CAPITOLO XXIV

LA REGRESSIONE LINEARE MODELLO II E LEAST-PRODUCTS. IL CONFRONTO TRA DUE METODI QUANTITATIVI. IL SEI-SIGMA NEL CONTROLLO DI QUALITA'

CAPITOLO XXIV

LA REGRESSIONE LINEARE MODELLO II E LEAST-PRODUCTS. IL CONFRONTO TRA DUE METODI QUANTITATIVI. IL SEI-SIGMA NEL CONTROLLO DI QUALITA'

24.1 I MODELLI I E II NELLA REGRESSIONE LINEARE; IL CASO DI BERKSON

Il modello di regressione lineare semplice fino a ora presentato, fondato sul principio dei **minimi quadrati** (***least-squares method***), è di uso abituale. Denominato anche ***Ordinary Least-Squares Regression*** (abbreviato in **OLR** su alcuni testi e **OLS** in altri), è l'unico riportato su quasi tutti i testi di statistica, anche a diffusione internazionale.

Un'altra denominazione della ***Least-Squares Regression***, preferita dai biologi, è ***Model I Regression***

- sia per analogia all'**ANOVA I** nelle **assunzioni di validità** e nel **modello additivo,**
- sia per distinguerlo dal **Model II Regression** proposto successivamente, che utilizza un approccio diverso.

La **Regression Model I** è fondata su **quattro assunzioni,** più volte ripetute discutendo le condizioni di validità e la trasformazione dei dati.

1 - I campioni lungo la retta di regressione devono essere **omoschedastici** (avere varianza uguale).

In altri termini, **la varianza reale** σ^2 lungo la retta deve essere **costante** e quindi

- **indipendente dalle dimensioni** sia della **variabile X** sia della **variabile Y**.

2 - Per ogni valore X_i della variabile X,

– **i valori della Y sono indipendenti e distribuiti in modo normale**, come richiede il modello additivo

$$Y_{ij} = \alpha + \beta X_i + \varepsilon_{ij}$$

dove si assume, altro modo per definire lo stesso concetto, che

- **i valori** ε_{ij} **siano distribuiti in modo normale** e con **media zero**.

3 - I **valori attesi per la variabile Y devono essere in accordo con una funzione lineare**. Quindi i valori medi per ogni X devono essere descritti

dal modello matematico

$$\mu_Y = \alpha + \beta X$$

4 - **La variabile indipendente X è misurata senza errore**.

In termini più tecnici, si dice che **la variabile X è fissa (*fixed*)**, è sotto il **controllo dello sperimentatore**, mentre la **variabile dipendente Y è casuale (*random*)**, affetta da errori casuali. L'esempio classico è quando

- la X indica la dose di un farmaco somministrato a un paziente,
- mentre la variabile Y fornisce la quantità della risposta biologica.

Nelle condizioni sperimentali effettive, sovente queste condizioni richieste dal modello statistico - matematico non sono rispettate. In particolare le prime tre.

Per applicare correttamente il test della regressione lineare, è allora richiesto di

- ricostruirle mediante trasformazione**,**

- di eluderle con l'uso di un test non parametrico, come ampiamente descritto nei paragrafi dedicati alla regressione lineare non parametrica, quali il **metodo di Theil** e il **metodo di Bartlett**.

In questo capitolo, l'attenzione soprattutto la condizione 4.

Sia nella raccolta dei dati in natura, sia negli esperimenti di laboratorio, in molte situazioni

- **la variabile X presenta una variazione naturale e/o è sottoposta a errori di misura**, non diversamente da quanto avviene per la variabile Y.

Non è quindi vero l'assunto che la variabile X è misurata senza errore ed è nota con precisione.

Gli esempi possono essere numerosi.

A) Un primo caso tipico è quando **la variabile X e la Y sono due variabili continue**, che formano una **distribuzione normale bivariata**, come quando si utilizza la **correlazione**.

Può essere il caso

- dell'ampiezza dell'ala sinistra e di quella destra in un campione di uccelli,
- dell'altezza in coppie di sposi o di fratelli,
- della massa muscolare e della prestazione atletica in un gruppo di individui.

Sono tutti casi in cui la variabile **X è sottoposta a diversi fattori di variabilità**, da quelli **genetici** e **ambientali** a quelli **culturali** (nel caso dell'uomo), oltre a quelli di **misura**.

B) Un secondo gruppo di casi è quando **la distribuzione non è bivariata**, per una scelta specifica dello sperimentatore. Per il calcolo della retta di regressione tra altezza e peso nell'uomo, è possibile non utilizzare un campione casuale, ma per motivi tecnici scegliere **un campione bilanciato**.

Accade, ad esempio, quando per **ogni gruppo di altezze** (variabile X) è **stato scelto** un **campione con lo stesso numero di dati**, che

- fornisce una informazione con errore standard costante per ogni gruppo di individui,

- ma non rispetta certamente la normalità della distribuzione dell'altezza e l'omogeneità della varianza tra i diversi raggruppamenti.

In questo caso, non è rispettata soprattutto la condizione di normalità della distribuzione.

C) Un terzo gruppo di casi avviene nelle discipline fisiche e chimiche, più che in quelle biologiche, come nelle misure di conducibilità dello stesso campione di metallo, a diverse temperature.

Entrambe le variabili non hanno una variabilità naturale.

In questo caso, non è rispetta la condizione richiesta per la variabile Y, poiché la conducibilità (Y) come la temperatura (X) sono entrambe soggette solamente all'errore dello strumento.

D) La regressione lineare parametrica è il metodo comunemente utilizzato fino ad ora, per confrontare i risultati ottenuti con **due metodologie differenti**, come già illustrato nel problema della **calibratura** o **calibrazione** (***calibration***). (Nell'***International Vocabulary of Basic and General Terms in Metrology***, ISO, GENEVA, Switzerland, 2nd ed. 1993, dicesi ***calibration***, la sequenza di operazioni necessarie a stabilire, in determinate condizioni sperimentali, la relazione tra i valori forniti da uno strumento o sistema di misurazione (per es. **assorbanza**) e i valori ad essi corrispondenti di una parametro (per es. **concentrazione**) di uno o più materiali di riferimento).

Secondo un approccio più recente e diffuso nelle metodologie cliniche, assumere che X sia noto senza errore quando si confrontano tra loro due metodi equivalenti è indubbiamente lontano dalla realtà sperimentale, in quanto

- il valore della variabile X è rilevato con lo stesso errore con il quale è misurata la variabile Y .

Si consideri la somministrazione a un organismo di dosi differenti di un ormone, per valutarne le conseguenze su qualche altro parametro. **Come si può affermare che la variabile indipendente X (dose) è applicata senza errore?**

Nella realtà, esistono molti fattori che per essa determinano tanti tipi di errore, quali

- la tecnica di somministrazione,
- la lettura strumentale della quantità di sostanza in cui l'ormone è diluito,
- la determinazione della sua concentrazione nel diluente.

Complessivamente, **la somma di questi fattori** diventa spesso una quantità importante, rispetto al valore rilevato della X. Essa quindi non è nota con precisione, senza errore.

A questa situazione, pure in un contesto di confronto tra due metodi, fa eccezione il **caso di Berkson** (***Berkson case***), citato come un caso di **apparente regressione Model II** da Robert R. **Sokal** e F. James **Rohlf** nel loro testo del 1995 ***Biometry. The principles and practice of statistics in biological research*** (3rd ed. W. H. Freeman and Company, New York, XIX, + 887 p.).

Secondo il modello descritto da J. **Berkson** nel 1950 nell'articolo ***Are there two regression?*** (su **Journal of the Statistical Association** Vol. 45, pp.: 164-180), anche quando si confrontano due metodi vi è un caso in cui la variabile esplicativa o indipendente X, chiamata **variabile controllata** (***controlled variable***),

- è sotto la diretta gestione dello sperimentatore ed è nota con precisione, come richiede la ***regression model I*** o ***least–squares regression***.

Avviene quando il campione che deve essere sottoposto a determinazione, in termini più tecnici l'**analita**,

- ha valori di X prefissati oppure prestabiliti con il metodo classico, (ad esempio, l'analita è stato preparato con una certa concentrazione),
- e con il metodo di confronto si ricava la stima di Y, ripetendo eventualmente l'analisi più volte per lo stesso valore di X.

In questa situazione, secondo **Berkson** si dovrebbe ugualmente applicare il **metodo ordinario di regressione lineare** (***ordinary least squares***), per stimare i coefficienti della retta, in quanto l'errore commesso nella stima di X può essere ignorato.

Questi concetti di **Berkson** nell'espressione matematica diventano:

- la misura X_i è data dalla quantità vera ξ_i più un errore δ_i:

$$X_i = \xi_i + \delta_i$$

- e la misura Y_i è determinata attraverso la regressione lineare con

$$Y_i = \alpha + \beta\xi_i + \varepsilon_i$$

dove

- la misura X_i (non importa se stocastica o non stocastica) è **controllata dallo sperimentatore**,
- gli **errori** ε_i, δ_i sono due sequenze di **variabili random**.

In questo caso, ξ_i **e** δ_i **sono tra loro dipendenti**. E' la differenza fondamentale rispetto alla situazione in cui si confrontano due metodi e si misurano le due variabili in modo indipendente.

Il modello può quindi essere scritto come

$$Y_i = \alpha + \beta\xi_i + (\varepsilon_i - \beta\delta_i)$$

dove

- la misura X_i non è correlata con $\varepsilon_i - \beta\delta_i$ cioè con l'errore della misura Y_i.

In questa situazione, secondo **Berkson**

- **è possibile utilizzare la retta parametrica** fondata sul principio dei minimi quadrati, in quanto **l'errore di misura** δ_i **può essere ignorato** poiché ξ_i **e** δ_i **sono tra loro dipendenti**;

- mentre quando ξ_i e δ_i sono tra loro indipendenti, quindi la misura X_i **non è controllata dallo sperimentatore,** ***the ordinary least squares method is not appropriate for the estimation of same parameters.***

Tuttavia, in molti dei prima casi citati, nei quali le **condizioni teoriche di validità chiaramente sono violate** almeno sotto l'aspetto teorico, è prassi raramente contesta in biologia
- **utilizzare ugualmente la procedura classica di regressione Model I.**

Quando allora considerare non corretta la regressione Model I e utilizzare la Model II?
Le indicazioni di **Sokal** e **Rohlf** (a pag. 543), l'unico tra i grandi testi internazionali che affronta il problema, sono molto vaghe: ***Research on and controversy over Model II regression continues, and definitive recommendations are difficult to make. Much depends on the intentions of the investigator***.

In linea generale, se la regressione è effettuata per **scopi predittivi**, la **Model I e senza dubbio sempre corretta**, anche se alcuni statistici pongono una ulteriore **distinzione** scolastica, più fine**, tra predizione e relazione funzionale**.
Quando invece si vuole una **interscambialità delle due variabili,** sarebbe richiesto **il Model II**.

I **metodi della Model II Regression** sono fondati essenzialmente su **due approcci**: (1) la **correlazione** e (2) la **regressione lineare non parametrica**.

1 – I concetti che stanno alla base della **correlazione** sono appropriati, perché in essa non si distingue tra le due variabili, per quanto attiene le condizioni di validità e la precisione con la quale la variabile è stata misurata. Fondamentalmente esistono due situazioni

a) quando **le due variabili hanno la stessa unità di misura**, si può utilizzare il **coefficiente angolare dell'asse maggiore** (***the slope of the major axis***) o **dell'asse principale (*principal axis*);**

b) quando **le due variabili hanno unità di misura differenti**, **la correlazione come indicatore della regressione diventa priva di significato specifico**, poiché essa è indipendente dal tipo di scala e varia sempre tra –1 e +1; il valore della correlazione ritorna a essere una indicazione non banale, quando le scale possono essere differenti, ma sono **predeterminate e non arbitrarie** e quindi sono noti i rapporti tra esse.

Il metodo del coefficiente angolare dell'asse maggiore utilizza anche scale trasformate. Spesso, una delle due variabili ha una trasformazione logaritmica o in radice.
Un altro metodo ancora è la **standardizzazione delle due variabili**, per cui ognuna assume media zero e deviazione standard uno, prima del calcolo del coefficiente angolare.

L'asse principale di queste variabili standardizzate ha vari nomi: ***reduced major axis***, ***standard major axis***, ***geometric mean regression***, mentre nei testi francesi si trova anche il termine ***relation d'allometrie***, utilizzato soprattutto da G. **Teissier** nell'articolo del 1948 ***La relation d'allometrie: sa signification statistique et biologique*** (pubblicato su **Biometrics** Vol. 4, pp.: 14-48).
Più recentemente, in particolare quando si confrontano due metodi clinici, biologici, chimici o fisici, impiegati per determinare la stessa sostanza o per misurare lo stesso fenomeno, sono proposti i metodi di **Deming**, di **Mantel** e di **York**, illustrati nel prosieguo del capitolo.

2 - **La regressione lineare non parametrica** può essere utilizzata in quasi tutte le condizioni sperimentali. Come tutti i test fondati sui ranghi e ***distribution free***, non richiede il rispetto dei quattro assunti di validità. Tra i metodi non parametrici, il più noto, diffuso e accettato è il **metodo di Theil,** la cui significatività è analizzata mediante la correlazione non parametrica **τ di Kendall**. Quindi è chiamato anche metodo di **Theil-Kendall** oppure **metodo robusto di Kendall (*Kendall's robust line-fit method*)**.
Tra questi test non parametrici d'inferenza sulla regressione lineare, è sovente citato anche il **metodo di Bartlett,** detto più estesamente **metodo dei tre gruppi di Bartlett** (***Bartlett's three-group method***).
Rispetto al metodo di **Theil**, presenta il vantaggio di essere ancora più semplice e rapido.
Ma ha lo svantaggio di essere più frequentemente criticato nei testi di statistica poiché, utilizzando le medie delle X e delle Y sia del primo sia dell'ultimo terzo dei dati della X, è molto sensibile ai valori anomali, spesso collocati agli estremi.

Tra i grandi testi a diffusione internazionale, il **test di Bartlett** è riportato solamente nel volume di Robert R. **Sokal** e F. James **Rohlf** del 1995 ***Biometry. The Principles and Practice of Statistics in Biological Research*** (3rd ed. W. H. Freeman and Company, New York, XIX, + 887 p.).
In questi ultimi anni, è citato con frequenza soprattutto nella letteratura chimica e clinica, per il confronto tra metodi.
La presentazione di questi due test non parametrici sulla regressione (**Theil** e **Bartlett**) è stata effettuata nel capitolo dedicato alla regressione non parametrica.

24.2 LA RETTA DEL COEFFICIENTE ANGOLARE DELL'ASSE MAGGIORE.

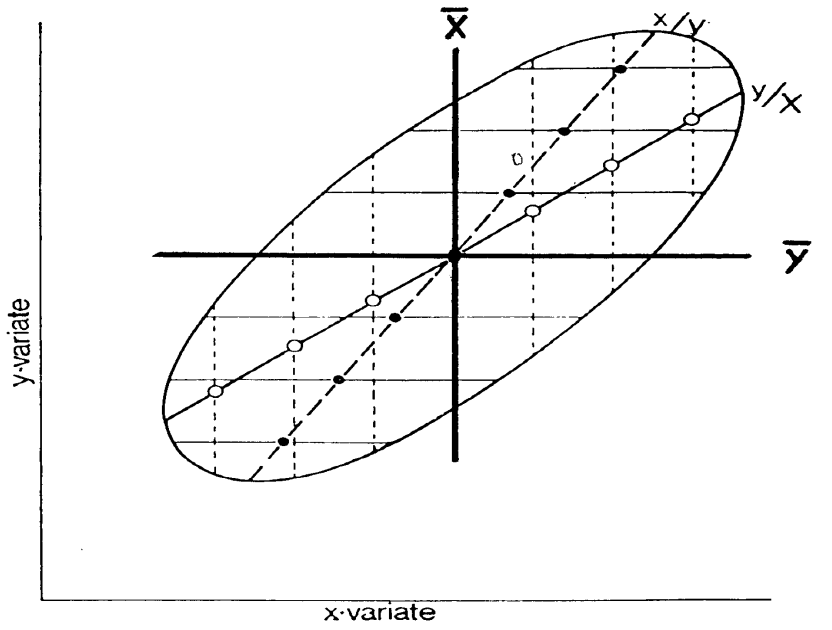
Come è stato sviluppato più ampiamente nel capitolo sulla correlazione,

quando si dispone di due variabili (X e Y),

- con **dati campionari** che sono rappresentati nel **diagramma di dispersione** come punti compresi entro un piano cartesiano circoscritto dall'**ellissoide**,

è possibile calcolare

- due **rette di regressione**, non più una sola.



La **prima** è la retta di regressione

$$\hat{Y}_i = a + bX_i$$

che, come nella prassi, assume la X come variabile indipendente o predittiva e la Y come variabile dipendente o predetta.

L'intercetta a e il coefficiente angolare b sono rappresentati rispettivamente con

- $a_{Y/X}$ oppure, meno frequentemente, a_{Y-X}
- $b_{Y/X}$ oppure, meno frequentemente, b_{Y-X}

per indicare che la Y è stata assunta come variabile dipendente e la X come variabile indipendente.

Se venisse calcolata con i dati rappresentati dall'ellissoide nel diagramma di dispersione, tale retta

- coinciderebbe con quella descritta dalla **serie dei punti vuoti rappresentata da Y/X.**

E' chiamata **retta di Y su X**, passa dal baricentro, il **punto d'incontro delle medie** ($\overline{X}$ e $\overline{Y}$) delle due variabili, ma non coincide con l'asse maggiore dell'ellissoide, essendo più spostata verso la media $\overline{Y}$ della variabile dipendente.

La **seconda retta** è ricavata scambiando tra loro le variabili X e Y:

$$\hat{X}_i = a + bY_i$$

Per distinguerla dalla precedente, l'intercetta a e il coefficiente angolare b sono rappresentati rispettivamente con

- $a_{X/Y}$ oppure, meno frequentemente, a_{X-Y}
- $b_{X/Y}$ oppure, meno frequentemente, b_{X-Y}

per indicare che la X è stata assunta come variabile dipendente e la Y come variabile indipendente.

Nella formula della retta, i valori di questa intercetta a e del suo coefficiente angolare b sono diversi da quelli della retta precedente, anche se qui per semplicità sono impiegati gli stessi simboli.

Rappresentata nello stesso grafico precedente, questa ultima retta

- coinciderebbe con quella descritta dalla **serie dei punti pieni rappresentata da X/Y.**

E' detta **retta di X su Y.**

Anch'essa attraversa il baricentro, ma non coincide con l'asse maggiore dell'ellissoide. In modo simmetrico alla prima, in questo caso è più spostata verso la $\overline{X}$, la media della nuova variabile dipendente.

Le due rette sono tra loro tanto più differenti, quanto più ampio è l'asse minore, perpendicolare all'asse principale.

Per trovare un **numero solo** che fornisca una **misura sintetica della relazione esistente tra le due variabili X e Y**, è possibile utilizzare la correlazione r

$$r = \sqrt{b_{Y/X} \cdot b_{X/Y}}$$

che concettualmente è fondata **sulla media geometrica dei due differenti coefficienti angolari,** calcolati con le due diverse rette.

Tale metodo è stato ampiamente discusso nel capitolo sulla correlazione.

Un altro metodo per ricavare **un indicatore sintetico delle due rette**, che è il problema della **Model II Regression** qui discusso,

- è **calcolare la retta che passa lungo l'asse maggiore**.
- per utilizzare il suo **coefficiente angolare** b_p, detto appunto **coefficiente angolare dell'asse maggiore** (***the slope of the major axis***) o **dell'asse principale (*principal axis*).**

Anche questa retta attraversa il baricentro della distribuzione.

In queste analisi sulla **Model II Regression**, le due variabili X e Y possono essere ritenute scambievolmente sia la causa sia l'effetto dell'altra.

Pertanto, per convenzione, su molti testi non sono più indicate con X (causa) e Y (effetto), ma con

- **X_1 riportata sull'asse delle ordinate**
- **X_2 riportata sull'asse delle ascisse**.

In questo caso sono scambiate rispetto alla correlazione.

Su altri testi, per meglio distinguere le formule che si riferiscono alle due variabili, continua a essere utilizzata la simbologia della regressione lineare X e Y.

L'**equazione** che individua i punti collocati sull'**asse principale ($\hat{X}$),** riferiti sempre agli **stessi assi cartesiani X_1 e X_2** è

$$\hat{X}_1 = a + b \cdot X_2$$

dove

- $a = \overline{X}_1 - b \cdot \overline{X}_2$
- b = **coefficiente angolare dell'asse principale** (***slope of the principal axis***)

ed è ricavato da

$$b = \frac{S_{12}^2}{\lambda - S_1^2}$$

con

- S_{12}^2 = covarianza delle variabili **X_1** e **X_2**

che nella formula euristica è

$$S_{12}^2 = \frac{\sum_{i=1}^{n} \left[\left(X_{1i} - \overline{X}_1 \right) \cdot \left(X_{2i} - \overline{X}_2 \right) \right]}{n-1}$$

e nella formula abbreviata diventa

$$S_{12}^2 = \frac{\sum_{i=1}^{n}(X_{1i} \cdot X_{2i}) - \frac{\left(\sum_{i=1}^{n} X_{1i}\right) \cdot \left(\sum_{i=1}^{n} X_{2i}\right)}{n}}{n-1}$$

- S_1^2 = varianza della X_1

che nella formula euristica è

$$S_1^2 = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2}{n-1}$$

e nella formula abbreviata diventa

$$S_1^2 = \frac{\sum_{i=1}^{n}(X_{1i}^2) - \frac{\left(\sum_{i=1}^{n} X_{1i}\right)^2}{n}}{n-1}$$

- S_2^2 = varianza della X_2

che nella formula euristica è

$$S_2^2 = \frac{\sum_{i=1}^{n}(X_{2i} - \overline{X}_2)^2}{n-1}$$

e nella formula abbreviata diventa

$$S_2^2 = \frac{\sum_{i=1}^{n}(X_{2i}^2) - \frac{\left(\sum_{i=1}^{n} X_{2i}\right)^2}{n}}{n-1}$$

- λ è la quantità nuova, che misura la variabilità dei punti campionari lungo l'asse maggiore; è definita in termini di varianza e covarianza, misurate sull'asse delle ascisse e delle ordinate originali,

$$\lambda = \frac{S_1^2 + S_2^2 + \sqrt{\left(S_1^2 + S_2^2\right)^2 - 4\left((S_1^2 \cdot S_2^2) - S_{12}^2\right)}}{2}$$

I criteri per individuare questa retta e sulla base dei quali è stata definita la formula sono tre:

- la retta deve passare attraverso il **punto d'incontro delle medie** ($\overline{X}_1$ e $\overline{X}_2$) delle due variabili,
- la **devianza dei punti da questa retta** deve essere **quella minima**,
- e **le deviazioni sono misurate**

a) **non sull'asse delle ordinate** (Y nella retta di regressione, X_1 in questa rappresentazione per l'asse principale),

b) ma **sull'asse delle ascisse** (X nella retta di regressione, X_2 in questa rappresentazione).

ESEMPIO. Per illustrare in modo semplice e operativo il metodo del **coefficiente angolare dell'asse maggiore** (***the slope of the major axis***) o **dell'asse principale (*principal axis*),** appositamente scelto come riferimento bibliografico internazionale e autorevole, è utile svolgere in tutti i passaggi logici l'esempio riportato nel testo di **Sokal** e **Rohlf** già citato (pag. 587-593).

Su un campione di 12 granchi della specie *Pachygrapsus crassipes* come coppie di variabili sono state rilevate

- il **peso delle branchie** (**X_1,** espresse in milligrammi)
- il **peso complessivo del corpo** (**X_2**, espresso in grammi)

X_1	159	179	100	45	384	230	100	320	80	220	320	210
X_2	14,40	15,20	11,30	2,50	22,70	14,90	1,41	15,81	4,19	15,39	17,25	9,52

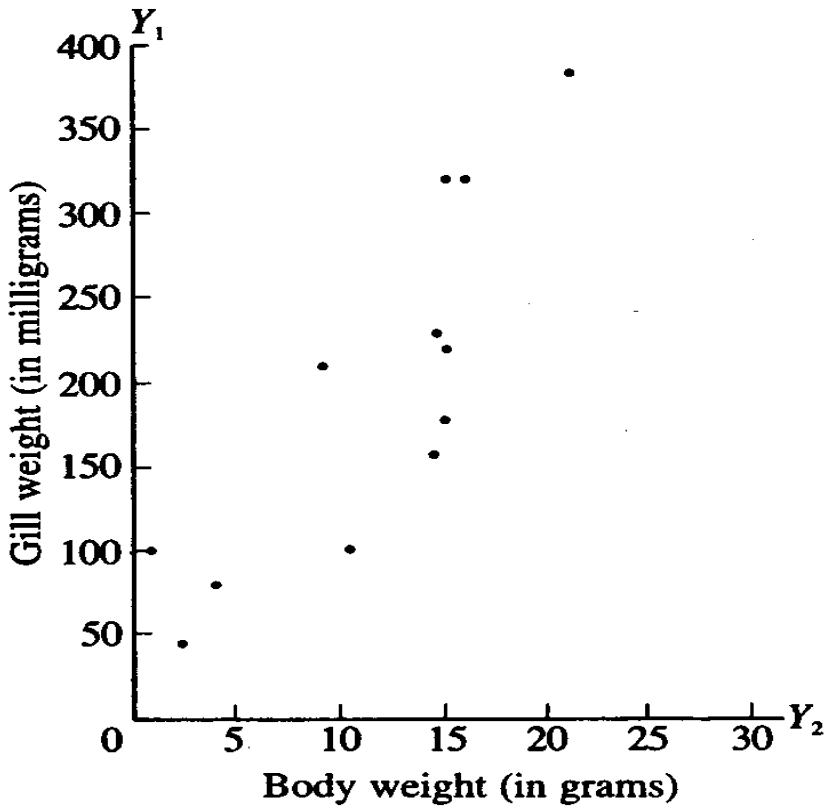
Il grafico mostra con evidenza la dispersione dei punti lungo l'asse principale.

(In esso, come in molte pubblicazioni, i due assi cartesiani sono indicati con **Y_1** e **Y_2** invece di **X_1** e **X_2** come nel testo e nelle formule).

