov: ne segue la procedura, ma utilizza una tavola di valori critici differente. Come in tutti i test di normalità, l'ipotesi nulla è che la popolazione dalla quale è stato estratto il campione non sia troppo differente dalla famiglia di distribuzioni che seguono la legge di Gauss, quindi che sia $\mathbf{N(\mu, \sigma^2)}$ con μ e σ qualsiasi ma $\gamma_1 = 0$ e $\gamma_2 = 0$, contro l'ipotesi alternativa che sia diversa dalla normale a causa di asimmetria e/o curtosi

Dopo aver stimato la funzione di ripartizione della legge normale ridotta **N (0, 1),** si calcolano

- la cumulata delle frequenze attese, nell'ipotesi che la distribuzione sia normale,
- la cumulata delle frequenze osservate,
- lo scarto massimo tra le due distribuzioni.

La distribuzione dei valori critici è differente da quella di Kolmogorov-Smirnov, poiché la distribuzione normale è calcolata a partire dalla media e dalla varianza campionarie. Oltre al numero di dati, uguale sia nella distribuzione osservata che in quella attesa, sono introdotti due vincoli ulteriori di similarità tra le due distribuzioni a confronto.

Quantili della statistica di Lilliefors

per verificare la normalità di una distribuzione campionaria

N	α				
	0.20	**0.15**	**0.10**	**0.05**	**0.01**
4	0,300	0,319	0,352	0,381	0,417
5	0,285	0,299	0,315	0,337	0,405
6	0,265	0,277	0,294	0,319	0,364
7	0,247	0,258	0,276	0,300	0,348
8	0,233	0,244	0,261	0,285	0,331
9	0,223	0,233	0,249	0,271	0,311
10	0,215	0,224	0,239	0,258	0,294
11	0,206	0,217	0,230	0,249	0,284
12	0,199	0,212	0,223	0,242	0,275
13	0,190	0,202	0,214	0,234	0,268
14	0,183	0,194	0,207	0,227	0,261
15	0,177	0,187	0,201	0,220	0,257
16	0,173	0,182	0,195	0,213	0,250
17	0,169	0,177	0,189	0,206	0,245
18	0,166	0,173	0,184	0,200	0,239
19	0,163	0,169	0,179	0,195	0,235
20	0,160	0,166	0,174	0,190	0,231
25	0,142	0,147	0,158	0,173	0,200
30	0,131	0,136	0,144	0,161	0,187
>30	0,736/√n	0,768/√n	0,805/√n	0,886/√n	1,031/√n

Si ricorre quindi alla **tavola dei quantili di Lilliefors**.

Se lo scarto massimo calcolato è superiore a quello riportato nella tabella, si rifiuta l'ipotesi nulla: il campione non è stato estratto da una popolazione distribuita secondo la legge di Gauss, ma ha distrosioni dovure ad asimmetria e/o curtosi.

ESEMPIO. Prima di applicare un test inferenziale sulla media delle seguenti 10 misure, si vuole verificare se esse siano state estratte da una popolazione distribuita in modo normale.
Per facilitare la procedura, fondata come il test di Kolmogorov-Smirnov (di cui rappresenta una evoluzione) sulla cumulata della distribuzione di frequenza, i valori sono già ordinati per rango

Individui	A	B	C	D	E	F	G	H	I	L
Dimensioni X_i	10	11	12	12	13	15	15	16	17	19

Dopo aver calcolato la media ($\overline{X}$) del campione e la deviazione standard (s), ottenendo

- $\overline{X} = 14$
- $s = 2{,}87$

per ogni misura campionaria (X_i) si stimano

- i valori di **Z** corrispondenti (riportati nella seconda colonna della tabella successiva)

$$Z_i = \frac{X_i - \overline{X}}{s}$$

- la ripartizione delle probabilità della normale ridotta corrispondente

X_i	Z_i	$P_{att.}$	$P_{oss.}$	Di
10	-1,39	0,083	0,000	0,083
11	-1,05	0,147	0,100	0,047
12, 12	-0,70	0,242	0,200	0,042
13	-0,35	0,363	0,400	-0,037
15, 15	0,35	0,637	0,500	**0,137**
16	0,70	0,758	0,700	0,058
17	1,05	0,853	0,800	0,053
19	1,74	0,959	0,900	0,059
---	---	---	1,000	---

(vedi: $P_{att.}$. riportati nella terza colonna, che rappresenta la cumulata delle frequenze in una distribuzione normale, procedendo dai valori bassi verso quelli alti)

Successivamente, si calcolano

- la cumulata delle probabilità per i valori osservati X_i (vedi $P_{oss.}$ riportata nella 4 colonna: poiché i valori sono 10, ognuno di essi ha una probabilità pari a 1/10 = 0.1 e la loro cumulata è la somma delle frequenze fino a quel valore); nelle righe 3 e 5, nelle quali sono presenti due valori identici, la cumulata delle probabilità include un solo valore; se i dati per ogni classe fossero numerosi, si cumulerebbero le frequenze fino al valore medio della classe;
- e differenze $D_i = P_{att.} - P_{oss.}$ (quinta colonna).

Per esempio,

- la prima **D** (0,083) è data da 0,083 - 0,000;
- la quarta **D** (-0,037) da 0,363 – 0,400

La differenza massima tra le due distribuzioni è **D** = 0,137 (nella quinta riga).

Nella tabella dei valori critici di Lilliefors, per **n = 10**

- alla probabilità α **= 0.05** il valore riportato è **0,258**
- alla probabilità α **= 0.20** è uguale a **0,215**.

Il valore **D** calcolato è inferiore anche a questo ultimo. Non è possibile rifiutare l'ipotesi nulla. Inoltre, poiché la probabilità α è maggiore di 0.20, è possibile affermare che lo scostamento della distribuzione campionaria da quella normale; con stessa media e stessa varianza, è trascurabile.

Il test di Lilliefors utilizza la metodologia di Kolgorov-Smirnov. I vincoli, cioè i parametri stimati dal campione sulla base dei quali sono stati calcolati i valori attesi, sono tre

- il numero totale di osservazioni,
- la media,
- la deviazione standard.

Non potendo ridurre i gdl come nel χ^2, si ricorre a valori critici differenti.

Per n = 10 (il caso dell'esempio), il semplice confronto tra le due serie di valori critici alle stesse probabilità α mostra come il valore di Lilliefors sia minore di quello corrispondente di Kolmogorov-Smirnov.

Valori critici per $n = 10$	α				
	0.20	**0.15**	**0.10**	**0.05**	**0.01**
Kolmogorov-Smirnov	0,322	0,342	0,368	0,410	0,490
Lilliefors	0,215	0,224	0,239	0,258	0,294

Il **test** proposto da Ralph **D'Agostino** nel 1971 (vedi articolo ***An omnibus test of normality for moderate and large size sample***, pubblicato su **Biometrika**, vol. 58, pp.: 341-348), chiamato anche **test di D'Agostino–Pearson**, per l'articolo di Ralph **D'Agostino** e E. S. **Pearson** del 1973 (vedi ***Test for departure from normality. Empirical results for the distributions of*** b_2 ***and*** $\sqrt{b_1}$, pubblicato su **Biometrika**, vol. 60, pp. 613-622), appare uno dei test più potenti. (E. S. Perason non deve essere confuso con il più famoso **Karl Pearson,** che pubblicò nei primi decenni del Novecento)

Per l'illustrazione di questo metodo, è stato seguito l'esempio riportato nel volume di Jarrold **Zar** del 1999 ***Biostatistical Analysis*** (4^{th} ed. Prentice Hall, Upper Saddle River, Nee Jersey), uno dei testi classici più diffusi; ad esso si rimanda per approfondimenti.

L'**ipotesi nulla bilaterale sulla normalità di un campione** può essere verificata

mediante la statistica

$$K^2 = Z_{g1}^2 + Z_{g2}^2$$

dove

- Z_{g1} e Z_{g2} sono ricavati rispettivamente dall'indice di simmetria g_1 e di curtosi g_2

(poiché possono essere sia positivi che negativi, permettono di sommare i diversi tipi di asimmetria e curtosi solo se elevati al quadrato).

- K^2 è un χ^2 con 2 gradi di libertà, ricordando la relazione

$$\chi_{(n)}^2 = \sum_{i=1}^{n} Z_i^2$$

Il valore di K^2 deve quindi essere confrontato con la tabella

α					
.25	*.10*	*.05*	*.025*	*.01*	*.005*
2.773	**4.605**	**5.991**	**7.378**	**9.210**	**10.597**

che riporta i valori critici del χ_2 per df = 2 nella coda destra della distribuzione (vedi cap. 3).

La procedura di D'Agostino, a partire da una distribuzione di dati, permette di

- calcolare $\mathbf{g_1}$ e $\mathbf{g_2}$,
- ricavare da essi $\sqrt{\mathbf{b_1}}$ e $\mathbf{b_2}$
- valutare la **normalità**sia in complesso, sia indipendentemente gli indici di simmetria e curtosi.

Per illustrare la procedura proposta nel testo di **Zar** in tutti i suoi passaggi, è stata ripresa la distribuzione di frequenza già utilizzata per la rappresentazione grafica, nella quale non si evidenziava un particolare scostamento dalla normale.

Dai valori delle classi (X_i) e dalle loro frequenze osservate (f_i)

Altezza X_i	Freq. Oss. f_i	$f_i \cdot X_i$	$f_i \cdot X_i^2$	$f_i \cdot X_i^3$	$f_i \cdot X_i^4$
63	2	126	7.938	500.094	31.505.922
64	2	128	8.192	524.288	33.554.432
65	3	195	12.675	823.875	53.551.875
66	5	330	21.780	1.437.480	94.873.680
67	4	268	17.956	1.203.052	80.604.484
68	6	408	27.744	1.886.592	128.288.256
69	5	345	23.805	1.642.545	113.335.605
70	8	560	39.200	2.744.000	192.080.000
71	7	497	35.287	2.505.377	177.881.767
72	7	504	36.288	2.612.736	188.116.992
73	10	730	53.290	3.890.170	283.982.410
74	6	444	32.856	2.431.344	179.919.456
75	3	225	16.875	1.265.625	94.921.875
76	2	152	11.552	877.952	66.724.352
Totale	70	4.912	345.438	24.345.130	1.719.341.106

- si ricavano i totali di colonna

$$\sum f_i = n = 70 \qquad h_1 = \sum f_i X_i = 4.912 \qquad h_2 = \sum f_i X_i^2 = 345.438$$

$$h_3 = \sum f_i X_i^3 = 24.345.130 \qquad h_4 = \sum f_i X_i^4 = 1.719.341.106$$

Da essi si ottengono:

- la **devianza** (SQ) che con la formula abbreviata

$$SQ = \sum f_i X_i^2 - \frac{\left(\sum f_i X_i\right)^2}{n} = 345.438 - \frac{4.912^2}{70} = 755{,}9429$$

risulta uguale a 755,9429;

- la **varianza** (s^2)

$$s^2 = \frac{SQ}{n-1} = \frac{755{,}9429}{69} = 10{,}9557$$

che risulta uguale a 10,9557;

- il **momento terzo intorno alla media** (qui indicato con k_3, utile per calcolare direttamente g_1; è analogo a m_3, che serve per ricavare direttamente $\sqrt{b_1}$); per una distribuzione campionaria è

$$k_3 = \frac{n\sum\left(X_i - \overline{X}\right)^3}{(n-1)\cdot(n-2)}$$

e con la formula abbreviata è calcolato con

$$k_3 = \frac{nh_3 - 3h_1h_2 + 2\frac{h_1^3}{n}}{(n-1)\cdot(n-2)}$$

risultando con i dati dell'esempio

$$k_3 = \frac{70\cdot(24.345.130) - 3\cdot(4.912)\cdot(345.438) + 2\frac{(4.912)^3}{70}}{(69)\cdot(68)}\backslash$$

$$k_3 = \frac{1.704.159.100 - 5.090.374.368 + 3.386.156.529}{4.692} = \frac{-58.739}{4.692} = -12{,}519$$

k_3 = -12,519;

- il **momento quarto intorno alla media** (indicato con k_4, utile per calcolare direttamente g_2; è analogo a m_4 che serve per ricavare b_2); per una distribuzione campionaria è

$$k_4 = \frac{\dfrac{\sum\left(X_i - \overline{X}\right)^4 n(n+1)}{n-1} - 3\left[\sum\left(X_i - \overline{X}\right)^2\right]^2}{(n-2)\cdot(n-3)}$$

e con la formula abbreviata diventa

$$k_4 = A - B$$

dove

$$A = \frac{(n+1)\cdot\left(nh_4 - 4h_1h_3 + \dfrac{6h_1^2h_2}{n} - \dfrac{3h_1^4}{n^2}\right)}{(n-1)\cdot(n-2)\cdot(n-3)}$$

$$B = \frac{3(SQ)^2}{(n-2)\cdot(n-3)}$$

risultando con i dati dell'esempio

$$A = \frac{71\left[70(1.719.341.106) - 4(4.912)(24.345.130) + \dfrac{6(4.912)^2(345.438)}{70} - \dfrac{3(4.912)^4}{70^2}\right]}{(69)\cdot(68)\cdot(67)}$$

$$A = \frac{91.185.300}{314.364} = 290{,}0628$$

$$B = \frac{3(755{,}9429)^2}{(68)\cdot(67)} = \frac{1.714.349}{4.556} = 376{,}2838$$

$$k_4 = A - B = 290{,}0628 - 376{,}2838 = -86{,}221$$

k_4 = -86,221.

Infine si ricavano $\mathbf{g_1}$ e $\mathbf{g_2}$

con

$$g_1 = \frac{k_3}{s^3} = \frac{k_3}{\sqrt{\left(s^2\right)^3}} = \frac{-12{,}519}{\sqrt{(10{,}9557)^3}} = \frac{-12{,}519}{36{,}2627} = -0{,}3452$$

e

$$g_2 = \frac{k_4}{s^4} = \frac{k_4}{\left(s^2\right)^2} = \frac{-86{,}221}{(10{,}9557)^2} = -0{,}7183$$

ottenendo g_1 = -0,3452 e g_2 = -0,7183.

Da queste stime si possono ricavare $\sqrt{\mathbf{b_1}}$ e $\mathbf{b_2}$, (che sarebbe stato possibile ricavare direttamente dai dati attraverso m_3 e m_4).

- Da $\mathbf{g_1}$ mediante

$$\sqrt{b_1} = \frac{(n-2)\cdot g_1}{\sqrt{n\cdot(n-1)}}$$

e con i dati dell'esempio

$$\sqrt{b_1} = \frac{(70-2)\cdot(-0{,}3402)}{\sqrt{70\cdot(70-1)}} = \frac{-23{,}4736}{69{,}4982} = -0{,}3378$$

si ottiene $\sqrt{\mathbf{b_1}}$ **= - 0,3378.**

- Da $\mathbf{g_2}$ mediante

$$b_2 = \frac{(n-2)\cdot(n-3)\cdot g_2}{(n+1)\cdot(n-1)} + \frac{3\cdot(n-1)}{n+1}$$

e con i dati dell'esempio

$$b_2 = \frac{(70-2)\cdot(70-3)\cdot(-0{,}7183)}{(70+1)\cdot(70-1)} + \frac{3\cdot(70-1)}{70+1}$$

$$b_2 == \frac{-3272{,}5748}{4899} + \frac{207}{71} = -0{,}6680 + 2{,}9155 = 2{,}2475$$

si ottiene $\mathbf{b_2}$ **= 2,2475**.

TAVOLA DEI VALORI CRITICI DI SIMMETRIA g_1
APPROSSIMATI ALLA NORMALE PER IL TEST DI D'AGOSTINO

n	**α bil.**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.005**	**0.002**
	α uni.	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0025**	**0.001**
9		0,907	1,176	1,416	1,705	1,909	2,103	2,351
10		0,866	1,125	1,359	1,643	1,846	2,041	2,290
11		0,830	1,081	1,309	1,587	1,787	1,981	2,230
12		0,799	1,042	1,264	1,536	1,733	1,924	2,171
13		0,771	1,007	1,223	1,490	1,682	1,871	2,115
14		0,747	0,976	1,186	1,447	1,636	1,820	2,061
15		0,724	0,948	1,153	1,407	1,592	1,773	2,010
16		0,704	0,922	1,122	1,370	1,551	1,729	1,961
17		0,685	0,898	1,093	1,336	1,513	1,687	1,915
18		0,668	0,875	1,066	1,304	1,477	1,648	1,871
19		0,652	0,855	1,041	1,274	1,444	1,611	1,829
20		0,638	0,836	1,018	1,246	1,412	1,576	1,790
21		0,624	0,818	0,997	1,220	1,383	1,543	1,753
22		0,611	0,801	0,976	1,195	1,355	1,512	1,717
23		0,599	0,786	0,957	1,171	1,328	1,482	1,684
24		0,588	0,771	0,939	1,149	1,303	1,454	1,652
25		0,577	0,757	0,922	1,128	1,279	1,427	1,621
26		0,567	0,744	0,906	1,108	1,256	1,401	1,592
27		0,558	0,731	0,891	1,089	1,235	1,377	1,564
28		0,549	0,719	0,876	1,071	1,214	1,354	1,538
29		0,540	0,708	0,862	1,054	1,194	1,332	1,512
30		0,532	0,697	0,849	1,037	1,175	1,311	1,488
40		0,467	0,611	0,742	0,905	1,024	1,140	1,290
50		0,422	0,550	0,668	0,813	0,917	1,019	1,151
60		0,387	0,505	0,612	0,743	0,837	0,929	1,047
70		0,361	0,469	0,568	0,688	0,775	0,858	0,965
80		0,339	0,440	0,532	0,644	0,724	0,801	0,899
90		0,320	0,416	0,502	0,607	0,681	0,753	0,845
100		0,305	0,396	0,477	0,576	0,646	0,713	0,799
120		0,279	0,362	0,436	0,525	0,588	0,649	0,725
140		0,259	0,336	0,404	0,486	0,544	0,599	0,668
160		0,243	0,315	0,378	0,454	0,508	0,558	0,622
180		0,230	0,297	0,357	0,428	0,478	0,525	0,585
200		0,218	0,282	0,339	0,406	0,453	0,497	0,553
300		0,179	0,231	0,277	0,331	0,368	0,404	0,448
400		0,156	0,200	0,240	0,286	0,318	0,348	0,386
500		0,139	0,180	0,215	0,256	0,284	0,311	0,344
600		0,127	0,164	0,196	0,233	0,259	0,283	0,313
700		0,118	0,152	0,181	0,216	0,240	0,262	0,289
800		0,110	0,142	0,170	0,202	0,224	0,245	0,270
900		0,104	0,134	0,160	0,190	0,211	0,231	0,255
1000		0,099	0,127	0,152	0,181	0,200	0,219	0,241

Il **test per la simmetria** (***symmetry***) è **bilaterale** con ipotesi

$$\mathbf{H_0: \gamma_1 = 0} \qquad \text{contro} \qquad \mathbf{H_1: \gamma_1 \neq 0}$$

oppure l'equivalente

$$\mathbf{H_0: \sqrt{\beta_1} = 0} \qquad \text{contro} \qquad \mathbf{H_1: \sqrt{\beta_1} \neq 0}$$

quando di vuole verificare **se la distribuzione dei dati raccolti è simmetrica**, almeno approssimativamente.

A questo scopo, è sufficiente il semplice confronto del $\mathbf{g_1}$ calcolato con **i valori critici riportati nella tabella**.

Con $\mathbf{n = 70}$ e $g_1 = \mathbf{-0{,}3452}$ il valore critico alla probabilità $\alpha = \mathbf{0.20}$ per il test **bilaterale** è **0,723**.

La stima ottenuta dai dati in valore assoluto è minore; di conseguenza, si può affermare che la distribuzione è in sostanziale accordo con la normale, per quanto riguarda la simmetria

Ma per

- **dimensioni campionarie non riportate nella tabella** (ma sempre per $\mathbf{n \geq 9}$), sebbene sia possibile un calcolo rapido di interpolazione, e/o
- per una **stima precisa** della probabilità α di ottenere casualmente $\mathbf{H_0}$, cioè per non limitarsi a verificare se è maggiore o minore di una probabilità α prefissata,

si deve ricavare $\mathbf{Zg_1}$, cioè **il valore della normale standizzata Z** per il valore di $\mathbf{g_1}$ calcolato.

A questo scopo, dopo aver ripreso il valore di $\mathbf{\sqrt{b_1} = -\ 0{,}337758}$ già stimato, poiché le formule proposte sono state impostate su di esso, si deve ricorrere a vari passaggi (nei quali è importante avere valori molto precisi, almeno 6 cifre dopo la virgola):

- da $\mathbf{\sqrt{b_1}}$ e **n** si stima **A**

$$A = \sqrt{b_1} \cdot \sqrt{\frac{(n+1)\cdot(n+3)}{6\cdot(n-2)}} = 0{,}337758 \cdot \sqrt{\frac{(71)\cdot(73)}{6\cdot(68)}} = -1{,}203833$$

ottenendo A = 1,203833;

- da **n** si calcola **B**

$$B = \frac{3\cdot\left(n^2+27n-70\right)\cdot(n+1)\cdot(n+3)}{(n-2)\cdot(n+5)\cdot(n+7)\cdot(n+9)} = \frac{3\cdot\left[70^2+27\cdot(70)-70\right]\cdot(71)\cdot(73)}{(68)\cdot(75)\cdot(77)\cdot(79)} = 3{,}368090$$

ottenendo B = 3,368090;

- da **B** si ricava **C**

$$C = \sqrt{2 \cdot (B-1)} - 1 = \sqrt{2 \cdot (3{,}368090 - 1)} - 1 = 2{,}176277 - 1 = 1{,}176277$$

ottenendo C = 1,176277;

- da **C** si ricava **D**

$$D = \frac{1}{\sqrt{\ln \sqrt{C}}} = \frac{1}{\sqrt{\ln \sqrt{1{,}176277}}} = \frac{1}{\sqrt{\ln 1{,}084563}} = \frac{1}{0{,}284916} = 3{,}509806$$

ottenendo D = 3,509806;

- da **A** e **C** si ricava **E**

$$E = \frac{A}{\sqrt{\frac{2}{C-1}}} = \frac{-1{,}203833}{\sqrt{\frac{2}{1{,}176377 - 1}}} = \frac{-1{,}203833}{\sqrt{11{,}339347}} = \frac{-1{,}203833}{3{,}367395} = -0{,}357497$$

ottenendo E = -0,357497.

Infine da **D** e **E** si ottiene $\mathbf{Z_{g1}}$ con

$$Z_{g1} = D \cdot \ln\left(E + \sqrt{E^2 + 1}\right)$$

$$Z_{g1} = 3{,}509806 \cdot \ln\left(-0{,}357497 + \sqrt{(-0{,}357497)^2 + 1}\right)$$

$$Z_{g1} = 3{,}509806 \cdot \ln\left[(-0{,}357497) + \sqrt{(-0{,}357497)^2 + 1}\right] = 3{,}509806 \cdot \ln\left(-0{,}357497 + \sqrt{1{,}127804}\right)$$

$$Z_{g1} = 3{,}509806 \cdot \ln 0{,}704484 = 3{,}509806 \cdot (-0{,}350290) = -1{,}2294$$

ottenendo $\mathbf{Z_{g1}}$ = -1,2294.

Approssimato a **Z = -1,23** in una **distribuzione normale bilaterale** corrisponde ad una probabilità α = 0,219 o 21,9%. E' una probabilità alta: non solo non permette di rifiutare l'ipotesi nulla, ma autorizza a sostenere ragionevolmente che lo scostamento dalla normale è molto ridotto.

Il **test per la simmetria** (***symmetry***) è **unilaterale** con ipotesi

$$\mathbf{H_0: \gamma_1 \geq 0} \qquad \text{contro} \qquad \mathbf{H_1: \gamma_1 < 0}$$

oppure l'equivalente

$$\mathbf{H_0: \sqrt{\beta_1} \geq 0} \qquad \text{contro} \qquad \mathbf{H_1: \sqrt{\beta_1} < 0}$$

quando si vuole verificare

- **se la distribuzione dei dati raccolti ha una asimmetria sinistra o negativa.**

Il calcolo ha una procedura identica a quella prima illustrata; ma per rifiutare l'ipotesi nulla il valore di g_1 deve essere negativo e, in valore assoluto, essere superiore a quello critico.

Si ricorre a un test **unilaterale** con ipotesi

$$\mathbf{H_0: \gamma_1 \leq 0} \qquad \text{contro} \qquad \mathbf{H_1: \gamma_1 > 0}$$

oppure l'equivalente

$$\mathbf{H_0: \sqrt{\beta_1} \leq 0} \qquad \text{contro} \qquad \mathbf{H_1: \sqrt{\beta_1} > 0}$$

quando si vuole verificare

- **se la distribuzione dei dati raccolti ha una asimmetria destra o positiva.**

Per rifiutare l'ipotesi nulla, il valore di g_1 deve essere positivo e, in valore assoluto, essere superiore a quello critico. Se si ricorre al calcolo di $\mathbf{Z_{g1}}$, per rifiutare l'ipotesi nulla la probabilità α stimata in una distribuzione normale unilaterale deve essere minore di quella prefissata.

Il **test per la curtosi** (***kurtosis***) è **bilaterale** con ipotesi

$$\mathbf{H_0: \gamma_2 = 0} \qquad \text{contro} \qquad \mathbf{H_1: \gamma_2 \neq 0}$$

oppure l'equivalente

$$\mathbf{H_0: \beta_2 = 3} \qquad \text{contro} \qquad \mathbf{H_1: \beta_2 \neq 3}$$

quando si vuole verificare

- **se il campione è stato estratto da una popolazione mesocurtica** (**normale**).

Il metodo più semplice è il confronto con la tabella dei valori critici (pagina successiva). Ad esempio, con $\mathbf{n = 70}$ e $g_2 = \mathbf{-0,7183}$ come stimato in precedenza, non è possibile rifiutare l'ipotesi nulla, poiché il valore è minore di quello critico corrispondente alla probabilità $\alpha = 0.05$.

TAVOLA DEI VALORI CRITICI DI CURTOSI g_2
APPROSSIMATI ALLA NORMALE PER IL TEST DI D'AGOSTINO

n	**α bil.**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.005**	**0.002**
	α uni.	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0025**	**0.001**
20		1,241	1,850	2,486	3,385	4,121	4,914	6,063
21		1,215	1,812	2,436	3,318	4,040	4,818	5,967
22		1,191	1,776	2,388	3,254	3,963	4,727	5,835
23		1,168	1,743	2,343	3,193	3,889	4,639	5,728
24		1,147	1,711	2,300	3,135	3,818	4,555	5,624
25		1,127	1,681	2,260	3,080	3,751	4,474	5,524
26		1,108	1,653	2,222	3,027	3,686	4,397	5,427
27		1,090	1,626	2,185	2,976	3,624	4,322	5,335
28		1,074	1,601	2,150	2,928	3,565	4,251	5,245
29		1,057	1,576	2,117	2,882	3,508	4,182	5,159
30		1,042	1,553	2,085	2,838	3,453	4,116	5,075
32		1,014	1,509	2,025	2,574	3,350	3,990	4,917
34		0,988	1,469	1,971	2,677	3,254	3,874	4,769
36		0,964	1,432	1,919	2,606	3,165	3,765	4,631
38		0,942	1,398	1,872	2,539	3,081	3,663	4,502
40		0,921	1,366	1,828	2,476	3,003	3,568	4,380
42		0,902	1,337	1,787	2,418	2,930	3,478	4,266
44		0,884	1,309	1,748	2,363	2,861	3,394	4,158
46		0,868	1,282	1,711	2,311	2,796	3,314	4,057
48		0,852	1,258	1,677	2,262	2,735	3,239	3,961
50		0,837	1,234	1,644	2,216	2,677	3,168	3,870
60		0,773	1,135	1,504	2,017	2,428	2,862	3,480
70		0,723	1,055	1,394	1,859	2,230	2,620	3,171
80		0,681	0,990	1,303	1,730	2,069	2,423	2,921
90		0,646	0,935	1,227	1,622	1,934	2,259	2,714
100		0,617	0,889	1,162	1,531	1,820	2,121	2,538
110		0,590	0,848	1,105	1,452	1,722	2,002	2,389
120		0,567	0,813	1,056	1,383	1,637	1,898	2,259
140		0,529	0,753	0,974	1,268	1,494	1,727	2,045
160		0,497	0,704	0,907	1,175	1,380	1,590	1,875
180		0,470	0,663	0,851	1,098	1,287	1,478	1,737
200		0,447	0,628	0,804	1,034	1,208	1,384	1,621
220		0,428	0,599	0,764	0,979	1,141	1,305	1,524
240		0,410	0,572	0,729	0,931	1,083	1,236	1,440
300		0,368	0,510	0,645	0,819	0,948	1,077	1,247
400		0,320	0,439	0,551	0,694	0,798	0,902	1,038
500		0,287	0,391	0,488	0,610	0,700	0,787	0,902
600		0,262	0,355	0,442	0,550	0,629	0,706	0,805
700		0,243	0,328	0,406	0,504	0,575	0,643	0,732
800		0,227	0,305	0,378	0,468	0,532	0,594	0,675
900		0,214	0,287	0,355	0,438	0,497	0,555	0,628
1000		0,203	0,272	0,335	0,412	0,486	0,521	0,590

Anzi, poiché il $\mathbf{g_2}$ calcolato è minore, in valore assoluto, di quello riportato nella tabella per la probabilità bilaterale $\alpha = 0.20$, per quanto rigurda la curtosi si può sostenere che lo scostamento da una perfetta normalità è minimo: la distribuzione è in buon accordo con la normale.

Anche in questo caso, per

- **dimensioni campionarie non riportate nella tabella** (ma sempre per **n ≥ 20**), sebbene sia possibile un calcolo rapido di interpolazione, e/o
- per una **stima precisa** della probabilità α di ottenere casualmente $\mathbf{H_0}$, cioè per non limitarsi a verificare se è maggiore o minore di una probabilità prefissata,

si deve ricavare $\mathbf{Zg_2}$, cioè **il valore della normale standizzata Z** per il valore di $\mathbf{g_2}$ calcolato.

A questo scopo, utilizzando **n = 70** e g_2 **= -0,7183** si deve ricorrere a vari passaggi (nei quali è ancora importante avere valori molto precisi, almeno 6 cifre dopo la virgola):

- utilizzando **n** si calcola **A**

$$A = \frac{24 \cdot n \cdot (n-2) \cdot (n-3)}{(n+1)^2 \cdot (n+3) \cdot (n+5)} = \frac{24 \cdot (70) \cdot (68) \cdot (67)}{(71)^2 \cdot (73) \cdot (75)} = \frac{7.654.080}{27.599.475} = 0{,}277327$$

ottenendo A = 0,277327;

- da **A** e $\mathbf{g_2}$ si ricava **B**

$$B = \frac{(n-2) \cdot (n-3) \cdot |g_2|}{(n+1) \cdot (n-1) \cdot \sqrt{A}} = \frac{(68) \cdot (67) \cdot (0{,}7183)}{(71) \cdot (69) \cdot \sqrt{0{,}277327}} = \frac{3.272{,}574800}{2.579{,}903825} = 1{,}268487$$

ottenendo B = 1,268487;

- utilizzando **n** si ricava **C**

$$C = \frac{6 \cdot (n^2 - 5n + 2)}{(n+7) \cdot (n+9)} \cdot \sqrt{\frac{6 \cdot (n+3) \cdot (n+5)}{n \cdot (n-2) \cdot (n-3)}} = \frac{6 \cdot [70^2 - 5 \cdot (70) + 2]}{(77) \cdot (79)} \cdot \sqrt{\frac{6 \cdot (73) \cdot (75)}{70 \cdot (68) \cdot (67)}}$$

$$C = \frac{27.312}{6.083} \cdot \sqrt{\frac{32.850}{318.920}} = (4{,}489890) \cdot (0{,}320942) = 1{,}440994$$

ottenendo C = 1,440994;

- da **C** si ricava D

$$D = 6 + \frac{8}{C} \cdot \left(\frac{2}{C} + \sqrt{1 + \frac{2^2}{C^2}} \right) = 6 + \frac{8}{1{,}440994} \cdot \left(\frac{2}{1{,}440994} + \sqrt{1 + \frac{2^2}{1{,}440994^2}} \right)$$

$$D = 6 + 5{,}551723 \cdot (1{,}387931 + 1{,}710658) = 6 + 17{,}202508 = 23{,}202508$$

ottenendo D = 23,202508;

- da **B** e da **D** si ricava **E**

$$E = \frac{1 - \frac{2}{D}}{1 + B \cdot \sqrt{\frac{2}{D-4}}} = \frac{1 - \frac{2}{23{,}202508}}{1 + 1{,}268487 \cdot \sqrt{\frac{2}{23{,}202508 - 4}}}$$

$$E = \frac{1 - 0{,}086206}{1 + 0{,}409376} = \frac{0{,}913794}{1{,}409376} = 0{,}648368$$

ottenendo E = 0,648368.

Infine da **D** e da **E** si ricava $\mathbf{Z_{g2}}$

$$Z_{g2} = \frac{1 - \frac{2}{9D} - \sqrt[3]{E}}{\sqrt{\frac{2}{9D}}} = \frac{1 - \frac{2}{9 \cdot (23{,}202508)} - \sqrt[3]{0{,}648368}}{\sqrt{\frac{2}{9 \cdot (23{,}202508)}}}$$

$$Z_{g2} = \frac{1 - 0{,}009578 - 0{,}865514}{\sqrt{0{,}009578}} = \frac{0{,}124908}{0{,}097867} = 1{,}2763$$

ottenendo $\mathbf{Z_{g2} = 1{,}2763}$.

Approssimato a **Z = 1,27** in una **distribuzione normale bilaterale** corrisponde ad una probabilità α = 0,204 o 20,4%. E' una probabilità alta: non solo non permette di rifiutare l'ipotesi nulla, ma autorizza a sostenere ragionevolmente che lo scostamento dalla normale è molto ridotto.

E' sempre opportuno che l'arrotondamento del valore di Z a due cifre dopo la virgola, come richiesto dalla tabella dei valori critici, avvenga per difetto. Il valore di α è maggiore e il test risulta più cautelativo.

Anche il **test per la curtosi** può essere **unilaterale** con ipotesi

$$\mathbf{H_0: \gamma_2 \leq 0} \qquad \text{contro} \qquad \mathbf{H_1: \gamma_2 > 0}$$

oppure l'equivalente

$$\mathbf{H_0: \beta_2 \leq 3} \qquad \text{contro} \qquad \mathbf{H_1: \beta_2 > 3}$$

quando si vuole verificare specificatamente **se la distribuzione dei dati raccolti è platicurtica.**
Per rifiutare l'ipotesi nulla, il valore di $\mathbf{g_2}$ deve essere positivo e, in valore assoluto, essere superiore a quello critico. Con il calcolo di $\mathbf{Z_{g1}}$, per rifiutare l'ipotesi nulla la probabilità α, stimata in una distribuzione normale unilaterale, deve essere minore di quella prefissata.