Calcolare il coefficiente angolare b della retta $\hat{X}_1$ che rappresenta l'asse principale e

che può essere scritta come

$$\hat{X}_1 = a + b \cdot X_2$$



Risposta. Ricordando di dover utilizzare più cifre decimali,

dopo aver calcolato

- le **due medie**

$$\overline{X}_1 = \frac{\sum_{i=1}^{n} X_{1i}}{n} = \frac{2.347}{12} = 195{,}5833$$

$$\overline{X}_2 = \frac{\sum_{i=1}^{n} X_{2i}}{n} = \frac{144{,}57}{12} = 12{,}0475$$

- le **due varianze**

$$S_1^2 = \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)^2}{n-1} = \frac{124.368,92}{11} = 11.306,26515$$

$$S_2^2 = \frac{\sum_{i=1}^{n}\left(X_{2i} - \overline{X}_2\right)^2}{n-1} = \frac{462,48}{11} = 42,04347$$

- e la **covarianza**

$$S_{12}^2 = \frac{\sum_{i=1}^{n}\left[\left(X_{1i} - \overline{X}_1\right)\cdot\left(X_{2i} - \overline{X}_2\right)\right]}{n-1} = \frac{6561,62}{11} = 596,51068$$

- si **ricava λ**

$$\lambda = \frac{S_1^2 + S_2^2 + D}{2} \quad \text{con } D = \sqrt{\left(S_1^2 + S_2^2\right)^2 - 4\cdot\left((S_1^2 \cdot S_2^2) - S_{12}^2\right)}$$

$$\lambda = \frac{11.306,26515 + 42,04347 + D}{2}$$

dove

$$D = \sqrt{\left(11.306,26515 + 42,04347\right)^2 - 4\cdot\left(11.306,26515\cdot 42,04347 - 596,51068\right)} = 11.327,22290$$

Da essi si ottiene

$$\lambda = \frac{11.348,30862 + 11.327,22290}{2} = \frac{22.675,53152}{2} = 11.337,76576$$

il risultato $\lambda = 11.337,76576$

e infine **il coefficiente** b

$$b = \frac{S_{12}^2}{\lambda - S_1^2} = \frac{596,51068}{11.337,76576 - 11.306,26515} = \frac{596,51068}{31,50061} = 18,9365$$

ottenendo b = 18,9365

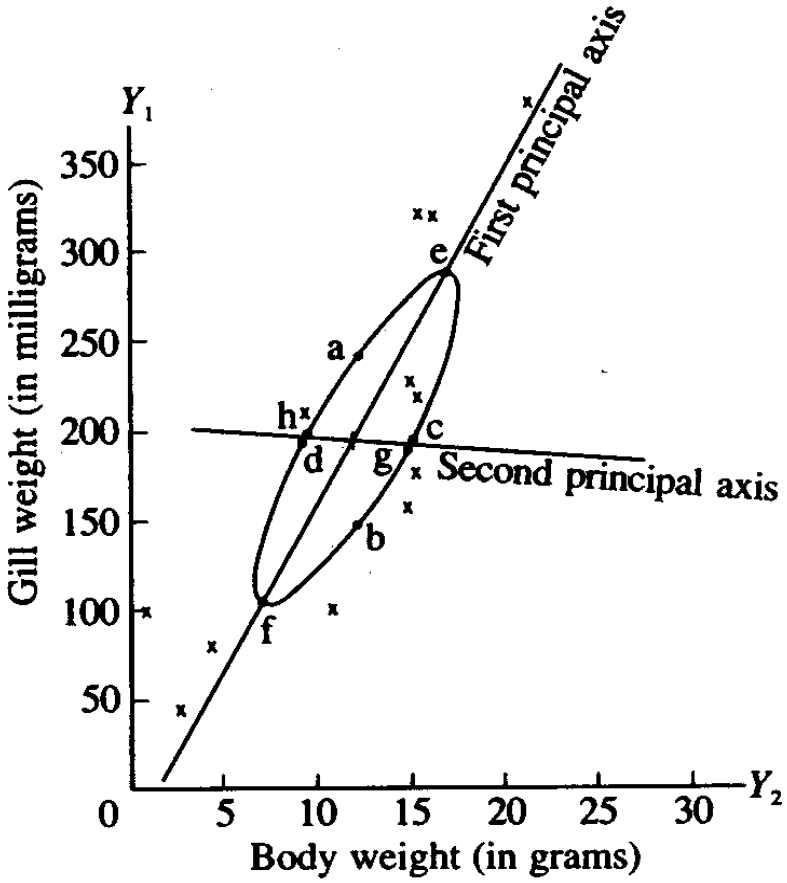
con il quale si ricava **l'intercetta** a

$$a = \overline{X}_1 - b \cdot \overline{X}_2 = 195{,}5833 - 18{,}9365 \cdot 12{,}0475 = 195{,}5833 - 228{,}1375 = -32{,}5542$$

e infine con entrambi si definisce **la retta dell'asse principale**

$$\hat{X}_1 = a + b \cdot X_2 = -32{,}5542 + 18{,}9365 \cdot X_2$$

La sua rappresentazione grafica è



(Come nel precedente diagramma di dispersione,

- le due variabili sono indicate con $\mathbf{Y_1}$ e $\mathbf{Y_2}$ invece di $\mathbf{X_1}$ e $\mathbf{X_2}$;
- i simboli x corrispondono ai punti della figura precedente)

Per approfondimenti sul calcolo del secondo asse e soprattutto

- sulla **trasposizione dei punti su questi assi,**
- sugli **autovalori** (***eighenvalues***, ***latent roots***, ***characteristic roots)***, che stanno alla base di molte **tecniche di statistica multivariata,**

si rimanda al testo citato di **Sokal** e **Rohlf** del 1995 ***Biometry*** dal titolo ***The Principles and Practice of Statistics in Biological Research*** (3rd ed. W. H. Freeman and Company, New York, XIX, + 887 p.).

24.3. IL PLOT DELLE DIFFERENZE E DELLE MEDIE; IL TEST DI BLAND-ALTMAN, PER IL CONFRONTO TRA METODI E PER LA RIPETIBILITA' DI UN METODO.

Con **due modi differenti di analisi o di determinazione quantitativa**, vengono effettuate le misure

- **su** n **campioni o casi indipendenti**,
- per verificare se **esista una relazione di tipo lineare tra le coppie di risultati ottenuti su gli stessi campioni**.

Quasi sempre, a questo scopo in letteratura è utilizzata la regressione lineare semplice. Ma le critiche sono numerose, in quanto **non sono rispettate almeno due condizioni di validità della regressione parametrica**:

1 - è del tutto **arbitrario e ingiustificato identificare un metodo come variabile indipendente** X e l'altro come **variabile dipendente** Y;

2 - le **due misure presentano lo stesso tipo di errore** e quindi **non è vero** che la **X sia fissa**, poiché **entrambe le variabili sono di tipo random**.

Ad esse, spesso deve essere aggiunta una terza e un quarta condizione che non rispettano quanto richiesto dal modello matematico:

3 – **L'errore commesso nella misura della variabile scelta come dipendente** Y **non è costante**, poiché spesso è proporzionale rispetto al suo valore o a quello della X, **anche quando i dati del campione (quasi sempre pochi) non risultano significativamente eteroscedastici ai test specifici**.

In termini più tecnici, questo concetto può essere espresso in due modi: la varianza σ^2 non è costante, **poiché ad essere costante** è

- il coefficiente di variazione CV

$$CV = \frac{\sigma}{\mu} \cdot 100$$

- il rapporto λ tra le due varianze

$$\lambda = \frac{\sigma_X^2}{\sigma_Y^2}$$

che (a) non è infinito e (b) non è zero, lungo il campo di variazione dei dati.

4 – La presenza di punti cosiddetti

- **dispersi** o **spuri** (***stragglers*** o ***spurious data***), definiti come misure o risultati analitici classificabili come sospetti al livello di fiducia del 95%, ma non al livello fiduciale del 99%),
- anche se essi non possono essere chiamati **dati anomali** (***outliers***), definiti come le misure o i risultati analitici che sono anormalmente diversi dai valori plausibili e che con un test statistico possono essere rigettati a un certo livello fiduciale.

E' possibile, ma dovrebbe essere evitata, anche la presenza di **errori grossolani** (***gross error*** o ***blunder***), un termine che identifica l'errore inaccettabile, quello che impone l'abbandono dell'analisi oppure di essere eliminato per mezzo di un controllo di qualità che sia efficiente.

Sono accettati solamente gli **errori casuali** (***random error***), quelli determinati da variazioni indefinite dei parametri sperimentali, quelli che sono caratterizzati quel livello di incertezza che è sempre collegata agli strumenti di misura. Gli errori casuali hanno una dispersione intorno al valore medio che è tanto più simmetrica quanto maggiore è il numero di osservazioni. Il livello di errore può essere ridotto, con strumenti più precisi e operazioni più accurate; ma non può mai essere totalmente azzerato.

Con il **metodo classico dei minimi quadrati** (***least-squares regression analysis***), nella regressione i **dati spuri** e gli **outliers** generano quadrati degli errori molto grandi. Quindi la retta calcolata è fortemente attratta verso di essi, allontanandosi sensibilmente dall'insieme di tutti gli altri punti.

Quando si deve decidere **se un metodo di misurazione è migliore di un altro**, è possibile utilizzare sia test, sia metodi grafici per la descrizione e l'analisi dei dati.

Tuttavia, l'indicazione fondamentale di James O. **Westgard** (attualmente ritenuto il maggior esperto di metodi clinici), nell'articolo del 1988 ***Points of care in using statistics in methods comparisons studies*** (editoriale della rivista **Clinical Chemistry**, Vol. 44, No. 11, pp.: 2240-2242) è:

- **Chi decide è l'individuo esperto della disciplina; la statistica con i metodi grafici e i test d'inferenza aiuta a decidere, ma non può mai sostituirsi all'esperto**.

Per chiarire ulteriormente questi concetti, è utile leggere quanto (a pag. 2240) scrive: ***The statistics do not directly tell you whether the method is acceptable; rather they provide estimates of errors that allow you to judge the acceptabilility of a method. You do this by comparing the amount of error observed with the amount of error that be allowable without compromising the medical use and interpretation of the test result. Methods performance is judged acceptable when the observed error is smaller than the defined allowable error. Method performance is not acceptable when the***

observed error is larger the allowable error. The decision-making process can be facilitated by mathematical criteria or by graphic tools.

IL TEST DI BLAND-ALTMAN

Per confrontare due metodi di misurazione, nei programmi informatici da alcuni anni è diffuso il metodo di J. M. **Bland** e D. G. **Altman,**

- proposto sinteticamente nell'articolo del 1986 ***Statistical methods for assessing agreement between two methods of clinical measurement*** (su **Lancet**, i, pp.: 307-310),
- successivamente ampliato nella lunga review del 1999 ***Measuring agreement in method comparison studies*** (su **Statistical Methods in Medical Research** Vol. 8, pp.: 135-169).

Su supponga di aver misurato n campioni, ottenendo per ognuno due dati quantitavi:

- la misura con il sistema 1 (X_1),
- la misura con il sistema 2 (X_2).

Bland e Altman propongono:

A – dapprima di ricavare da queste n coppie di misure altre due quantità:

- **la media ($\overline{X}$) delle due misure** per ognuno degli n campioni ($\overline{X}_i = \frac{X_{1i} + X_{2i}}{2}$),
- **la differenza (d) tra le due misure** per ognuno degli n campioni ($d_i = X_{1i} - X_{2i}$);

B – successivamente, di costruire **un grafico dei punti identificati da queste nuove coppie di valori**, nel quale

- **sull'asse delle ascisse è riportata la media X_i di ogni coppia**,
- **sull'asse delle ordinate la differenza d_i tra i valori della stessa coppia**.

Questo metodo risulta appropriato quando le differenze restano costanti. Ma spesso non lo sono.

Questa proposta è stata quindi integrata da altre due varianti, che cambiano il valore da riportare sull'asse delle ordinate. Con esse è diventato possibile scegliere tra tre opzioni, per il valore da riportare nelle ordinate:

1 - le differenze d_i tra le coppie di misure**:**

$$d_i = X_{1i} - X_{2i}$$

2 - le differenze d_i trasformate in percentuale p_i delle medie:

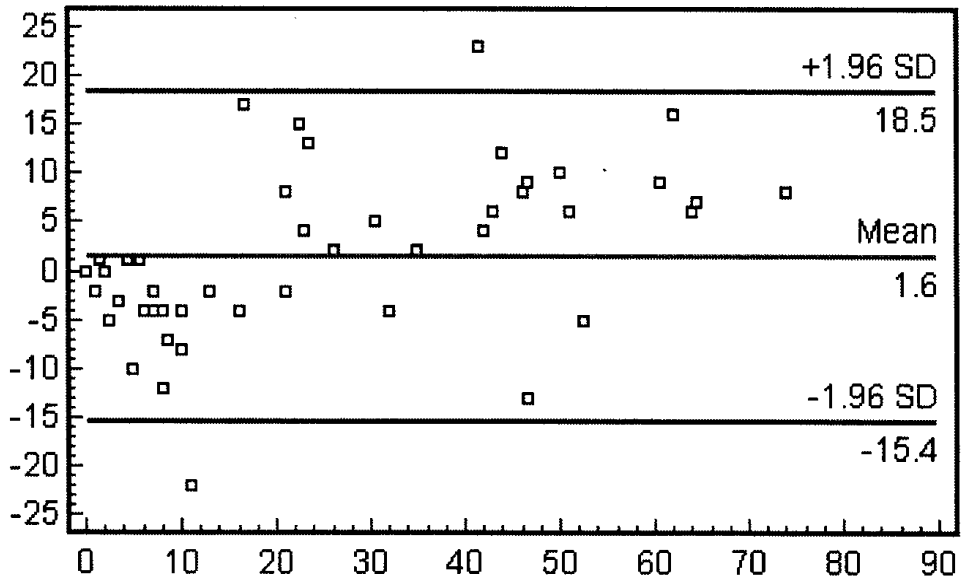
$$p_i = \frac{d_i}{\overline{X}_i} \cdot 100$$

3 - il rapporto r_i tra le due misure, trasformato in log:

$$r_i = \log\left(\frac{X_{1i}}{X_{2i}}\right)$$

Questa ultima formula richiede che non siano presenti valori uguali a zero, in nessuno dei due sistemi di misurazione. Se sono presenti, è necessario che nella trasformazione sia aggiunta una costante (il concetto è sviluppato nel capitolo sulle trasformazioni dei dati univariati).

La figura successiva rappresenta un esempio con $n = 46$.
Nel diagramma



Caso 1

- sull'**asse delle ascisse** è riportata **la media** $\overline{X}_i = \frac{X_{1i} + X_{2i}}{2}$,

- sull'**asse delle ordinate** è riportata **la differenza** $d_i = X_{1i} - X_{2i}$

L'analisi descrittiva dei dati richiede che per **le differenze** $d_i = X_{1i} - X_{2i}$ siano calcolate:

1 - il numero : $n = 46$

2 - la media aritmetica delle n differenze: $\overline{X}_d = 1{,}565$

3 - la deviazione standard: $s = 8{,}661$

4 - l'errore standard: $es = \frac{s}{\sqrt{n}} = \frac{8{,}661}{\sqrt{46}} = \frac{8{,}661}{6{,}782} = 1{,}277$

5 - l'intervallo di confidenza al 95% **della media delle differenze**: $IC = \overline{X}_d \pm 1{,}96 \cdot \frac{s}{\sqrt{n}}$.

Con i dati della figura, l'**intervallo di confidenza della media** delle differenze è

$$IC = 1{,}565 \pm 1{,}96 \cdot 1{,}277$$

quindi come

- limite inferiore (***lower limit***) ha $1{,}565 - 2{,}503 = -0{,}938$
- limite superiore (***upper limit***) ha $1{,}565 + 2{,}503 = +4{,}068$

6 – l'**intervallo di confidenza delle differenze**: $IC = \overline{X}_d \pm 1{,}96 \cdot s$.

Con i dati della figura, l'**intervallo di confidenza delle differenze è**

$$IC = 1{,}565 \pm 1{,}96 \cdot 8{,}661$$

quindi come

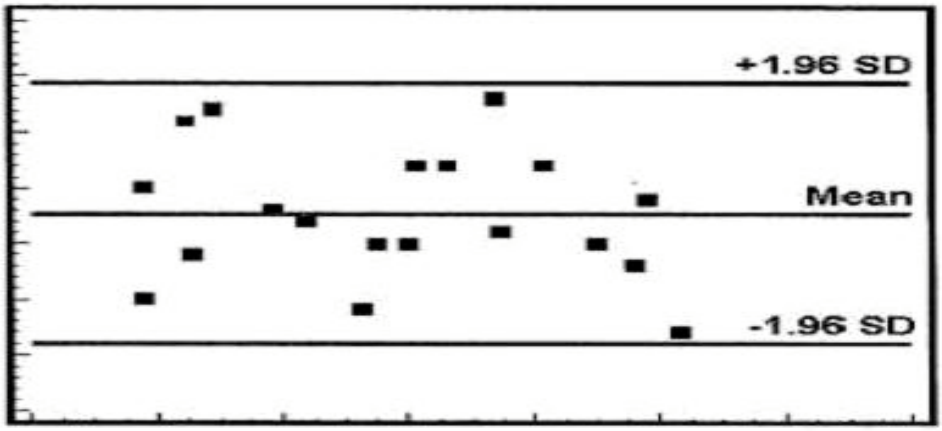
- limite inferiore (***lower limit***) ha $1{,}565 - 16{,}976 = -15{,}411$
- limite superiore (***upper limit***) ha $1{,}565 + 16{,}976 = +18{,}541$

L'informazione più importante del grafico è fornita congiuntamente

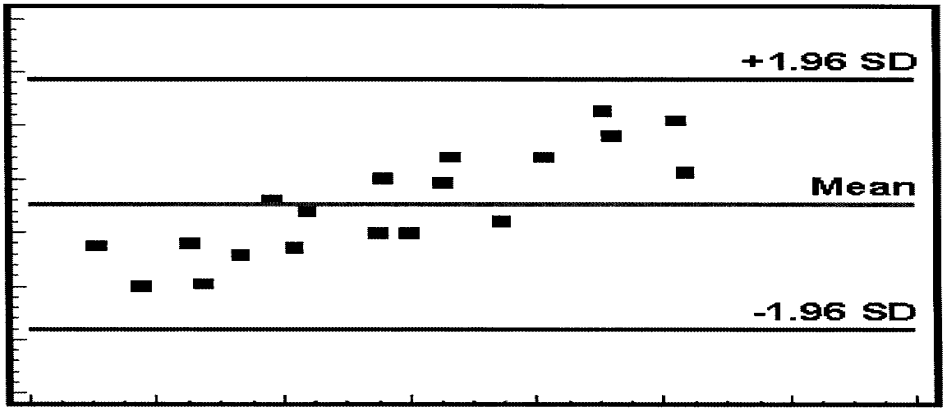
- dal **valore medio** $Mean = 1{,}6$ e
- dall'**intervallo di confidenza delle differenze** al 95% che varia tra + 18,5 e –15,4.

Il test di Bland-Altman consiste nel giudizio del ricercatore: se la variazione della media entro l'intervallo di confidenza non è clinicamente importante, i due metodi possono essere considerati intercambiabili.

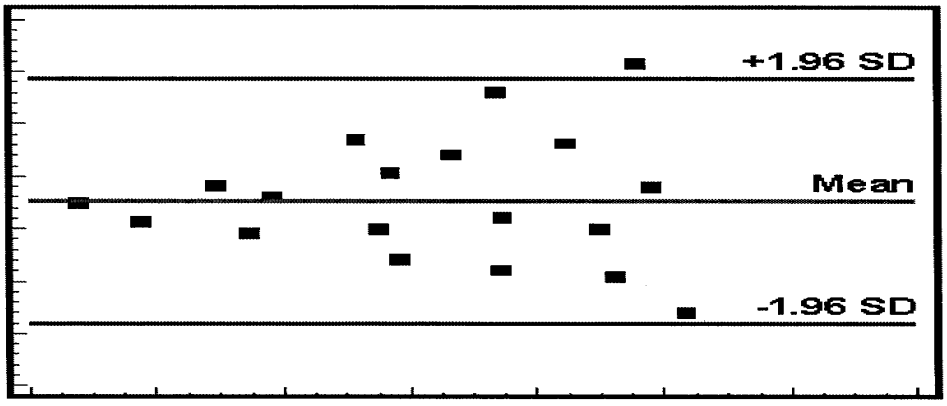
Non è quindi fondato su valori critici, ma sul giudizio dell'esperto della disciplina, come per la significatività della capacità predittiva di R^2 nella regressione least-squares.



Caso 2



Caso 3



Caso 4

Altre informazioni importanti sulla corrispondenza tra i due metodi di misurazione sono fornite dalla disposizione dei punti intorno alla media ed entro i limiti dell'intervallo di confidenza.

Le quattro figure riportate nel paragrafo sono rappresentative di altrettanti casi tipici.
- Nel caso 1, la disposizione dei punti è casuale: i due metodi possono essere ritenuti equivalenti, se l'analisi precedente sull'intervallo di variazione della media è positivo.

- Nel caso 2, i punti presentano un alternarsi periodico sopra e sotto la media: gli errori hanno una distribuzione non casuale, ma sistematica in valore assoluto, per cui i due metodi forniscono misure differenti.
- Nel caso 3, si ha un errore proporzionale: le differenze sono negative per valori piccoli, e positive per valori grandi.
- Nel caso 4, le differenze tra i due metodi non sono costanti, ma dipendono dal valore. Questo ultimo è un caso classico di **non uniformità della varianza** (***heteroscedasticity***).

Ne consegue che l'errore standard è una misura inadeguata o errata (***bias***) della variabilità, poiché è un valore medio di tutti gli errori e quindi sovrastima la variabilità quando i valori sono piccoli e la sottostima quando i valori sono grandi.
In questa situazione, è conveniente verificare sperimentalmente se si ottengono risultati migliori, modificando i valori da porre sull'asse delle ordinate. Le trasformazioni che con frequenza maggiore si dimostrano adeguate sono
- le differenze trasformate in percentuale p_i delle medie,
- il rapporto r_i tra le due misure, trasformato in log.

Il **test di Bland e Altman**, sia come proposto nella formula originaria, sia nelle sue varianti, può essere utilizzato anche quando si valuta
- la **ripetibilità** (***repeatability***) **di un metodo.**

Per ***repeatability*** (in italiano chiamata anche **ripetibilità ristretta)** si intende la bontà dell'accordo tra i risultati di misurazioni successive dello stesso misurando, condotte nelle stesse condizioni di misurazione. L'analisi dovrebbe essere condotta nello stesso laboratorio, dal medesimo operatore, con l'identico strumento, su campioni uguali, nelle stesse medesime operative e a breve distanza di tempo.
Quando uno o più di queste condizioni sono volutamente modificate, allo scopo di valutarne gli effetti mediante le differenze che determinano, si parla di **ripetibilità intermedia**.
Nel linguaggio internazionale, si adopera il termine ***intermediate precision***, tradotto in italiano anche con **precisione intermedia**.

Quando si impiega lo stesso metodo su una serie di n campioni, la media delle differenze dovrebbe essere zero.

E' possibile calcolare anche un **Coefficiente di Ripetibilità** (***Coefficient of Repeatability***) CR

$$CR = 1{,}96 \cdot \sqrt{\frac{\sum_{i=1}^{n} (X_1 - X_2)^2}{n-1}}$$

dove

- X_1 e X_2 sono le due misure ottenute nelle due condizioni sullo stesso campione,
- $1{,}96$ è il valore di Z alla probabilità $\alpha = 0.05$ bilaterale.

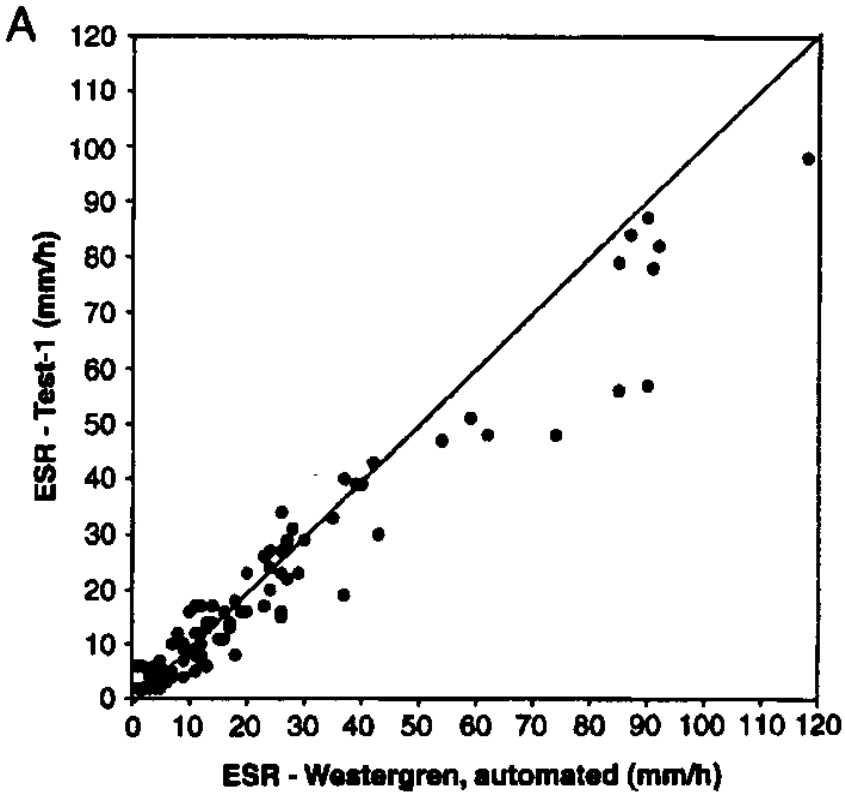
Benché i metodi proposti da Bland e Altman risultino generalmente più adatti dei metodi fondati sulla correlazione e sulla **regressione dei minimi quadrati** (***least-squares regression***), anche essi in alcune situazioni presentano limiti. Attualmente sono giudicate soluzioni ottimali quelle ottenute con

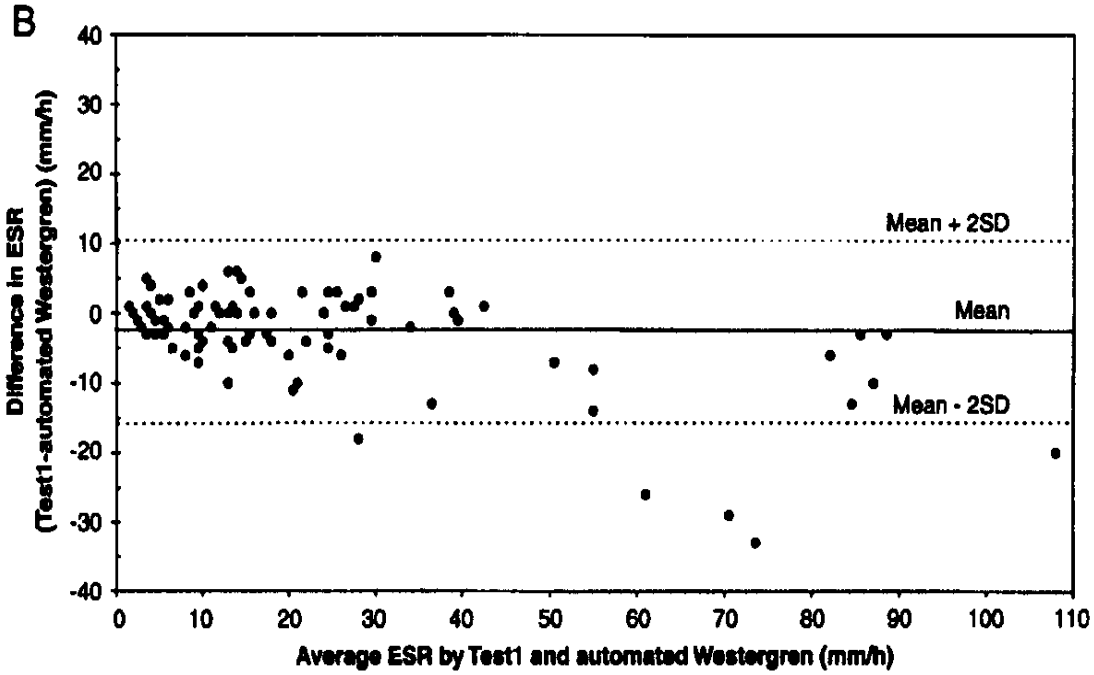
- la **regressione dei minimi prodotti** (***least-products regression***).

Il **test di Bland e Altman** è spesso impiegato in associazione alla stima della regressione lineare, ottenuta con il metodo ***least-squares*** oppure con il metodo ***least-products,*** illustrato nei paragrafi successivi, per valutare se due metodi analitici sono intercambiabili.
Confrontare il **plot di Bland-Altman** con il **diagramma di dispersione dei dati**, ottenuto con le misure originarie forniti dai due metodi a confronto, permette di meglio comprendere le relazioni presenti tra essi.

Le due figure rappresentate nella pagina successiva sono tratte dall'articolo dell'anno 2000 ***Erythrocyte Sedimentation Rate by the Test-1 Analyzer*** (pubblicato su **Clinical Chemistry** Vol. 46, No.6, pp.: 881 - 882). In esse sono riportati i risultati di 105 misure campionarie del tasso di sedimentazione degli eritrociti misurato in mm/h (indicato con **ESR** da ***Erythrocyte Sedimentation Rate***) ottenute su campioni di sangue con

- l'**automated Westergren method** (Starrsed, Charles Goffin Mwdical System), che rappresenta il metodo di confronto, già riconosciuto come valido,
- il metodo **Test-1, tripotassium ADTA-anticoagulated**, che rappresenta il metodo nuovo, di cui si vuole verificare la validità o l'intercambialità con il precedente.


A
ESR - Test-1 (mm/h)
ESR - Westergren, automated (mm/h)
0
10
20
30
40
50
60
70
80
90
100
110
120


B
Difference in ESR
(Test1-automated Westergren) (mm/h)
Average ESR by Test1 and automated Westergren (mm/h)
Mean + 2SD
Mean
Mean - 2SD
40
30
20
10
0
-10
-20
-30
-40
0
10
20
30
40
50
60
70
80
90
100
110

La figura A rappresenta l'analisi della regressione di **Passing-Bablock** (illustrata in un paragrafo successivo).

Come in tutti i diagrammi di dispersione, costruiti con i valori ottenuti nel confronto tra due metodi,

- **sull'asse delle ascisse (** X **)** sono riportati i valori ottenuti con il **metodo storico**, la cui validità è già riconosciuta,
- **sull'asse delle ordinate** (Y) sono riportati i valori ottenuti con il **metodo nuovo**, di cui si vuole verificare la validità o intercambialità con il precedente, a motivo dei vantaggi che può offrire (ad esempio: costi minori, tempi più brevi, prodotto nuovo di altra ditta, ecc. …)

La figura B rappresenta il **plot di Bland-Altman** degli stessi dati. L'intervallo di confidenza è 2SD e non 1,96SD come in precedenza. E' una approssimazione impiegata in molte pubblicazioni e molti test in cui si usa la distribuzione normale.