Per verificare l'ipotesi che la curva sia **leptocurtica**, cioè con ipotesi **unilaterale**

$$\mathbf{H_0: \gamma_2 \geq 0} \qquad \text{contro} \qquad \mathbf{H_1: \gamma_2 < 0}$$

oppure l'equivalente

$$\mathbf{H_0: \beta_2 \geq 3} \qquad \text{contro} \qquad \mathbf{H_1: \beta_2 < 3}$$

il valore di g_2 deve essere negativo e, in valore assoluto, essere superiore a quello critico. Se si ricorre al calcolo di $\mathbf{Z_{g1}}$ per rifiutare l'ipotesi nulla, la probabilità α calcolata deve essere minore di quella prefissata.

Il **test per la normalità**, come già presentato all'inizio di questo paragrafo, permette di **verificare solo l'ipotesi nulla bilaterale:** se il campione è in accordo con la corrispondente distribuzione normale, costruita con stessa media e varianza uguale.
Tale test è fondato sulla statistica

$$K^2 = Z_{g1}^2 + Z_{g2}^2$$

e il valore di $\mathbf{K^2}$ calcolato deve essere confrontato con la tabella del **tabella χ^2 con df = 2**, qui riportata solo nella coda destra della distribuzione:

α					
.25	*.10*	*.05*	*.025*	*.01*	*.005*
2.773	**4.605**	**5.991**	**7.378**	**9.210**	**10.597**

Con i dati dell'esempio, in cui

$$\mathbf{Z_{g1}} = 1{,}2294 \quad \text{e} \quad Z_{g2} = 1{,}2763$$

mediante

$$K^2 = Z_{g1}^2 + Z_{g2}^2 = (-1{,}2294)^2 + (1{,}2763)^2 = 1{,}5041 + 1{,}6289 = 3{,}133$$

si ottiene $K^2 = 3{,}133$.

Il valore calcolato è nettamente inferiore a quello critico per $\alpha = 0.10$ (uguale a 4,605). Di conseguenza, la probabilità che l'ipotesi nulla sia vera è alta: c'è sostanziale accordo tra la distribuzione osservata e quella normale corrispondente.

13.8. CENNI DEL TEST DI CRAMER-VON MISES PER UN CAMPIONE E PER DUE CAMPIONI INDIPENDENTI

Il test di **Cramér e von Mises,** riportato in alcuni programmi informatici e spesso citato in varie pubblicazioni per cui è utile che sia conosciuto almeno nei suoi aspetti principali, è fondato su una logica del tutto simile a quella del **test di Kolmogorov-Smirnov.** La metodologia è stata proposta alla fine degli anni '20 con l'articolo di H. **Cramér** del 1928 ***On the composition of elementary errors*** (pubblicato sulla rivista **Skandinavisk Aktuarietidskrift**, Vol. 11, pp. 13-74 e pp. 141-180) e all'inizio degli anni '30 nel volume di R. **von Mises** del 1931, pubblicato in tedesco, ***Wahrscheinlichkeitsrechnung und Ihre Anwendung in der Statistik und Theoretischen Physik*** (edito da F. Deutiche, Leipzig).

E' stata perfezionata da N. V. **Smirnov** nel 1936 per quanto riguarda le caratteristiche della distribuzione dei valori critici con l'articolo, in francese, ***Sur la distribution de W^2*** (criterium de M. R. v. Mises) (pubblicato su **Comptes Rendus,** Paris, Vol. 202, pp. 449-452).

Il test di Kolmogorov-Smirnov e molto più frequentemente utilizzato e è riportato in quasi tutte le librerie informatiche. Il test di Cramér von Mises gode del vantaggio di essere più semplice.

Come il test di Kolmogorov-Smirnov, può essere applicato

- nel caso di **un solo campione**, per verificare la **bontà dell'adattamento**,
- nel caso di **due campioni** indipendenti, per verificare **se appartengono alla stessa popolazione** o comunque a **popolazioni identiche**.

Per verificare l'accordo tra una distribuzione campionaria e una distribuzione attesa di qualsiasi forma, è necessario che **la variabile casuale** sia **continua.** Come in tutti i test per la bontà dell'adattamento, l'ipotesi riguarda tutti i parametri della distribuzione (media, varainza, simmetria, curtosi): quando il test risulta significativo, **la distribuzione osservata si differenzia da quella attesa per almeno un parametro**, senza alcuna informazione su quale esso sia.

Ovviamente il test è utile quando il parametro non è noto; soprattutto quando la differenza può essere determinata da un concorso di più parametri.

E' un test generalista, in cui l'ipotesi nulla è che il campione osservato appartenga alla popolazione teorica indicata. Appunto perché dipendente da più fattori, l'ipotesi alternativa è quasi sempre bilaterale.
Limitando anche in questo caso la spiegazione alla comprensione dei programmi informatici, quindi senza entrare nel dettaglio delle procedure di calcolo,

1 - dopo aver costruito le ***n*** classi della distribuzione osservata e della distribuzione attesa, sulla base della legge matematica o statistica prescelta

2 - si stima il valore di un indicatore W_n^2 che è uguale a

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F(x_i) - \frac{2i-1}{2n}\right]^2$$

dove

- $F(x_i)$ è lo scarto tra osservato ed atteso nella classe i.

3 - Si rifiuta l'ipotesi nulla, quindi c'è disaccordo tra distribuzione osservata e distribuzione attesa, quando W_n^2 supera il valore critico $\mathbf{C_\alpha}$, riportato nella tabella seguente.

Con **n > 10**, si possono usare i seguenti valori critici $\mathbf{C_\alpha}$ per la probabilità α prefissata.

α	0.10	0.05	0.01	0.001
$\mathbf{C_\alpha}$	0,347	0,461	0,743	1,168

Quando la distribuzione attesa è costruita sulla base di uno o più parametri (media, varianza, simmetria, curtosi) calcolati nella distribuzione osservata, il valore di W_n^2 è inferiore.
Sono stati stimati valori critici anche per queste analisi più specifiche che considerano contemporanemamente **k** parametri.

La metodologia per il caso di due campioni indipendenti, può essere illustrata con la presentazione di un esempio.
Si supponga di avere il campione A

186	191	217	220	255	270	300	380

con un numero di osservazioni m = 8

e il campione B

104	115	120	150	171	175	188	210	215	220	260	300

con un numero di osservazioni n = 12.

Esiste una differenza significativa tra le due distribuzioni?

Risposta.

1 – Delle due serie di dati, si forma una distribuzione unica, mantenendo l'informazione sul gruppo di appartenza di ogni valore

(1)	(2)	(3)	(4)	(5)	(6)
A	B	Cum. A	Cum B	D	D^2
---	104	0,0	0,083	-0,083	0,00689
---	115	0,0	0,167	-0,167	0,02789
---	120	0,0	0,250	-0,250	0,06250
---	150	0,0	0,333	-0,333	0,11089
---	171	0,0	0,416	-0,416	0,17316
---	175	0,0	0,500	-0,500	0,25000
186	---	0,125	0,500	-0,375	0,14062
---	188	0,125	0,583	-0,458	0,20976
191	---	0,250	0,583	-0,333	0,11089
---	210	0,250	0,667	-0,417	0,17389
---	215	0,250	0,750	-0,500	0,25000
217	---	0,375	0,750	-0,375	0,14062
220	220	0,500	0,833	-0,333	0,11089
255	---	0,625	0,833	-0,208	0,04326
---	260	0,625	0,917	-0,292	0,08526
270	---	0,750	0,917	-0,167	0,02789
---	300	0,750	1,000	-0,250	0,06250
330	---	0,875	1,000	-0,125	0,01563
380	---	1,000	1,000	0,000	0,00000
				$\sum D^2$ =	2,00254

come risulta nella tabella, leggendo insieme la colonna del campione A e la colonna del campione B

2 - Poiché il numero di osservazioni, con spesso nel caso di due campioni indipendenti, è diverso, si trasformano i valori nelle rispettive proporzioni e si costruiscono sia la cumulata del campione A sia la cumulata del campione B (colonna 3 e 4)

3 – Si costruisce la serie delle differenze D (colonna 5),
dove

$$D = \text{Cum. A} - \text{Cum.B}$$

4 – Si costruisce la serie delle differenze al quadrato (D^2 della colonna 6) e se ne ricava la somma (S), ottenendo

$$\sum D^2 = S = 2{,}00254$$

5 – Attraverso la relazione

$$W^2 = \frac{m \cdot n \cdot S}{(m+n)^2} = \frac{8 \cdot 12 \cdot 2{,}00254}{(8+12)^2} = \frac{192{,}24}{400} = 0{,}48$$

con i dati dell'esempio di trova $W^2 = 0{,}48$.

6 – In un **test bilaterale**, come di solito nel confronto generico tra due distribuzioni di dati,
- alla probabilità $\alpha = 0.05$ il valore critico è 0,461
- alla probabilità $\alpha = 0.01$ il valore critico è 0,743

7 – Poiché il valore calcolato (0,48) si colloca tra la probabiltà 0.05 e 0.01 è possibile rifiutare l'ipotesi nulla con probabilità $P < 0.05$ di commettere un errore di Tipo I.
Le due distribuzioni non appartengono alla stessa popolazione.

13.9. L'OUTLIER: DATO ANOMALO O DATO SBAGLIATO? DEFINIZIONI DI OUTLIER

Un **outlier**, in italiano tradotto con i termini **dato anomalo** o **valore fuori limite**, è una osservazione che **appare differente** dalle altre dello **stesso gruppo**. Il concetto spesso è limitato a un solo dato; ma può essere esteso a più valori contemporaneamente, rispetto al gruppo più ampio di osservazioni raccolte nelle stesse condizioni. In termini più tecnici,
- **un dato è outlier quando non appare consistente con gli altri,**

cioè quando **altera uno o più parametri** contemporaneamente tra media, varianza e simmetria.

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.), in seguito utilizzato anche da altri dipartimenti della difesa americana, sono chiamati ***gross errors***. Sono definiti come quei valori che possono apparire troppo grandi oppure troppo piccoli allo sperimentatore, tali da suscitargli il timore che la loro presenza alteri i risultati reali.
Ma qual è il risultato reale? Quello che li comprende oppure quello che li esclude?

In tempi più recenti, vari autori di testi di statistica applicata forniscono una determinazione ancor più sfumata di **outlier**, definendoli **quelli inconsistenti rispetto all'ambiente** nel quale sono stati osservati. Non è solamente il confronto con gli altri, ma il contesto delle analisi di laboratorio o quello naturale in cui sono stati rilevati, che li fa definire outlier.

E' una impostazione che altri criticano, se non altro per la carenza di criteri condivisi.
B. E. **Rodda** nel 1990 in ***Bioavailability: design and analysis*** (pubblicato su **Statistical Methology in the Pharmaceutical Sciences**, **Berry** D. A. ed., Marcel **Dekker**, New York, p.78) afferma: ... ***they are very difficult to define and are only called outlier because they are inconsistent with the environment in which they are observed***.

Queste definizioni legate al "concetto dell'**apparire differente"**, implicano una **valutazione soggettiva**. Nel loro volume ***Outliers in statistical data*** del 1994 (3rd ed. Chichester, John Wiley & Sons), V. **Barnett** e T. **Lewis** enfatizzano questa idea, affermando che
- **un outlier è tale quando suscita una sorpresa genuina nell'osservatore.**

Di conseguenza, dipende dalla **conoscenza personale** del fenomeno. Come enunciazione di un principio, questa indicazione può essere accettabile, quando si tratta di un fenomeno che rientra nella cultura generale; ma persone differenti possano classificare diversamente lo stesso dato.

Esempi di conoscenza comune nella individuazione di un dato anomalo possono essere tratti, ad esempio, dalla misure di (a) peso e di (b) altezza di un gruppo di maschi adulti.

A) Si osservi il peso in Kg di un gruppo di maschi adulti, come nella seguente serie già ordinata

65	69	75	78	80	81	85	130

La presenza dell'ultimo valore, una persona che pesa 130 Kg, non è tale da suscitare meraviglia in molte culture, anche se è il doppio del peso minore: in molte popolazioni non è raro trovare individui adulti di quel peso.

B) Nella serie successiva di altezze in cm sempre di maschi adulti

160	166	170	172	180	182	195	230

si pone un problema di credibilità per l'ultimo dato: 230 cm di altezza sono un valore eccezionale nell'esperienza di ogni persona. E, in questo caso, è solo il 44% in più dell'altezza minima rilevata in quel campione, mentre il peso di 130 Kg era il 100% in più di quello minimo.

In fenomeni meno noti, quali la quantità di colesterolo in maschi adulti oppure in giovani donne, il livello di ozono in una città, quello di radioattività emessa naturalmente da una roccia, la valutazione non è altrettanto semplice, per persone non sufficientemente esperte. Inoltre, per decidere scientificamente, si pone il problema

- di ricavare dai dati una stima espressa in **termini probabilistici**,
- che, a partire dagli stessi dati, induca i ricercatori alle **medesime conclusioni**,
- sulla base di **concetti e metodi condivisi**.

Prima di avviare **la discussione se un dato è anomalo**, sulla base della loro lunga esperienza di elaborazione di dati statistici per gruppi o intere strutture di ricovero e di ricerca, molti statistici suggeriscono di chiedersi se **il dato** è semplicemente **sbagliato**.

Geoff R. **Norman** e David L. **Streiner** nel loro testo del 1998 ***Biostatistics: The Bare Essentials*** (pubblicato da B. C. Decker, Inc. Hamilton, Ontario, Canada), tradotto in italiano da Giovanni **Capelli** e Giovanni **D'Abramo** nell'anno 2000 con il titolo ***Biostatistica. Quello che avreste voluto sapere*** (edito da Casa Editrice Ambrosiana, Milano XII + 260 p.) con il loro stile intelligente e scherzoso di presentare i problemi della statistica (a pag. 202 - 203) scrivono: *Idealmente, questo paragrafo avrebbe dovuto chiamarsi "Scovare i dati sbagliati", perché è questo il nostro obiettivo: trovare i dati che sono riusciti a eludere tutti i nostri sforzi per identificare gli errori prima che arrivassero all'archivio finale del nostro studio (aggiungendo, nella nota: Ci sono parecchi metodi per fare*

qualcosa del genere, a cominciare dal doppio inserimento dei dati seguito da una verifica delle differenze tra i due archivi. Ma se state leggendo questo libro per cercare altri metodi finalizzati a questo, avete preso quello sbagliato: andate a cercare su qualcun altro).
Ad esempio, se avete passato la notte in bianco e oggi vi si incrociano gli occhi, è possibile che inseriate sul computer 42 anni invece di 24 per l'età di uno dei soggetti: potreste non riuscire mai ad accorgervi dell'errore. Infatti, tutti e due questi valori stanno molto probabilmente all'interno dell'intervallo dei valori ammessi nel vostro studio e non avete indizi che possano far sospettare che voi (o qualche vostro collaboratore) abbiate fatto confusione. Il meglio che si può fare è andare a verificare i dati che si trovano fuori dell'intervallo di valori attesi o se sono presenti incongruenze nell'ambito di un singolo soggetto.

In questo settore, ogni disciplina ha i suoi metodi per scoprire gli errori. Poiché le cause sono diverse, in un testo generalista sono inutili i dibattiti statistici su di essi e le discussioni tecniche sui fattori che li possono determinare. Quelli più frequenti, comuni a molte discipline, sono:
1) un funzionamento errato, anche se temporaneo, dello strumento;
2) una contaminazione in analisi di laboratorio;
3) il tecnico che ha interpretato male il risultato;
4) un errore nella lettura e/o successiva scrittura del dato, in uno dei vari passaggi dalla prima rilevazione all'analisi statistica.

Ma può essere un valore reale, che è solamente molto grande o molto piccolo entro i valori estremi della distribuzione delle osservazioni.

La ricerca dei **dati anomali,** che possono non essere giudicati **valori sbagliati**, può avere **varie finalità**:
- la **stima della media o della varianza reale** di un fenomeno, una volta che siano stati eliminati gli outlier;
- la **identificazione degli outlier per distinguerli dai valori normali** e **studiare le cause che li hanno generati**;
- la motivazione per **passare dalla statistica parametrica a quella non parametrica** nella scelta del test;
- la giustificazione per una **trasformazione dei dati**;
- a volte è la **eliminazione dell'outlier** dal gruppo, per **effettuare un test parametrico** che rispetti le condizioni di validità.

Secondo alcuni autori, la individuazione degli **outliers** per questa ultima finalità, cioè l'applicazione corrette di un test parametrico, è solo un problema teorico. Nella pratica sperimentale, può essere ignorato ai fini dell'inferenza statistica: è sufficiente ricorrere a **procedure robuste**, cioè a metodi i cui risultati non sono alterati dal mancato rispetto delle condizioni di normalità, omoschedasticità e simmetria.
Nel suo testo del 1996 ***Applied multivariate statistics for the social sciences*** (3rd. Ed. Mahwah, NJ, Lawrence Erlbaum Associates, Publishers), J. **Stevens** condivide questa impostazione per la statistica multivariata. A maggior ragione, può essere applicata alla statistica univariata e a quella bivariata. Ad esempio, è nota la **robustezza del test t di Student** per due **campioni indipendenti,** rispetto alle assunzioni di normalità.

Ma non tutti gli studiosi sono dello stesso parere. Molti ritengono necessario decidere se nei dati sono presenti degli outlier, ai fini della **scelta più corretta del test,** pure rilevando le **difficoltà di tale operazione** di identificazione, che sono attribuite a tre cause principali.

1) Nel suo volume del 1998 ***Data driven statistical methods*** (London, Chapman & Hall), P. **Sprent** annota con ironia che molti test, proposti per evidenziare la presenza di uno o più outliers in un campione di dati, perdono potenza quando essi sono presenti. Si parla di **masking effect** o di **influential observation**:
- **la potenza di un test per evidenziare un outlier è compromessa, quando esso è presente**.

2) La seconda serie di problemi deriva dall'**effetto delle dimensioni del campione**, quando il giudizio è di tipo ***occhiometrico***, per usare ancora il linguaggio di **Norman** e **Streiner**, secondo i loro traduttori.
Lo **stesso dato**
- può **apparire outlier,** se il campione è formato da **poche unità**;
- **non risultare outlier,** se il campione è formato da almeno **due decine di osservazioni.**

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. 530-531) con un esempio mette in evidenza come lo **stesso dato**
- possa apparire **anomalo in un campione di 6** osservazioni
- mentre **non risulti più un outlier se le osservazioni diventano 12**.

Si assuma che nell'analisi chimica di una concentrazione e riportati nella tabella seguente, siano stati ottenuti
- prima i 6 dati della prima riga

- successivamente gli altri 6 della seconda riga della tabella sottostante

97	98	98	95	86	99
98	98	97	99	98	95

Il numero 86 della prima riga, a molti può apparire un outlier. Ma lo è?

Applicando il principio della **consistenza del dato**, vale a dire **se la presenza o meno del dato altera le statistiche** del campione, è stato misurato l'impatto del potenziale outlier sulle misure di tendenza centrale e di variabilità dei due campioni. Nella tabella successiva sono stati analizzati

- solo la prima riga, senza il valore 86;
- solo la prima riga, con il valore 86;
- i 12 dati delle due righe.

Misure	Riga 1		Riga 1 + 2
	Senza 86	Con 86	
Dimensioni n	5	6	12
Media $\overline{X}$	97,4	95,5	96,5
Deviazione st. S	1,5	4,8	3,5
Range o Diff. Massima	4	11	11
Mediana	98	97,5	98

I risultati mostrano come

- analizzando solo la prima riga, le statistiche differiscono sensibilmente se si comprende oppure si esclude il valore 86;
- mentre analizzando tutti i 12 dati, le statistiche (media e deviazione standard) si avvicinano molto a quelle calcolate per la prima riga, quando non comprendono il valore 86.

Con tale analisi,

- se disponessero solo dei primi 6 dati, probabilmente molti affermerebbero che **86 è un outlier**;
- ma, disponendo dei 12 dati, altrettanto facilmente direbbero che **86 non è un outlier**.

Il numero minimo di dati per decidere se è presente un outlier è tre. Naturalmente, più dati sono presenti più è probabile che un dato sia identificato come tale. I metodi di identificazione fondati sulla logica statistica, come tutti i test, sono impostati in modo tale che

- con un campione piccolo la discrepanza deve essere molto grande, quando i dati sono pochi,
- mentre essa si riduce, quando i dati aumentano.

Di conseguenza, diversamente dal masking effect, l'effetto del numero è più apparente che reale, quando l'analisi è condotta seguendo i metodi statistici

3) Il terzo gruppo di difficoltà deriva dal tipo di **distribuzione ipotizzata**. Nel 1943, R. A. **Fisher** in un articolo in collaborazione con A. S. **Corbet** e C. B. **William,** dal titolo ***The relation between the number of species and the number of individuals in a random sample of a animal population*** (pubblicato su **Journal of Animal Ecology** Vol. 12, pp. 42 – 57), nel conteggio di insetti raccolti con una trappola riporta la seguente distribuzione di 15 osservazioni, qui ordinata in modo crescente per meglio evidenziarne le caratteristiche statistiche:

3	3	4	5	7	11	12	15	18	24	51	54	84	120	560

Il valore 560 è un outlier?

Riprendendo questi dati, V. **Barnett** e T. **Lewis** nel loro volume del 1994 ***Outliers in statistical data*** (3rd ed., Chichester, John Wiley and Sons) affermano che **quasi tutti i test statistici** portano ad affermare che **560 è un outlier**. Questo a causa del fatto che molti test ipotizzano la distribuzione normale dei dati, sia nel caso in cui il valore sospettato è escluso, sia quando è compreso. Anche test di statistica non parametrica indurrebbero ad affermare che 560 è un outlier.

In realtà non lo è. La distribuzione di un conteggio di questi insetti è fortemente asimmetrica, come possono facilmente capire coloro che sanno che queste specie vivono in sciami.

Sciami

Oltre a queste incertezze sulla necessità e sui metodi per identificare un outlier, il dibattito verte su come effettuare test che siano condivisi, almeno per compiere una scelta sulla base di una probabilità corretta, non su quella di una semplice impressione. Nel loro volume già citato, V. **Barnett** e T. **Lewis** riportano **48** metodi statistici per identificare uno o più outlier, solamente rispetto all'assunzione di **normalità.** In letteratura esistono proposte anche per altre distribuzioni, quali la binomiale, la poissoniana, la gamma, … .

Una prima serie di metodi, di uso molto semplice, sono derivati dalle rappresentazioni grafiche delle distribuzioni.

Esse evidenziano visivamente **la distanza di un valore dalla media del gruppo** e/o dalla sua **distribuzione complessiva**; ma non forniscono la probabilità di errore nella decisione di considerarlo un outlier. Esistono vari programmi informatici che evidenziano la **presenza potenziale di un outlier**, sulla base della sua collocazione rispetto a tutti gli altri dati del gruppo.

13.10. IDENTIFICAZIONE DEGLI OUTLIER CON IL METODI GRAFICI: IL BOX-AND-WHISKERS DI TUKEY.

Tra questi metodi grafici già riportati nel capitolo I sulla statistica descrittiva, quali gli **stem-and-leaf plots** e gli **istogrammi**, il più diffuso è il **diagramma a scatola** (***box-plot***), chiamato anche **diagramma a scatola e baffi** (***Box-and-Whiskers***), presentato in modo organico da John W. **Tukey** nel suo testo del 1977 ***Exploratory Data Analysis*** (pubblicato da Addison-Wesley, Reading, Mass.).

Serve per rappresentare visivamente le tre caratteristiche fondamentali di una distribuzione statistica:

- il grado di dispersione o variabilità dei dati, rispetto alla mediana e ai quartili;
- la simmetria;
- la presenza di **valori anomali**.

La preferenza attribuita a questo metodo rispetto agli altri metodi grafici deriva dal fatto che con gli altri metodi *l'identificazione di una osservazione anomala è basata soltanto sul nostro occhio, mentre nel caso dei diagrammi a scatola le osservazioni anomale sono definite su base statistica: questo può aiutarci a beccare anche quelle che altrimenti l'avrebbero passata liscia. Dunque i diagrammi a scatola combinano la ricerca visiva degli anomali con un po' di statistica* (ancora Geoff R. **Norman** e David L. **Streiner** a pag. 203).

Anche se la serie di concetti implicati è lunga, partendo dal centro nella figura successiva è semplice osservare che

1 - la linea interna alla **scatola** (***box***) rappresenta la **mediana**;

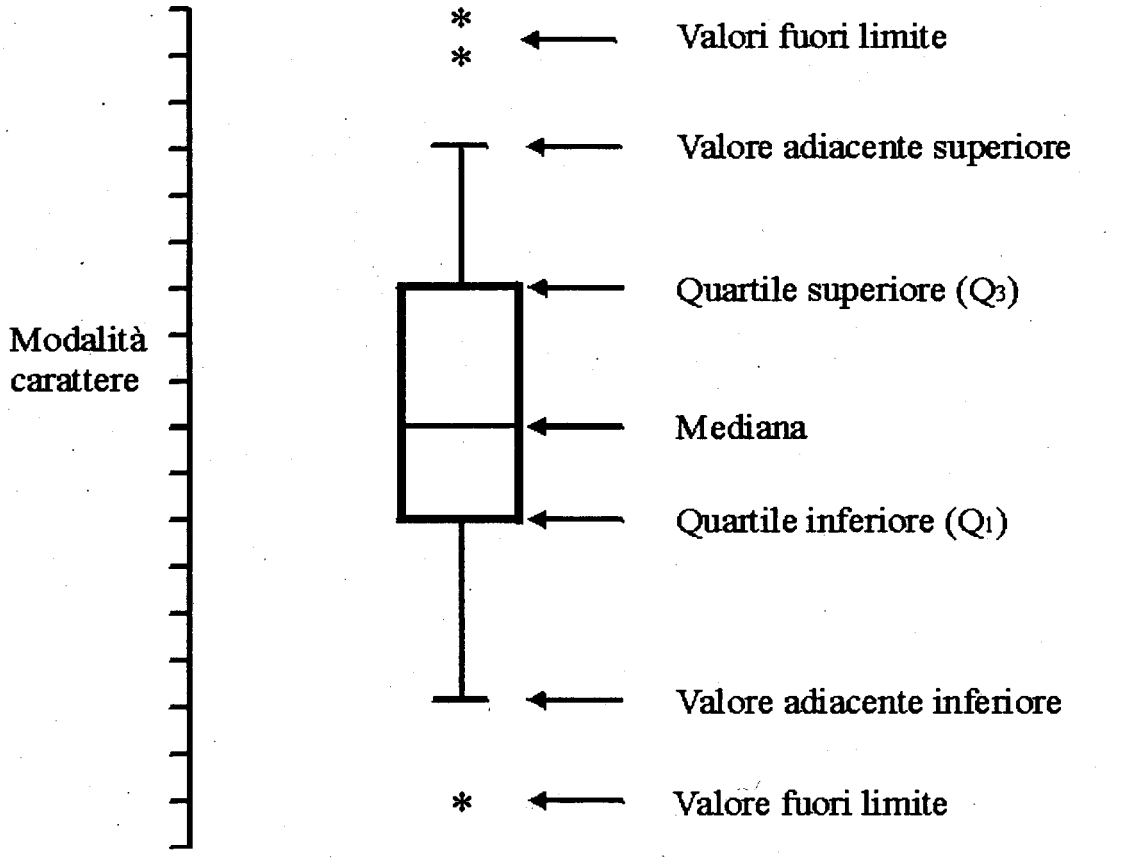
2 - mentre le **due linee orizzontali** rappresentano i bordi della scatola e identificano

- il **primo quartile** (**Q_1**, nella parte inferiore o ***Q lower***) e
- il **terzo quartile** (**Q_3**, nella parte superior o ***Q upper***);

di conseguenza, entro esse è compreso il 50 % delle osservazioni, quelle "più normali", collocate nella parte centrale della distribuzione.

3 - La distanza r tra i due quartili ($r = Q_3 - Q_1$), chiamata **distanza interquartile**, è assunta come misura di riferimento per valutare e testare la variabilità dei dati.

4 - A partire dai bordi della scatola, si allungano due linee verticali o **baffi** (***whiskers***) di lunghezza uguale (o minore) a **1,5** r (chiamato ***step*** da **Tukey**) e che terminano con un tratto orizzontale:
- la linea che parte dal **quartile inferiore Q_1** e si prolunga verso il basso è il **Valore Adiacente Inferiore** (in italiano **VAI;** in inglese ***lower fence***);
- quella che parte dal **quartile superiore Q_3** e si prolunga verso l'alto è il **Valore Adiacente Superiore** (in italiano **VAS**, in inglese ***upper fence***).



5 - In particolare quando le osservazioni sono poche, i due valori limiti che distano dal quartile **1,5** r **(uno *step*)**, non coincidono con nessun dato; di conseguenza, come nel caso della figura riportata, al loro posto per ognuno dei due baffi è stato preso quello più vicino, che è compreso nella parte interna della distribuzione.

6 - Il valore adiacente inferiore **VAI è il più piccolo dei valori osservati che risultano maggiori o uguali al limite prima prefissato;**

7 - mentre il valore adiacente superiore **VAS è il più grande dei valori osservati che risultano minori o uguali al limite prima prefissato.**

8 - Ne deriva che le due linee verticali, come nel grafico, quando sono calcolate su dati reali possono avere **lunghezza inferiore a 1,5** r e **diversa** per i due baffi; essi individuano un confine detto **cinta interna** (***inner fence***).

9 - I **dati esterni a questa cinta** sono rappresentati **individualmente** (perché quasi sempre sono pochi) e negli output informatici di norma sono indicati con un asterisco (come nella figura riportata): sono gli **outlier** o **valori anomali** detti anche **valori fuori limite**.

10 – Seppure senza la presenza di alcun tratto grafico e quindi **non visibile**,
- nel disegno è presente un **ulteriore confine,** più esterno,
- detto **cinta esterna** (***outer fence***) e che dista **2 step (3** r **)** dal quartile di riferimento:

esso delimita un **ultimo limite,** oltre il quale **gli outlier** da **anomali** diventano **estremamente anomali**;

Molti pacchetti informatici che disegnano il **box plot**, **distinguono queste due categorie di outlier** utilizzando **simboli differenti** per i **punti vicini** e **quelli lontani**. Ad esempio, nelle versioni in commercio nell'anno 2000,
- SPSS/PC usa **O** per gli **outlier** vicini e **E** per gli outlier **estremi,**
- Minitab usa un *** (asterisco)** per gli outlier **vicini** e una **O** per gli **outlier** estremi.

Spesso il disegno è riportato in modo verticale, come nella figura precedente. I concetti sono del tutto identici quando la rappresentazione è orizzontale; ma a molti il primo modo appare di più facile lettura, in particolare quando si confrontano due distribuzioni.

Da questa presentazione del **box-whisker plot**, si evidenzia che le **due misure fondamentali** che permettono di **individuare singolarmente** tutti gli **outlier** di una distribuzione di dati sono

- la **distanza interquartile (** $r = Q_3 - Q_1$ **)**
- e la scelta della **lunghezza dei due tratti, che come valore unitario ha 1,5** r **.**

Ma mentre il valore $r = Q_3 - Q_1$ è **oggettivo** e dipende dalla dispersione dei dati, **il coefficiente 1,5 è soggettivo**. Alla domanda perché avesse fissato una distanza di **1,5** r e non un altro coefficiente**, John Tukey** rispose:**" Perché 1 è poco e 2 è troppo".**

Si ritorna quindi alla **soggettività nei criteri di individuazione degli outlier**, anche se fondati sulla esperienza e competenza di uno statistico come John **Tukey** (1915 - 2000), nato nel Massachusettes, fondatore del Dipartimento di Statistica di Princeton, con contributi fondamentali sulle time series, il concetto di robustezza di un test, l'analisi della varianza, i confronti multipli e inventore della **exploratory data analysis**, entro la quale è da collocare questo metodo grafico.

A questo grafico è imputato il limite di **non essere fondato direttamente sulla stima di una probabilità precisa per ogni valore**.

E' tuttavia importante sottolinearne il concetto di base: se **i dati sono distribuiti in modo normale**, approssimativamente perché quasi sempre sono pochi,

- **oltre la cinta interna cade complessivamente il 5%** delle osservazioni,
- **oltre la cinta esterna l'1%** delle osservazioni più estreme.

ESEMPIO 1. Si supponga di avere raccolto la seguente serie di 20 valori, in una scala di rapporti o ad intervalli:

61, 69, 28, 51, 112, 80, 73, 103, 40, 47, 58, 58, 74, 56, 64, 68, 56, 54, 63, 59

Organizzare i valori in modo da costruire il **box-plot** e individuare gli eventuali **outlier**.

Risposta. Dopo aver ordinato le osservazioni per rango,

28, 40, 47, 51, 54, 56, 56, 58, 58, 60, 60, 63, 64, 68, 69, 73, 74, 80, 103, 112

con 20 dati

- la **mediana** è tra i valori di rango 10 (60) e di rango 11 (60): **mediana = 60**
- il **quartile inferiore** cade tra i valori di rango 5 (54) e rango 6 (56): $\mathbf{Q_1 = 55}$
- il **quartile superiore** cade tra i valori di rango 15 (69) e rango 16 (73): $\mathbf{Q_3 = 71}$
- la **distanza interquartile** è $r = Q_3 - Q_1$**:** r **= 71 – 55 = 16**

- il **valore adiacente inferiore** (VAI) è il più piccolo dei valori osservati che risultano maggiori o uguale al limite ottenuto con

$$Q_1 - 1{,}5r = 55 - (1{,}5 \cdot 16) = 55 - 24 = 31$$

ma poiché nessun valore è uguale a 31 e quello minore tra i maggiori di 31 è 40: VAI = 40

- **il valore adiacente superiore** (VAS) è il più grande dei valori osservati che risultano minori o uguali al limite ottenuto con

$$Q_3 + 1{,}5r = 71 + (1{,}5 \cdot 16) = 71 + 24 = 95$$

ma poiché nessun valore è uguale a 31 e quello maggiore tra i minori di 95 è 80: **VAS = 80**

Da queste due ultime indicazioni si ricavano gli **outlier**:

- nella parte inferiore è 28, in quanto inferiore a 40;
- nella parte superiore sono outlier 103 e 112, in quanto superiori a 80

Se si rappresentano graficamente questi dati, si ottiene una figura molto simile a quella riportata in precedenza.