24.4. LA REGRESSIONE MODELLO II O LEAST-PRODUCTS DI DEMING, PER IL CONFRONTO TRA DUE METODI ANALITICI.

Quando si confrontano i risultati delle determinazioni quantitative di due metodi differenti,

- **gli errori di misura sono uguali,** sia per la variabile indicata con X , sia per quella con Y .

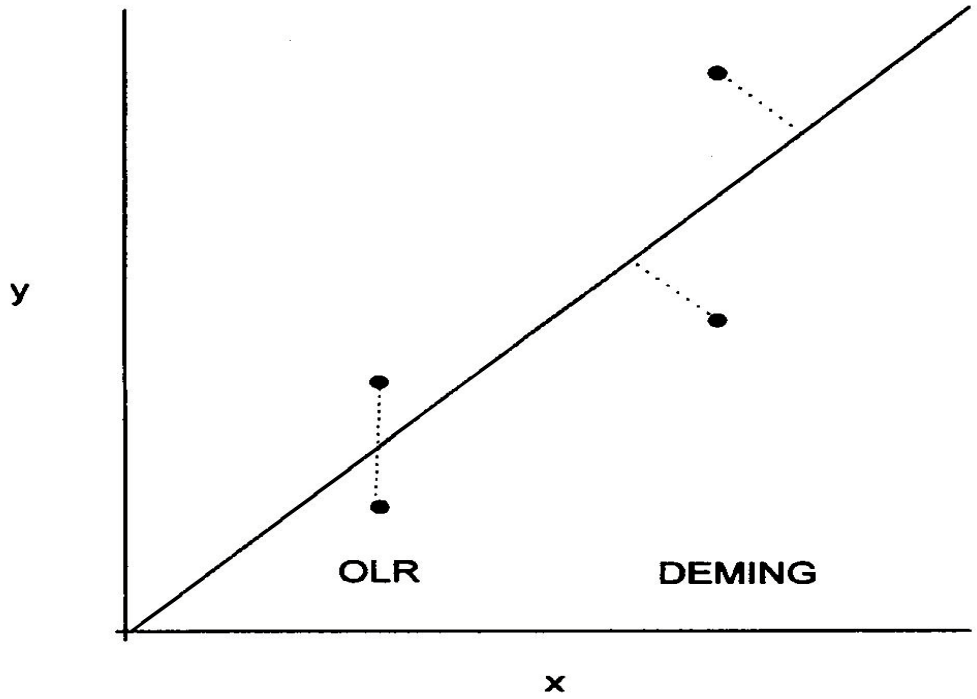


Diagramma di dispersione dei punti ottenuti con due metodi analitici, analizzati con il metodo dei minimi quadrati e il metodo di Deming

La retta di regressione classica,

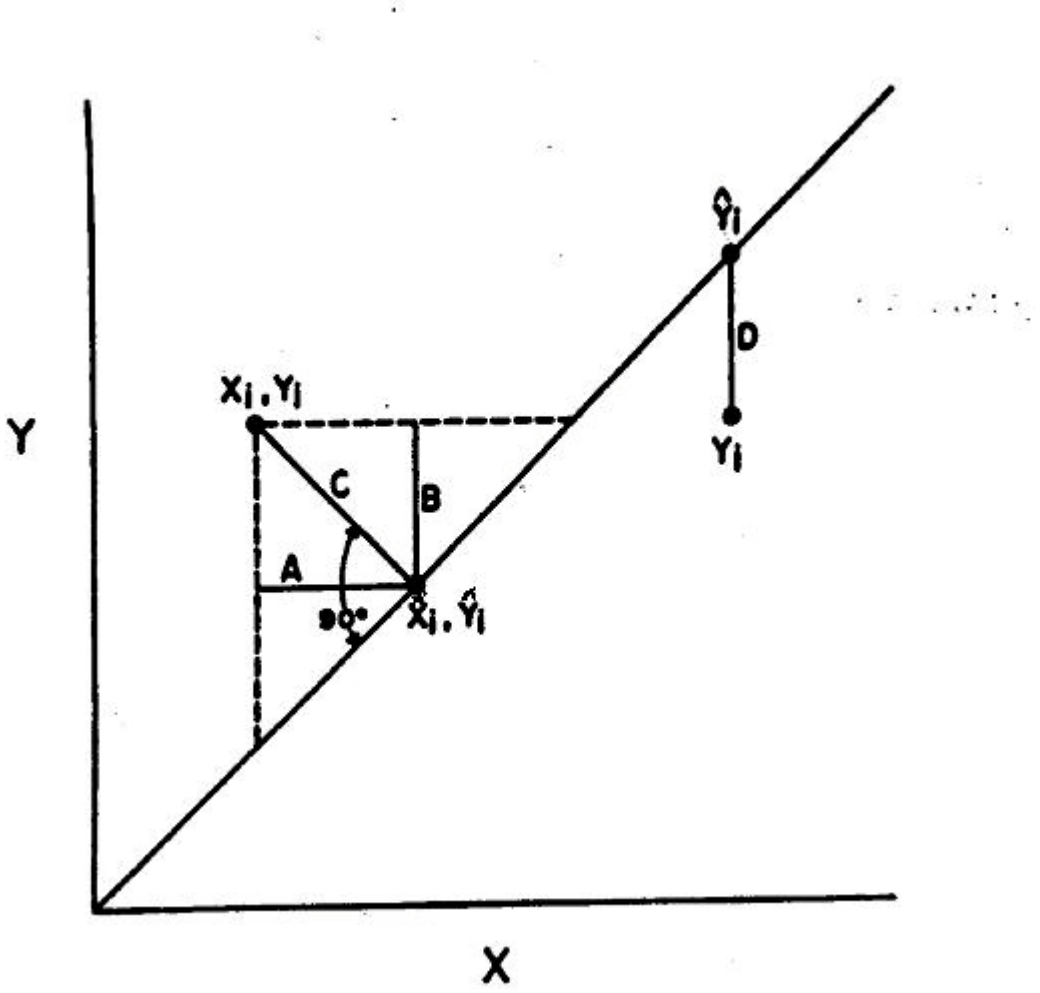
- chiamata ***Ordinary Least-Squares Regression*** (abbreviata in **OLR** oppure in **LSR**)
- e fondata sul **quadrato degli errori della sola variabile Y**, non è più adeguata.

Già nell'anno 1943, W. E. **Deming** nel volume ***Statistical Adjustement of Data*** (John Wiley and Sons, New York, NY) suggerisce una **alternativa statistica tecnicamente corretta,** per calcolare la **relazione lineare** esistente tra i due metodi di misurazione. Essa

- è fondata sul principio dei **minimi quadrati della distanza perpendicolare alla retta**, data dal prodotto (***least-products***) della distanza simultanea lungo l'asse della variabile X e l'asse della variabile Y (***minimizing the sum of the squares of the residuals in both the*** X ***and*** Y ***directions simultaneously***).

La **retta di Deming** o ***least-products*** è quella che

- **rende minima la somma dei quadrati delle distanze perpendicolari tra i punti e la retta** (come illustrato nella figura precedente e, in modo più dettagliato, in quella successiva).



Modello di regressione di Deming (a sinistra) e dei minimi quadrati (a destra)

La **retta di Deming** considera la somma dei quadrati dei residui tra il punto e la retta

- sia lungo l'asse delle ascisse X con $A^2 = \left(X_i - \hat{X}_i\right)^2$,

- sia lungo l'asse delle ordinate Y con $B^2 = \left(Y_i - \hat{Y}_i\right)^2$,

e minimizza la distanza $C^2 = A^2 + B^2$.

Il suo coefficiente angolare b su molte riviste di statistica applicata è indicato con

$$b_{Y \bullet X}.$$

mentre quello least-squares è indicato spesso con

$$b_{Y/X}$$

Trascurata per diversi decenni, questa metodologia statistica è oggi proposta in molti programmi informatici, scritti appositamente per le analisi cliniche e chimiche. Sulle riviste di biologia e ecologia è chiamata **Model II Regression**, alternativa corretta alla **Model I Regression**, discussa nei capitoli precedenti.

Tuttavia questa classificazione dicotomica, diffusa da Robert R. **Sokal** e F. James **Rohlf** già nel 1969 nella prima edizione del loro testo ***Biometry. The principles and practice of statistics in biological research*** (3rd ed. W. H. Freeman and Company, New York, XIX, + 887 p.) e confermata nelle due edizioni successive del 1981 e del 1995, è criticata in quanto appare poco utile alla esatta comprensione delle differenze tra due metodi di regressione. E' quanto sostiene, ad esempio, anche Brian H. **McArdle** nell'articolo del 2003 ***Lines, models, and errors: Regression in the field*** (pubblicato sulla rivista della The American Society of Limnology and Oceanography, **Limnol. Oceanogr**. Vol. 48 (3), pp.: 1363 - 1366). Egli dichiara di preferire la classificazione in

- metodi con **relazioni asimmetriche** (***asymmetric relationships***) tra le due variabili, come la **least-squares regression**,
- metodi con **relazioni simmetriche** (***symmetric relationships***) tra le due variabili, come la **least-products regression**.

Una presentazione dettagliata del **metodo di Deming** è stata effettuata da P. Joanne **Cornbleet** e Nathan **Gochman** nell'articolo del 1979 ***Incorrect Least-Squares Regression Coefficients in Method-Comparison Analysis*** (pubblicato su **Clinical Chemistry**, Vol. 25, No. 3, pp.: 432-438).

La loro pubblicazione comprende il confronto con il metodo di **Mandel** (illustrato nella pagine successive) e quello non parametrico di **Bartlett** (presentato nel capitolo della regressione lineare non parametrica).

Per questo articolo, **Cornblett** e **Gochman** sono indicati come coloro che hanno avuto il merito di portare il **metodo di Deming** all'attenzione definitiva della collettività scientifica, che finalmente in larga parte lo considera

- **l'alternativa corretta all'impostazione classica, nel caso del confronto tra due metodi di misurazione**.

Da questo articolo sono ripresi alcune figure e molti concetti di questo paragrafo.

L'analisi della regressione di Deming non attribuisce pesi differenti ai valori (*the unweighted form of Deming regression analysis*) e

- **è appropriata quando l'errore analitico è costante**.

In altri termini, è da ritenere corretta quando la deviazione standard è indipendente dalle dimensioni delle misure.

Quando invece

- l'**errore è costante** come **percentuale del valore** oppure come **coefficiente di variazione** rispetto al valore (i due concetti sono analoghi),

è opportuno utilizzare una delle modifiche successive. Alcune di esse, le più diffuse nella ricerca applicata e più frequentemente citate nelle riviste di analisi cliniche e chimiche, sono presentate nei paragrafi successivi.

Le **condizioni di validità della retta di Deming**, con una formulazione matematica più appropriata, affermano che

- il valore osservato X_i è

$$X_i = x_i + \varepsilon_i$$

- il valore osservato Y_i è

$$Y_i = y_i + \eta_i$$

con $i = 1,2,...,n$.

Le assunzioni su questi **errori random** sono quattro:

1 - ε e η sono entrambi distribuiti in modo **normale** e con **media zero**;

2 - le coppie $\varepsilon_i, \varepsilon_j$ e η_i, η_j sono distribuite in modo indipendente per $i \neq j$: **le covarianze entro metodi** (***within-method covaraince***s) sono uguali a **zero**;

3 - le coppie ε_i, η_i sono distribuite in modo indipendente per ogni i: **la covarianza tra metodi** (***between-method covariance***) è uguale a zero;

4 - le varianze σ_ε^2 e σ_η^2 non sono costanti, ma **è costante il loro rapporto** $\lambda = \dfrac{\sigma_\varepsilon^2}{\sigma_\eta^2}$

Con n coppie di misure, ottenute con i due metodi X e Y che si intende confrontare, per ricavare la **retta di Deming,** come al solito si devono prima calcolare

- la Devianza delle X **:** $DevX = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$

- **la Devianza delle** Y **:** $DevY = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$

- la **Codevianza** $X \cdot Y$: $CodevXY = \sum_{i=1}^{n} \left(X_i - \overline{X}\right) \cdot \left(Y_i - \overline{Y}\right)$

- e dalle due devianze il loro **rapporto** λ

$$\lambda = \frac{DevX}{DevY}$$

Da queste, si ricava il **coefficiente angolare** b

$$b = \left(\lambda \cdot DevY - DevX\right) + \frac{\sqrt{\left(DevX - \lambda \cdot DevY\right)^2 + \left(4\lambda \cdot \left(codXY\right)^2\right)}}{2\lambda \cdot CodXY}$$

e con esso l'intercetta a mediante le medie:

$$a = \overline{Y} - b\overline{X}$$

La retta di **regressione di Deming** è

$$\hat{Y}_i = a + b\left(X_i - \overline{X}\right)$$

L'errore standard dell'intercetta a **e del coefficiente angolare** b **sono stimate con la procedura Jackknife** (alla quale si rimanda, per la presentazione del metodo).

In alcuni esperimenti, i valori di X_i **e di** Y_i **per lo stesso campione misurando sono ripetute due volte**. In questo caso, nelle formule precedenti,

- il valore X_i è la media delle due repliche

$$X_i = \frac{X_{1i} + X_{2i}}{2}$$

- il valore Y_i è la media delle due repliche

$$Y_i = \frac{Y_{1i} + Y_{2i}}{2}$$

Le **deviazioni standard analitiche** (***analitical standard deviations***) dei metodi x e y possono essere calcolate rapidamente come differenza tra le due misure:

$$SD_{ax}^2 = \frac{\sum_{i=1}^{n} (X_{2i} - X_{1i})^2}{2n}$$

e

$$SD_{ay}^2 = \frac{\sum_{i=1}^{n} (Y_{2i} - Y_{1i})^2}{2n}$$

mentre il coefficiente λ diventa

$$\lambda = \frac{SD_{ax}^2}{SD_{ay}^2}$$

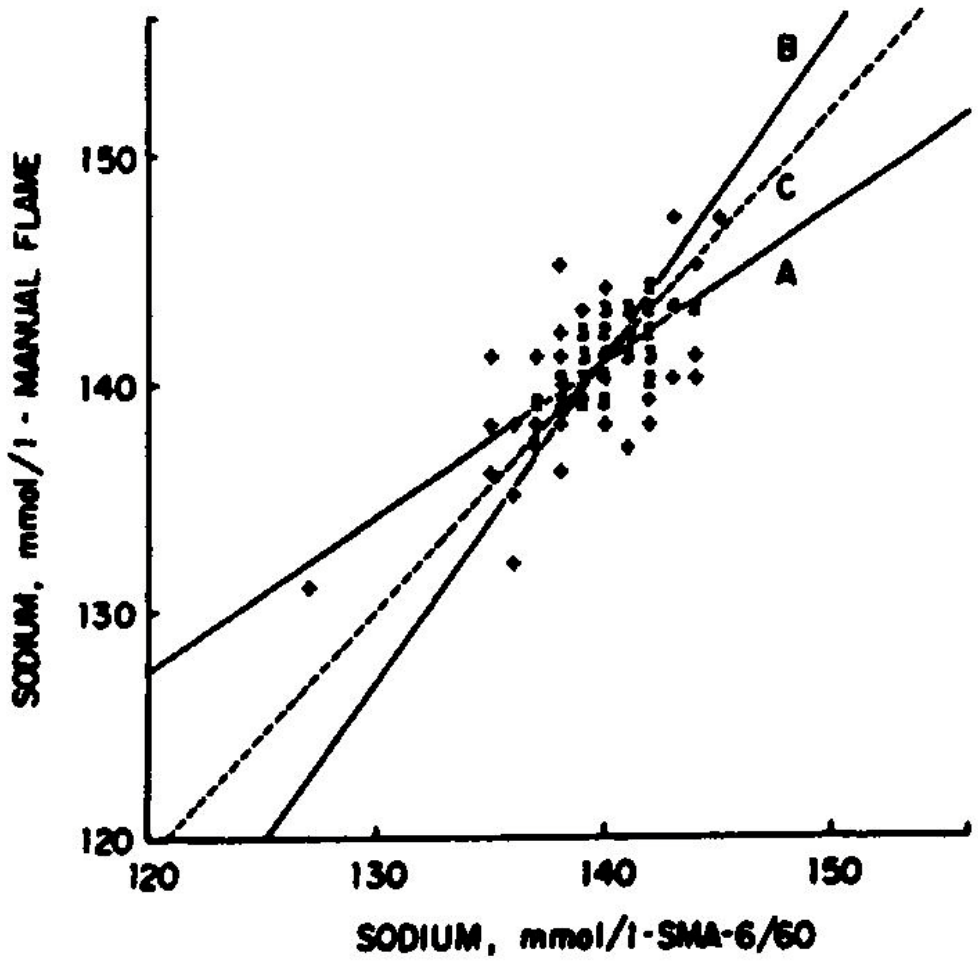
ESEMPIO. (LE DUE RETTE LEAST-SQUARES E LA RETTA LEAST-PRODUCTS DI DEMING, con S_X PICCOLO). Il numero di dati abitualmente raccolti in queste analisi e la lunghezza dei calcoli richiedono l'uso di programmi informatici. Pertanto, dopo aver presentato il metodo nei suoi passaggi logici, è utile un esempio che illustri i risultati.
Tratta dall'articolo di **Cornbleet** e **Gochman**, la figura della pagina successiva mostra il **plot** di **87 determinazioni di sodio (mmol/L)**, ottenuti con **due metodi flame-photometric**:

- il **metodo di riferimento**, riportato sull'asse delle X ,

ha media $\overline{X}$ = 139,8 e deviazione standard S_x = 2,67

- il **metodo da testare**, riportato sull'asse delle Y ,

ha media $\overline{Y}$ = 140,7 e deviazione standard S_x = 2,60



Tratta da P. Joanne **Cornbleet** e Nathan **Gochman** del 1979 ***Incorrect Least-Squares Regression Coefficients in Method-Comparison Analysis*** (pubblicato su **Clinical Chemistry**, Vol. 25, No. 3, pp.: 432-438) a pag. 436.

Con $n = 87$ (i punti sono in numero minore perché alcuni hanno dati uguali),
A – Regressione **Least-Squares**, con X come variabile indipendente:

$$\hat{Y}_i = 47{,}6 + 0{,}667 \cdot X_i$$

B – Regressione **Least-Squares**, con Y come variabile indipendente:

$$\hat{Y}_i = -59{,}2 + 1{,}43 \cdot X_i$$

C – Regressione di **Deming**:

$$\hat{Y}_i = -11{,}7 + 1{,}09 \cdot X_i$$

Benché la logica indichi che

- **il coefficiente angolare della retta di regressione dovrebbe essere** b **= 1,0**

(se fosse vera l'ipotesi nulla che i due metodi sono equivalenti)

- la **regressione least-squares** fornisce $b_{Y/X}$ = 0,66.

Se si scambiano le due variabili X e Y, rendendo Y la variabile indipendente,

- la **regressione least-squares** fornisce la stima $b_{X/Y}$ = 1,43.

La **retta di Deming** fornisce la stima $b_{X \bullet Y}$ = 1,09.

Essa quindi appare una stima nettamente migliore, rispetto alle due precedenti, della relazione tra X e Y che viene stimata dai dati.

La sua significatività può essere verificata con vari metodi, tra cui principalmente

- il test di Bland-Altman (già illustrato in un paragrafo precedente)
- il test di Passing-Bablok (illustrato in un paragrafo successivo).

24.5. EFFETTI DEGLI OUTLIER SULLA RETTA LEAST-SQUARES E INDICAZIONI OPERATIVE PER IL CALCOLO DELLA RETTA DI CONFRONTO TRA DUE METODI ANALITICI.

L'**individuazione degli outlier** nella statistica bivariata richiede metodi specifici che sono già stati descritti separatamente.

La loro presenza determina problemi non trascurabili, quando si deve calcolare la retta di regressione. Quindi sono stati presentati metodi che permettono di identificarli con facilità, quando si richiede una **stima più precisa della retta che deve descrivere la relazione vera tra due metodi di misura**.

Nel 1966 N. R. **Draper** e H. **Smith** nel volume ***Applied Regression Analysis*** (John Wiley and Sons, New York, NY, pp.: 44-103) suggeriscono:

- i punti che generano **residui maggiori di** $4 \cdot S_{Y/X}$
- possono essere **eliminati nell'analisi della regressione least-squares**.

L'analisi dei residui intorno alla retta di regressione può essere un criterio

- **per individuare** e successivamente **eliminare i dati spuri**
- e quindi **annullare i loro effetti sul coefficiente angolare della retta least-squares**,

che in tal modo, con $a \cong 0$ e $b \cong 1$, diventa utilizzabile anche nel confronto tra metodi.

ESEMPIO (ELIMINAZIONE DI DATI SPURI NELLA RETTA LEAST-SQUARES). La Figura riportata nella pagina successiva è il **plot di 169 campioni,** sui quali è stata fatta la determinazione (mg/dl) del calcio con due metodi differenti

- per assorbimento atomico o AAS (asse X),
- e per SMA 12-60 (asse Y).

Si tratta del confronto tra due metodi.

Perché essi possano essere considerati **intercambiabili** (ipotesi nulla), la retta **dovrebbe avere**:

- **intercetta** $a = \mathbf{0},0$
- **coefficiente angolare** $b = \mathbf{1,0}$.

Utilizzando **tutti i punti** (nel grafico il numero appare minore di 169 in quanto molti sono sovrapposti), si calcola la retta di Regressione **Least-Squares**, con X come variabile indipendente e indicata nel grafico con A:

$$\hat{Y}_i = 1{,}96 + 0{,}78 \cdot X_i$$

Essa ha un valore $a = 1{,}96$ e un valore $b = 0{,}78$ che sono lontani dall'atteso.

Inoltre, e ciò può essere la causa dello scostamento dall'atteso, il punto indicato con 1 risulta anomalo rispetto alla retta, in quanto

- dista più di $4 \cdot S_{Y/X}$ dal punto corrispondente sulla retta.

E' quindi possibile e conveniente

- eliminare la coppia di dati corrispondente a questo punto
- e calcolare una nuova retta **Least-Squares** (con $n = 168$), ottenendo

$$\hat{Y}_i = 0{,}39 + 0{,}95 \cdot X_i$$

Questa seconda retta **Least-Squares**, indicata con B è sensibilmente più vicino all'atteso.

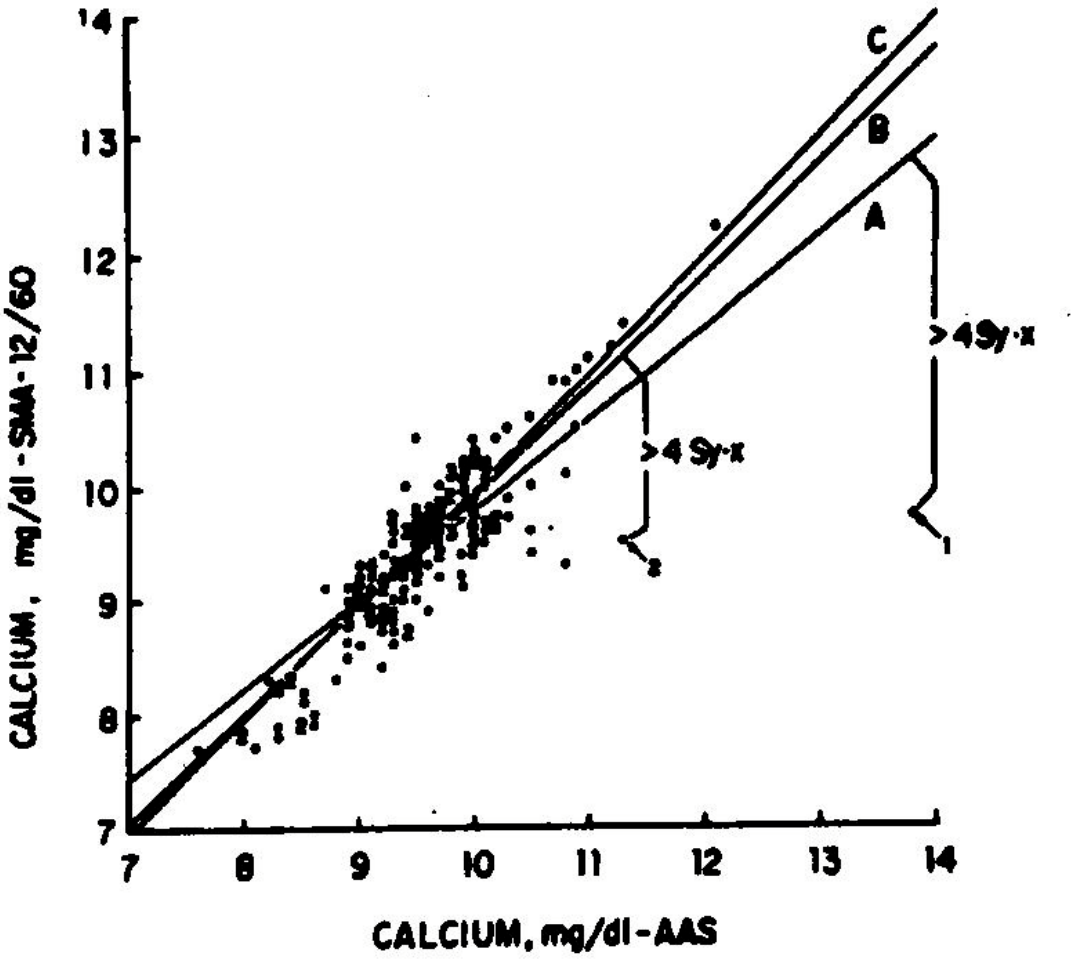
Ma essa evidenzia un altro punto che ora è diventato spurio rispetto ad essa.

Eliminando i due dati corrispondenti a questo punto 2 (con $n = 167$),

- si calcola una nuova retta **Least Squares**, ottenendo

$$\hat{Y}_i = 0{,}04 + 0{,}99 \cdot X_i$$

E' la retta indicata con C, che non evidenzia più valori spuri, in quanto nessun punto dista più di $4 \cdot S_{Y/X}$ dal punto verticale sulla retta.



Tratta da P. Joanne **Cornbleet** e Nathan **Gochman** del 1979 ***Incorrect Least-Squares Regression Coefficients in Method-Comparison Analysis*** (pubblicato su **Clinical Chemistry**, Vol. 25, No. 3, pp.: 432-438) a pag. 437.

Con n = 169 (i punti sono in numero minore perché alcuni hanno dati uguali),
A – Regressione **Least-Squares** calcolata con **tutti i punti**, con X come variabile indipendente: il punto 1 è anomalo, in quanto dista più di $4 \cdot S_{Y/X}$ dal punto corrispondente sulla retta

$$\hat{Y}_i = 1{,}96 + 0{,}78 \cdot X_i$$

B – Regressione **Least-Squares**, con X come variabile indipendente, **senza il punto 1**; il punto 2 è anomalo, in quanto dista più di $4 \cdot S_{Y/X}$ dal punto corrispondente sulla retta

$$\hat{Y}_i = 0{,}39 + 0{,}95 \cdot X_i$$

C – Regressione **Least-Squares**, con X come variabile indipendente, **senza i punti 1 e 2**:

$$\hat{Y}_i = 0{,}04 + 0{,}99 \cdot X_i$$

Con un valore dell'intercetta a molto vicino a 0 e un coefficiente angolare b prossimo a 1,

- **può essere una buona stima della relazione lineare attesa tra i due metodi**.

Nell'articolo di P. Joanne **Cornbleet** e Nathan **Gochman** del 1979 ***Incorrect Least-Squares Regression Coefficients in Method-Comparison Analysis*** (pubblicato su **Clinical Chemistry**, Vol. 25, No. 3, pp.: 432-438) la presentazione del **metodo di Deming** e la **dimostrazione degli effetti degli outliers sulla regressione lineare** si concludono (a pag. 437) con quattro gruppi di indicazioni operative (***guidelines for linear regression analysis***).

1 - Fare sempre il **diagramma di dispersione** (***plot***) dei dati e applicare l'analisi della regressione least-squares s**olamente nella regione di linearità**. Nel grafico, i valori sospettati di essere **outlier** risultano sempre con **evidenza maggiore** di quanto possa apparire alla lettura dei valori.

2 – E' sempre importante stimare l'errore nelle misure. Per un calcolo rapido, è sufficiente il **rapporto**

$$\frac{S_{eX}}{S_X}$$

dove

- S_{eX} rappresenta la precisione di una singola misura X vicino alla media $\overline{X}$.