Non esiste alcun outlier estremo. Infatti

- per il valore minimo la cinta esterna è

$$Q_1 - 3r = 55 - (3 \cdot 16) = 55 - 48 = 7$$

uguale a 7, mentre il minimo del campione è 28

- per il valore massimo la cinta esterna è

$$Q_3 + 3r = 71 + (3 \cdot 16) = 71 + 48 = 119$$

uguale a 119, mentre il massimo del campione è 112.

13.11. METODI STATISTICI PER GRANDI CAMPIONI: LA DISTRIBUZIONE DI CHEBYSHEV E LA DISTRIBUZIONE NORMALE; THE HUGE RULE

Quando il **campione è grande**,

- se la forma della **distribuzione non è nota** e ancor più se è certo che **non è normale**, un metodo statistico è la **disuguaglianza di Chebyshev** (scritto in inglese; in francese è **Cebicev;** in tedesco è **Tchebysheff;** il cognome reale è in cirillico, trattandosi di un russo);
- se la distribuzione della popolazione dei dati è normale, almeno in modo approssimato, si utilizza la distribuzione Z.

Il metodo statistico più generale per stimare la probabilità di appartenenza di un dato a una popolazione, che non richiede alcuna conoscenza o ipotesi sulla forma della distribuzione dei dati, è l'uso della **disuguaglianza di Chebyshev** (già presentata nel capitolo sulle distribuzioni teoriche).
Indicando con

- **k** il numero di deviazioni standard (σ) che separano il valore (**X**) dalla media (μ) della popolazione,

$$k = \frac{X - \mu}{\sigma}$$

si ricava la percentuale di osservazioni (P) che cadono tra l'osservazione e **k** deviazioni standard dalla media attraverso la relazione

$$P > \left(1 - \frac{1}{k^2}\right) \cdot 100$$

dove **k** deve essere superiore a 1.

Ad esempio,

- entro un intervallo compreso tra due deviazioni standard (K = 2) dalla media

$$P > \left(1 - \frac{1}{2^2}\right) \cdot 100 = 75$$

è compreso almeno il 75% dei dati;

- entro un intervallo compreso tra quattro deviazioni standard (K = 4) sopra e sotto la media

$$P > \left(1 - \frac{1}{4^2}\right) \cdot 100 = (1 - 0{,}0625) \cdot 100 = 93{,}75$$

è compreso almeno il 93,75% dei dati.

Quindi, da quei due limiti verso i valori più estremi, è compreso meno del 6,25% dei dati.
Se un dato dista 6,3 deviazioni standard dalla media, ha una probabilità

$$P < \frac{1}{6{,}3^2} \cdot 100 = \frac{100}{39{,}69} = 2{,}52$$

P < 0,0252 o 2,52% di appartenere alla stessa popolazione, in quanto collocato oltre quegli estremi.

Tale legge è utile quando la **forma della distribuzione** dei dati **è ignota.** Quindi può essere necessaria quando si ritiene che i dati abbiano una asimmetria fortissima. La **disuguaglianza di Chebyshev**

- offre il **vantaggio** di essere applicata a **qualsiasi distribuzione** di dati,
- ma è **molto meno potente dei metodi che ricorrono alla distribuzione normale Z**
- oppure a distribuzioni che la presuppongono almeno approssimativamente, come la **t di Student**.

Se la distribuzione è normale almeno in modo approssimato, si utilizza la distribuzione Z mediante

$$Z = \frac{X - \mu}{\sigma}$$

in un test che può essere

- sia **unilaterale**, quando a priori è nota la coda nella quale si può trovare l'**outlier**,
- sia **bilaterale**, quando a priori questa informazione non è disponibile.

I valori critici si uso più comune alle varie probabilità sono

α	0.05	0.01	0.005	0.001
Z unilaterale	1,645	2,33	2,58	3,09
Z bilaterale	1,96	2,58	2,81	3,28

Il concetto fondamentale è lo scarto di un dato ipotizzato outlier dalla media, in rapporto alla deviazione standard.

Con una distribuzione di dati campionari, nella quale si presume siano presenti outlier, la formula diventa

$$Z = \frac{X - \overline{X}}{S}$$

dopo aver calcolato la media $\overline{X}$ e la deviazione standard S, usando tutti gli n dati, **compreso quello sottoposto a verifica per essere giudicato outlier**.

Se il test è bilaterale, si calcola Z utilizzando come valore X il dato che è più distante dalla media, qualunque sia la coda in cui è collocato.

Se il test è unilaterale, si utilizza il valore più estremo nella coda prescelta.

Con quale valore di Z si può affermare che il dato è un outlier?

La soglia tra che cosa è atteso e che cosa è anomalo è certo arbitraria, ma generalmente quando si trova sopra le +3,00 deviazioni standard o sotto le –3,00 deviazioni standard di distanza dalla media va guardato quantomeno con sospetto (ancora Geoff R. **Norman** e David L. **Streiner** a pag. 203).

Se si decide che è un outlier,

- si elimina il dato,

- e si effettua una seconda analisi con i **rimanenti** $n-1$ **dati**, verificando se il **nuovo dato più distante dalla media** è anch'esso un **outlier**, nel suo **nuovo contesto**.

A questo scopo, con i rimanenti $n-1$ dati si calcolano nuovamente la media e la deviazione standard, che ovviamente risultano leggermente modificati, rispetto alle precedenti. Ne consegue che potrebbe diventare outlier un valore che prima non lo era.

E' possibile ripetere l'operazione più volte, eliminando ogni volta l'osservazione ritenuta anomala, **finché il valore Z risulta inferiore al limite prestabilito**. Da quel momento, **nessun altro dato sarà anomalo**. E' un **principio di cautela**, anche se effettivamente, modificando appunto media e varianza, potrebbero comparirne altri, nei gruppo di dimensioni minori.

Proposto in vari manuali di statistica applicata e utilizzato da molti ricercatori, perché semplice concettualmente e rapido, più recentemente **questo uso della Z** allo scopo di evidenziare la presenza di **uno o più outlier è criticato**, in quanto può condurre a **conclusioni errate più facilmente di altri metodi**.

Nel 1988 R. **Shiffler** con l'articolo ***Maximum z scores and outliers*** (pubblicato su **American Statistician** Vol. 42, pp. 79 –80) dimostra che

- **il valore massimo assoluto di** z **dipende da** n, le dimensioni del campione;

- e che questo **limite massimo è**

$$\frac{n-1}{\sqrt{n}}$$

Di conseguenza, ad esempio

- con n **= 10**

$$\frac{10-1}{\sqrt{10}} = \frac{9}{3{,}162} = 2{,}846$$

il **limite massimo di z è 2,846**

- con n **= 20**

$$\frac{20-1}{\sqrt{20}} = \frac{19}{4{,}472} = 4{,}289$$

il **limite massimo di z sale a 4,289**

- con n = 100

$$\frac{100-1}{\sqrt{100}} = \frac{99}{10} = 9{,}9$$

il **limite massimo di z diventa 9,9**.

Più grande è il campione, più alto è il limite del valore di z che un outlier può raggiungere.
Quindi con un campione piccolo, è più facile dichiarare che un valore X non è un outlier, con la giustificazione che lo Z calcolato è piccolo. Ma era basso anche il suo limite massimo possibile.
In realtà poteva essere un outlier, se fossero state considerate anche le dimensioni del campione.

Tra le numerose proposte per evidenziare la presenza di un outlier, sui manuali specialistici si è affermata anche un'altra strategia,

- la ***the Huge Rule***, riportata nel testo di L. A. **Marascuilo** del 1971 ***Statistical Methods for Behavioral Science Research*** (edito da McGraw-Hill, New York, 578 p. a pag. 199), che **utilizza un metodo del tutto analogo alla distribuzione Z,** ma si differenzia dalla metodologia precedente per tre aspetti:

- il valore, chiamato **M,** è fondato su uno scarto preso in valore assoluto (il test è sempre bilaterale)

$$M = \frac{\left|X_i - \overline{X}\right|}{S}$$

- la media $\overline{X}$ e la deviazione standard S sono calcolati su $n-1$ dati, **escludendo quello giudicato outlier,**
- il valore di **M deve essere maggiore di 4 (M > 4)**.

E una regola empirica, che utilizza **un valore "enorme"** al quale corrisponde una probabilità molto piccola, più esattamente P < 0.00005, se la distribuzione della popolazione dei dati dalla quale è stato estratto il campione è perfettamente normale.
Il **limite maggiore** di questo metodo è di **non stimare una probabilità precisa** per l'outlier analizzato.

Per meglio comprendere questa proposta e deciderne correttamente l'eventuale applicazione al proprio settore di ricerca, è importante ricordare che la **regola empirica** (***rule of thumb***) sulla quale è fondata (the ***huge rule***) è riportata da **Marascuilo** tra i metodi adatti alle **scienze del comportamento**, caratterizzate sempre da una **variabilità estremamente grande**.

Ma essa è utilizzata anche 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. 533) dal quale è tratto l'esempio successivo.

ESEMPIO. (THE HUGE RULE). Considerando le seguenti 15 osservazioni

99,3	**99,7**	**98,6**	**99,0**	**99,1**	**99,3**	**99,5**	**98,0**	**98,9**	**99,4**	**99,0**	**99,4**	**99,2**	**98,8**	**99,2**

il valore 98,0 può essere considerato un outlier?

Risposta. **<u>Escludendo</u> il potenziale outlier** (98,0), con gli altri n = 14 dati

- si ottiene $\overline{X}$ = 99,17 e S = 0,29;

da essi con

$$M = \frac{|X_i - \overline{X}|}{S} = \frac{|99{,}17 - 98{,}0|}{0{,}29} = \frac{1.17}{0{,}29} = 4{,}03$$

si ricava M = 4,03.

Ne deriva che, sulla base del ***Huge Role***, il valore 98,0 **è un outlier,** nel contesto degli altri 14 valori .

Le procedure illustrate in precedenza con l'uso della distribuzione Z rimangono invariate, quando per la stima della probabilità si utilizza la **disuguaglianza di Chebyshev.** Sono differenti la stima della probabilità e le assunzioni sulla forma della distribuzione dei dati.

Confrontando i risultati ottenuti con la distribuzione normale e quelli con la diseguaglianza di **Chebyshev,** nel caso di un test bilaterale

- con almeno due deviazioni standard dalla media (Z = 2) un dato ha una probabilità P < 4,6% di appartenere alla popolazione, mentre con la **disuguaglianza di Chebyshev** è P < 25% (come risulta nei calcioli precedenti);
- con 4 deviazioni standard (Z = 4) tale probabilità scende a circa 2 su diecimila, mentre con **Chebyshev** la stima era P < 6,25%**.**

La stima fornita dalla distribuzione di **Chebyshev** certamente ha il vantaggio rilevante di **non richiedere la normalità della distribuzione** dei dati e quindi di poter essere teoricamente **utilizzata in una varietà di situazioni molto più ampia**, senza **essere mai sospettata di invalidità**.

Ma, nella pratica sperimentale, in molte situazioni fornisce risposte troppo generiche. E' troppo cautelativa, per essere realmente utile.

Per scoprire **outlier univariati**, vale a dire sempre quando si utilizza **una sola variabile** indipendente, in letteratura sono proposti anche

- il **test di Grubb** (***Grubbs test for Outlying Observations***) e il **test Q di Dixon** (***Dixon Q test*****),** che sono illustrati nel paragrafo successivo;
- il **Youden's test for outliers** e il **Cochran's test for extreme values of variance**, utilizzati per misure chimiche e rintracciabili nel testo di J. K. **Taylor** del 1987 ***Quality Assurance of Chemical Measures*** (edito da Lewis Publishers, Chelsea)
- il metodo **studentized deleted residuals**, descritto nel testo di R. L. **Mason**, R. F. **Gunst**, J. L: **Hess** del 1989 ***Statistical Design and Analysis of Experiments*** (editi da John Wiley and Sons, New York)

13.12. VERIFICA DEGLI OUTLIER O GROSS ERROR PER CAMPIONI PICCOLI CON DISTRIBUZIONE NORMALE: IL TEST DI GRUBBS O EXTREME STUDENTIZED RESIDUAL; IL TEST Q DI DIXON.

Sempre nel casi di dati

- che siano stati estratti da una **popolazione distribuita in modo normale**,
- ma con **campioni piccoli** ($n < 25 - 30$), anche se (**come il test t di Student**) il metodo è ugualmente applicabile a campioni grandi,
- e per una **individuazione degli outlier** fondata su una **probabilità definita**,

è possibile utilizzare la **procedura T** o **metodo T** (***T procedure*** or ***T method***) **di Grubbs** proposto appunto da F. E. **Grubbs** nel 1969 con l'articolo ***Procedures for detecting outlying observations in samples*** (pubblicato su **Technometrics** Vol. 11, pp. 1 – 21).

In alcuni testi questo metodo è chiamato anche ***extreme Studentized residual*** o, più frequentemente, ***extreme Studentized deviate*** (**ESD**). Rappresenta una evoluzione del test di Grubbs e può essere applicata sia alla ricerca di un solo outlier sia alla ricerca di più outlier, con modifiche lievi.

Secondo la presentazione di 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. 533), da cui è tratto l'esempio successivo,

la **procedura di Grubbs** richiede

- di ordinare per ranghi i dati del campione
- e, in rapporto al fatto che il dato ritenuto anomalo sia il primo oppure l'ultimo,
- di calcolare il valore T

con

$$T = \frac{\overline{X} - X_1}{S} \qquad \text{oppure} \qquad T = \frac{X_n - \overline{X}}{S}$$

Il risultato deve essere confrontato con i valori critici riportati nella pagina seguente, validi per un **test unilaterale**, vale a dire quando a priori è nota la cosa nella quale è collocato il potenziale outlier.

I valori della tabella, la cui versione completa è rintracciabile nelle tavole statistiche di Robert R. **Sokal** e F. James **Rohlf** del 1995 (3rd ed. W. H. Freeman and Company, New York, XIV + 199 p.), rappresentano una elaborazione di quanto pubblicato da F. E. **Grubbs** e G. **Beck** nel 1972 con ***Extension of Sample Size and Percentage Points for Significance Tests of Outlying Observations*** (su **Technometrics** Vol. 14, pp. 847 – 854).

ESEMPIO 1 (MODIFICATO DA DE MUTH; STESSI DATI DI THE HUGE RULE). Considerando le seguenti 15 osservazioni

99,3	**99,7**	**98,6**	**99,0**	**99,1**	**99,3**	**99,5**	**98,0**	**98,9**	**99,4**	**99,0**	**99,4**	**99,2**	**98,8**	**99,2**

il valore **98,0** può essere considerato un outlier?

Risposta. **<u>Comprendendo</u> il potenziale outlier** (98,0), con tutti gli n = 15 dati

- si ottiene $\overline{X}$ = 99,09 e S = 0,41;

Poiché **98,0 è il valore minore** della serie,

con

$$T = \frac{\overline{X} - X_1}{S} = \frac{99{,}09 - 98{,}00}{0{,}41} = \frac{1{,}09}{0{,}41} = 2{,}66$$

si ottiene T = 2,66.

Nella tabella dei valori critici, per n = 15 il valore calcolato (T = 2,66) risulta

- **maggiore** di quello critico (T = 2,549) alla **probabilità α = 0.025**
- **minore** di quello critico (T = 2,705) alla **probabilità α = 0.01**

Valori critici per il test di Grubbs (test unilaterale)

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$	n
3	1.148	1.153	1.155	1.155	1.155	3
4	1.425	1.463	1.481	1.492	1.496	4
5	1.602	1.672	1.715	1.749	1.764	5
6	1.729	1.822	1.887	1.944	1.973	6
7	1.828	1.938	2.020	2.097	2.139	7
8	1.909	2.032	2.126	2.221	2.274	8
9	1.977	2.110	2.215	2.323	2.387	9
10	2.036	2.176	2.290	2.410	2.482	10
11	2.088	2.234	2.355	2.485	2.564	11
12	2.134	2.285	2.412	2.550	2.636	12
13	2.175	2.331	2.462	2.607	2.699	13
14	2.213	2.371	2.507	2.659	2.755	14
15	2.247	2.409	2.549	2.705	2.806	15
16	2.279	2.443	2.585	2.747	2.852	16
17	2.309	2.475	2.620	2.785	2.894	17
18	2.335	2.504	2.651	2.821	2.932	18
19	2.361	2.532	2.681	2.854	2.968	19
20	2.385	2.557	2.709	2.884	3.001	20
21	2.408	2.580	2.733	2.912	3.051	21
22	2.429	2.603	2.758	2.939	3.060	22
23	2.448	2.624	2.781	2.963	3.087	23
24	2.467	2.644	2.802	2.987	3.112	24
25	2.486	2.663	2.822	3.009	3.135	25
26	2.502	2.681	2.841	3.029	3.157	26
27	2.519	2.698	2.859	3.049	3.178	27
28	2.534	2.714	2.876	3.068	3.199	28
29	2.549	2.730	2.893	3.085	3.218	29
30	2.563	2.745	2.908	3.103	3.236	30
31	2.577	2.759	2.924	3.119	3.253	31
32	2.591	2.773	2.938	3.135	3.270	32
33	2.604	2.786	2.592	3.150	3.286	33
34	2.616	2.799	2.965	3.164	3.301	34
35	2.628	2.811	2.979	3.178	3.316	35
36	2.639	2.823	2.991	3.191	3.330	36
37	2.650	2.835	3.003	3.204	3.343	37
38	2.661	2.846	3.014	3.216	3.356	38
39	2.671	2.857	3.025	3.228	3.369	39
40	2.682	2.866	3.036	3.240	3.381	40
50	2.768	2.956	3.128	3.336	3.483	50
60	2.837	3.025	3.199	3.411	3.560	60
70	2.893	3.082	3.257	3.471	3.622	70
80	2.940	3.130	3.305	3.521	3.673	80
90	2.981	3.171	3.347	3.563	3.716	90
100	3.017	3.207	3.383	3.600	3.754	100

Di conseguenza, con probabilità di errare P < 0.025, si può affermare che il valore 98,0 è un outlier, rispetto al gruppo complessivo di osservazioni.
E' la stessa conclusione alla quale si era pervenuti con il metodo Huge Rule. Ma ora è stata stimata una probabilità abbastanza precisa di commettere un errore di Tipo I.

Il confronto tra i due risultati mostra che, con campioni piccoli, **il valore M > 4 del metodo Huge Rule non è un risultato così estremo**. In questo caso, corrisponde a una probabilità minore di 0,025 ma maggiore di 0.01.

Il test di Grubbs può essere utilizzato anche quando si sospetta che vi sia più di un outlier.

Su testi o manuali di statistica applicata differenti, il test **Q di Dixon** è presentato con due modalità **apparentemente diverse**. Esse fanno riferimento all'articolo del 1951 oppure a quello del 1953 di W. J. **Dixon** su gli outlier. Di seguito, sono presentati entrambi i metodi, ricordando che
a) **il primo è per un test bilaterale**,
b) **il secondo per un test unilaterale**

A) Il manuale della Marina Militare Americana del 1960 (***Statistical Manual*** by Edwin L. **Crow**, Frances A. **Davis**, Margaret W. **Maxfield**, Research Department U. S: Naval Ordnance Test Station, Dover Pubblications, Inc., New York, XVII + 288 p.) per la identificazione degli outlier consiglia il ***test ratios for gross errors*** proposto da W. J. **Dixon** nel 1951 nell'articolo ***Ratios involving extreme values*** (pubblicato su **Annals of Mathematichal Statistics**, Vol. 22, pp. 68 – 78).
Il test
- che verifica l'ipotesi nulla $\mathbf{H_0}$ che il dato sospettato appartenga alla **stessa popolazione degli altri dati**, contro l'ipotesi alternativa $\mathbf{H_1}$ che provenga da una popolazione differente,
- sembra limitato a un **solo outlier**, anche se tale limite non è espressamente dichiarato,
- **è bilaterale** e quindi a priori non è determinata la coda nella quale occorre verificare l'esistenza dell'outlier,
- può essere applicato a un **campione piccolo** ($n \leq 30$),
- tratto da una **popolazione distribuita in modo normale**.

La procedura è molto semplice e rapida, **non richiedendo né il calcolo della media** $\overline{X}$ **né quello della deviazione standard** S **campionarie.**

Test ratios for gross errors

Valori critici del test di Dixon (1951)

Quando è ignota la coda, prima di osservare i valori (test unilaterale)

Rapporti	n	Valori critici		
		$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$r_{10} = \frac{X_2 - X_1}{X_n - X_1}$	3	0.941	0.970	0.994
	4	0.765	0.829	0.926
	5	0.642	0.710	0.821
	6	0.560	0.628	0.740
	7	0.507	0.569	0.680
$r_{11} = \frac{X_2 - X_1}{X_{n-1} - X_1}$	8	0.544	0.608	0.717
	9	0.503	0.564	0.672
	10	0.470	0.530	0.635
	11	0.445	0.502	0.605
	12	0.423	0.479	0.579
$r_{22} = \frac{X_3 - X_1}{X_{n-2} - X_1}$	13	0.563	0.611	0.697
	14	0.539	0.586	0.670
	15	0.518	0.565	0.647
	16	0.500	0.546	0.627
	17	0.483	0.529	0.610
	18	0.469	0.514	0.594
	19	0.457	0.501	0.580
	20	0.446	0.489	0.567
	21	0.435	0.478	0.555
	22	0.426	0.468	0.544
	23	0.418	0.459	0.535
	24	0.410	0.451	0.526
	25	0.402	0.443	0.517
	26	0.396	0.436	0.510
	27	0.389	0.429	0.502
	28	0.383	0.423	0.495
	29	0.378	0.417	0.489
	30	0.373	0.412	0.483

Disponendo di una serie di dati,

- prima i valori devono essere **ordinati per rango**,

- in modo crescente oppure decrescente, in funzione della **coda nella quale è collocato il dato, ma individuata dalla lettura dei dati e non specificata in anticipo**;

- successivamente, utilizzando solo i valori estremi, la scelta dei quali dipende anche dalle dimensioni del campione, si calcola un rapporto r.

La scelta dei dati per calcolare il **rapporto** r dipende dal numero n di dati del campione:

- per **campioni molto piccoli**, fino a $n = 7$ dati,

è

$$r_{10} = \frac{X_2 - X_1}{X_n - X_1}$$

- per **campioni intermedi** da $n = 8$ e fino a $n = 12$ dati,

è

$$r_{11} = \frac{X_2 - X_1}{X_{n-1} - X_1}$$

- per **campioni maggiori**, da $n = 13$ e fino a $n = 30$ dati,

è

$$r_{22} = \frac{X_3 - X_1}{X_{n-2} - X_1}$$

I valori critici sono riportati nella **tabella precedente**.

ESEMPIO 2. (TRATTO DAL TESTO **STATISTICAL MANUAL** CITATO). Da una distribuzione normale, sono stati estratti i sei valori seguenti

0,505	0,511	0,519	0,478	0,357	0,506

Dalla lettura dei dati risulta che valore 0,357 è nettamente minore degli altri.

Può essere considerato un outlier?

Risposta. Da come è stata impostata la domanda si deduce che **il test è bilaterale**.
Dopo avere ordinato tutti i dati del campione in modo crescente

X_1	X_2	X_3	X_4	X_5	X_6
0,357	0,478	0,505	0,506	0,511	0,519

poiché il valore sospettato è il minore di un gruppo con $n = 6$
mediante

$$r_{10} = \frac{X_2 - X_1}{X_6 - X_1} = \frac{0{,}478 - 0{,}357}{0{,}519 - 0{,}357} = \frac{0{,}121}{0{,}162} = 0{,}747$$

si ottiene il rapporto $r_{10} = 0{,}747$.

Nella tabella dei valori critici, con $n = 6$ il valore per $\alpha = 0.01$ (la probabilità minore riportata) è uguale a 0,740.
Di conseguenza, poiché il valore calcolato è maggiore, si rifiuta l'ipotesi H_0 con probabilità $P < 0.01$.

B) Il testo di James E. De **Muth** del 1999 ***Basic Statistics and Pharmaceutical Statistical Applications*** (edito da Marcel Dekker, Inc. New York, XXI + 596 p. a pag. 534-536) presenta anch'esso un **test Q di Dixon**, ma rifacendosi all'articolo di W. J. **Dixon** del 1953 ***Processing data for outliers*** (pubblicato su **Biometrics** Vol. 1, pp. 74 - 89) e nella versione di **test unilaterale.**

Il test
- verifica l'ipotesi nulla **H_0** che il dato sospettato appartenga alla **stessa popolazione degli altri dati**, contro l'ipotesi alternativa **H_1** che provenga da una popolazione differente,
- quando **a priori è noto in quale coda** della distribuzione si dovrà verificare l'esistenza dell'**outlier**,
- disponendo di un **campione piccolo** ($n \leq 30$),
- tratto da una **popolazione distribuita in modo normale**.

Come nella versione precedente, **non si deve calcolare né la media $\overline{X}$ la deviazione standard S** del campione. Quindi rispetto ad altri metodi, il **Q di Dixon offriva un vantaggio pratico rilevante**, quando tutti i calcoli dovevano essere svolti manualmente.
Nonostante questo non è possibile affermare che sia un test non-parametrico, in quanto ipotizza che i dati siano distribuiti in modo normale.
La procedura è identica a quella precedente, per quanto riguarda l'ordinamento dei dati.

La scelta dei dati per calcolare il **rapporto** τ dipende

- dal numero n di dati del campione
- e **dalla coda** nella quale a priori si è ipotizzato che il valore più estremo sia un outlier.

Nella serie successive di formule, tra ogni coppia si sceglie la prima oppure la seconda formula, se nella serie dei valori ordinati in modo crescente l'outlier è il primo oppure l'ultimo dato:

- per **campioni molto piccoli**, da n = 3 e fino a n = 7 dati,

si utilizza

$$\tau_{10} = \frac{X_2 - X_1}{X_n - X_1} \qquad \text{oppure} \qquad \tau_{10} = \frac{X_n - X_{n-1}}{X_n - X_1}$$

- per **campioni** da n = 8 e fino a n = 10 dati,

si utilizza

$$\tau_{11} = \frac{X_2 - X_1}{X_{n-1} - X_1} \qquad \text{oppure} \qquad \tau_{11} = \frac{X_n - X_{n-1}}{X_n - X_2}$$

- per **campioni** da n = 11 e fino a n = 13 dati,

si utilizza

$$\tau_{21} = \frac{X_3 - X_1}{X_{n-1} - X_1} \qquad \text{oppure} \qquad \tau_{21} = \frac{X_n - X_{n-2}}{X_n - X_2}$$

- per **campioni maggiori**, da n = 14 e fino a n = 30 dati,

si utilizza

$$\tau_{22} = \frac{X_3 - X_1}{X_{n-2} - X_1} \qquad \text{oppure} \qquad \tau_{22} = \frac{X_n - X_{n-2}}{X_n - X_3}$$

I valori critici sono riportati nella tabella successiva

ESEMPIO 3 (MODIFICATO DA DE MUTH; STESSI DATI DI THE HUGE RULE E DEL TEST DI GRUBBS DELL'ESEMPIO 1). Considerando le seguenti 15 osservazioni

99,3	**99,7**	**98,6**	**99,0**	**99,1**	**99,3**	**99,5**	**98,0**	**98,9**	**99,4**	**99,0**	**99,4**	**99,2**	**98,8**	**99,2**

il valore minore 98,0 può essere considerato un outlier?

Test Q di Dixon (1953) per gli outlier

Se è sospetto **Primo**	n	Valori critici $\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$	n	Se è sospetto **Ultimo**
$\tau_{10} = \frac{X_2 - X_1}{X_n - X_1}$	3	0.941	0.988	0.994	3	$\tau_{10} = \frac{X_n - X_{n-1}}{X_n - X_1}$
	4	0.765	0.889	0.926	4	
	5	0.642	0.780	0.821	5	
	6	0.560	0.698	0.740	6	
	7	0.507	0.637	0.680	7	
$\tau_{11} = \frac{X_2 - X_1}{X_{n-1} - X_1}$	8	0.554	0.683	0.725	8	$\tau_{11} = \frac{X_n - X_{n-1}}{X_n - X_2}$
	9	0.512	0.635	0.677	9	
	10	0.477	0.597	0.639	10	
$\tau_{21} = \frac{X_3 - X_1}{X_{n-1} - X_1}$	11	0.576	0.679	0.713	11	$\tau_{21} = \frac{X_n - X_{n-2}}{X_n - X_2}$
	12	0.546	0.642	0.675	12	
	13	0.521	0.615	0.649	13	
$\tau_{22} = \frac{X_3 - X_1}{X_{n-2} - X_1}$	14	0.546	0.641	0.674	14	$\tau_{22} = \frac{X_n - X_{n-2}}{X_n - X_3}$
	15	0.525	0.616	0.647	15	
	16	0.507	0.595	0.624	16	
	17	0.490	0.577	0.605	17	
	18	0.475	0.561	0.589	18	
	19	0.462	0.547	0.575	19	
	20	0.450	0.535	0.562	20	
	21	0.440	0.524	0.551	21	
	22	0.430	0.514	0.541	22	
	23	0.421	0.505	0.532	23	
	24	0.413	0.497	0.524	24	
	25	0.406	0.489	0.516	25	

Risposta. E' un test unilaterale, con il quale si verifica

- l'ipotesi nulla $\mathbf{H_0}$ che il dato minore appartenga alla **stessa popolazione degli altri dati**,
- contro l'ipotesi alternativa $\mathbf{H_1}$ che esso provenga da una popolazione differente.

A questo scopo, si ordinano i valori del campione attribuendo i ranghi:

98,0	**98,6**	**98,8**	**98,9**	**99,0**	**99,0**	**99,1**	**99,2**	**99,2**	**99,3**	**99,3**	**99,4**	**99,4**	**99,5**	**99,7**
X_1	X_2	X_3	X_4	---	---	---	---	---	---	---	X_{n-3}	X_{n-2}	X_{n-1}	X_n

Poiché n = 15 e l'eventuale outlier è il primo nella serie ordinata,

si utilizza la formula

$$\tau_{22} = \frac{X_3 - X_1}{X_{n-2} - X_1}$$

Con i dati dell'esempio

$$X_3 = \mathbf{98,8} \qquad X_1 = \mathbf{98,0} \qquad X_{n-2} = \mathbf{99,4}$$

si ottiene

$$\tau_{22} = \frac{98,8 - 98,0}{99,4 - 98,0} = \frac{0,8}{1,4} = 0,57$$

il risultato τ = 0,57.

Poiché nella tabella del **test Q di Dixon (1953) per gli outlier** sono riportati

- per α = 0.05 il valore critico τ = 0.525
- per α = 0.01 il valore critico τ = 0.616

si rifiuta l'ipotesi nulla con probabilità P > 0.05 di commettere un errore di Tipo I.

Dal confronto degli esempi 1 e 3 che sono stati applicati agli stessi dati, e dalle informazioni precedenti a conclusione si può dedurre che:

- il test di **Dixon è più semplice**, mentre il test di **Grubbs richiede più calcoli**;
- ma il test di **Grubbs è più potente**;
- inoltre il test di **Grubbs può essere ripetuto**, per **individuare più outlier** negli stessi dati.

Entrambi i test richiedono che **la distribuzione dei dati sia normale**.

Verificato che il valore sospettato è statisticamente un outlier, per eliminare il dato dalle analisi successive **il dubbio sulla correttezza del valore non può derivare solamente dalla semplice osservazione** che esso è nettamente minore o maggiore degli altri. Il dubbio deve essere giustificato esternamente all'analisi statistica, come possono essere la condizione particolare dell'esperimento con

cui quel dato è stato ottenuto, un errore strumentale, una trascrizione sbagliata del risultato reale. Condizioni che eventualmente possono essere verificate a posteriori, dopo il suggerimento della probabile rarità del dato sospettato fornita dal test.

13.13. LA EXTREME STUDENTIZED DEVIATE E LA MEDIAN ABSOLUTE DEVIATION.

I **metodi attualmente più diffusi,** per **identificare gli outliers** in un campione di dati, sono

A - un **metodo parametrico**, la **extreme Studentized deviate** (**ESD**) chiamata anche **extreme Studentized residuals**, che utilizza la **media** e della **deviazione standard** del campione,

B - un **metodo non parametrico**, **la median absolute deviation (MAD),** che utilizza la **mediana** e della **deviazione mediana assoluta**.

A - La procedura di **statistica parametrica** detta **Extreme Studentized Deviate** (acronomimo **ESD**) è un altro nome del **test di Grubbs,** già presentato.

Una dimostrazione elementare di tale corrispondenza è data sia dalle formule, sia dall'uguaglianza dei valori critici, anche se in questo caso il test spesso è presentato come bilaterale bilaterale, mentre in Grubbs era unilaterale. In questo paragrafo sono stati approfonditi i concetti precedenti e **il test è stato esteso al caso in cui nello stesso campione siano presenti più outlier**.

Il problema della identificazione degli outlier è teoricamente semplice. E' sufficiente rispondere alla domanda: "Quanto deve distare un valore per essere ritenuto **outlier** rispetto al **campione**?"

La risposta deve prendere in considerazione tre fattori:

- la distanza del dato i dalla media ($X_i - \overline{X}$),
- la deviazione standard del campione (S),
- il numero di dati del campione (n).

Per definizione, la **Extreme Studentized Deviate**

è

$$ESD = \max_{i=1,\ldots,n} \frac{\left|X_i - \overline{X}\right|}{S}$$

considerando che

- in un **campione di** n **dati**,
- nel quale **non siano presenti uno o più outlier**,
- **il valore massimo** approssimativamente deve corrispondere

al percentile

$$\frac{N}{N+1} \cdot 100$$

Ad esempio,

- in un campione di 60 **dati**

- estratti da una popolazione distribuita normalmente

- e quindi **senza outlier**,

- **il valore più alto** non dovrebbe distante dalla media più di quanto lo sia

all'incirca

$$\frac{60}{60+1}\cdot 100 = 98{,}36$$

il percentile 98,36.

Per usare la distribuzione normale, il percentile deve essere tradotto in unità di deviazioni standard σ dalla media:

- in una distribuzione normale **bilaterale** (vedi tavola della normale bilaterale),

- dal percentile 98,36 si ricava la probabilità dell'area sottesa (0,9836);

- ad essa nelle due code corrisponde la probabilità bilaterale P = 0,0167 (ricavato da **1 - 0,9836);**

- arrotondata a **0.017** determina il valore Z = 2,39.

Pertanto, per essere considerato outlier,

- **se è grande** deve un dato (X) essere

$$X > \overline{X} + Z \cdot \sigma$$

maggiore della media di almeno 2,39 volte la deviazione standard.