Se questo **rapporto** eccede 0,2 si deve dedurre che

- nella stima del **coefficiente angolare** $b_{Y/X}$ **ottenuto con i minimi quadrati** è presente un **errore significativo**
- e quindi è più appropriato impiegare il **coefficiente angolare** $b_{Y\bullet X}$ **della retta di Deming.**

Sempre nel dibattito sulla correttezza della regressione least-squares, quando i dati sono marcatamente asimmetrici, come indica il fatto che

- la **deviazione standard** delle X (S_X) sia **maggiore della media** $\overline{X}$

$$S_X > \overline{X}$$

- e l'**errore nelle misure è proporzionale al valore** di X,

un **rapporto** uguale o maggiore di 0,15 indica che

- il coefficiente angolare $b_{Y/X}$ della retta **least-squares ha un errore significativo.**

3 - Il calcolo del coefficiente angolare $b_{Y\bullet X}$ **della retta di Deming**

richiede che sia calcolato il rapporto

$$\frac{S_{eX}}{S_{eY}}$$

Queste stime possono essere ottenute,

- sia dalla precisione dell'analisi di un singolo campione che sia vicino alla media dei dati,
- sia, quando si dispone di due repliche per ogni campione,

con

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}\left(differenza\ \ tra\ \ repliche\right)^2}{2n}}$$

Se, per calcolare la retta di regressione, invece dei singoli valori delle due repliche(X_{i1}, X_{i2} e Y_{i1}, Y_{i2}) ottenute dallo stesso campione, si vuole impiegare solamente la media ($\overline{X}_i, \overline{Y}_i$), è utile ricordare che

- la **deviazione standard della misura medi** è uguale alla **deviazione standard calcolata sui singoli valori divisa per** $\sqrt{2}$.

4 – E' sempre necessario calcolare l'**errore standard della regressione** S_{YX} **.**.

Sia per la retta **il coefficiente angolare** $b_{Y/X}$ **least-squares**, sia per **quello** $b_{Y\bullet X}$ **last-products** di **Deming**, questa statistica può essere ricavata dalle statistiche abitualmente calcolate su dati bivariati, con

$$S_{YX} = \sqrt{\frac{n-1}{n-2}\left(SY^2 - b_{YX}\cdot r\cdot S_X\cdot S_Y\right)}$$

L'errore standard della regressione S_{YX} deve essere interpretata come l'errore standard del valore medio atteso di Y per un dato valore di X , collocato vicino alla media $\overline{X}$.

E' una misura della dispersione dei punti intorno alla retta di regressione.

24.6. LA FORMULA RAPIDA DI MANDEL E LA REGRESSIONE LEAST-PRODUCTS DI YORK.

Nell'anno 1964 J. **Mandel** nel volume ***The Statistical Analysis of Experimental Data*** (pubblicato da John Wiley and Sons, New York, NY) a pag. 290-291 propone due formule abbreviate per passare direttamente

- dal coefficiente angolare $b_{Y/X}$ **least-squares** al coefficiente angolare $b_{Y \cdot X}$ **least-products** di **Deming**

$$b_{Y \bullet X} = b_{Y/X} \cdot \left(1 + \frac{S_{eX}^2}{S_X^2 - S_{eX}^2}\right) = \frac{b_{Y/X}}{1 - \frac{S_{eX}^2}{S_X^2}}$$

dove

- S_{eX}^2 = varianza d'errore di un singolo valore di X, vicino alla media X
- S_X^2 = varianza delle X.

Per questo passaggio da un coefficiente angolare all'altro, è sufficiente calcolare la varianza (S_{eX}^2) di un singolo valore di X, una misura che si ricava rapidamente dal campione di osservazioni.

La **formula di Deming** è fondata

- **non** sul presupposto che la varianza del metodo X e quella del metodo Y **siano costanti,**
- ma sul concetto più generale che **il rapporto** λ tra esse sia costante:

$$\lambda = \frac{DevX}{DevY} = \frac{\sigma^2(\varepsilon_i)}{\sigma^2(\eta_i)}$$

Tuttavia anche questa ipotesi, in varie situazioni sperimentali, è lontana dalla realtà. Le due varianze hanno **un rapporto che si modifica** entro il campo di variazione delle osservazioni.

Nell'anno 1966 J. **York** nell'articolo ***Least squares fitting of a straight-line*** (pubblicato su **Can. J. Phys**. Vol. 44, pp.: 1079 – 1086) ha proposto una **soluzione** che consiste nel

ridurre al minimo la funzione

$$\sum_{i=1}^{n}\left[\frac{(x_i - X_i)^2}{\sigma^2(\varepsilon_i)} + \frac{(y_i - Y_i)^2}{\sigma^2(\eta_i)}\right]$$

dove

- x_i, y_i sono i punti collocati sulla retta
- X_i, Y_i sono i punti osservati,

con $X_i = x_i + \varepsilon_i$ e $Y_i = y_i + \eta_i$ per $i = 1,2,...,n$.

In questa formula, le due varianze al denominatore sono calcolate entro ogni gruppo e sono in realtà dei fattori pesati. E' un metodo che **generalizza la precedente proposta di Deming.**
La soluzione è raggiunta attraverso un processo iterativo, per il quale occorre disporre del programma informatico.

24.7 LA REGRESSIONE LINEARE E IL TEST PER L'EQUIVALENZA TRA DUE METODI ANALITICI DI PASSING-BABLOK

Nella ricerca di laboratorio e nelle applicazioni industriali, è sempre forte l'esigenza di ridurre i tempi e i costi delle analisi quantitative, con metodi nuovi e meno costosi. Nelle determinazioni chimiche e cliniche, sovente si propongono metodi più raffinati o più rapidi oppure basati su un principio differente, per valutare la quantità di principio attivo presente. Si tratta di decidere se i risultati sono equivalenti e quindi se i due metodi di misurazione sono uguali.
Per questi problemi di chimica analitica e chimica clinica, H. **Passing** e W. **Bablok** con gli articoli
- del 1983 ***A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry***. ***Part I***. (pubblicato sulla rivista **Journal of Clin. Chem. Clin. Biochem**., Vol. 21, pp.: 709 - 720),
- del 1984 ***Comparison of several regression procedures for method comparison studies and determination of sample size. Application of linear regression procedures for method comparison srudies ib clinical chemistry. Part II*** (pubblicato sulla rivista **Journal of Clin. Chem. Clin. Biochem**., Vol. 22, pp.: 431 - 445),
propongono un **test statistico** per verificare se i due metodi forniscono le stessa misura.
Il problema è da essi presentato (1983, pag. 710) in questi termini:
- ***There are 2 different methods (instruments) which measure the same chemical analyte in a given medium (e.g. serum, plasma, urine, …). The question is: Do the methods measure the same concentration of the analyte or is there a systematic difference in the measurements? (For simplicity, we only refer to concentration but our statements are also valid for any other quantity.)***

Quando, in n campioni indipendenti estratti da una popolazione, un dato analita è misurato con valore X_i e Y_i per ogni campione i, si ottengono

$$X_i = x_i + \varepsilon_i$$

e

$$Y_i = y_i + \eta_i$$

dove

- x_i e x_i sono i valori stimati dalla retta per lo stesso campione i,

- ε_i e η_i sono gli errori delle due misure.

Con questa impostazione logica,

- **ognuno dei due metodi ha il suo valore atteso** X_i e Y_i **per il campione** i

- e se esiste una **relazione strutturale tra i due metodi**, essa può essere **descritta dalla** generica **equazione lineare**

$$\hat{Y}_i = \alpha + \beta \hat{X}_i$$

Al momento della pubblicazione del primo articolo di **Passing** e **Bablok** (anno 1983), la **relazione strutturale tra i due metodi** era ricavata con 4 metodi statistici:

1 - la **regressione lineare** Y / X

$$\hat{Y}_i = a + bX_i$$

2 - la **regressione lineare** X / Y

$$\hat{X}_i = a + bY_i$$

3 - la procedura di **Deming**, che **Passing** e **Bablok** chiamano **principal component analysis**,

4 - la **standardized principal component analysis.**

Essi hanno condizioni di validità differenti e forniscono rette differenti (già illustrate nei paragrafi e nei capitoli precedenti).

Nella **impostazione concettuale di Passing e Bablok**, per un vera analisi di ***comparability*** **tra due metodi** mediante la regressione lineare, è necessario raggiungere **quattro obiettivi**:

1 – calcolare l'**intercetta** a e il **coefficiente angolare** b;

2 – verificare con un test se **sussistono le assunzioni per la linearità**.

Solamente dopo **aver dimostrato il punto due,**

3 – **verificare l'ipotesi** $\beta = 1$,

4 - **verificare l'ipotesi** $\alpha = 0$.

I due ultimi test sono fondati sul concetto che, se i **due metodi sono equivalenti** e senza errori, **i punti dovrebbero collocarsi lungo la retta che ha** $\beta = 1$ e $\alpha = 0$**.** Ovviamente occorre considerare la presenza di errori random, ma scoprire **errori costanti** (**differenze sistematiche)** tra i due metodi.

Il metodo di **Passing** e **Bablok** non ha **una dimostrazione teorica**, ma solo verifiche empiriche fondate su simulazioni. E' utilizzato su campioni grandi, che hanno varie decine di coppie di osservazioni, per cui già nella presentazione da parte degli autori sono stati proposti programmi informatici, da essi scritti in PASCAL e BASIC, che rendono possibile e semplice l'applicazione del metodo.
Da alcuni anni, questi programmi per il test di equivalenza tra due metodi sono rintracciabili nelle librerie informatiche per analisi cliniche e chimiche, in quanto la procedura ora è ritenuta la prassi corretta per valutare l'equivalenza tra due metodi.

I concetti di base della **regressione lineare di Passing e Bablok** sono simili a quelli della **retta non parametrica di Theil** (vedi capitolo relativo), che ovviamente sviluppò il metodo senza pensare al confronto tra metodi. Quindi anche la **regressione di Passing e Bablok** è classificata tra i **metodi non parametrici**.
Nell'articolo del 1983, i due autori (a pag.713) scrivono: ***The basic concept of our regression procedure is due to Theil who developed this idea without reference to the problem of method comparison.***

La differenza concettuale più importante tra questi due metodi è che mentre la retta non parametrica di **Theil** assume, come avviene per la retta parametrica **least-squares**, che la X sia fissa, **Passing-Bablok** assumono (come evidenziato all'inizio del paragrafo) che i due metodi abbiano lo stesso tipo di errore (non che esso abbia la stessa dimensione, nelle misure ottenute con i due metodi).
Quindi differenziano il loro procedimento da quello di **Theil**, introducendo un parametro K, che misura il numero di coefficienti angolari collocati dalle due parti rispetto a $\beta = 1$.
Successivamente, costruiscono l'intervallo di confidenza del coefficiente angolare b **calcolato, per il rischio** α o la probabilità P prestabiliti (quasi sempre il 95%),
- sulla base del **numero di osservazioni** e utilizzando **la distribuzione normale** (come in quasi tutti i test non parametrici fondati sui ranghi, poiché la loro distribuzione diventa rapidamente normale).

Con logica simile a quella utilizzata per il coefficiente angolare, la stima dell'**intercetta** a richiede che una metà dei punti sia collocata sopra la retta e l'altra metà sotto.
Il successivo **test di linearità** prende in considerazione
- la **posizione dei punti** (valori osservati) rispetto alla loro **proiezione sulla retta** (valori attesi).

Con un test analogo e con gli stessi valori critici del **test di Kolmogorov-Smirnov**, si valuta l'accordo tra la distribuzione osservata e quella attesa.

Si passa quindi al **test per la verifica dell'ipotesi nulla** $\beta = 1$**,** che è effettuata ricorrendo all'intervallo di confidenza di β**.** L'ipotesi $\beta = 1$ è accettata, se il valore 1 è compreso nell'intervallo. Il rifiuto dell'ipotesi nulla dimostra che un metodo fornisce un valore non uguale all'altro, ma al massimo proporzionale. (***This hypothesis is accepted if the confidence interval for*** b ***contains the value 1. If the hypothesis is rejected, then it is concluded that*** b ***is significant different from 1 and there is at least a proportional difference between the two methods***.)

In modo analogo, è **verificata l'ipotesi nulla** $\alpha = 0$.
L'ipotesi è accetta, se 0 è compreso nell'intervallo.
Se si rifiuta l'ipotesi nulla, tra i due metodi esiste almeno una differenza costante. (***If the hypothesis is rejected, then it is concluded that*** a ***is significant different from 0 and both methods differ at least by a constant amount***.)

Infine la conclusione.
Se sono accettate entrambi le ipotesi $\beta = 1$ e $\alpha = 0$**, si può inferire che**
- $X_i = Y_i$
- e quindi i due metodi sono identici.

In un programma informatico standard, i risultati delle analisi statistiche e dei test vengono presentati mediante due grafici:
- il **diagramma di dispersione** dei valori X_i e Y_i rappresentati da punti, nel quale sono riportate la **retta calcolata** (eventualmente anche **quella teorica** con $\beta = 1$ e $\alpha = 0$**,** per meglio evidenziarne le differenze o l'uguaglianza) e i **due intervalli di confidenza**, di solito alla probabilità del 95%;
- il **plot dei residui**, rispetto alla retta calcolata che viene rappresentata come una media

Inoltre sono fornite le statistiche:
- **dimensione del campione** o numero di coppie di dati
e, per o per ogni variabile o metodo,
- valore **massimo e minimo**, **media**, **mediana**, **deviazione standard** dei valori e **errore standard** della media.
Relativamente ai parametri della retta, sono riportati:
- il valore dell'**intercetta** a e del **coefficiente angolare** b, entrambi con il **loro intervallo di confidenza al 95%**.

ESEMPIO (TRATTO DA CLINICAL CHEMISTRY). Come dimostrazione della presentazione di un'analisi della regressione effettuata con il metodo di Passing-Bablok, sono riportate le due figure tratte dall'articolo di Niels **de Jonge** e alii nell'anno 2000 ***Erythrocyte Sedimentation Rate by the Test-1 Analyzer*** (pubblicato in **Letters** della rivista **Clinical Chemistry** Vol. 46, No.6, pp.: 881 - 882).

L'analisi

- della **intercambialità di due metodi** analitici (***Automated Westergren method*** di riferimento e un test **manuale rapido**, chiamato ***Test-1***),

con la quale si vuole verificare

- se il test metodo manuale, più rapido, **può sostituire** validamente il tradizionale metodo automatico di Westergren, per valutare l'***Erythrocyte Sedimentation Rate*** o **ESR** (in mm/h),

è presentata con i seguenti dati:

1 - le **dimensioni del campione**: $n = 105$

2 – il **coefficiente di correlazione di Pearson**: $r = 0{,}97$

3 - la **retta di regressione**:

$$\hat{Y}_i = -0{,}48 + 0{,}91 \cdot X_i$$

4 - **il coefficiente angolare** $b = 0{,}91$ con l'**intervallo di confidenza al 95%** : 0,86 - 0,97

(quindi, seppure di poco, esclude il valore $b = 1{,}0$)

5 - l'**intercetta** $a = -0{,}48$ con l'**intervallo di confidenza al 95%** : -0,87 - 0,29

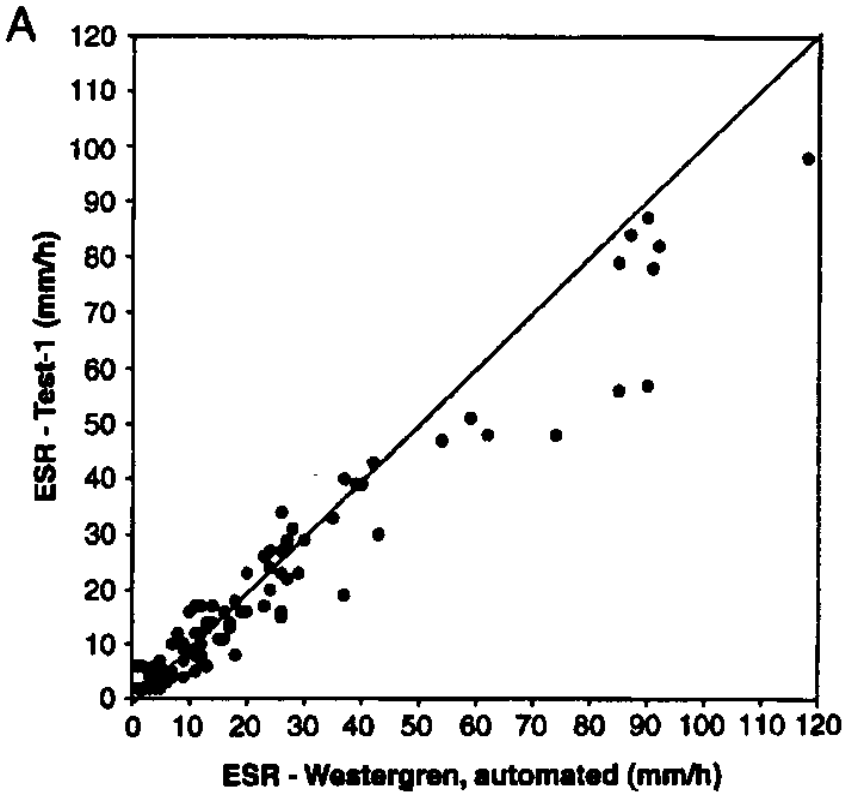
(quindi include il valore $a = 0{,}0$)

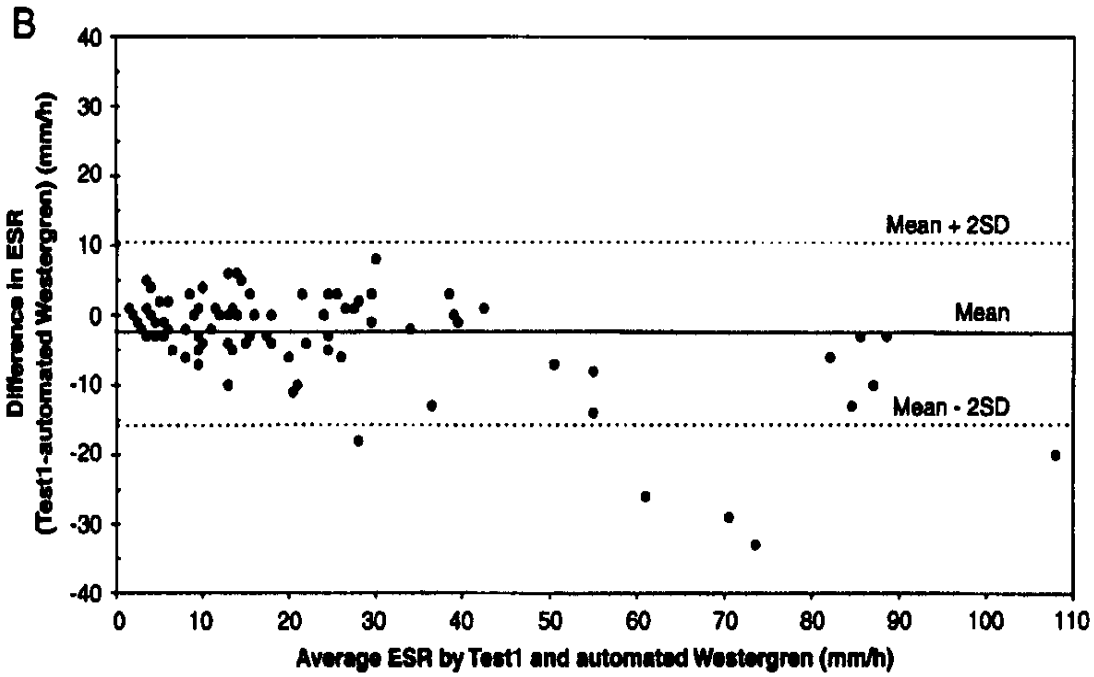
6 – la **deviazione standard** della misura Y (metodo manuale), sulla misura X (metodo di riferimento): $S_{Y/X} = 5{,}06$

7 - il **plot di Bland-Altman**, nel quale in ordinata si legge:

- la **media delle differenze** tra le due serie di valori: **–2,7**;
- il valore delle **differenze calcolate per ogni coppia di valori** è compreso tra il massimo di **+8** e il minimo di **–33**;
- **i limiti al 95% della media delle differenze** (con 2DS = 13,2) sono **+10,3** e **–15,7**.

Dall'**analisi di questi risultati** sono tratte le seguenti **conclusioni**:


A
ESR - Test-1 (mm/h)
ESR - Westergren, automated (mm/h)
0
10
20
30
40
50
60
70
80
90
100
110
120


B
Difference in ESR
(Test1-automated Westergren) (mm/h)
Average ESR by Test1 and automated Westergren (mm/h)
Mean + 2SD
Mean
Mean - 2SD
40
30
20
10
0
-10
-20
-30
-40
0
10
20
30
40
50
60
70
80
90
100
110

1 – il **test di Passing Bablok** evidenzia che il test manuale, chiamato Test-1, offre una determinazione rapida di ESR, con una variabilità tra i due metodi che è accettabile e con una buona correlazione con il tradizionale metodo Westergen (***acceptable intraassay variability and good correlation with traditional method***);

2 – il **test di Bland-Altman** mostra che l'errore sistematico (**bias**) e le differenze massime dell'accordo (***agreement limits***) tra i due metodi sono delle stesse dimensioni di quelli rilevati con altri metodi.

Si arriva alla conclusione finale che, seppure con l'aiuto dell'informazione fornita da vari test, è **il giudizio del clinico esperto** il criterio fondamentale che permette di dichiarare

- **accettabile l'errore commesso con il metodo nuovo,**
- e che quindi esso **può sostituire validamente il metodo precedente.**

Ovviamente colui che giudica deve porre particolare attenzione al differente significato dei parametri a e b nella **regressione di Passing-Bablok**:

- a dipende da un **errore analitico costante**, un errore di **specificità,** definita come l'abilità di un metodo a misurare solamente quello che si intende misurare;
- b è determinato da un **errore proporzionale costante**, un errore di calibrazione, derivante dalla relazione reale che esiste tra i due strumenti o metodi.

24.8. DIBATTITO SUL CONFRONTO TRA DUE METODI DI ANALISI CLINICHE ED ESEMPI DI TEST

Prima della presentazione del test di **Bland-Altman**, del rilancio della **regressione lineare last-products di Deming** e della proposta **di Passing-Bablok**, che in letteratura avvengono all'inizio degli anni '80 e nella ricerca applicata si diffondono circa 10 anni dopo, il confronto tra metodi presentava una serie di incongruenze tecniche. Erano applicati molti test, ognuno dei quali analizzava solo un aspetto e spesso in modo parziale o non congruo.

L'articolo di James O. **Westgard** e Marian R. **Hunt** del 1973 ***Use and Interpretation of Common Statistical Test in Method-Comparison Studies*** (su **Clinical Chemistry** Vol. 19, No. 1 pp.: 49 – 57) e quelli in cui **Bland-Altman** e **Passing-Bablok** presentano i loro metodi denunciano

- **i limiti delle numerose statistiche classiche**, quando esse sono applicate al caso specifico del **confronto tra due metodi di misura**, per valutarne la **precisione** e l'**accuratezza**.

1 – Il **test** t **di Student** per due campioni dipendenti è applicato alle coppie di valori X_i e Y_i per valutare se esiste una differenza tra le medie dei due metodi.

Se il test **risulta significativo**, i due **metodi non sono** giudicati **equivalenti**, a meno che gli errori casuali non siano molto piccoli. Tuttavia, quando il test **non risulta significativo** e in particolare quando il valore di t è vicino a 0, occorre ricordare che si possono avere due medie uguali, anche se le singole differenze nelle coppie di misure X_i e Y_i sono grandissime.

Come evidenzia la formula del test t di Student per due campioni dipendenti

$$t_{(n-1)} = \frac{\overline{d}}{\frac{s_d}{\sqrt{n}}}$$

i parametri utilizzati forniscono informazioni

- sull**a grandezza relativa dell'errore sistematico** ($\overline{d}$ = media delle differenze tra coppie di misure) rispetto all'**errore random** dei termini (s_d).

Ma la vera informazione importante e necessaria all'analisi di equivalenza tra i due metodi deriva dalle **singole differenze** ($d_i = Y_i - X_i$), non dalla loro media.

Infine, utilizzando il test t sarebbe importante conoscere sia $\overline{d}$, sia s_d, non semplicemente il valore t ottenuto.

2 – Il **test T di Wilcoxon** ha le stesse finalità e la stessa grave incongruenza logica. E' utilizzato in **sostituzione del** t **di Student**, quando la distribuzione delle differenze tra coppie di misure non è normale. E' riportato **insieme con il test** t**,** quando la non normalità è solamente sospettata. Oltre ai limiti del t, il test di Wilcoxon presenta anche quello di non utilizzare né la media delle differenze né la misura della loro variabilità e quindi fornisce una quantità di informazioni ancora minore.

3 – Il **coefficiente di correlazione** r **di Pearson** tra le coppie di valori X_i e Y_i, per misurare il loro grado di **accordo lineare** (***linear agreement***), ha indubbiamente la capacità di valutare se le misure campionarie aumentano o diminuiscono insieme. Ma, per verificare l'equivalenza tra due metodi, presenta almeno tre limiti:

a) il coefficiente r può essere **molto vicino a 1** o addirittura **uguale a 1,** anche se tra i due metodi esiste un rilevante **errore sistematico**, come possono essere (ad esempio, con + 12) coppie di valori 1 e 13, 2 e 14, 3 e 15, ecc. … ;

b) la dimensione del coefficiente r è influenzato dal **campo di variazione delle misure**, per cui tende a essere piccolo se i campioni analizzati hanno una distanza ridotta tra il valore minimo e quello massimo; inoltre, non fornisce informazioni sulle dimensioni delle differenze tra coppie di misure;

c) il test per la correlazione verifica l'ipotesi nulla $\rho = 0$, mentre in questo caso è applicato per valutare l'ipotesi se ρ è vicino 1 e, idealmente, se è $\rho = 1$ (ma in pratica non lo raggiunge mai).

4 – I coefficienti di correlazione **non parametrici** ρ **di Spearman e** τ **di Kendall** hanno gli stessi limiti del coefficiente parametrico r di Pearson, ma con il vantaggio di non essere influenzati ugualmente influenzati dalla presenza di valori anomali.

5 – Il **test** F ottenuto dal rapporto

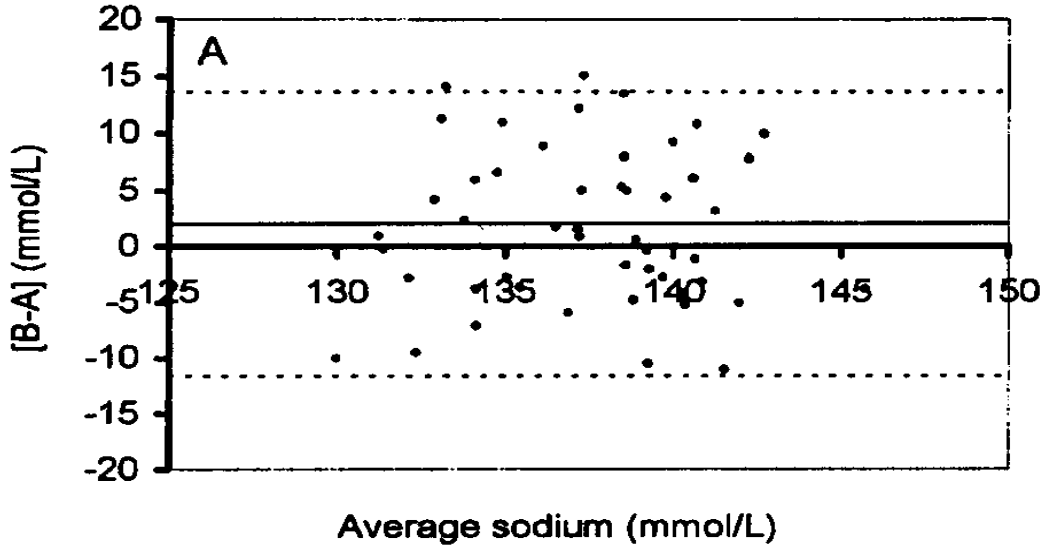
$$F = \frac{s^2_{\max}}{s^2_{\min}}$$

tra la varianza maggiore e quella minore, calcolate indipendentemente nelle due serie di misure X_i e Y_i (ma ora quasi tutti i programmi informatici adoperano il test di Levene, più potente) valuta se esiste una **precisione (variabilità) differente** tra i due metodi; ma come il test t sulle differenze, questo rapporto **confronta il livello complessivo di errore** nei due diversi metodi e **non è un indicatore dell'accettabilità dell'errore tra e entro** ognuna delle due serie di misure.

Rispetto a tutti questi test statistici, con il vantaggio ulteriore di essere **semplice da costruire e da capire** (***simple both to do and to interpret***), il test di **Bland-Altman** ha il pregio di essere finalizzato esplicitamente alla verifica se **due tecniche di misura sono comparabili**.