- mentre **se è piccolo** un dato (X) deve essere

$$X < \overline{X} - Z \cdot \sigma$$

minore della media di almeno 2,39 volte la deviazione standard

Sempre nell'ipotesi che i dati siano distribuiti in modo normale, per una cautela maggiore e per ottenere una stima più precisa con campioni piccoli, invece della distribuzione Z si può utilizzare la distribuzione **t di Student**, che tuttavia è specifica solo per la media.

Si ricorrere ai valori critici della tabella successiva (identica alle due precedenti), proposti nel 1961 da C. P. **Quesenberry** e H. A. **David** (nell'articolo ***Some tests for outliers*** su **Biometrika** Vol. 48, pp. 379-399) e successivamente modificati.

Le procedure per **identificare gli outlier** si differenziano sulla base del numero di outlier da verificare:

A_1 - un s**ingolo outlier,**

A_2 - **più outlier**.

VALORI CRITICI PER

L'**EXTREME STUDENTIZED DEVIATE** (ESD)

IN **OUTLIER** STATISTICI PER **TEST BILATERALI**

n	$\alpha = 0.05$	$\alpha = 0.01$		*n*	$\alpha = 0.05$	$\alpha = 0.01$
5	1,72	1,76		25	2,82	3,14
6	1,89	1,97		26	2,84	3,16
7	2,02	2,14		27	2,86	3,18
8	2,13	2,28		28	2,88	3,20
9	2,21	2,39		29	2,89	3,22
10	2,29	2,48		30	2,91	3,24
11	2,36	2,56		35	2,98	3,32
12	2,41	2,64		40	3,04	3,38
13	2,46	2,70		45	3,09	3,44
14	2,51	2,75		50	3,13	3,48
15	2,55	2,81		60	3,20	3,56
16	2,59	2,85		70	3,26	3,62
17	2,62	2,90		80	3,31	3,67
18	2,65	2,93		90	3,35	3,72
19	2,68	2,97		100	3,38	3,75
20	2,71	3,00		150	3,52	3,89
21	2,73	3,03		200	3,61	3,98
22	2,76	3,06		300	3,72	4,09
23	2,78	3,08		400	3,80	4,17
24	2,80	3,11		500	3,86	4,23

A_1 - Per **<u>un solo outlier</u>**, utilizzando tutto il campione e quindi comprendendo anche il valore sospettato (X), si calcola la media ($\overline{X}$) e la deviazione standard (S).
Successivamente, per verificare l'ipotesi

H_0: **non è presente alcun outlier**

contro l'ipotesi

H_1: **è presente un valore outlier**

si calcola il valore **ESD**, che deve essere confrontato con la tabella dei valori critici.

ESEMPIO 1. In un campione di 50 dati (n = 50), la media è $\overline{X}$ = 56,2 e la deviazione standard è risultata S = 12,3. Il dato più distante dalla media è X = 14,1. Può essere considerato un outlier?

Risposta. Per verificare l'ipotesi

H_0: **non è presente alcun outlier**

contro l'ipotesi

H_1: **è presente un valore outlier**

in un test bilaterale in quanto a priori non era nota in quale coda potesse trovarsi un outlier,
si calcola il valore di **Extreme Studentized Deviate**

$$ESD = \frac{|X - \overline{X}|}{S} = \frac{14{,}1 - 56{,}2}{12{,}3} = 3{,}42$$

che risulta ESD = 3,42.
Con N = 50, il valore critico riportato nella tabella
- alla probabilità $\alpha = 0.05$ è 3,13
- alla probabilità $\alpha = 0.01$ è 3,48.

Poiché il valore calcolato (3,42) è maggiore di 3,13 e minore di 3,48 si può affermare che il valore **X è un outlier**, con **probabilità** di sbagliare $P < 0.05$

La procedura illustrata nell'esempio, che non si discosta da quella dei paragrafi precedenti, è corretta quando si ipotizza la presenza di **un solo outlier**. Ma quando **gli outlier sono due o più**, la loro presenza amplia notevolmente il valore della deviazione standard S e quindi diventa poco probabile individuare anche un solo valore outlier, poiché con S grande si riduce il valore ESD calcolato.
Questo effetto degli outlier di nascondere la loro presenza è noto come ***masking problem.***

A_2 - In considerazione di questo problema e della probabilità implicata in confronti multipli, nel caso di **<u>più outlier</u>**, la procedura è più lunga:

- deve essere applicata quella precedente varie volte, quanti sono gli **outlier da verificare**,
- dopo aver prestabilito il loro **numero massimo** k .

Per definire questo numero k di potenziali outlier, il primo problema è che, in rapporto al numero totale di osservazioni del campione, **il numero massimo di outlier non deve essere troppo alto**, altrimenti si determinano due conseguenze indesiderate:

- la distribuzione si allontana eccessivamente dalla normalità, quindi il modello utilizzato non è più credibile,
- aumenta eccessivamente la varianza, generando il **masking effect** ricordato.

Una stima **giudicata ragionevole** (da esperti, ma sempre soggettiva) del **numero massimo** k **di outlier** in un campione di n dati deve rispettare due limiti:

1 – è $k = n/10$, arrotondando per difetto la parte intera, quando il campione è formato da poche decine;

ad esempio, con $n = 7$ sia avrà $k = 7/10 = 0{,}7$ quindi $k = 1$

2 – anche se il campione è grande**,** k **non deve mai superare 5**, a meno che il campione non sia molto grande, in questo caso superiore almeno a un centinaio di osservazioni;

ad esempio, con $n = 67$ sia avrà $k = 67/10 = 6{,}7$ quindi $k = 5$.

Come sempre, questi confini non sono definiti in modo preciso, essendo appunto fondati sul "buon senso statistico" o "esperienza statistica".

Nel caso di k **outlier** (con $k > 1$), il test serve per verificare

l'ipotesi nulla

H_0: **non è presente alcun outlier**

contro l'ipotesi alternativa

H_1: **sono presenti da** 1 **a** k **outlier**

La procedura statistica richiede vari passaggi logici, che per comodità didattica sono schematizzati in nove punti.

1 – Dopo aver prestabilito k ,

2 – sul campione totale di n dati, si calcolano la media $\overline{X}$ e la deviazione standard S .

Indicando con $X^{(n)}$ **il valore più distante dalla media** degli n dati, qualunque sia la coda in cui è collocato, si calcola la sua

Extreme Studentized Deviate con

$$ESD^{(n)} = \frac{\left| X^{(n)} - \overline{X}^{(n)} \right|}{S^{(n)}}$$

3 – Se si rifiuta l'ipotesi nulla, poiché il valore $ESD^{(n)}$ calcolato è significativo, dal campione complessivo di n dati, si toglie il valore che è risultato statisticamente un outlier; pertanto il campione diventa di dimensioni $n-1$.

4 - In questo campione successivo di $n-1$ dati, si calcolano nuovamente

- la media $\overline{X}^{(n-1)}$
- e la deviazione standard $S^{(n-1)}$.

Identificato il nuovo estremo $X^{(n-1)}$, cioè il valore più distante dalla media in uno dei due estremi della distribuzione, si calcola la sua

Extreme Studentized Deviate con

$$ESD^{(n-1)} = \frac{\left| X^{(n-1)} - \overline{X}^{(n-1)} \right|}{S^{(n-1)}}$$

5 – Se anche questo ESD risulta significativo, dopo aver tolto questo secondo valore, si continua la procedura fino all'ultimo outlier prefissato, che avrà un campione con $n-k+1$ dati**;**

pertanto si ottengono k valori ESD **(al massimo 5)**

che saranno

$$ESD^{(n)},\ ESD^{(n-1)},\ ESD^{(n-2)}, ESD^{(n-k+1)}, ESD^{(n-k+1)}$$

6 – Successivamente si confronta $ESD^{(n-k+1)}$, cioè **l'ultimo ESD calcolato**, con il suo ESD critico alla probabilità α prefissata e per il numero n di dati del campione.

Se l'ultimo ESD calcolato risulta significativo, tutti i valori testati **sono outlier**.

7 - Se invece questo $ESD^{(n-k+1)}$ **non risulta significativo,** si confronta il valore ESD precedente, cioè $ESD^{(n-k+2)}$, con lo stesso valore critico precedente per n dati

Se il penultimo ESR risulta significativo, tutti i **k-1** valori testati fino a quello (cioè dal primo al penultimo) sono outlier.

8 – **Se non risulta significativo,** si prosegue fino al primo test che risulta significativo. Si dichiareranno outlier sia quel valore, sia tutti i valori precedenti a quello che è risultato significativo.

9 – Se anche il **primo ESD** calcolato, cioè $ESD^{(n)}$, **non risultasse significativo,** si conclude che nel campione **non sono presenti outlier**.

ESEMPIO 2. (Tratto, con modifiche, dal testo di Bernard **Rosner** dell'anno 2000, ***Fundamentals of Biostatistics***, 5th ed. Duxbury, Pacific Grove, CA, USA, XIX + 792 p.).

In campione di **64** dati, ordinati in modo crescente e di cui sono riportati solo gli estremi nelle due code

13, 23, 26, 30, 31, , 70, 72, 73, 79, 84

individuare gli eventuali outliers.

Risposta. Con n = **64**, il numero massimo di outlier identificabili sarebbe k = 64/10 = **6**.

Ma poiché la parte intera del rapporto $n/10$ è maggiore di 5, si determina k = **5**.

Di conseguenza, supponiamo di voler verificare se nei 64 dati sono compresi 5 outlier, come massimo possibile.

Successivamente,

1 – sul campione totale di 64 dati,

- si calcolano la media $\overline{X}$ = 54,4 e la deviazione standard S = 12,1
- e si individua il valore più estremo X = 13 (in quanto 13 - 54,4 è lo scarto massimo in valore assoluto di tutta la distribuzione).

In questo caso il potenziale outlier è collocato nella coda sinistra della distruzione dei valori, ordinati per rango. Sui 64 dati si calcola la

prima **Deviata Estrema Studentizzata**

$$ESD^{(64)} = \frac{\left|X^{(64)} - \overline{X}^{(64)}\right|}{S^{(64)}} = \frac{|13 - 54{,}4|}{12{,}1} = 3{,}42$$

2 - Eliminato il valore estremo 13, il campione resta con n = 63 dati. Su di essi

- si calcolano la nuova media $\overline{X}$ = 55,1 e la nuova deviazione standard S = 10,9
- e si individua il nuovo valore più estremo, che in questo caso è X = 23 in quanto dista dalla media di questo secondo campione ($\overline{X}$ = 55,1) più del valore estremo (84) collocato nell'altra coda (ricordare che **è un test bilaterale**).

Su questi 63 dati, si calcola la

seconda **Deviata Estrema Studentizzata**

$$ESD^{(63)} = \frac{\left|X^{(63)} - \overline{X}^{(63)}\right|}{S^{(63)}} = \frac{|23 - 55{,}1|}{10{,}9} = 2{,}94$$

3 – Si procede nello stesso modo per gli altri 3 possibili outlier, con il numero n di osservazioni che progressivamente scende da 62 a 60.

I risultati dei vari passaggi, per i $k = 5$ potenziali outlier (X), sono riportati nella tabella:

N	X	$\overline{X}$	S	ESD	P
64	13	54,4	12,1	3,42	< 0.05
63	23	55,1	10,9	2,94	NS
62	26	55,6	10,2	2,90	NS
61	84	56,1	9,6	2,91	NS
60	79	55,6	8,9	2,63	NS

I 5 valori **ESD**, nell'ordine con il quale sono stati calcolati, sono: 3,42 2,94 2,90 2,91 2,63. Si tratta di valutare la loro significativà

4 – Nella tabella dei valori critici riportata in precedenza, si individuano i valori teorici massimi per i vari n . Ma sono riportati solamente di valori critici per n = 60 e n = 70; essi sono

- per $\alpha = 0.05$ con n = 70 il valore ESD critico = 3,26
- per $\alpha = 0.05$ con n = 60 il valore ESD critico = 3,20.

Mediante interpolazione tra questi due estremi, è possibile calcolare i valori critici per i cinque valori n , da 64 a 60.

Per semplicità e come scelta prudenziale, si può assumere come valore critico ESD = 3,26.

Dal confronto emerge che gli ultimi 4 valori ESD sono nettamente minori (anche di 3,20).

Di conseguenza, per essi non si può rifiutare l'ipotesi nulla: nessuno dei 4 valori estremi corrispondenti (79, 84, 26, 23) può essere considerato un outlier.

Risulta significativo solamente il primo valore ESD, quello calcolato per N = 64.

In conclusione, **l'unico vero outlier individuato dal test è il valore 13 con probabilità P < 0.05** di commettere un errore di Tipo I.

Nella tabella precedente che sintetizza i risultati, tali concetti sono esposti con **P < 0.05** per il primo outlier (X = 13) e con **NS** (per **Non Significativo**) per gli altri 4 valori.

B - **Una procedura statistica non parametrica**, quindi più robusta della precedente ma meno potente, è la **Median Absolute Deviation** (acronimo **MAD**) illustrata anche da P. **Sprent** nel suo volume del 1998 ***Data driven statistical methods*** (London, Chapman & Hall). E' un metodo che egli giudica **semplice** e **ragionevolmente robusto** (***a simple and reasonably robust test***).

Come il precedente ESD, questo metodo MAD è valido per la scoperta sia di **uno solo** sia di **più outlier**.

Per la verifica dell'ipotesi nulla

H_0: **non è presente alcun outlier**

contro l'ipotesi alternativa

H_1: **sono presenti k valori outlier**

si rifiuta l'ipotesi nulla per ogni specifico outlier

se

$$Max < \frac{|X - M|}{MAD}$$

dove

- X è il valore ritenuto **outlier**,
- M è la **mediana** del campione di dati, comprendendo l'outlier,
- MAD è la **deviazione mediana assoluta** (in inglese ***median absolute deviation***)
- Max è il **valore critico**, che nella proposta di **Sprent** è prefissato **sempre uguale a 5**.

MAD è una misura non parametrica di **dispersione** o **variabilità** di una distribuzione di dati, analoga alla **deviazione standard** S. E' nota da tempo, tanto da essere citata già nell'Ottocento da Johann Karl Friedrich **Gauss**, il matematico tedesco al quale è attribuita la distribuzione normale. Come caratteristiche statistiche, **MAD** è ritenuta uno **stimatore meno efficiente** (sinonimo di meno potente) della **deviazione standard** S. Ma di essa è più robusta, soprattutto con dati distribuiti in modo non normale, benché la sua validità sia crescente all'aumentare della normalità.

Da una distribuzione campionaria di dati, **MAD** è ricavata calcolando

- prima la **mediana M**,
- successivamente tutte le differenze (D) in valore assoluto di ogni dato (**X**) dalla loro mediana (**M**)

$$D = |X - M|$$

- Si ottengono n differenze D.
- A loro volta, esse devono essere ordinate in modo crescente, per ricavare la MAD, che è appunto la **mediana di questa serie di differenze**.
- Per la ricerca di k **outlier**,

$$Max < \frac{|X - M|}{MAD}$$

il calcolo di Max deve essere ripetuta altrettante volte.

In questa procedura,

- varia il valore X **che identifica l'outlier**,

- mentre restano costanti sia **la mediana** M **sia la** MAD**,** utilizzando sempre quelle **calcolate su tutto il campione di** n **osservazioni.**

Non esiste una tabella di valori critici, collegati alla probabilità α e al numero n di dati.

Il valore critico di Max **è uno solo, prefissato uguale a 5 (**$Max = 5$**),** secondo l'indicazione di **Sprent**. Tale scelta deriva

- dalla **relazione empirica** che esiste tra MAD e **deviazione standard** S: $5MAD = 3S$

- e dal fatto che se **una distribuzione dei dati è approssimativamente normale**, **senza gli outlier,** è ragionevole assumere che un dato che dista dalla sua media più di 3 deviazioni standard sia un outlier.

Nel testo di statistica non parametrica del 2001 ***Applied Nonparametric Statistical Methods*** (3rd ed. Chapman & Hall/CRC, London, XII + 461), a pag. 409 P. **Sprent** e N. C. **Smeeton** scrivono: ***The choice of 5 as a critical value is motivated by the reasoning that if the observations other than outliers have an approximately normal distribution, it picks up as an outlier any observations more than about three standard deviations from the means***.

Quando i dati hanno **una distribuzione lontana dalla normalità e di forma ignota**, è utile la **disuguaglianza di Chebyshev**, ripresa nei paragrafi precedenti.

Con la relazione

$$1 - \frac{1}{k^2}\left(1 - \frac{1}{k^2}\right) \cdot 100$$

essa permette di stimare

- **la percentuale di osservazioni che cadono entro** k **deviazioni standard dalla media.**

Il confronto tra la **percentuale** ottenuta con la distribuzione normale e questa percentuale ottenuta per i vari k da 2 a 5 (con 1 non si può calcolare Chebyshev)

k	2	3	4	5
Normale	97,72	99,87	> 99,990	> 99,999
Chebyshev	75,00	88,89	93,75	96,00

permette di vedere che con $k = 5$

- quando la distribuzione dei dati è molto lontana dalla normalità, l'errore nel definire un outlier è uguale al 4% (100 - 96);
- mentre se la distribuzione fosse normale, l'errore è minore di 1 uno su centomila.

ESEMPIO 3. (Tratto, con modifiche, dal testo di P. **Sprent** e N. C. **Smeeton** del 2001 ***Applied nonparametric statistical methods*****,** 3rd ed. Chapman & Hall/CRC, London, IX + 461 p.).
Nella seguente serie di 11 osservazioni

8,9	6,2	7,2	5,4	3,7	2,8	22,2	12,7	6,9	3,1	29,8

verificare se esistono outlier.

Risposta. Per verificare l'ipotesi nulla

H_0: **non è presente alcun outlier**

contro l'ipotesi alternativa

H_1: **sono presenti k valori outlier**

1 - dapprima si ordinano gli n dati (X) in modo crescente

2,8	3,1	3,7	5,4	6,2	**6,9**	7,2	8,9	12,7	22,2	29,8

per individuare la mediana (M): si ottiene M = 6,9 (in grassetto al centro degli 11 dati ordinati).

2 – Successivamente, **si calcola la differenza** (in valore assoluto) di ognuna delle n osservazioni dalla mediana, cioè D = $|X - M|$, ottenendo la seguente serie di 11 differenze

4,1	3,8	3,2	1,5	0,7	**0,0**	0,3	2,0	5,8	15,3	22,9

Esempio: la prima D = $|2,8 - 6,9| = 4,1$ e l'ultima D = $|29,8 - 6,9| = 22,9$

3 - Dopo aver ordinato a sua volta questa serie di D in modo crescente, come nella tabella successiva

0,0	0,3	0,7	1,5	2,0	**3,2**	3,8	4,1	5,8	15,3	22,9

si individua la loro mediana (3,2), definita appunto come la mediana delle differenze, prese in valore assoluto.

E' il valore della MAD. In questo caso $MAD = 3{,}2$

4 – Infine si ritorna all'analisi statistica dei dati originali, per verificare se tra essi esistono outlier.

Poiché l'osservazione più distante dalla mediana ($M = 6{,}9$) è X = 29,8 si inizia la ricerca da essa; per tale dato, si stima

$$Max = \frac{|X - M|}{MAD} = \frac{|29{,}8 - 6{,}9|}{3{,}2} = 7{,}156$$

ottenendo un valore $Max = 7{,}15$.

E' superiore al valore critico, prefissato in $Max = 5$.

Di conseguenza, si rifiuta l'ipotesi nulla e si conclude che il valore X = 29,8 rappresenta un outlier, con probabilità molto piccola di commettere un errore di Tipo I.

Se il primo valore Max **calcolato è inferiore a 5, si deve concludere che non esistono outlier**.

5 – Trovato il primo outlier, si passa a verificare **la seconda osservazione più distante dalla mediana** ($M = 6{,}9$); nel campione degli $n = 11$ dati, risulta X = 22,2.

Utilizzando sempre **la mediana e la MAD precedenti**, calcolata su tutto il campione di 11 dati, questo secondo test

$$\frac{|X - M|}{MAD} = \frac{|22{,}2 - 6{,}9|}{3{,}2} = 4{,}78$$

stima $Max = 4{,}78$. E' inferiore a 5.

Se ne deve dedurre che per l'osservazione X = 22,2 non esiste evidenza sufficiente per ritenerla un outlier.

6 - La ricerca termina con il primo valore più estremo non significativo.

Con i dati di questo esempio, termina a questo punto e si traggono le conclusioni generali:

- il gruppo di 11 osservazioni contiene un solo outlier, esattamente il valore X = 29,8.

13.14. TRATTAMENTO DEGLI OUTLIER: ELIMINARLI O UTILIZZARLI? COME?

Se esistono outlier, la distribuzione del campione non dovrebbe avere forma normale. Di conseguenza, sarebbe possibile **utilizzare i test di normalità,** anche per una verifica della possibile **esistenza di outlier**. Se gli **outlier** sono solo in una **coda della distribuzione**, come in medicina quando gli individui ammalati sono caratterizzati da valori molto più alti oppure molto più bassi della norma, **possono essere utilizzati i test di simmetria**. Ma, come tutti quelli per la normalità, questi test sono poco potenti. Di conseguenza, per scoprire gli outlier è vantaggioso utilizzare i metodi proposti in questo capitolo.

La difficoltà di individuare gli outlier in molte analisi di laboratorio ha suggerito l'utilizzazione di **protocolli standard**. In un testo di statistica non è possibile una loro presentazione generale, perché sono specifici di ogni singola disciplina e entro esse di ogni tipo di analisi.
Una volta che sia stato dimostrato che un dato probabilmente è un outlier, nella letteratura si apre un altro dibattito interessante **sul suo uso** e sull'**importanza** che gli deve essere attribuita, per le analisi statistiche:
*L'**outlier è il dato meno importante**, quindi da eliminare, oppure è **quello più importante**, da analizzare con particolare attenzione e dal quale dipendono le decisioni?*
La risposta non è univoca: *Dipende dal contesto*.

Come esempio, nei testi di statistica applicata sono riportati due casi estremi, che possono essere frequenti nell'analisi chimica e più in generale nelle misure di laboratorio. Sono due casi tra loro identici come impostazione metodologica, ma che hanno scopi contrastanti e quindi conducono a decisioni opposte.
Come primo caso, si supponga di voler misurare quale è la concentrazione di una sostanza presente in un prodotto industriale attraverso 5 campioni, per ottenere la media e la varianze più vicine alla realtà. Se tra essi è presente un outlier, il risultato di quel dato è interpretato come un probabile errore nella conduzione dell'esperimento, eventualmente determinato da variazioni ambientali indesiderate, delle quali non ci si è accorti: è eliminato e sia la media sia la varianza sono calcolate solamente sulle altre 4 misure.
Come secondo caso, si ipotizzi di voler valutare l'attendibilità dello strumento nelle analisi precedenti. La presenza anche di un solo outlier lo rende inaffidabile: l'outlier diventa l'informazione più importante. Il rapporto di verifica sarà fondato su di esso e lo strumento sarà rifiutato perché poco attendibile.

Accertata l'esistenza di uno o più outlier, come comportarsi nell'analisi statistica?

La prima risposta è la **trasformazione dei dati**, per ricostruire la **normalità della distribuzione** e la condizione di **omoschedasticità,** se i gruppi sono almeno due. Ma esistono altre scelte.

Se le due risposte estreme, quali

- eliminarli dal campione,
- accettarli come gli altri ed effettuare l'analisi come se i dati fossero tutti corretti,

sono considerate poco logiche, anche se applicate da molti, esistono vari altri modi per raggiungere un compromesso,

- che da una parte **non li elimini**, perché esistono,
- ma che dall'altra **riduca il loro peso** sull'informazione fornita da tutti gli altri dati della distribuzione.

I metodi più diffusi sono quattro:

1 – l'uso della **mediana al posto della media**, come misura di tendenza centrale,

2 – passare da una **scala di rapporti** o **a intervalli** a una **scala di tipo ordinale**,

3 – ricorrere al **Trimming**,

4 – ricorrere alla **Winsorization**.

I primi due metodi sono già stati presentati in varie situazioni e sono fondamentalmente la scelta di un test non parametrico.

Analoghi a essi, sono il **jackknife e il bootstrap**, descritti in modo dettagliato in un capitolo successivo. Soprattutto il primo è in grado di evidenziare l'effetto del valore anomalo sulle statistiche della distribuzione, fornendo la statistica con esso e l'intervallo di confidenza della statistica senza di esso.

Il **Trimming data,** o semplicemente **Trimming,** è l'eliminazione di una percentuale fissa di valori estremi. Può essere fatta in entrambe le code o in una coda sola della distribuzione dei dati, sulla base delle caratteristiche del fenomeno. Anche la quota di estremi da eliminare è molto variabile, potendo essere

- solo il valore più alto e quello più basso,
- il primo e l'ultimo cinque per cento,
- il primo e l'ultimo quartile (25%),
- altre quote tra il minimo di un dato e il massimo di un quarto dei dati.

E' relativamente frequente la scelta di prendere in considerazione solamente il 50% dei valori centrali, come appunto si ottiene eliminando il primo e l'ultimo quarto. La media di questa distribuzione è

chiamata **media interquartile** e viene utilizzata quando la proporzione di outlier in entrambe le code è molto alta.

La **Winsorization** (la tecnica è chiamata **winsorizing)** presentata da vari autori di testi di statistica applicata, tra i quali W. J. **Dixon** e F. J. **Massey** con il testo del 1969 ***Introduction to Statistical Analysis*** (edito da McGraw-Hill, New York, a pagg. 330-332) non elimina i valori più estremi, ma li sostituisce con altri meno estremi.

E' una tecnica semplice, che serve **per attenuare l'effetto di possibili outlier**, quando i dati raccolti servono per il calcolo delle statistiche del campione o per test successivi (***A simple technique to soften the influence of possible outliers***).
Il numero di valori da prendere in considerazione ovviamente dipende
- da n , il numero di dati,
- e dalle caratteristiche della distribuzione.

Ad esempio, si supponga di avere ottenuto la seguente serie di 13 valori, qui ordinata

0	1	12	13	15	16	18	20	22	25	26	154	322

e la cui media è $\overline{X}$ = 49,5.
E semplice rilevare dalla lettura dei dati, quindi a posteriori, che sono presenti due valori molto differenti da tutti gli altri, in entrambi gli estremi (i valori 0 e 1 nella coda sinistra; 154 e 322 nella coda destra). Può essere utile costruire una nuova distribuzione, sempre di n dati; quindi senza diminuire le dimensioni del campione. Questi estremi in entrambe le direzioni sono sostituiti dal terzo valore, quello a loro più vicino, ottenendo la seguente serie di dati

12	12	12	13	15	16	18	20	22	25	26	26	26

la cui media è $\overline{X}$ = 18,7.
La mediana delle due distribuzioni dei 13 valori è 18. Si osservi come la seconda media (18,7) sia molto vicina alla mediana (18), che ovviamente è rimasta immutata, mantenendo n costante.

Questo metodo è da utilizzare soprattutto quando sono presenti valori indefiniti (come < 1 oppure > 100). Sono misure che si riscontrano quando la variabilità delle quantità presenti nei campioni è nettamente inferiore oppure superiore al campo di misura dello strumento, che è preciso solo per valori intermedi.
Il **trimming** può essere **simmetrico**, come in questo caso; ma può anche essere **asimmetrico**, quando l'operazione coinvolge un numero di dati differenti nelle due code della distribuzione.

Come nella scelta del test quando i metodi alternativi sono numerosi, anche in questa situazione dopo che sono state presentate varie metodologie si pone un problema: "Quale è l'operazione più appropriata, sia per identificare l'outlier, sia per effettuare test corretti in presenza di outlier?".
La risposta è data solo da una conoscenza della statistica che sia congiunta a una competenza ancora maggiore nella disciplina. Da questa ultima infatti dipendono
- il valore da attribuire all'outlier,
- la frequenza con la quale il fenomeno è atteso,
- la scelta del tipo di scala nella sua misurazione.

Dalla **competenza disciplinare** dipende la decisione sulla **esistenza stessa dell'outlier**, poiché l'analisi statistica fornisce solo una probabilità. Infatti è sempre possibile avere un valore che è anomalo solo apparentemente, perché raro.
Si ritorna alla **soggettività della scelta**, all'**esperienza come fattore prevalente nella decisione**, pure in presenza di tanti metodi rintracciabili nella letteratura statistica. Anche in questa serie di problemi, tra i meno schematizzati dell'analisi statistica, che vanno dalla individuazione degli outlier a quello della scelta del test più adatto per identificarli e infine alla decisione se eliminarli, si ritorna al problema più generale dell'**interpretazione dei risultati** dei test inferenziali. Essa dipende quasi totalmente dalla conoscenza della disciplina alla quale la statistica è applicata; il risultato del test assume vero significato solo nella successiva interpretazione disciplinare.
La statistica fornisce solo un contributo di informazioni. Compete al ricercatore decidere sulla significatività o meno di una media, della varianza o di un qualsiasi altro parametro.

Si può concludere la discussione sugli outlier, rispondendo alle domande precedenti "Quale è l'operazione più appropriata, sia per identificare l'outlier, sia per effettuare test corretti in presenza di outlier?" con un'altra domanda: "La rilevazione e l'eventuale rimozione degli outlier per applicare un test parametrico è importante o trascurabile?
Robert R. **Sokal** e F. James **Rohlf,** autori di uno dei testi internazionali più diffusi a livello di preparazione post-laurea, ***Biometry. The principles and practice of statistics in biological research***

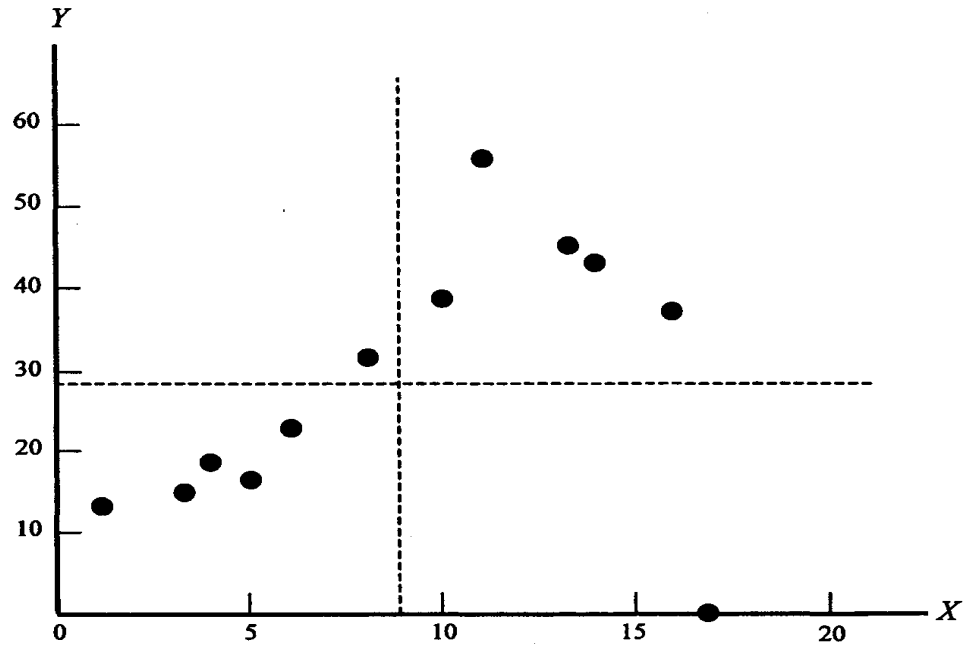
(3rd ed. W. H. Freeman and Company, New York, XIX + 887 p.) scrivono (a pag. 407): *Le conseguenze della non normalità degli errori (gli scarti dei valore dalla media) non sono molto gravi (**The consequences of nonnormality of error are not too serious**), poiché le medie hanno una distribuzione più vicina alla normale della distribuzione delle singole osservazioni, come conseguenza del teorema del limite centrale. Solamente distribuzioni fortemente asimmetriche possono avere effetti rilevanti sul livello di significatività di un test F o sull'efficienza del disegno sperimentale. Il modo migliore per correggere la perdita di normalità è effettuare una trasformazione che renda normale la distribuzione. Se la trasformazione non da risultati soddisfacenti, utilizzare test non parametrici.*
Tuttavia, spesso il problema è più complesso.

La scelta **tra mantenere** oppure **eliminare l'outlier** nelle analisi statistiche **dipende dalla teoria** che **si vuole verificare,** poiché ovviamente si desidera che il risultato del test coincida con la teoria che si vuole dimostrare.
Come esempio, assumiamo un problema di psicologia: gli studenti migliori in matematica sono i migliori anche nell'apprendimento delle lingue?

C'è chi afferma che gli studenti migliori in matematica sono tali perché più diligenti, logici e studiosi; quindi, con poche eccezioni, sono anche i migliori in tutte le altre discipline, tra cui lo studio della lingua.
Ma appare ugualmente convincente anche la teoria opposta.
Chi è portato alla logica matematica ha poca attitudine per l'apprendimento alle lingue; inoltre la conoscenza delle lingue straniere richiedono attività e impegni, come i viaggi, i soggiorni all'estero e i contatti con le persone, che male si conciliano con lo studio e la riflessione richiesti dalla matematica.
Un esperimento con 12 studenti, che hanno svolto un compito di Matematica e una prova scritta di Lingua straniera, ha dato i seguenti risultati nel conteggio degli errori:

Studente		A	B	C	D	E	F	G	H	I	L	M	N
Matematica	(X)	1	3	4	5	6	8	10	11	13	14	16	17
Lingua	(Y)	13	15	18	16	23	31	39	56	45	43	37	0

La rappresentazione grafica facilita la lettura del risultato complessivo ed evidenzia la presenza di un outlier.



E' un problema di statistica bivariata, che sarà discussa in capitoli successivi. Ma i concetti sull'uso del'outlier sono identici.

Si osserva che per undici giovani all'aumentare del numero di errori in matematica (X) aumentano anche quelli lingua. Il dodicesimo giovane è un outlier: di madre lingua parla meglio del docente, ma ha dovuto cambiare spesso scuola e in matematica è quello che ha commesso più errori.

Se si analizzano solamente i primi dati con una correlazione parametrica (come l'**r di Pearson**) o meglio ancora una correlazione non parametrica (**come il tau di Kendall** o il **rho di Sperman)** si dimostra che esiste una correlazione positiva.

Se l'outlier viene lasciato e si analizzano insieme i 12 dati, è tale il peso del punto anomalo che la teoria potrebbe essere rovesciata. Anche in questo caso, è importante la scelta del test, che può variare da una correlazione non parametrica classica al **test della mediana di Blomqvist**.

Quale la scelta adeguata?

Tutte. Sia separatamente, sia insieme.

Purché adeguatamente motivate, sotto l'aspetto disciplinare e di metodologia statistica.