In un loro articolo divulgativo del 1986 (J. M. **Bland** e D. G. **Altman**, ***Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement,*** pubblicato su **The Lancet**, February 8, 1986, pp.: 307-310) presentano il problema con la domanda: ***Do the two methods of measurement agree sufficently closely?***

Resta da comprendere cosa si intenda con il termine **metodi comparabili** e quale sia la definizione di **differenze accettabili tra i due metodi**. La risposta non deve provenire dal tecnico che effettua le misure, ma dall'**esperto della disciplina che se ne serve**. Ad esempio, nelle analisi cliniche potrebbe significare che la diagnosi e la prescrizione non cambiano, se l'analisi biologica effettuata al paziente è stata condotta in laboratorio con un metodo oppure con un altro.



Per presentare e interpretare in modo corretto i**l plot di Bland-Altman**, è utile seguire le indicazioni di

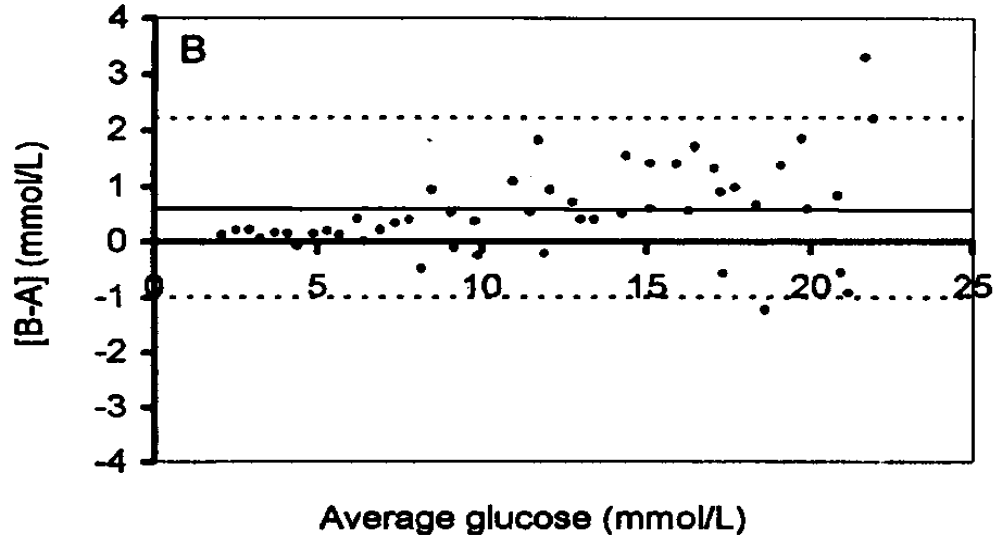
- Katy **Dewitte** et alii nell'articolo del 2002 ***Application of the Bland-Altman Plot for Interpretation of Method-Comparison Studies: A Critical Investigation of Its Practice*** (su **Clinical Chemistry** Vol. 48, No. 5, pp.: 799 - 801) e commentate da
- Douglas G. **Altman** e J. Martin **Bland** in ***Commentary on Quantifying Agreement between Two Methods of Measurement*** (su **Clinical Chemistry** Vol. 48, No. 5, pp.: 801 - 802).

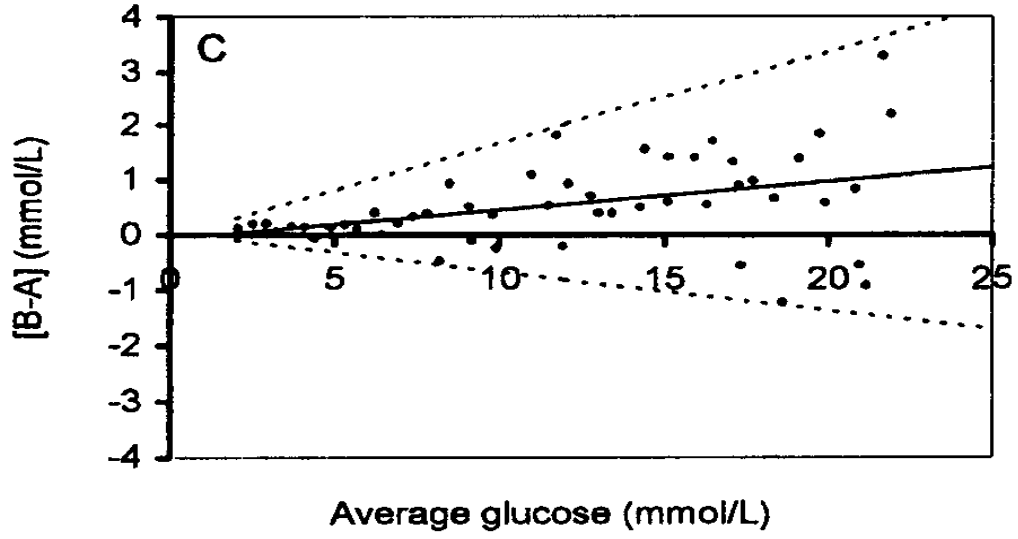
Riportando

- sull'asse delle ascisse la **media** dei risultati **dei due metodi** $\frac{X_i + Y_i}{2}$ sullo **stesso campione** i
- sull'asse delle ordinate la **differenza in valore assoluto** dei risultati **dei due metodi** $Y_i - X_i$ sullo **stesso campione** i

è diffuso come sostituto dell'analisi della regressione lineare.

Come nella precedente **figura A**, quando le analisi cliniche utilizzano concentrazioni che **variano in un campo limitato** e per valori proporzionalmente vicini (nell'esempio, hanno valori medi da 125 a 150), le **differenze** si mantengono **costanti**. E' quindi corretta la loro rappresentazione e analisi mediante i valori assoluti. La media delle differenze (linea continua più sottile) e i limiti $2S$ (le due linee tratteggiate, a distanza simmetrica dalla media) descrivono correttamente la distribuzione delle differenze.





Ma in molte situazioni sperimentali, la deviazione standard delle misure aumenta con la concentrazione e quindi anche le differenze tra i due metodi.

Nella **figura B**, i valori medi (asse delle ascisse) presentano una estensione molto grande: variano approssimativamente da 2 a 22. Le differenze in valore assoluto (asse delle ordinate) indicano una crescita proporzionale. Esse non sono descritte in modo corretto

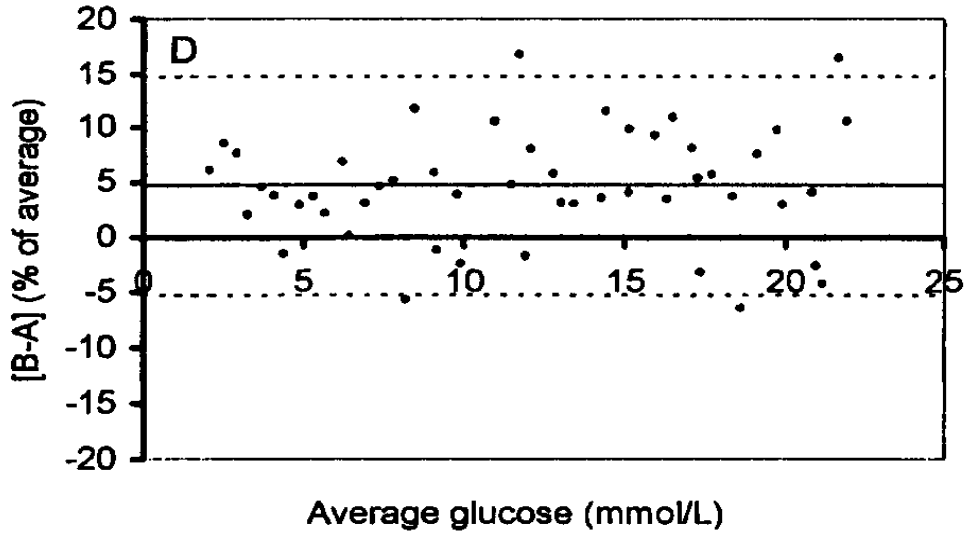
- né dalla media (le prime differenze sono inferiori e le ultime sono prevalentemente superiori)

- né tanto meno dai limiti $2S$ (le prime differenze sono piccole, le ultime sono grandi).

Una loro descrizione corretta è fornita dalla **figura C**, nella quale la media delle differenze in valore assoluto(linea continua più sottile) e l'intervallo di confidenza $2S$ (linee tratteggiate) indicano la quantità effettiva di variazione all'aumentare della concentrazione. Tuttavia è preferibile non utilizzare questa rappresentazione delle differenze assolute, ma ritornare a una grafico con una media e a un intervallo $2S$ costanti, mediante la trasformazione delle differenze.
Ma quale trasformazione è migliore?

Quando la deviazione standard S aumenta con la concentrazione,
- **Bland e Altman** hanno raccomandato di riportare sulle ordinate il **logaritmo delle differenze**
- altri propongono la loro trasformazione in **percentuale**,

come M. A. **Pollock** e alii nell'articolo del 1992 ***Method comparison – A different approach*** (su **Annals of Clinical Biochemistry** Vol. 26, pp.: 556 - 560) e Katy **Dewitte** nell'articolo citato del 2002.



Nella **figura D** è riportata la rappresentazione corretta, che usa la trasformazione in percentuale
Benché generalmente non vi sia molta differenza nei risultati grafici tra la trasformazione logaritmica e quella in percentuale, è preferibile il plot della percentuale, eccetto quando le concentrazioni si estendono su un campo di rilevazioni molto ampio, che varia di diversi ordini di grandezza:
- i numeri possono esser **letti direttamente sul plot,** senza richiedere una retro-trasformazione,

- il plot include sia la linea della **media delle differenze**, sia le due linee dei **limiti ($2S$) sperimentali osservati**.

Nel **test di Bland-Altman la significatività non è fornita da calcoli statistici, ma consiste essenzialmente nel confronto** (effettuato da un esperto della disciplina) **dei due limiti** ($2S$) **con la differenza clinicamente accettabile tra due metodi.**

Tuttavia, su molte riviste, in aggiunta al metodo di Bland-Altman, sono riportate

- l'analisi della **correlazione**,
- l'analisi della **regressione**,
- il test sulla **concordanza** (K di Cohen).

con la motivazione che il plot delle differenze è complementare, non sostitutivo.

E' di questa opinione anche il **National Committee for Clinical Laboratory Standard**, secondo le linee guida del 1995, pubblicate nel manuale ***Method comparison and bias estimation using patient samples, approved guideline*** (NCCLS publication EP9-A, Villanova, PA:NCCLS,1995).

Esso raccomanda di

- costruire il **diagramma di dispersione** dei dati X_i e Y_i,
- calcolare e riportare nel diagramma la **retta di regressione** (**Deming** o **Passing-Bablok**),
- analizzare con ottica disciplinare il **plot delle differenze di Bland-Altman.**

Un aspetto importante nell'analisi della equivalenza tra due metodi è l'utilizzazione di **misure ripetute**, che secondo Bland e Altman sarebbero **sempre necessarie**, per meglio capire le diverse origine degli errori. Come nell'analisi della varianza a due criteri, nel calcolo della differenza reale tra i due metodi permetterebbe di eliminare gli effetti di altri fattori ambientali (come le differenze tra operatori, tra strumenti, ecc.), che per un confronto corretto dovrebbero essere uguali. In questo caso, la retta di regressione e l'analisi di Bland-Altman si avvalgono solo delle due variabili quantitative.

A favore delle misure ripetute, Douglas G. **Altman** e J. Martin **Bland** nella discussione del 2002 in ***Commentary on Quantifying Agreement between Two Methods of Measurement*** (su **Clinical Chemistry** Vol. 48, No. 5, pp.: 801 - 802) scrivono:

- ***Another important issue is that the full comparison of the performance of two methods of measurement ougth to include repeated measurements. Such repeat data can be used to compare observers or instruments, or simply to assess random error.***

Ma è un metodo impiegato raramente. Un esempio, ma al quale è stato applicato il **test di Passing-Bablok**, è riportato nelle pagine successive.

Come indicazione delle **modalità di pubblicazione di un test per il confronto tra due metodi** con il test di **Passing e Bablok** e dell'uso della **regressione lineare per la calibrazione**, è vantaggioso

seguire l'articolo di Torsten **Arndt** e alii del 2004 ***Total Plasma Homocysteine Measured by Liquid Chromatography-Tandem Mass Spectrometry with Use of 96-Well Plates*** (su **Clinical Chemistry** Vol. 50, No. 4, pp.: 755 – 757) dal quale sono tratte anche le due figure seguenti.

L'aumento del plasma totale o siero omocisteina è considerato un fattore di rischio per malattie collegate all'occlusione delle arterie o delle vene.
Il problema tecnico da affrontare è che il metodo classico, denominato HPLC, richiede molto tempo e materiale costoso. Dagli autori dell'articolo è quindi proposto un metodo nuovo, denominato sistema LC-MS/MS, che presenta alcuni vantaggi pratici ma del quale deve essere preventivamente dimostrata la ***comparability,*** per affermare che può vantaggiosamente sostituire quello precedente.

L'analisi è stata condotta su 187 campioni di plasma.
Il **diagramma di dispersione**, riportato nella **figura A** successiva, è costruito ponendo
- sull'asse delle ascisse i valori ottenuti con il metodo vecchio e consolidato, cioè la quantità di omocisteina in $\mu mol / L$ con il metodo HPLC,
- sull'asse delle ordinate i valori ottenuti con il metodo nuovo, cioè la quantità di omocisteina in $\mu mol / L$ con il metodo LC-MS/MS.

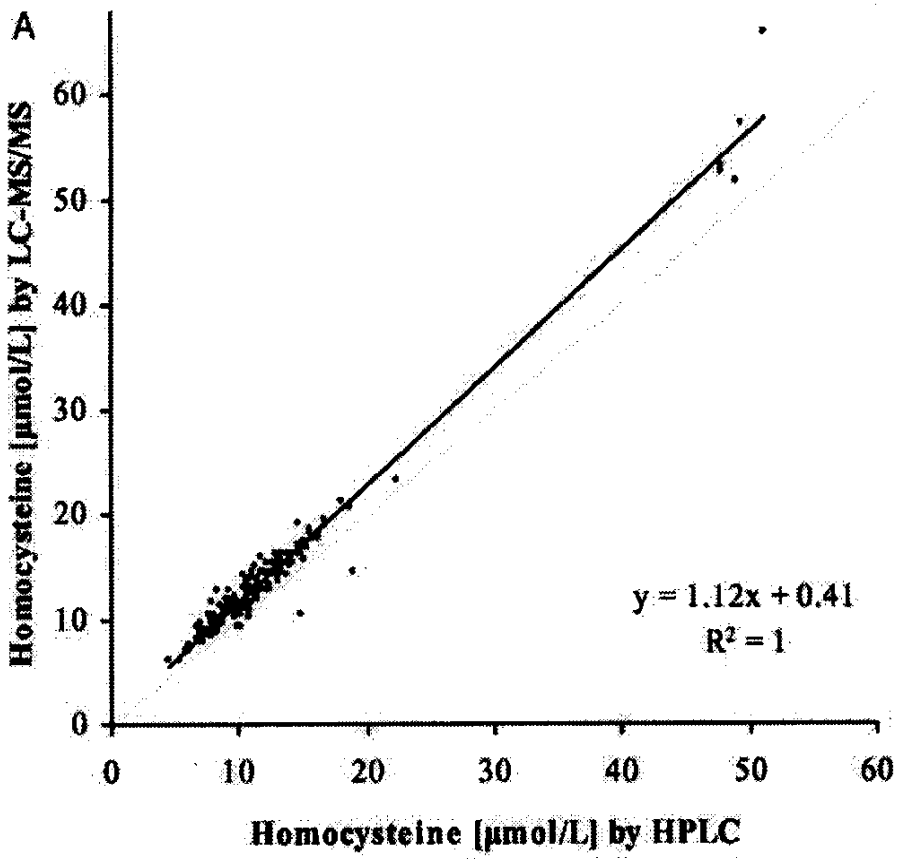
La **funzione di regressione di Passing-Bablok**
è

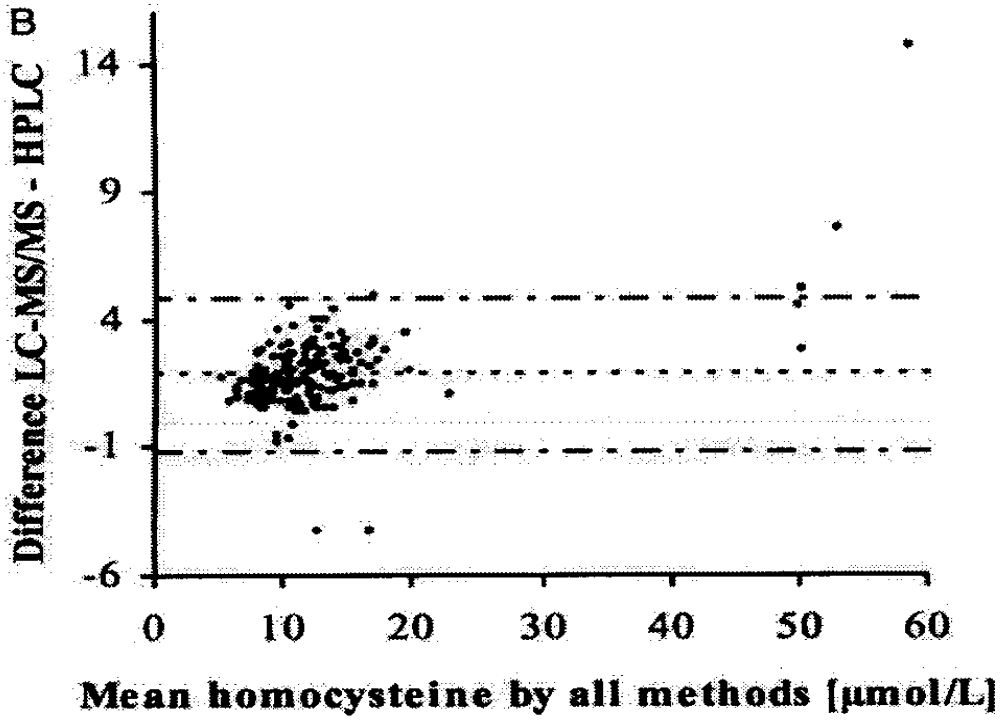
$$LC - MS / MS = 0{,}41 + 1{,}12 HPLC$$

espressi in $\mu mol / L$.

Per il **coefficiente angolare** $b = 1{,}12$ l'intervallo di confidenza al 95% è tra 1,075 e 1,162.
Per l'**intercetta** $a = 0{,}41$ l'intervallo di confidenza al 95% è tra –0,049 e + 0,856
Si può dedurre che
- mentre è accettabile l'ipotesi $\alpha = 0$
- il coefficiente angolare è statisticamente differente da $\beta = 1$.

L'**analisi di Bland-Altman**, riportata nel **figura B**, mostra che
- la media delle differenze tra le coppie di valori è 1,81
- intervallo $\pm 2S$ è compreso tra circa 5 e circa –1 (la lettura del grafico non permette di essere più precisi).


A
Homocysteine [µmol/L] by LC-MS/MS
Homocysteine [µmol/L] by HPLC
y = 1.12x + 0.41
R² = 1
0
10
20
30
40
50
60


B
Difference LC-MS/MS - HPLC
Mean homocysteine by all methods [µmol/L]
14
9
4
-1
-6
0
10
20
30
40
50
60

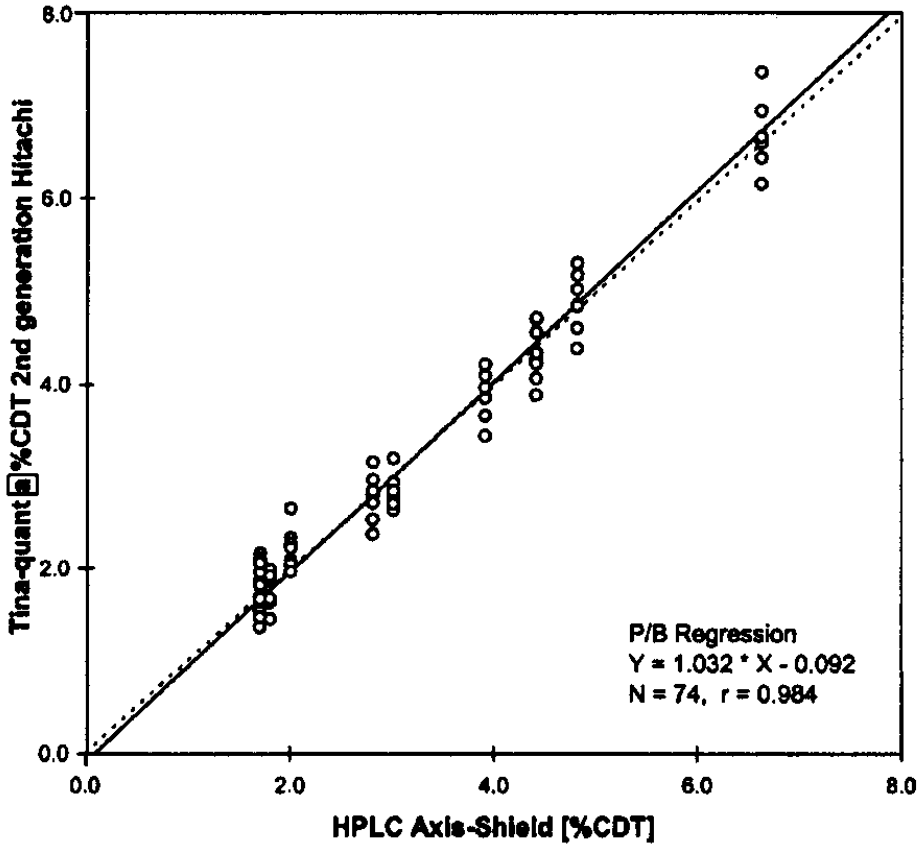
La conclusione degi autori è che tra i **due metodi** esistono **differenze** che sono

- statisticamente significative,
- ma trascurabili, l'aspetto biologico o clinico.

Infine, poiché sulla base dei loro esperimenti il metodo nuovo permette un risparmio del 90% per quanto riguarda il materiale e del 60% per quanto riguarda i tempi e quindi i costi del personale rispetto a quello classico, propongono come più vantaggioso l'uso del metodo nuovo.

Nell'articolo del 2003 di Markus J. **Schwarz** e alii dal titolo ***Multicentre evaluation of a new assay for determination of carbohydrate-deficient transferrin*** (su **Alcohol & Alcoholism** Vol. 38, No.3 pp.: 270 – 275) è riportato un esempio di **misure ripetute** allo scopo di valutare anche le differenze tra metodi e laboratori. Ma in questo caso è preentata solamente la parte che riguaeda tutti i dati.

Il Carbohydrate-Deficient Transferrin (CDT) è ritenuto un eccellente marcatore biochimico di un consumo eccessivo di alcol. Esiste consenso internazionale che il CDT sia misurato in modo adeguato mediante il metodo High-Performance Liquid Chromatography (HPLC). Recentemente, è stato sviluppato un metodo chiamato Tina-quant$[a]$%CDT 2nd generation che viene confrontato con il precedente, mediante l'analisi della regressione di Passing-Bablok.



La retta risulta

$$Y_i = -0{,}092 + 1{,}032 X_i$$

e il coefficiente di correlazione è

$$r = 0{,}984$$

Sulla base di queste due sole analisi, gli autori concludono che l'accordo tra i due metodi è molto buono: ***Very good agreement, with a slope of 1.03, a negligeable intercept of –0,1% CDT and a correlation coefficient (r) of 0,984, was obtained***.

In molte pubblicazioni, la **verifica della *comparability*** non è fondata solamente sul test di **Passing-Bablok** e sul test di **Bland-Altman**, ma si serve anche di misure di **correlazione** r o misure della capacità predittiva r^2, oltre che del test t di Student per la significatività delle differenze tra le due serie di misure appaiate.
Questi **ulteriori test sulla corrispondenza tra due metodi analitici** in realtà possono **condurre in errore coloro che non sono esperti della disciplina**. E' solamente la conoscenza degli effetti biologici o clinici che permette di decidere quando **due metodi sono in realtà bioequivalenti** e di sapere quale sia il margine di errore accettabile, in quella situazione specifica.

Infatti per campioni grandi, e in questi confronti quasi sempre si supera il centinaio di osservazioni, i test tendono a essere significativi, anche quando le differenze sono biologicamente trascurabili.
Dominique **Gerbet**, Phlippe **Richardot** e **alii** nell'articolo del 1983 ***New Statistical Approach in Biochemical Method-Comparison Studies by Using Westlake's Procedure, and Its Application to Continous-Flow, Centrifugal Analysis, and Multilayer Film Analysis Techniques*** (su **Clinical Chemistry** Vol. 29, No. 6, pp.: 1131 – 1136) a pag. 1135 scrivono:
- ***The joint testing procedure provided significant differences ($\alpha < 0.01$) for all pairwise comparisons. The reason is that, when available experimetal data are numerous, variances of $\hat{a}$ and $\hat{b}$ are very small. Hence, the smallest deviation from the point ($a = 0$, $b = 1$) is regarded as statistically significant, even though not significant biologically.***

Quando le misure fornite dai due metodi di analisi sono di **tipo qualitativo binario** oppure a **più livelli e ordinabili per rango**, l'accordo tra i due metodi deve essere stimato e verificato mediante il t**est K di Cohen** (al quale si rimanda).

24.9 IL CONFRONTO CON IL GOLD STANDARD: UTILIZZARE IL METODO DELLA CALIBRATION OPPURE QUELLO DELLA COMPARABILITY?

Per **valutare l'accordo** (***agreement***) tra **due serie di misure**, determinate con metodi differenti, è possibile impiegare

- sia le tecniche di ***calibration*** (già illustrate nel capitolo sulla regressione lineare),
- sia i metodi di ***comparability*** (illustrati in questo capitolo).

Si ricorre alla ***comparability***, quando entrambi i metodi hanno la stessa quantità di errore (***reproducibility***). Nessuno dei due è più preciso dell'altro.

Si ricorre alla ***calibration***, quando un metodo è preciso e l'altro presenta una variabilità (deviazione standard) elevata. Il caso classico di **calibration** è rappresentato dalla relazione tra la dose di un farmaco (X nota con precisione) e la risposta biologica (Y con variabilità o errore) indotta sul paziente.

Quando per effettuare la stessa analisi esistono almeno due metodi, spesso avviene che

- un metodo (chiamato **C**, da ***crude***) di norma sia più economico, ma meno preciso, mentre
- l'altro (chiamato **P**, da ***precise***) sia costoso (in denaro o tempo richiesto), ma molto preciso.

Considerazioni di convenienza economica e di praticità,

- rendono il metodo meno preciso (C) più attraente di quello più preciso (P)
- e quindi spingono a utilizzare il primo (C) per stimare i valori del secondo metodo (P).

Ma prima deve esserne dimostrata l'equivalenza.

Con quale metodo?

Con la **calibrazione** si stima il valore del metodo **precise**, partendo dal valore del metodo **crude**, più facilmente disponibile.

Il calcolo avviene con due passaggi:

1 - dalla variabile precise (P), in cui la X ha un errore piccolo, **si ricava la variabile Y crude** (C) attraverso la retta di regressione least-squares:

$$P = a + bC$$

2 – dalla stima della variabile **crude Y** (C)**,** si ricava il valore della variabile **precise X** (P), con il percorso inverso, ma sempre

con la regressione least-squares

$$C = a + bP$$

che invertita diventa

$$P = \frac{C - a}{b}$$

Soprattutto nel caso del confronto tra un metodo nuovo e un **gold-standard**, da molti studiosi di statistica sono ritenuti corretti la **regressione least-squares** e la **calibration**. Ad esempio, Alan M. **Batterman** nell'articolo del 2004 ***Commentary on Bias in Bland-Altman but not RegressionValidity Analyses*** (su **Sportscience** Vol. 8 pp.: 47-49) afferma che le **tecniche classiche di regressione least-squares** possono essere **impiegate correttamente**:

- per la **calibratura**, quindi per la **conversione** da un metodo all'altro;
- negli studi di **validità di un metodo**, particolarmente nelle situazioni in cui il test di confronto è **conosciuto** o è **gold-standard**;
- infine, quando l'**errore di misura della X e della Y è piccolo,** rispetto al campo di variazione dei dati e al loro effetto biologico.

Altri contestano questa scelta e suggeriscono ugualmente l'uso del plot di Bland-Altman.
In particolare J. Martin **Bland** e Douglas G. **Altman**, che nel 1995 in ***Comparing methods of measurement: why plotting differences against standard method is misleading*** (su **Lancet** Vol. 346, pp.: 1085-1087) sostengono come anche in questo caso sia corretto solamente utilizzare il loro metodo, soprattutto nel confronto tra misure cliniche, in cui la equivalenza tra i metodi dipende in larga parte dall'esperienza dello specialista.
Quando si confrontano due metodi e uno dei due è considerato lo **standard**,

- è fuorviante il concetto che tale termine sia sinonimo di **misura vera o corretta;**
- inoltre, la situazione non è differente quando il metodo di riferimento è definito **gold standard**.

Le conseguenze statistiche che il **gold-standard** possa rappresentare la misura di **riferimento senza errore** sono state analizzate in profondità da **Bland** e **Altman**. Accettando tale idea, sull'asse delle ascisse dovrebbe essere riportato il valore dello standard, non più la media dei due metodi.
Ma a pag. 1085 essi scrivono: ***it is sometimes argued that when one method may be regarded as a "gold standard", it is presumably more accurate than the other method and so we should plot the difference against the gold standard. We think that this idea is misguided and is likely to lead to misinterpretation. Here we will show why, and that the plot of difference against average is almost always preferable.***

La dimostrazione tecnica di tale affermazione è fondata sulla stima del**la correlazione tra le differenze e le medie delle coppie di valori,** rilevati con i due metodi a confronto.
Indichiamo

- con T il metodo da testare e con S lo standard,
- con ρ il coefficiente di correlazione tra le due serie di misure,
- con σ_T^2 e σ_S^2 le varianze rispettivamente della serie di valori T e S.

Quando l'analisi di due metodi include un **campo di variazione** dei valori che sia grande e le due varianze σ_T^2 e σ_S^2 **sono simili**, il coefficiente di correlazione tra le coppie di misure di solito risulta alto ($\rho > 0.7$), a meno che l'accordo (***agreement***) tra i due metodi sia estremamente basso.

Allorché i risultati sono rappresentati nel **plot** di **Bland-Altman**,

tra la differenza $T - S$ e la media $\frac{T+S}{2}$ si ha una **correlazione apparente**

che è

$$Corr\left(T - S, \frac{T+S}{2}\right) = \frac{\sigma_T^2 - \sigma_S^2}{\sqrt{\left(\sigma_T^2 + \sigma_S^2\right)^2 - 4\rho^2\sigma_T^2\sigma_S^2}}$$

Come evidenzia questa formula, la correlazione tra la differenza $T - S$ e la media $\frac{T+S}{2}$

- è zero, se le due **varianze sono uguali**;
- è **sempre piccolo**, eccetto quando esiste una **differenza grande tra le due varianze**.

Se il **plot** venisse costruito in modo diverso, collocando sull'**asse delle ascisse il valore dello standard,** si avrebbe che la correlazione tra la differenza $T - S$ e il valore dello standard S

è

$$Corr(T - S, S) = \frac{\rho\sigma_T - \sigma_S}{\sqrt{\sigma_T^2 + \sigma_S^2 - 2\rho\sigma_T\sigma_S}}$$

Di norma, questa correlazione risulta negativa (poiché $\rho < 1$ e $\sigma_T^2 = \sigma_S^2$).

Ne consegue che, quando le due varianze sono uguali e inoltre non esiste correlazione tra le differenze e la grandezza dei valori,

- il **plot delle differenze**, che sull'**asse delle ascisse abbia riportato i valori dello standard**, mostra una correlazione che tende alla quantità

$$r = -\sqrt{\frac{1-\rho}{2}}$$

Questa **correlazione spuria** è piccola, quando i due metodi sono altamente correlati (quindi $\rho \cong 1$ tra le coppie di osservazioni).

Si ha il fenomeno opposto, allorquando sull'asse delle ascisse è riportato il valore del metodo T. Si determina sempre una **correlazione spuria** della stessa quantità, ma di segno positivo.

I concetti sono illustrati con una serie di esempi, tratti dall'articolo di J. Martin **Bland** e Douglas G. **Altman** del 1995, già citato.

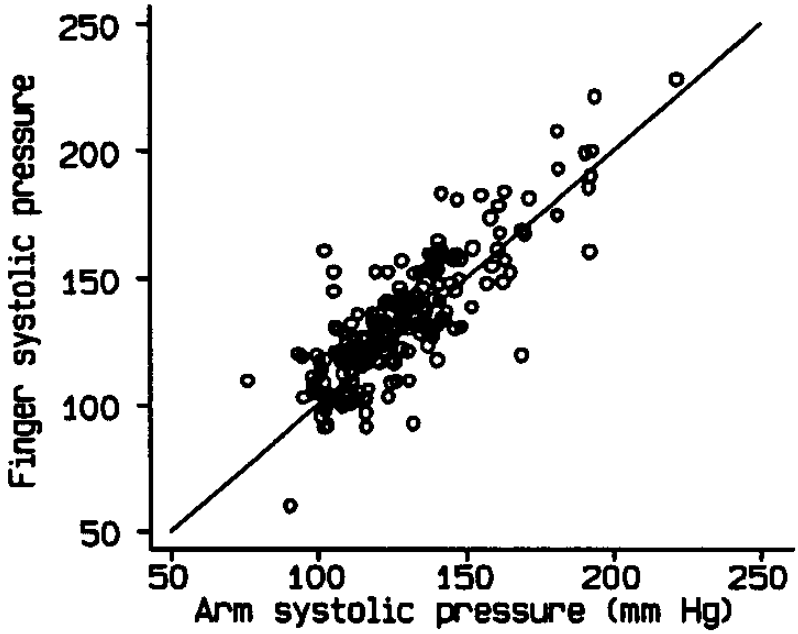


Figura 1. Diagramma di dispersione con il valore della misura standard (X) al braccio e il valore al dito (Y). La linea continua rappresenta la retta teorica della uguaglianza ($a = 0$ e $b = 1$).

Il metodo classico o **standard** per determinare la pressione sistolica (mm Hg) di un individuo è la misura presa al braccio. Un metodo alternativo o **test**, del quale si vuole analizzare la corrispondenza, è la pressione al dito: ***We saw the aim of such studies as to determine wheter two methods agreed sufficiently well for them to be used interchangeably***.

A questo scopo sono state prelevate le misure su

- un **campione abbastanza numeroso** di 200 pazienti,
- che presentano una **variabilità grande**.

Infatti nella figura 1 è possibile osservare che i livelli minori hanno approssimativamente valore 100 e quelli maggiori valori prossimi a 200, anche escludendo le osservazioni più estreme.

La **rappresentazione grafica classica** (Figura 1), vale a dire il **diagramma di dispersione** nel quale

- sull'asse delle ascisse è riportato il **metodo standard**
- sull'asse delle ordinate il **metodo nuovo,**

si dimostra **inefficiente**, poiché

- **i punti tendono a essere aggregati lungo la linea dell'uguaglianza** ($a = 0$ e $b = 1$)

soprattutto se, come atteso in questo caso, i due metodi forniscono misure tra loro collegate.

Il valore del coefficiente di correlazione tra pressione al braccio (**standard**) e pressione al dito (**test**) risulta $r = 0{,}87$.

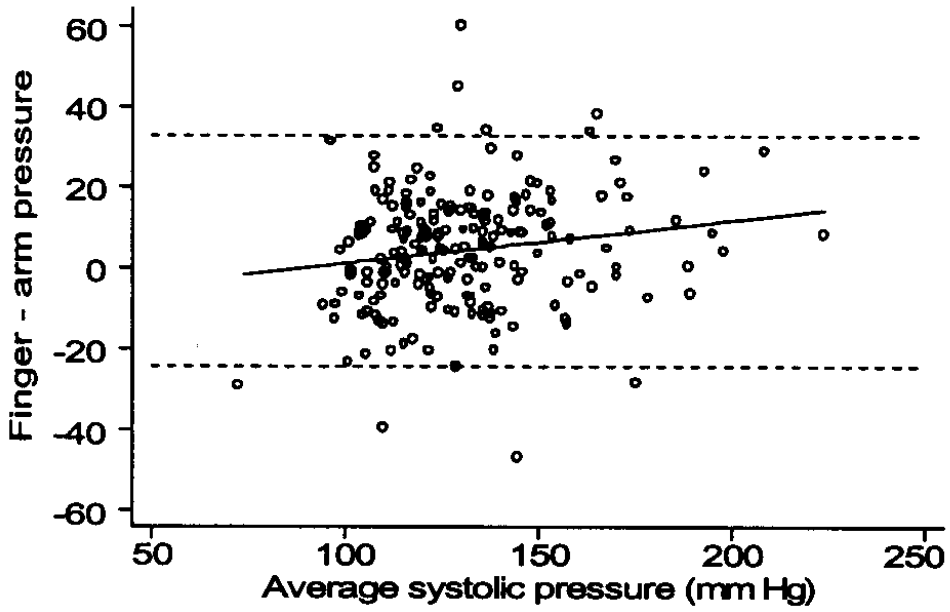


Figura 2. Diagramma di Bland-Altman con la media dei due metodi (X) e la differenza tra i due metodi (Y): test-standard. La linea continua rappresenta la retta di regressione tra la media delle coppie di valori e la loro differenza. Le linee tratteggiate sono la media delle differenze $\pm 1{,}96SD$

Il plot di Bland-Altman si dimostra più utile del diagramma di dispersione e del calcolo della retta di regressione, per l'analisi dell'equivalenza tra i due metodi.

La media delle differenze (test – standard **o** dito meno braccio**) risulta +4,3 (mm Hg) e la deviazione standard di queste differenze è** SD **= 14,6 (mm Hg). Da questi dati deriva che**

- **il limite inferiore è 4,3 – 1,96 x 14,6 = -24 mm Hg**

- il limite superiore è 4,3 + 1,96 x 14,6 = +33 mm Hg

Inoltre nel grafico è riportata

- la retta di regressione della differenza ($test - standard$) sulla media ($\frac{test + standard}{2}$)
- e il coefficiente di correlazione tra media e differenza è $r = 0{,}17$.

Si evidenzia anche che il **dito** (**test**) tende a dare valori più alti del **braccio** (**standard**), all'aumentare della pressione.

E' quindi presente un **errore sistematico** (***bias***), che in precedenza la retta di regressione, anche nella condizione teorica ($a = 0$ e $b = 1$), non riusciva a evidenziare.

Il confronto del **test** (metodo nuovo) con il **gold-standard** non modifica i risultati emersi da questo confronto, per quanto riguarda la retta di regressione e il metodo di **Bland-Altman**.

Infatti il diagramma delle differenze, costruito non più sui valori medi delle coppie di osservazioni ma sul valore dello **standard** (Figura 3), ritenuto il **valore vero** e quindi **senza errore**, determina in realtà un errore sistematico, rispetto al **bias reale** precedente, in accordo con la teoria presentata.

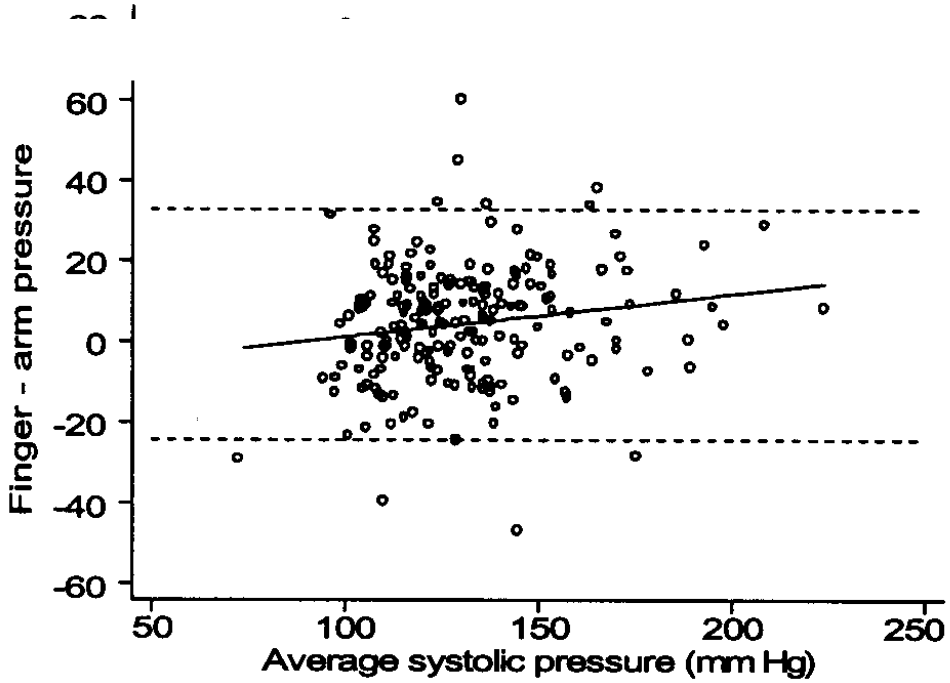


Figura 3. Diagramma di Bland-Altman con il valore dello standard (X) e la differenza tra i due metodi (Y): test-standard. La linea continua rappresenta la retta di regressione tra il valore dello standard e la differenza. Le linee tratteggiate sono la media delle differenze $\pm 1{,}96SD$

Nel grafico precedente è riportata

- la retta di regressione della differenza (*test* – *standard*) sul valore dello standard
- e il coefficiente di correlazione, che tra il valore dello standard e la differenza è r = - 0,14.

E' un **valore negativo**, come predetto dalla formula, ma per una quantità inferiore all'atteso

$$r = -\sqrt{\frac{1-\rho}{2}} = -\sqrt{\frac{1-0{,}83}{2}} = -\sqrt{0{,}085} = -0{,}29$$

che è $r = -0{,}29$.

E' la dimostrazione sperimentale che quando sull'asse delle ascisse viene riportato il valore del **test**, si ha ugualmente un errore sistematico, ma di segno opposto.

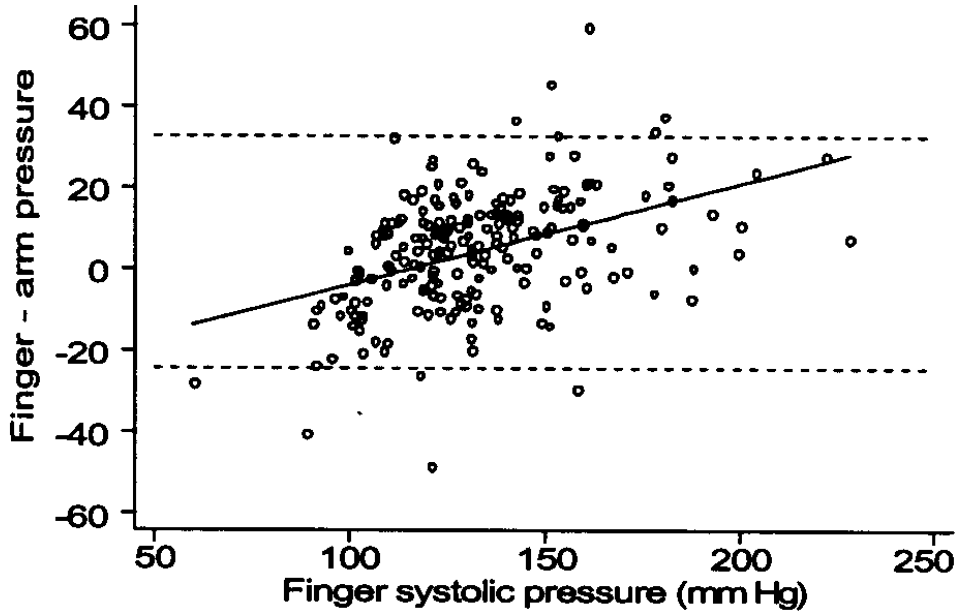


Figura 4. Diagramma di Bland-Altman con il valore del test (X) e la differenza tra i due metodi (Y): test-standard. La linea continua rappresenta la retta di regressione tra il valore del test e la differenza. Le linee tratteggiate sono la media delle differenze $\pm 1{,}96SD$

Nel grafico è riportata

- la retta di regressione della differenza (*test* – *standard*) sul valore del test
- e il coefficiente di correlazione, che tra il valore del test e la differenza è r = + 0,44.

Le tre correlazioni (quella del grafico esatto e le due calcolate con il valore dello standard oppure del test riportato sull'asse delle ascisse), risultano tutte significative.

La conclusione di **Bland e Altman** è che le ultime due forniscono informazioni sbagliate, come dimostrano le formule matematiche e la loro applicazione ai dati sperimentali. Inoltre sono tutte significative e tra loro contraddittorie: ***Thus we get significant correlations in different directions!***

Su quale sia la **tecnica statistica migliore**, in quanto più appropriato e semplice, per **confrontare due metodi di misura**, **Bland e Altman** non hanno dubbi: **il loro**. E' un concetto che riportano in forme leggermente diverse in molti loro articoli.

Ad esempio, in ***Statistical methods for assessing agreement between two methods of clinical measurement*** (su **Lancet** 1986; i, pp.: 307-310),

- nella summary scrivono:

In clinical measurement comparison of a new measurement technique with an established one is often needed to see whether they agree sufficiently for the new replace the old. Such investigations are often analysed inappropriately, notably by using correlation coefficients. The use of correlation is misleading. An alternative approach, based on graphical techniques and simple calculations, …

- Nella conclusione scrivono

In the analysis of measurement method comparison data, neither the correlation coefficient … nor techniques such as regression analysis are appropriate. We suggest replacing these misleading analyses by a method that is semple both to do and to interpret.

24.10 IL TEST DI BLAND-ALTMAN PER IL CONFRONTO TRA DUE METODI, CON MISURE RIPETUTE PER OGNI METODO SULLO STESSO SOGGETTO

Per valutare l'accordo tra due metodi di misurazione mediante il test di Bland Altman, come essi stessi suggeriscono è vantaggioso utilizzare **misure ripetute**. Il caso più semplice, soprattutto per l'analisi statistica, è che le misure siano effettuati due volte nelle stesse condizioni sperimentali, come nell'esempio successivo:

Campione	Metodo X		Metodo Y	
	Prova 1	Prova 2	Prova 1	Prova 2
1	557	586	601	589
2	417	422	432	424
3	657	641	672	684
i	---	---	---	---
n	178	163	151	172

Il primo passo è calcolare separatamente le medie dei due metodi

Campione	Metodo X	Metodo Y
	Medie	Medie
1	571,5	595
2	419,5	428
3	649	678
i	---	---
n	170,5	161,5

Il secondo è calcolare la media generale di ogni **campione** e **la differenza tra le medie** dei due metodi

Campione	Medie $\frac{X+Y}{2}$	Differenze $X-Y$
1	583,25	-23,5
2	423,75	-8,5
3	663,5	-29
i	---	---
n	166	+9

Infine **costruire il plot**, ponendo

- sull'asse delle ascisse la media $\frac{X+Y}{2}$

- sull'asse delle ordinate la differenza $X-Y$.

La **media delle differenze** e la **sua deviazione standard** (SD) per ottenere l'**intervallo** $\pm 2SD$ sono calcolate su questa ultima serie di differenze. Il 95% di queste differenze sarà compreso entro l'intervallo $\pm 2SD$, se la distribuzione dei valori è normale e senza la presenza di outlier, che ne ampliano notevolmente il valore.

Ma quando il grafico viene costruito con i dati originari raccolti, riportando i punti individuati con la Prova 1 e la Prova 2, non le loro medie, la stima della deviazione standard SD prima ottenuta (che possiamo indicare con S_D) è troppo piccola. E' necessario impiegare

- la **deviazione standard corretta** (***corrected standard deviation***) S_c, che è ottenuta

con

$$S_c = \sqrt{S_D^2 + \frac{S_{dX}^2}{4} + \frac{S_{dY}^2}{4}}$$

dove

- S_{dX} e S_{dY} sono le **deviazioni standard delle differenze tra le misure ripetute entro ogni metodo** (X e Y) separatamente, vale dire

Campione	Metodo X	Metodo Y
	$d_X = P_1 - P_2$	$d_Y = P_1 - P_2$
1	-29	+12
2	-5	+8
3	+16	-8
i	---	---
n	+15	-21

la colonna delle differenze del metodo X e del metodo Y

Il plot di Bland Altman può essere costruito anche con le singole osservazioni, nelle quali per ogni campione si ha il punto della Prova 1 e quello della Prova 2.

Per calcolare i loro intervallo di confidenza non sempre i dati originali sono disponibili e quindi non è possibile ricavare le loro differenze e da esse le due deviazioni standard S_{dX} e S_{dY}.

Con la sola S_D è possibile ricavarne una stima approssimata.

La **deviazione standard corretta** (***corrected standard deviation***) o approssimata S_c può essere ottenuta con

$$S_c = \sqrt{2S_D^2}$$

24.11 LA RIPETIBILITA' E LA RIPRODUCIBILITA' DI UNO STRUMENTO O DI UN METODO: RANGE & AVERAGE METHOD

Quando si vogliono confrontare due metodi analitici, come nei paragrafi precedenti, è preliminarmente richiesto che lo **strumento** (in linguaggio tecnico, ***gage*** o ***device***) e l'**operatore** (***appraiser***) siano attendibili, vale a dire che nelle stesse condizioni forniscano sempre la stessa risposta o almeno risposte simili.

Lo strumento può essere un micrometro, un termometro, un manometro o qualsiasi dispositivo di analisi e misura. La quantità rilevata può riguardare la dimensione di un prodotto, la concentrazione di un principio attivo in un farmaco, il risultato di un'analisi clinica di un paziente.

La **valutazione dei dati raccolti** deve riguardare contemporaneamente sia l'**analisi statistica**, sia l'**importanza disciplinare** che le variazioni assumono nel problema affrontato, per decidere se sono importanti oppure trascurabili per i loro effetti pratici.

Inoltre, l'analisi della variabilità determinata da uno strumento o da un operatore è importante per se stessa. Quando si analizza una misura o una intera serie allo scopo di prendere decisioni, come può essere la prescrizione della cura per un paziente o la qualità di un prodotto da porre in vendita, occorre sapere quale credito attribuire ad esse. Quindi è fondamentale determinare la **quantità di variazione dovuta al processo**, che dipende sia dallo **strumento** impiegato che dal **tecnico**.

Affinché la misura sia credibile, è senso comune che la variazione strumentale deve essere contenuta entro valori chiamati **limiti di tolleranza** (ricordando che **i limiti di tolleranza riguardano le singole misure**, mentre i limiti di confidenza o limiti fiduciali riguardano le medie); pertanto

- le differenze tra strumenti devono essere trascurabili,
- operatori diversi devono fornire dati simili sullo stesso campione e quindi commettere errori di dimensioni analoghe ,quando utilizzano lo stesso strumento.

La variazione tra **strumenti** (***gage*** o ***device***) è chiamata **ripetibilità** (***repetability***).

La variazione tra **operatori** (***appraisers***) è chiamata **riproducibilità** (***riproducibility***).

Possono essere combinate insieme, formando un indicatore unico: la ***R&R***, che somma la variabilità dovuta allo strumento a quella determinata dall'operatore o attribuibile ad altra condizione ambientale, se non mantenuta costante.

La ***R&R*** **non comprende tutti i fattori di variazione.**

Insieme con la **ripetibilità** allo strumento possono essere attribuiti altri tre tipi di variazione:

- la **calibrazione** (***calibration***): lo strumento deve essere accurato, cioè essere tarato esattamente;
- la **stabilità** (***stability***): la taratura non deve cambiare nel tempo;
- la **linearità** (***linearity***): l'errore deve avere dimensioni uguali dai valori piccoli a quelli grandi, mantenendosi costante in tutto il campo di variazione delle misure effettuate.

Il fatto che l'analisi statistica *R&R* li trascuri, non significa che essi non siano importanti. Semplicemente, il loro impatto viene ritenuto meno significativo.

La **ripetibilità** (***repeatability***) è l'accordo tra i risultati di misure effettuate in **condizioni omogenee**, ripetute con lo stesso strumento e/o nelle **medesime situazioni sperimentali** sullo **stesso soggetto**. E' un parametro fondamentale **nel confronto tra metodologie** di analisi chimiche e/o cliniche, che è sempre necessario quantificare. Infatti,

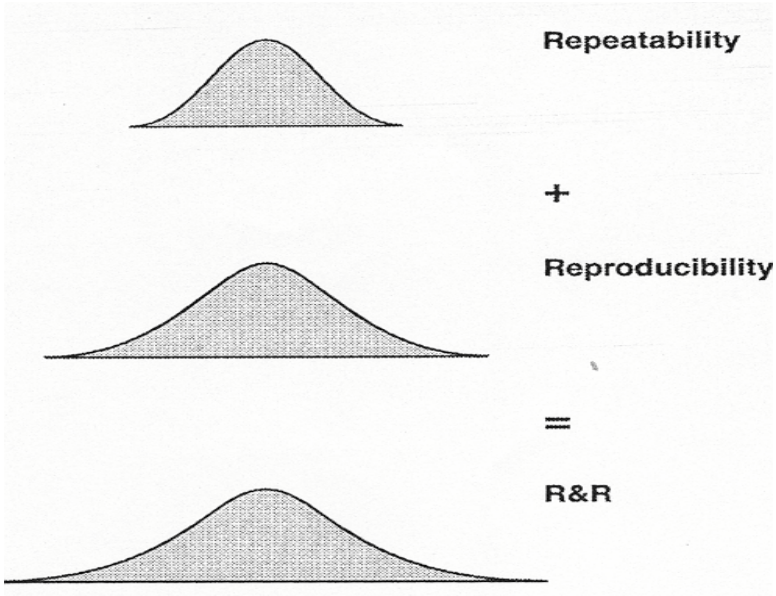
- **quando la ripetibilità di un metodo è bassa anche il livello di accordo (*agreement*) possibile tra due metodi è basso**.

Se poi entrambi i metodi hanno **varianza grande e quindi bassa ripetibilità**, il confronto tra metodi è realizzato nelle condizioni peggiori: è poco probabile che si riesca a dimostrare che tra essi esiste un grado di ***agreement*** accettabile.

La **riproducibilità** (***reproducibility***) è l'accordo tra i risultati di misure effettuate in **condizioni differenti**, quindi da un operatore diverso e/o con uno strumento e/o in un laboratorio e/o in tempi differenti. L'aumento della variabilità determina una diminuzione anche della riproducibilità.

Come verrà approfondito, mediante l'applicazione di un'ANOVA a un criterio,

- la **ripetibilità** è misurata dalla **varianza entro**,
- la **riproducibilità** è misurata dalla **varianza tra**.



Nei capitoli introduttivi, si è parlato della **accuratezza** e della **precisione** che una misura campionaria deve avere, per permettere di stimare il valore vero della variabile. Sono concetti differenti dalla ripetibilità e dalla riproducibilità; ma l'accuratezza (***accurate, calibrated gage***) è analoga alla taratura, mentre la precisione (***precise, capable operator***) è analoga alla ripetibilità.

Come evidenziato nella figura precedente, in termini matematici

- la **varianza *R&R*** è la somma della **varianza di ripetibilità** più la **varianza di riproducibilità:**

$$\sigma^2_{R\&R} = \sigma^2_{Repetibilità} + \sigma^2_{Riproducibilità}$$

- **la varianza della riproducibilità è maggiore di quella della riproducibilità**, come di norma nell'ANOVA la varianza tra è maggiore della varianza entro.

I modi statistici per valutare la **ripetibilità e la riproducibilità di un metodo analitico,** quando **la prova non è distruttiva del campione** e quindi può essere **ripetuta**, sono sostanzialmente **tre**:

1 – il metodo della media generale e della differenza (***average and range method***),

2 – il metodo della variazione parziale (***within part variation (WIV) method*),**

3 – l'analisi della varianza (***Analysis Of Variance***).

I primi due sono illustrati in questo paragrafo, il terzo in un paragrafo successivo.

Per misurare (A) la ***repeatability,*** (B) la ***reproducibility*** e (C) la ***R&R*** di uno strumento o di un metodo di analisi, in caso di **prove non distruttive** e quindi ripetibili sullo stesso oggetto, l'**esperimento standard raccomandato** dagli organismi internazionali addetti al Controllo di Qualità richiede

- un numero di campioni (***parts, units***) compreso tra 5 e 10,
- un numero di operatori (***appraisers***) costante e pari a 3,
- un numero di prove o ripetizioni (***trials, replications***) da 2 a 3.

Metodo o strumento X						
Campione	Operatore A		Operatore B		Operatore C	
	Prova 1	Prova 2	Prova 1	Prova 2	Prova 1	Prova 2
1	65,2	60,1	62,9	56,3	71,6	60,6
2	85,8	86,3	85,7	80,5	92,0	87,4
3	100,2	94,8	100,1	94,5	107,3	104,4
4	85,0	95,1	84,8	90,3	92,3	94,6
5	54,7	65,8	51,7	60,0	58,9	67,2
6	98,7	90,2	92,7	87,2	98,9	93,5
7	94,5	94,5	91,0	93,4	95,4	103,3
8	87,2	82,4	83,9	78,8	93,0	85,8
9	82,4	82,2	80,7	80,3	87,9	88,1
10	100,2	104,9	99,7	103,2	104,3	111,5

Per illustrare questa procedura statistica, si assuma che un metodo analitico sia stato sperimentato da **tre operatori**, ognuno dei quali ha effettuato **due prove**, sugli stessi **dieci campioni**, ottenendo i valori riportati nella tabella precedente.

La prima elaborazione richiede di

- calcolare la **differenza**, che misura l'ampiezza della risposta (***range***) tra le **due prove eseguite dallo stesso operatore**, per ognuno dei 10 campioni (le tre colonne $|R|$, in grassetto e corsivo).

Si ottengono 30 differenze $|R|$, che devono essere considerate in **valore assoluto**.

Metodo o strumento X									
Campione	Operatore A			Operatore B			Operatore C		
	Prov 1	Prov 2	$\|R\|$	Prov 1	Prov 2	$\|R\|$	Prov 1	Prov 2	$\|R\|$
1	65,2	60,1	***5,1***	62,9	56,3	***6,6***	71,6	60,6	***11,0***
2	85,8	86,3	***0,5***	85,7	80,5	***5,2***	92,0	87,4	***4,6***
3	100,2	94,8	***5,4***	100,1	94,5	***5,6***	107,3	104,4	***2,9***
4	85,0	95,1	***10,1***	84,8	90,3	***5,5***	92,3	94,6	***2,3***
5	54,7	65,8	***11,1***	51,7	60,0	***8,3***	58,9	67,2	***8,3***
6	98,7	90,2	***8,5***	92,7	87,2	***5,5***	98,9	93,5	***5,4***
7	94,5	94,5	***0,0***	91,0	93,4	***2,4***	95,4	103,3	***7,9***
8	87,2	82,4	***4,8***	83,9	78,8	***5,1***	93,0	85,8	***7,2***
9	82,4	82,2	***0,2***	80,7	80,3	***0,4***	87,9	88,1	***0,2***
10	100,2	104,9	***4,7***	99,7	103,2	***3,5***	104,3	111,5	***7,2***
Media	85,5		***50,4***	82,9		***48,1***	88,9		***57,0***

Su esse,

- calcolare la media generale o totale $\overline{\overline{R}}$ che, con i dati dell'esempio,

è

$$\overline{\overline{R}} = \frac{50{,}4 + 48{,}1 + 57{,}0}{10 + 10 + 10} = \frac{155{,}5}{30} = 5{,}1833$$

A – L'**intervallo di *Repeatability***

è

$$Repeatability = 5{,}15 \cdot \frac{\overline{\overline{R}}}{d_2}$$

dove

- 5,15 è la costante del Processo Sigma, legata alla distribuzione normale delle differenze e al livello di probabilità del 99%: -2,575 < Z < +2,575
- $\overline{\overline{R}}$ è la media generale di tutte le differenze tra coppie di prove
- d_2 è il valore riportato nella tabella sottostante dopo aver definito (1) Z e (2) W, con

1) Z = numero di differenze = **campioni x operatori** (***parts x appraisers***)
2) W = numero di prove (nel caso più semplice, come in questo, è uguale a 2)

Valori di d_2

	W													
Z	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1,41	1,91	2,24	2,48	2,67	2,83	2,96	3,08	3,18	3,27	3,35	3,42	3,49	3,55
2	1,28	1,81	2,15	2,40	2,60	2,77	2,91	3,02	3,13	3,22	3,30	3,38	3,45	3,51
3	1,23	1,77	2,12	2,38	2,58	2,75	2,89	3,01	3,11	3,21	3,29	3,37	3,43	3,50
4	1,21	1,75	2,11	2,37	2,57	2,74	2,88	3,00	3,10	3,20	3,28	3,36	3,43	3,49
5	1,19	1,74	2,10	2,36	2,56	2,73	2,87	2,99	3,10	3,19	3,28	3,36	3,42	3,49
6	1,18	1,73	2,09	2,35	2,56	2,73	2,87	2,99	3,10	3,19	3,27	3,35	3,42	3,49
7	1,17	1,73	2,09	2,35	2,55	2,72	2,87	2,99	3,10	3,19	3,27	3,35	3,42	3,48
8	1,17	1,72	2,08	2,35	2,55	2,72	2,87	2,98	3,09	3,19	3,27	3,35	3,42	3,48
9	1,16	1,72	2,08	2,34	2,55	2,72	2,86	2,98	3,09	3,19	3,27	3,35	3,42	3,48
10	1,16	1,72	2,08	2,34	2,55	2,72	2,86	2,98	3,09	3,18	3,27	3,34	3,42	3,48
11	1,15	1,71	2,08	2,34	2,55	2,72	2,86	2,98	3,09	3,18	3,27	3,34	3,41	3,48
12	1,15	1,71	2,07	2,34	2,55	2,72	2,85	2,98	3,09	3,18	3,27	3,34	3,41	3,48
13	1,15	1,71	2,07	2,34	2,55	2,71	2,85	2,98	3,09	3,18	3,27	3,34	3,41	3,48
14	1,15	1,71	2,07	2,34	2,54	2,71	2,85	2,98	3,09	3,18	3,27	3,34	3,41	3,48
15	1,15	1,71	2,07	2,34	2,54	2,71	2,85	2,98	3,08	3,18	3,26	3,34	3,41	3,48
>15	**1,128**	**1,693**	**2,059**	**2,326**	**2,534**	**2,704**	**2,847**	**2,970**	**3,078**	**3,173**	**3,258**	**3,336**	**3,407**	**3,472**

Con i dati dell'esempio, dove

- $\overline{\overline{R}}$ = 5,1833
- d_2 = 1,128 (per Z > 15 e W = 2)

si ottiene l'**intervallo al 99% della Ripetibilità**:

$$Repeatability = 5{,}15 \cdot \frac{5{,}1833}{1{,}128} = 23{,}66$$

Concettualmente, la **ripetibilità** è riferita **alla variazione presente nelle misure,** effettuate da uno **stesso operatore,** con lo stesso strumento o metodo analitico sul medesimo campione.
Come attuata nell'esempio,

- la stima della **ripetibilità** è fondata sulla media delle differenze di misure ripetute in condizioni identiche, adoperando lo stesso strumento o metodo su più campioni.
Se, come in questo caso, sono impiegati più operatori, la **media delle differenze** R è calcolata come media degli operatori.

Con tre operatori (A, B, C),

$$\overline{\overline{R}} = \frac{\overline{R}_A + \overline{R}_B + \overline{R}_C}{3}$$

La **deviazione standard** S_e **della ripetibilità** è

$$S_e = \frac{\overline{\overline{R}}}{d_2}$$

L'**intervallo al 99% della ripetibilità** (-2,575 <Z < +2,575, ricordando che per convezione si fa riferimento sempre a questa percentuale del 99% e quindi al valore 5,15)
è

$$5{,}15 \cdot \frac{\overline{\overline{R}}}{d_2} = 5{,}15 \cdot S_e$$

B – L'**intervallo di *Reproducibility*** del sistema
è

$$Reproducibility = \sqrt{\left(5{,}15 \cdot \frac{\overline{X}_R}{d_2}\right)^2 - \frac{(Repeatability)^2}{nr}}$$

dove

- $\overline{X}_R$ = media delle differenze tra l'operatore con la media maggiore e quello con la media minore in tutte le misure eseguite (per capire correttamente la procedure è utile vedere l'esempio successivo),
- d_2 è fornito dalla tabella precedente, con Z = 1 (costante) e W = numero di operatori (***appraisers***),
- n = numero di campioni (***parts***),
- r = numero di prove o repliche (***trials***)

Per calcolare $\overline{X}_R$, dapprima si devono confrontare le medie dei tre operatori, già calcolate nella prima tabella di elaborazione dei dati:

A = 85,5 B = 82,9 C = 88,9

Con i dati dell'esempio, emerge che

- la media minore è quella dell'operatore B
- la media maggiore è quella dell'operatore C

quindi tra essi esiste la differenza maggiore.

Successivamente, si devono

- calcolare, in valore assoluto, tutte le **differenze** (***range***) $|X_R|$ tra le nr misure effettuate da questi due operatori:

Campione	Prova	Operatore B	Operatore C	$\|X_R\|$
1	1	62,9	71,6	***8,7***
2	1	85,7	92,0	***6,3***
3	1	100,1	107,3	***7,2***
4	1	84,8	92,3	***7,5***
5	1	51,7	58,9	***7,2***
6	1	92,7	98,9	***6,2***
7	1	91,0	95,4	***4,4***
8	1	83,9	93,0	***9,1***
9	1	80.7	87,9	***7,2***
10	1	99,7	104,3	***4,6***
1	2	56,3	60,6	***4,3***
2	2	80,5	87,4	***6,9***
3	2	94,5	104,4	***9,9***
4	2	90,3	94,6	***4,3***
5	2	60,0	67,2	***7,2***
6	2	87,2	93,5	***6,3***
7	2	93,4	103,3	***9,9***
8	2	78,8	85,8	***7,0***
9	2	80,3	88,1	***7,8***
10	2	103,2	111,5	***8,3***
			$\sum\|X_R\| =$	***140,3***

- calcolare la loro media $\overline{X}_R = \dfrac{\sum|X_R|}{nr} = \dfrac{140,3}{10x2} = 7,015$

- individuare il valore d_2 = 1,91 (vedi tabella, con Z = 1 e W = 3)

Con questi dati, si calcola la ***Reproducibility*** o **intervallo di riproducibilità** del sistema, che è

$$Reproducibility = \sqrt{\left(5{,}15 \cdot \frac{\overline{X}_R}{d_2}\right)^2 - \frac{(Repeatability)^2}{nr}}$$

$$Reproducibility = \sqrt{\left(5{,}15 \cdot \frac{7{,}015}{1{,}91}\right)^2 - \frac{(20{,}66)^2}{10x2}} = \sqrt{(18{,}9148)^2 - 21{,}3418} = \sqrt{336{,}43} = 18{,}34$$

Concettualmente, **la riproducibilità valuta la differenza tra le medie di misure ottenute da operatori differenti**, ma che impiegano lo stesso strumento o metodo sui medesimi campioni.
Se l'operatore che utilizza lo strumento è solamente uno, non vi è riproducibilità, poiché è ovvio che non si ha variazione tra operatori. Una stima della riproducibilità è ricavata dalla media $\overline{X}$ di tutte le misure X fatte da ogni operatore.

Con tre operatori (A, B, C) la media
è

$$\overline{\overline{X}} = \frac{\overline{X}_A + \overline{X}_B + \overline{X}_C}{3}$$

L'intervallo di variazione R_O di un operatore è ottenuto
- sottraendo la media minima $\overline{X}_{\min}$ alla media massima $\overline{X}_{\max}$

La **deviazione standard della riproducibilità** S_O è

$$S_O = \frac{R_O}{d_2}$$

L'**intervallo al 99% della riproducibilità** (-2,715 <Z < +2,715)
è

$$5{,}15 \cdot \frac{\overline{\overline{R}}}{d_2} = 5{,}15 \cdot S_O$$

Nel calcolo dell'intervallo di variazione della **riproducibilità,** come evidenziato nella formula precedente, è compresa la quota

$$- \frac{(Repeatability)^2}{nr}$$

dove

- n = numero di campioni (***parts***),
- r = numero di prove o repliche (***trials***)

Il motivo logico di questa correzione (sinonimo di sottrazione) è che per stimare la riproducibilità, che è fondata sulle differenze tra le medie di due o più operatori, è necessario utilizzare tutte le misure da essi effettuate. Pertanto, la **stima della riproducibilità** deve essere diminuita della parte che dipende dalla variazione dovuta a ogni operatore (ripetabilità).

C - Da questi due parametri, si ricava un terzo indicatore:

- la **System** ***Repeatability and Reproducibility R&R***

con

$$R\,\&\,R = \sqrt{(Repeatability)^2 + (Reproducility)^2}$$

Con i dati dell'esempio, si ottiene

$$R\,\&\,R = \sqrt{(23{,}66)^2 + (18{,}34)^2} = \sqrt{559{,}78 + 336{,}36} = \sqrt{896{,}14} = 29{,}94$$

D - Un quarto indicatore è la ***variability***, che comprende

1 - la ***part variability*** V_p

$$V_p = 5{,}15 \cdot \frac{R_p}{d_2}$$

dove

- R_p = è differenza tra la media maggiore e la media minore di tutti gli n campioni (***parts***)
- d_2 è fornito dalla tabella precedente, con Z = 1 e W = numero di campioni (***parts***),

2 - la ***total variability*** V_T

$$V_T = \sqrt{(R\,\&\,R)^2 + (V_p)^2}$$

Con i dati dell'esempio (come evidenziato nella tabella successiva),

- dapprima si ricavano le medie delle 6 analisi (3 operatori x 2 prove) per tutti i 10 campioni impiegati nell'esperimento.

Successivamente si individuano

- quella maggiore = 103,97 (la media generale del campione 10)
- quella minore = 59,72 (la media generale del campione 5)
- per ottenere la **loro differenza** R_p

$$R_p = 103{,}97 - 59{,}72 = 44{,}25$$

Metodo o strumento X							
	Operatore A		Operatore B		Operatore C		
Campione	Prova 1	Prova 2	Prova 1	Prova 2	Prova 1	Prova 2	***Medie***
1	65,2	60,1	62,9	56,3	71,6	60,6	***62,78***
2	85,8	86,3	85,7	80,5	92,0	87,4	***86,28***
3	100,2	94,8	100,1	94,5	107,3	104,4	***100,22***
4	85,0	95,1	84,8	90,3	92,3	94,6	***90,35***
5	54,7	65,8	51,7	60,0	58,9	67,2	***59,72***
6	98,7	90,2	92,7	87,2	98,9	93,5	***93,53***
7	94,5	94,5	91,0	93,4	95,4	103,3	***95,35***
8	87,2	82,4	83,9	78,8	93,0	85,8	***85,18***
9	82,4	82,2	80,7	80,3	87,9	88,1	***83,60***
10	100,2	104,9	99,7	103,2	104,3	111,5	***103,97***

Successivamente, con d_2 = 3,18 (vedi tabella, con Z = 1 e W = 10)

si ricavano

- la ***part variability***

$$V_p = 5{,}15 \cdot \frac{R_p}{d_2} = 5{,}15 \cdot \frac{44{,}25}{3{,}18} = 71{,}66$$

- la ***total variability*** che somma gli effetti di ***R&R*** e della ***part variability***

$$V_T = \sqrt{(R\,\&\,R)^2 + (V_p)^2} = \sqrt{(29{,}94)^2 + (71{,}66)^2} = \sqrt{6031{,}56} = 77{,}66$$

24.12. LA CAPABILITY CON IL SEI SIGMA NORMALE E MOTOROLA.

La **ripetibilità** e la **riproducibilità** sono misure della variabilità, quando calcolata su dati **campionari**, come nell'esperimento preso a riferimento nel paragrafo precedente. In altre condizioni, come nei prodotti industriali,

- **la variabilità dei dati campionari deve essere rapportata ai limiti di tolleranza, definiti per il prodotto**, per valutare se esso rispetta la qualità richiesta.

E' un problema che si pone sempre più frequentemente nella professione del biologo, del chimico e dell'ambientalista, quando si effettua un controllo di qualità secondo le varie norme ISO.

Ad esempio, in farmacologia si supponga che una fiala debba contenere 57 mg di principio attivo, con un campo di variazione che nelle specifiche dell'azienda e per motivi clinici debba essere compreso tra il limite inferiore $L_1 = 55$ e il limite superiore $L_2 = 59$ mg.

Se su un gruppo di dati campionari la sua R&R è risultata uguale a 5,89, si può ricavare

- il **% di Tolleranza**

con

$$\%Tolerance = \frac{R\,\&\,R}{L_2 - L_1} \cdot 100 = \frac{5{,}89}{59 - 55} \cdot 100 = 147\%$$

che può essere espresso anche come **Tolleranza**

con

$$Tolerance = \frac{R\,\&\,R}{L_2 - L_1} = \frac{5{,}89}{59 - 55} = 1{,}47$$

Tuttavia, le misure attualmente più diffuse sono gli indici di **Capability**, nati nell'industria automobilistica. Da alcuni anni, sono richieste per molti prodotti sottoposti a certificazioni internazionali.

I concetti di base possono esser spiegati in modo semplice con un esempio. Per essere assemblato in modo funzionale, ogni pezzo di un'auto non può essere né troppo grande, né troppo piccolo. Quindi deve essere controllata sia media, sia la varianza della produzione, estraendo da essa un campione random. Gli **indici di Capability** riportati con frequenza maggiore sono

1 - il **Cp** (più raramente indicato con **Cm**) che è una misura di **variabilità,** determinata da **errori random:**

$$Cp = \frac{To - Tu}{6 \cdot \sigma} \quad oppure \quad \frac{To - Tu}{3 \cdot \sigma}$$

dove

- To e Tu sono il valore inferiore e superiore dei limiti di tolleranza,

2 - il **Cpk** (più raramente **Cmk**) che è una misura di **centratura**,

determinata da **errori sistematici**:

$$Cpk = \frac{\overline{X} - Tn}{6 \cdot \sigma} \quad oppure \quad \frac{\overline{X} - Tn}{3 \cdot \sigma}$$

dove

- Tn è il limite di tolleranza (To oppure Tu) che più vicino alla media effettiva della produzione.

Scomposizioni del **Cpk,** per analisi più particolareggiate, sono

3 - il **Cpu**

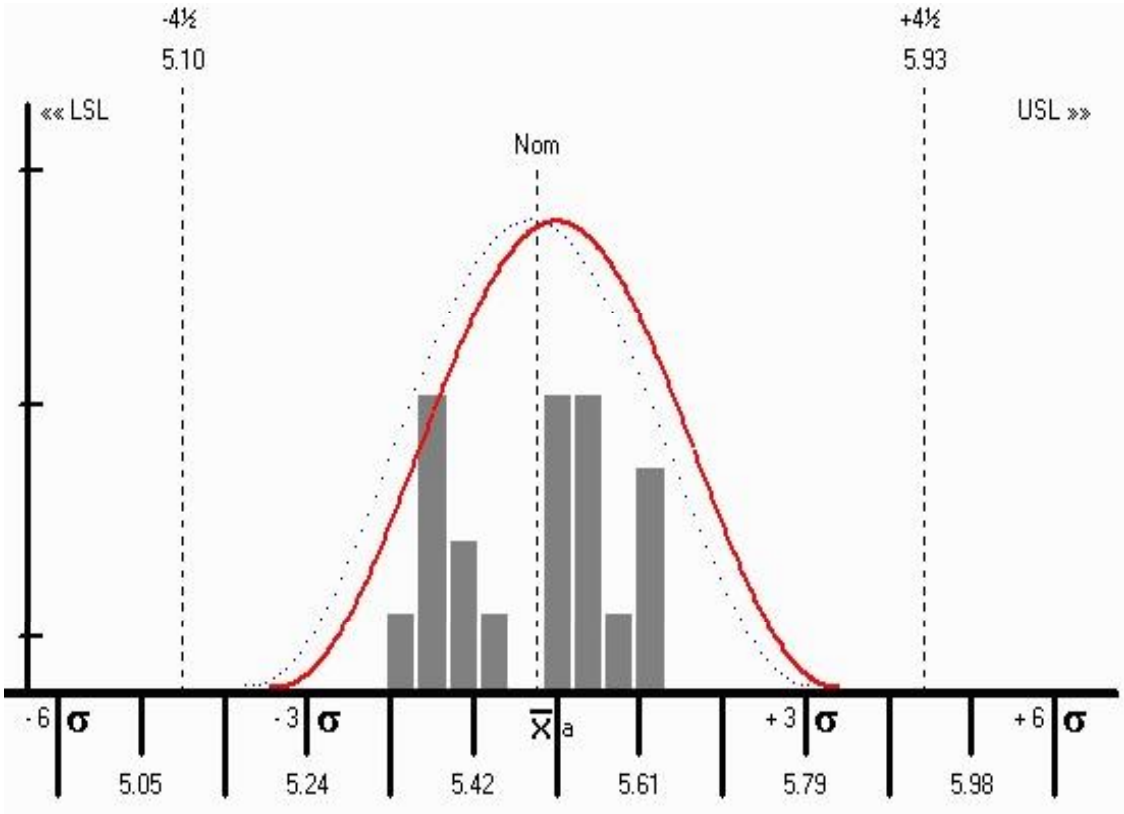
$$Cpu = \frac{\overline{X} - Tu}{6 \cdot \sigma} \quad oppure \quad \frac{\overline{X} - Tu}{3 \cdot \sigma}$$

4 - il **Cpo**

$$Cpo = \frac{To - \overline{X}}{6 \cdot \sigma} \quad oppure \quad \frac{To - \overline{X}}{3 \cdot \sigma}$$

che servono per valutare se gli errori effettivi sono più frequenti verso il limite di tolleranza inferiore oppure quello superiore.

Per comprendere esattamente i concetti, è utile analizzare la figura sottostante:



Come in qualsiasi processo industriale, prima di iniziare la produzione deve essere decisa

- la **dimensione ideale** e la **variabilità casuale del prodotto (**la distribuzione normale, con linea continua che è definita da una media μ e dalla sua deviazione standard σ).

Questa curva teorica deve essere sempre rapportata

- ai **limiti di tolleranza T_u e T_o**, che rappresentano i valori entro i quali è assicurata la piena funzionalità di ogni pezzo.

Tuttavia, la dimensione effettiva del prodotto**, analizzata su un campione e descritta dalla** curva normale tratteggiata**, potrà discostarsi dalla distribuzione teorica o programmata per due motivi:**

- gli errori casuali o random**, dovuti a fattori non prevedibili o comunque non controllati, con il risultato che la variabilità del prodotto è maggiore di quella programmata;**

- gli errori sistematici o bias**, dovuti alla perdita di taratura o centratura della macchina se non del sistema produttivo in genere.**

Nel caso rappresentato dalla figura precedente, la dimensione effettiva del prodotto **(descritta dalla curva tratteggiata),**

- ha una deviazione standard non diversa **da quella ipotizzata, poiché la curva ha la stessa forma o dispersione di quella programmata (linea continua);**

- ma la centratura si è spostata**, come indica il fatto che sia la curva tratteggiata sia la media campionaria $\overline{X}$ sono più vicine al limite inferiore che a quello superiore.**

Come interpretare questi risultati sotto l'aspetto della produzione?

Gli **errori sistematici non sono gravi**: è sufficiente ricalibrare la macchina o il sistema. E' una operazione tecnica che non determina problemi di investimenti; quasi sempre, lo strumento o il processo produttivo possono essere corretti con facilità.

Gli **errori casuali** sono quelli che richiedono il **controllo di qualità**. Con questo procedimento, si vuole determinare un miglioramento continuo della produzione, in modo che la variabilità del prodotto sia sempre minore e ogni pezzo sia sempre più vicino al valore ideale richiesto. E' un'operazione costosa, che richiede una organizzazione del lavoro accurata e/o la sostituzione della macchina o comunque a un suo sostanziale miglioramento in parti fondamentali.

Nella figura precedente, la **produzione è sotto controllo** (***in control***), in quanto

- solamente **una quota infinitesima del prodotto esce dai limiti di tolleranza** e quindi una parte minima del prodotto è da scartare in quanto **non conforme alle specifiche** (teoricamente non si raggiunge mai lo zero, poiché la curva normale arriva sull'asse delle ascisse solo in modo asintotico, non descrivibile nella rappresentazione grafica).

Gli indici **Cp** e **Cpk** al denominatore hanno $n \cdot \sigma$, vale a dire

- il numero n di volte in cui la deviazione standard σ del prodotto è compresa entro i limiti di tolleranza.

Questo valore è importante, poiché definisce la quantità relativa di scarti.

Nella figura precedente, dalla media della produzione **i limiti di tolleranza sono collocati a** $4{,}5 \cdot \sigma$**.**

Di norma, sono utilizzati valori che vanno da $3 \cdot \sigma$ a $6 \cdot \sigma$ come nelle due figure successive.

Quando l'indice **Cp**

$$Cp = \frac{To - Tu}{6 \cdot \sigma} \quad oppure \quad \frac{To - Tu}{3 \cdot \sigma}$$

è

- **Cp < 1** significa che i limiti di tolleranza sono minori della dispersione del prodotto; quindi si avrà una quota di scarti, che potrà essere più o meno elevata in funzione di $3 \cdot \sigma$ oppure $6 \cdot \sigma$**;**
- **Cp = 1** significa che, alla probabilità prefissata con $3 \cdot \sigma$ oppure $6 \cdot \sigma$, i limiti di tolleranza coincidono con quelli della produzione;
- **Cp > 1** significa che la produzione è entro i limiti di tolleranza e quindi la quota di scarto è totalmente trascurabile.

Nel figura precedente, l'indice **Cp** con $3 \cdot \sigma$ risulta **>1**. Infatti, seppure non esattamente centrata, la distribuzione reale dei pezzi è compresa ampiamente entro i limiti di tolleranza prestabiliti.

Quando l'indice **Cpk**

$$Cpk = \frac{\overline{X} - Tn}{6 \cdot \sigma} \quad oppure \quad \frac{\overline{X} - Tn}{3 \cdot \sigma}$$

è

- Cpk < 1 significa che la media reale della produzione, misurata su un campione, non è al centro rispetto ai due limiti di tolleranza, ma è più spostata verso un estremo. Ne deriva che gli scarti sono superiori alla quantità prevista.

Mediante gli indici **Cpu** e **Cpo** è possibile quantificare la quota di scarti verso i due estremi.

- Cpk > 1 significa che l'effetto è trascurabile, per quanto riguarda la quota di scarti.

Nel figura precedente, l'indice **Cpk** con la deviazione standard $3 \cdot \sigma$ risulta **>1**

Nel paragrafo precedente, per calcolare gli intervalli di confidenza della variabilità è stata utilizzata la **costante 5,15.** Essa è la somma del valore Z (± 2,575) della distribuzione normale unilaterale al 0,5% per ottenere la probabilità complessiva di errore di 1% in entrambe le code della distribuzione.

Negli indici di qualità industriale, con la stessa logica sono utilizzati $3 \cdot \sigma$ e $6 \cdot \sigma$.

Anche in questo caso, per comprendere esattamente i concetti sono utili le rappresentazioni grafiche Nella figura successiva, sono riportate **le specifiche** di un prodotto, con

Repeatability = 3-Sigma at $\tilde{n}$ 25(m

spesso scritte come

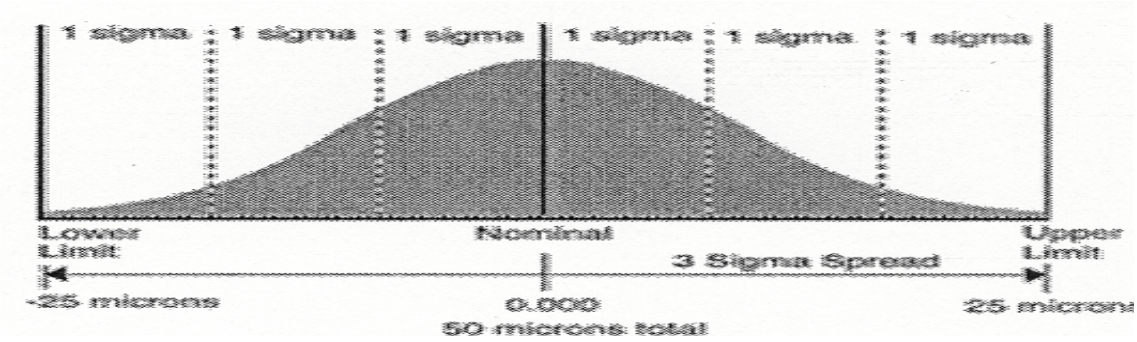
Repeatability = 3-Sigma @ $\tilde{n}$ 25(m

Questa simbologia indica che, rispetto al valore centrale prestabilito o **nominale** (***nominal***), la distanza tra limiti di tolleranza è di 25 unità di misura nelle due direzioni. In questo caso, corrisponde a 25 micron (millesimi di millimetro) e

- comprende il 99,973% dei pezzi prodotti.

Reciprocamente, gli scarti in percentuale sono solamente lo 0,027% o 2.700 pezzi per milione.

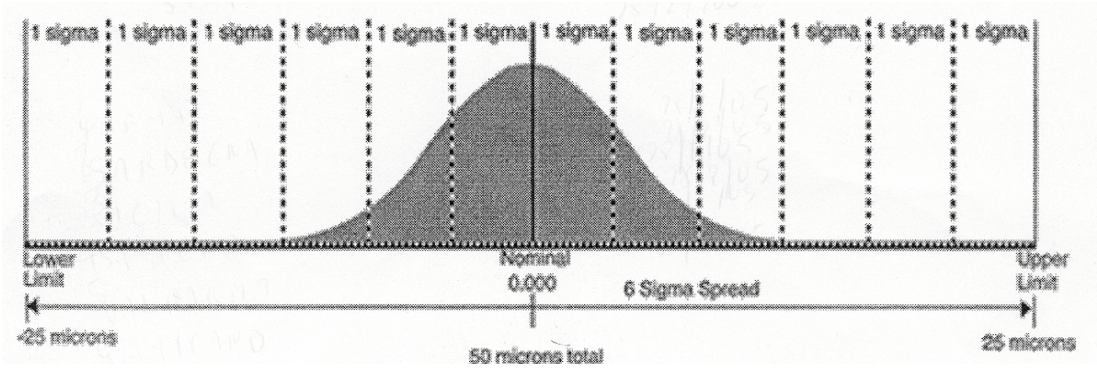
E' la quantità esclusa dai limiti di tolleranza $\pm 3 \cdot \sigma$, in una distribuzione normale.



Una qualità industriale nettamente migliore della precedente è assicurata dalle specifiche

Repeatability = 6-Sigma @ $\tilde{n}$ 25(m

come nella figura successiva.



Questa simbologia indica che, rispetto al **valore nominale** e entro una distanza di 25 unità di misura nelle due direzioni,

- è compreso il 99,9999998% dei pezzi prodotti
- e, reciprocamente, gli scarti sono solamente 2 ogni miliardo di unità.

E' la quantità compresa o esclusa dai limiti di tolleranza $\pm 6 \cdot \sigma$ di una distribuzione normale. In termini industriali, rappresenta la perfezione nella produzione. Gli scarti sono nulli e si parla di qualità totale. Anche macchine composte da migliaia di pezzi, come un aereo, un computer o apparecchi complessi, raramente avranno disfunzioni, se ogni suo pezzo non funzionerà correttamente solo con tale frequenza.

Livello – Sigma	Errori per milione
± un sigma	271.800
± due sigma	42.800
± tre sigma	2.700
± sei sigma	0,002

Il termine **6 Sigma** (***Six Sigma***) è quindi diventato sinonimo di produzione perfetta, nella quale non esistono errori o scarti. In realtà, occorre porre attenzione ai limiti che sono specificati. Ad esempio, se la specifica precedente fosse stata

Repeatability = 6-Sigma @ $\tilde{n}$ 50(m

quindi con un intervallo dei limiti di tolleranza doppio rispetto a quelli prima indicati,

avremmo avuto un prodotto del tutto uguale a

Repeatability = 3-Sigma @ $\tilde{n}$ 25(m

In conclusione, quando si descrive la **ripetibilità** di un processo, i concetti importanti sono

- ogni processo può essere definito 6-sigma, ma la sua **reale accettabilità industriale dipende dai limiti inferiori e superiori di variabilità**;
- **il termine 6-sigma indica che gli errori hanno frequenze bassissime**; ma esso deve sempre essere accompagnato dall'indicazione dei limiti di tolleranza;
- si determina un effettivo miglioramento nella ripetibilità di un processo, quale il passaggio da 3 a 6 sigma, quando i limiti non vengono modificati ma **si riduce la deviazione standard**.

Ad esempio, nelle due figure precedenti, si è passati da una qualità 3-sigma a una 6-sigma perché

- **i limiti di tolleranza sono identici**,
- mentre **la deviazione standard dei σ prodotti è stata dimezzata**.

I valori di riferimento di qualità non sono ovviamente solamente 3 e 6 sigma. Il valore 6 sigma rappresenta un obiettivo e quindi nella realtà produttiva la frequenza di errori può essere maggiore di quella desiderata, pure restando in un campo di alta accettabilità della qualità del prodotto.
In questo contesto, si comprende come in un processo **6 sigma** un **Cp compreso tra 1 e 2,** frequente nella pratica, possa indicare che si sta passando da una qualità 3-sigma alla qualità 6-sigma, senza averla ancora raggiunta.

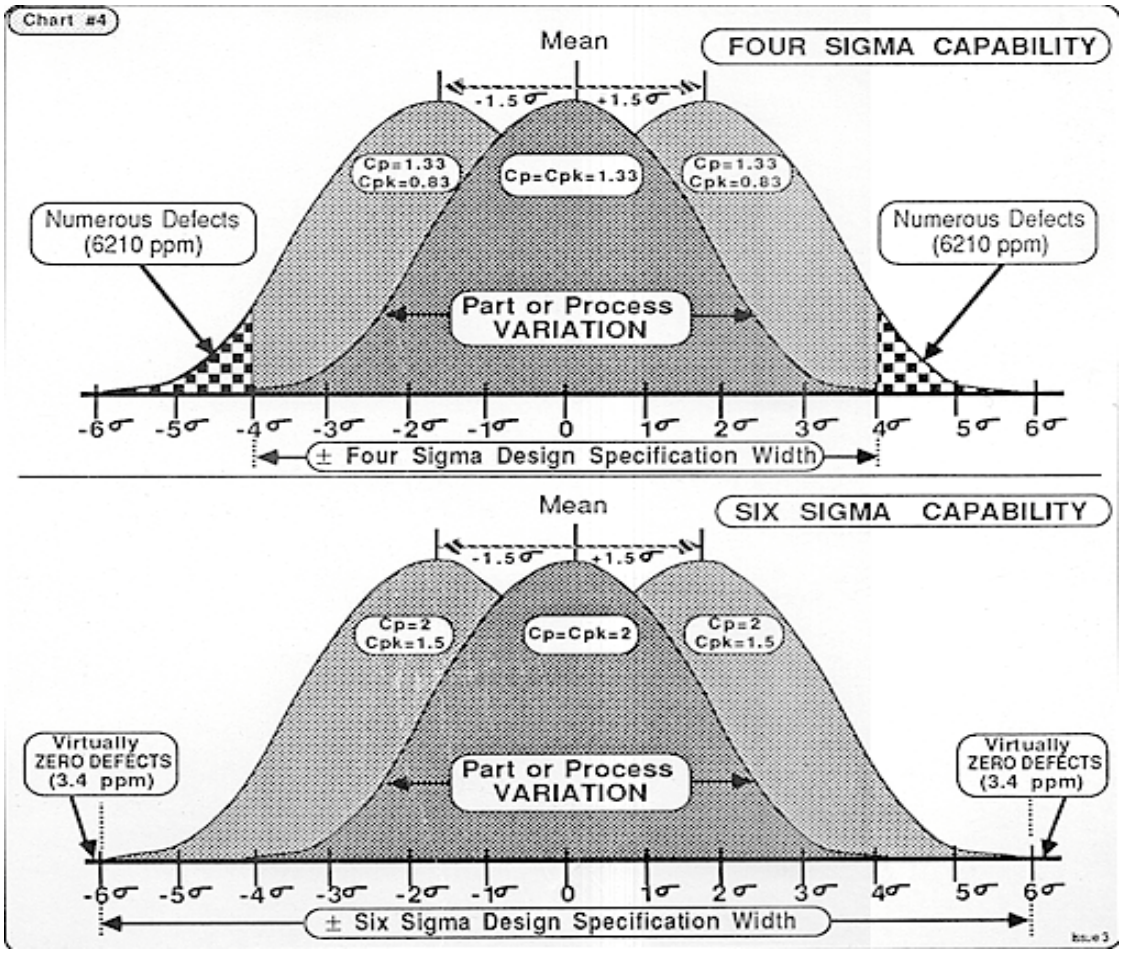
Nel linguaggio industriale, una produzione 6-sigma spesso indica una **frequenza di errori** non pari a 2 unità su un miliardo come indicato in precedenza, ma una frequenza nettamente più alta,
- pari a **3,4 per un milione di unità**.
E' il 6-sigma Motorola.

Sei Sigma Motorola		
Livello – Sigma	Errori per milione	Errori in percentuale
± un sigma	690.000	69,0 %
± due sigma	308.537	30,85 %
± tre sigma	66.807	6,68 %
± quattro sigma	6.210	0,621 %
± cinque sigma	233	0,0233 %
± sei sigma	3,4	0,00034 %

L'obiettivo industriale ***Six-Sigma*** è il miglioramento della qualità, la quasi perfezione del prodotto, per cui gli errori o difetti di produzione sono solamente 3,4 ogni milione di unità prodotte. In termini di perfezione è il 99,99966%.
Non è da confondere con le certificazioni del controllo di qualità come ISO-9001 o nelle discipline ambientali ISO-1401. E' solamente una metodologia, per ridurre la quantità di prodotti difettosi, che è fondato sul miglioramento del processo produttivo, anche se impiega metodi statistici per valutare se gli obiettivi di qualità sono stati raggiunti.
La metodologia **Sei Sigma Motorola** (così definita in quanto applicata la prima volta nel 1989 da questa azienda) è fondata sul concetto che in inglese è chiamato **DMAIC** (da ***Define***, ***Measure***, ***Analyze, Improve***, ***Control***) e indica le azioni di organizzazione produttiva che occorre seguire per raggiungere quel livello di qualità.

Il motivo per cui la quota di errori della procedura **6-sigma Motorola** non coincide con quella della distribuzione normale bilaterale è che, nella stima della frequenza di errori o scarti,
- è tenuta in considerazione anche la possibile perdita di centratura **o** accuracy della macchina**, un eventuale spostamento(***drift***) della media (non previsto, ma tollerabile nella produzione) pari a 1,5 sigma in entrambe le direzioni.**

E’ una stima della proporzione di errori che vuole essere più vicina alla realtà, in quanto la perdita di centratura è un evento relativamente frequente, comunque da tenere in valutazione attenta.
Nella figura sottostante, è rappresentata la distribuzione normale in un processo 4-sigma (distribuzione normale centrale). Se si tiene in considerazione che la produzione deve rispettare le specifiche anche se si perde la centratura per una quota pari a 1,5 sigma in entrambe le direzioni, si ha che
- la quota di errori diventa di 6.210 unità ogni milione di pezzi.



In un processo 6-sigma, nel quale ovviamente i limiti di tolleranza siano identici (la figura inferiore dovrebbe essere più stretta e i valori 6 sigma coincidere con i valori 4 sigma della figura superiore), si ha che
- la quota di errori diventa di 3,4 unità ogni milione di pezzi.

24.13 LA RIPETIBILITA' E LA RIPRODUCIBILITA' CON LE VARIANZE DELL'ANOVA, IN UN DISEGNO SPERIMENTALE A DUE CRITERI CON REPLICHE

L'analisi della varianza (ANOVA) rappresenta il metodo più accurato per quantificare la **ripetibilità** (***repeatability***) e la **riproducibilità** (***reproducibility***) di analisi cliniche o di un prodotto, con una procedura che nel linguaggio internazionale è chiamata ***gauge R&R studies***. Rispetto ai metodi statistici illustrati nei paragrafi precedenti, questa offre la possibilità di valutare anche l'**interazione tra operatori** (***appraiser***) **e campioni** (***parts***). (Vedere il concetto di interazione tra due fattori, nel capitolo relativo).

Il modello classico di raccolta dei dati, come nella tabella sottostante, prevede un'analisi della varianza a due criteri con repliche**, rappresentati rispettivamente da**

- un numero di **operatori** (***appraisers***) costante e quasi sempre pari a 3; anche qui $a = 3$;
- un numero di **campioni** (***parts, units***) compreso tra 5 e 10; in questo caso $b = 10$;
- un numero di **prove o ripetizioni** (***trials, replications***) da 2 a 3; in questo caso $n = 2$;

in un **disegno** (***design of experiment***) **bilanciato,** con N = 60 osservazioni ($N = a \cdot b \cdot n$)

Metodo o strumento X						
Campione	Operatore A		Operatore B		Operatore C	
	Prov 1	Prov 2	Prov 1	Prov 2	Prov 1	Prov 2
1	65,2	60,1	62,9	56,3	71,6	60,6
2	85,8	86,3	85,7	80,5	92,0	87,4
3	100,2	94,8	100,1	94,5	107,3	104,4
4	85,0	95,1	84,8	90,3	92,3	94,6
5	54,7	65,8	51,7	60,0	58,9	67,2
6	98,7	90,2	92,7	87,2	98,9	93,5
7	94,5	94,5	91,0	93,4	95,4	103,3
8	87,2	82,4	83,9	78,8	93,0	85,8
9	82,4	82,2	80,7	80,3	87,9	88,1
10	100,2	104,9	99,7	103,2	104,3	111,5

Con questo disegno sperimentale, il modello dell'ANOVA

è

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

dove

- X_{ijk} è la misura della prova k ottenuta dall'operatore i sul campione j,

- μ è la dimensione del fenomeno misurato, la cui stima migliore è fornita dalla media generale $\overline{\overline{X}}$;

- α_i è l'effetto dell'operatore i, la cui stima è fornita da $\overline{X}_i - \overline{\overline{X}}$;

- β_j è l'effetto dell'operatore j, la cui stima è fornita da $\overline{X}_j - \overline{\overline{X}}$;

- ε_{ijk} è l'errore che ogni operatore commette nell'analisi replicata dello stesso campione e rappresenta il **contributo dell'operatore alla ripetibilità**.

L'ANOVA a fattori fissi (nel capitolo relativo, vedi formule per ottenere le diverse quantità) fornisce il seguente risultato

Fonte di variazione	SQ	GDL	S^2	F	P
Totale	12,630,4	59	---	---	---
Operatori	502,5	2	251,3	13,8	0,0001
Campioni	11.545,5	9	1.282,8	70,4	0,0000
Operatori x Campioni	35,6	18	1,98	0,11	0,9999
Errore	546,8	30	**18,2**	**---**	**---**

Con questi dati, si ottengono i seguenti indici.

1 - Il valore della **Ripetibilità**:

è

$$Repeatability = 5{,}15 \cdot \sqrt{S_e^2} = 5{,}15 \cdot \sqrt{18{,}2} = 21{,}97$$

e risulta uguale a 21,97.

Nella stima precedente era risultata uguale a 23,66.

2 - Il valore della **Riproducibilità:**

è

$$Reproducibility = 5{,}15 \cdot \sqrt{\frac{S^2_{Operatori} - S^2_{Interazione}}{b \cdot n}} = 5{,}15 \cdot \sqrt{\frac{251{,}3 - 1{,}98}{10x2}} = 5{,}15 \cdot \sqrt{12{,}466} = 18{,}18$$

e risulta uguale a 18,18

Nella stima precedente era risultata uguale a 18,34.

3 – Il valore dell'**interazione tra operatori e campioni:**

è

$$Interazione = 5{,}15 \cdot \sqrt{\frac{S^2_{Interazione} - S^2_e}{n}} = 5{,}15 \cdot \sqrt{\frac{1{,}98 - 18{,}2}{10}} = immaginario$$

Con i dati di questo esempio, non può essere calcolata, poiché la radice quadrata di un numero negativo è un numero immaginario. Di norma, è una risposta diversa da questa, poiché

- la varianza d'interazione $S^2_{Interazione}$ è maggiore di quella d'errore S^2_e,

come avviene quando in un gruppo di operatori diversi alcuni forniscono sistematicamente valori medi maggiori e altri valori minori, analizzando lo stesso campione.

L'interazione ha una varianza uguale a quella d'errore e quindi determina un indice di ***Interazione* = 0**, quando non esistono differenze sistematiche tra le medie degli operatori.

In questo caso, per fornire comunque una risposta logica nonostante il valore sia immaginario, si assume che la varianza d'interazione sia uguale a quella d'errore: $S^2_{Interazione} - S^2_e = 0$.

4 - Il valore della **Ripetibilità e Riproducibilità** detto ***R&R*:**

è

$$R\,\&\,R = \sqrt{(Repeatability)^2 + (Reproducibility)^2 + (Interazione)^2}$$

$$R\,\&\,R = \sqrt{(21{,}97)^2 + (18{,}18)^2 + (0)^2} = \sqrt{813{,}19} = 28{,}52$$

e risulta uguale a 28,52.

Nella stima precedente, ***R&R*** era risultata uguale a 29,94.

5 - Il valore della **system part variation V_p**:

è

$$V_p = 5{,}15 \cdot \sqrt{\frac{S^2_{Campioni} - S^2_{Interazione}}{a \cdot n}} = 5{,}15 \cdot \sqrt{\frac{1282{,}8 - 1{,}98}{3x2}} = 5{,}15 \cdot \sqrt{213{,}47} = 75{,}24$$

e risulta uguale a 75,24.

Nella stima precedente, $\mathbf{V_p}$ era risultata uguale a 71,66.

6 - Il valore della **total measurement system variation V_T**

è

$$V_T = \sqrt{(R \,\&\, R)^2 + (V_p)^2} = \sqrt{(28{,}52)^2 + (75{,}24)^2} = \sqrt{6474{,}45} = 80{,}46$$

e risulta uguale a 80,46.

Nella stima precedente, $\mathbf{V_T}$ era risultata uguale a 76,66.

E' semplice osservare come, rispetto al paragrafo nel quale sono stati misurati questi indici con una metodologia differente, i risultati siano molto vicini anche se non identici. Tra i due metdodi, questo risulta più accurato e completo. Ha lo svantaggio di richiedere una quantità maggiore di calcoli, che è facilmente superabile con l'uso dei programmi informatici

24.14. STIMA DELLE DIMENSIONI MINIME DEL CAMPIONE, PER UN'ANALISI DELLA RIPETABILITA'.

Come già presentato in precedenza, negli studi di **ripetibilità** le dimensioni del campioni sono definite da **indicazioni standard**. In caso di **prove non distruttive**, l'**esperimento raccomandato** dagli organismi internazionali addetti al Controllo di Qualità è fondato su

- un numero di **campioni** (***parts, units***) compreso tra 5 e 10,
- un numero di operatori (***appraisers***) costante e pari a 3,
- un numero di **prove** o ripetizioni (***trials, replications***) da 2 a 3.

Metodo o strumento X												
	Operatore A				Operatore B				Operatore C			
Campione	P 1	P 2	P *j*	P *m*	P 1	P 2	P *j*	P *m*	P 1	P 2	P *j*	P *m*
1	---	---	---	---	---	---	---	---	---	---	---	---
2	---	---	---	---	---	---	---	---	---	---	---	---
---	---	---	---	---	---	---	---	---	---	---	---	---
i	---	---	---	---	---	---	---	---	---	---	---	---
---	---	---	---	---	---	---	---	---	---	---	---	---
n	---	---	---	---	---	---	---	---	---	---	---	---

E' quindi possibile scegliere entro uno spettro di possibili esperimenti, che è abbastanza ampio. Inoltre, la dimensione del campione può essere ulteriormente aumentata nella fase di programmazione dell'esperimento, soprattutto per quanto riguarda le repliche o prove.

Ma è utile capire come si devono stimare le dimensioni di un esperimento, sia per non costruire campioni eccessivi e quindi troppo costosi oppure troppo piccoli, sia perché esse sono strettamente legate alla varianza d'errore, il parametro fondamentale per una produzione di alta qualità.
Considerando il caso più semplice di un singolo operatore

Campione	Operatore A			
	P 1	P 2	P *j*	P *m*
1	---	---	---	---
2	---	---	---	---
---	---	---	---	---
i	---	---	---	---
---	---	---	---	---
n	---	---	---	---

dove i dati possono essere analizzati con l'analisi della varianza a due criteri, la precisione con la quale è possibile stimare la deviazione standard dell'errore S_e dipende

- dal numero n di campioni
- dal numero m di prove o repliche effettuate dallo stesso soggetto sullo stesso campione.

L'**intervallo di confidenza** alla probabilità α
è dato da

$$Z_\alpha \cdot \frac{\sigma_e}{\sqrt{2 \cdot n \cdot (m-1)}}$$

Per la probabilità del 95% ($\alpha = 0.05$) e quindi con Z = 1,96
è

$$1{,}96 \cdot \frac{\sigma_e}{\sqrt{2 \cdot n \cdot (m-1)}}$$

dalla quale si ricava

$$\sigma_e = \frac{1{,}96}{\sqrt{2 \cdot n \cdot (m-1)}}$$

E' una equazione con due incognite.

La soluzione è data da molte combinazioni di n e m, sempre per ottenere la stessa precisione richiesta, misurata dalla **deviazione standard d'errore** (σ_e).

Se il campione fosse solamente uno (n = 1), sarebbe semplice stimare m, il numero di analisi ripetute sullo stesso campione. Ma questa stima dell'**errore entro** (ricavata da un solo campione)

- vale a dire della **ripetibilità** delle analisi sullo stesso campione,
- **non è concettualmente corretta**, poiché campioni diversi possono avere concentrazioni molto differenti che spesso in analisi ripetute determinano una variazione molto alta del valore.

Nell'ANOVA la deviazione standard dell'errore σ_e è fondata su un'assunzione molto semplicistica: la variazione tra prove è uguale per ogni campione.

E' un'ipotesi adeguata per stimare la variazione media.

Ma quando serve una stima abbastanza precisa, è necessario **utilizzare campioni che hanno livelli di concentrazione molto differenti**: essi possono avere un errore diverso in valore assoluto, quello che determina la deviazione standard σ_e, poiché spesso l'errore nella misurazione è una percentuale costante del valore della concentrazione.

ESEMPI.

Si assuma di voler ottenere σ_e = 0,10 con α = 0.05 in una distribuzione bilaterale.

Se è fissato n, il valore di m è

$$m = 1 + \frac{(1{,}96)^2}{\sigma_e^2 \cdot 2 \cdot n}$$

Ad esempio, con n = 20, il valore di m

è

$$m = 1 + \frac{(1{,}96)^2}{(0{,}10)^2 \cdot 2 \cdot 20} = \frac{3{,}8416}{0{,}4} = 1 + 9{,}6 = 10{,}6$$

uguale a 10,6 che deve essere arrotondato all'unità superiore: m = 11

Servono quindi 11 repliche per ognuno dei 20 campioni.

Mantenendo costanti gli altri parametri, se per ogni campione si decide di effettuare solo due prove, quanti campioni servono?

Dalla formula

$$n = \frac{(1{,}96)^2}{\sigma_e^2 \cdot (m-1)}$$

con m = 2 si ricava

$$n = \frac{(1{,}96)^2}{(0{,}10)^2 \cdot (2-1)} = \frac{3{,}8416}{0{,}01} = 384{,}16$$

che servono 385 campioni.

E' un valore molto alto, che dipende dal fatto che si è voluta ottenere una **accuratezza molto alta**, vale a dire un errore σ_e molto piccolo.

Poiché esso è al denominatore ed è elevato al quadrato, è sufficiente raddoppiarlo (σ_e = 20), per ridurre a un quarto il campione necessario:

$$n = \frac{(1{,}96)^2}{(0{,}20)^2 \cdot (2-1)} = \frac{3{,}8416}{0{,}04} = 96{,}04$$

Con 4 repliche (m = 4)

$$n = \frac{(1{,}96)^2}{(0{,}20)^2 \cdot (3-1)} = \frac{3{,}8416}{0{,}12} = 32{,}01$$

il campione richiesto è n = 33 unità.

In questo modo, sono possibili tutte le combinazioni, per scegliere quella più adeguata a calcolare come la misurazione vari sia **tra operatori**, sia **tra repliche** effettuate dallo stesso operatore.

24.15 LE COMPONENTI DELLA VARIANZA NEGLI STUDI R&R, CON L'ANOVA A EFFETTI RANDON, FISSI E MISTI

I modi per analizzare gli stessi dati con l'ANOVA possono essere differenti. Dipende dalle modalità di scelta del campione, dalle **ipotesi che vengono formulate** e quindi dal **modello** che viene applicato: a **fattori random**, a **fattori fissi** oppure a **fattori misti** (vedi, nel capitolo relativo, concetti e metodi).

1 – Il **modello a fattori random** è quello più frequente negli studi ***gage R&R***.

L'obiettivo è verificare

- se **operatori diversi** (non importa quali) e **campioni differenti** (non importa quali) forniscono risposte simili oppure presentano una variabilità statisticamente significativa.

Teoricamente, operatori e campioni sono scelti casualmente da un pool più grande, per decidere se la loro varianza è significativamente differente da zero. Lo scopo non è valutare se le medie degli operatori sono statisticamente differenti, ma che tra esse esiste variabilità. Per analizzare l'interazione, è necessario che uno dei due fattori sia fisso. Per essa

si verifica

$$H_0 : \sigma^2_{\alpha\beta} = 0 \quad \text{contro} \quad H_1 : \sigma^2_{\alpha\beta} \neq 0$$

2 – Il modello a **fattori fissi** viene utilizzato quando il numero di operatori è piccolo e interessa solamente l'analisi di alcuni campioni, che hanno caratteristiche specifiche. E' da impiegare quando si è interessati a confrontare **le medie di due o più operatori chiaramente identificati** e di campioni che hanno caratteristiche particolari. Il risultato riguarda quegli operatori e quei livelli scelti e non può essere esteso o generalizzato ad altre situazioni.

3 – Il modello a **fattori misti** è applicato quando uno dei due fattori e fisso e l'altro è random. In una azienda, probabilmente il caso più frequente può essere quello di operatori fissi e di campioni random.

Nei tre modelli dell'ANOVA, le devianze con i loro gradi di libertà e quindi le varianze sono calcolati nello stesso modo. Nel test di significatività, i rapporti per calcolare F sono impostati in modo differente.

In un esperimento standard di analisi della varianza a due criteri con repliche,

indicando con

- S^2_A la varianza del fattore A,
- S^2_B la varianza del fattore B,
- S^2_{AB} la varianza d'interazione tra i fattori A e B,
- S^2_e la varianza d'errore,

i test F sono

Test F di Significatività	Model I (A e B fissi)	Model II (A e B random)	Model III (A fissa; B random)
A	S^2_A / S^2_e	S^2_A / S^2_{AB}	S^2_A / S^2_{AB}
B	S^2_B / S^2_e	S^2_B / S^2_{AB}	S^2_B / S^2_e
Interazione A x B	S^2_{AB} / S^2_e	S^2_{AB} / S^2_e	S^2_{AB} / S^2_e

Valutata la significatività, le varianze attese forniscono un modo per predire **le componenti della varianza** che è associata con ogni termine,

nel modello additivo

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

Si stima la **varianza attesa** per ogni componente di interesse, con le formule riportate nella tabella seguente:

Varianze Attese	Model I (A e B fissi)	Model II (A e B random)	Model III (A fissa; B random)
σ_A^2	$\sigma^2 + nb\dfrac{\sum \alpha_i^2}{a-1}$	$\sigma^2 + n\sigma_{\alpha\beta}^2 + nb\sigma_\alpha^2$	$\sigma^2 + n\sigma_{\alpha\beta}^2 + nb\dfrac{\sum \alpha_i^2}{a-1}$
σ_B^2	$\sigma^2 + na\dfrac{\sum \beta_j^2}{b-1}$	$\sigma^2 + n\sigma_{\alpha\beta}^2 + na\sigma_\beta^2$	$\sigma^2 + na\sigma_\beta^2$
σ_{AB}^2	$\sigma^2 + n\dfrac{\sum\sum (\alpha\beta)_{ij}^2}{(a-1)(b-1)}$	$\sigma^2 + n\sigma_{\alpha\beta}^2$	$\sigma^2 + n\sigma_{\alpha\beta}^2$
σ_e^2	σ^2	σ^2	σ^2

Il metodo statistico può essere illustrato nei suoi passaggi logici in modo semplice e operativo, presentando un esempio.

ESEMPIO 1. (MODEL II, A EFFETTI RANDOM).

Metodo o strumento X									
Campione	Operatore A			Operatore B			Operatore C		
	Prov 1	Prov 2	Prov 3	Prov 1	Prov 2	Prov 3	Prov 1	Prov 2	Prov 3
1	1019	1017	1018	1031	1031	1025	990	991	986
2	977	980	992	1001	1007	1010	962	966	952
3	992	1004	1001	1010	1025	1019	1015	1020	1013
4	988	991	982	1018	1018	1024	1023	1019	1027
5	967	981	971	997	992	1002	980	990	976

Dati e commenti sono tratti, con modifiche, dall'articolo di informazioni tecniche della General Electric Company, più esattamente dalla relazione di T. A. **Early** e R. **Neagu** del 1999 ***Random and Fixed Factor. ANOVA Models: Gauge R&R Studies*** (GE Research & Development Center, 99CRD094, Class 1, Technical Information Series, pp.: 1-10).

Si assuma che per valutare la ripetibilità di uno strumento che misura la glicemia in campioni di sangue, 3 tecnici di laboratorio analizzino gli stessi 5 campioni, effettuando ognuno 3 analisi indipendenti nelle stesse condizioni, con i risultati della tabella precedente.
Lo scopo è valutare se, utilizzando sempre la stessa procedura, esiste una variabilità significativa sia tra tecnici sia entro tecnici, per fornire una stima delle **componenti della *gage variance***.
E' una **ANOVA model II** o a **effetti random**.

L'analisi della varianza, della quale sono riportati i risultati,

Fattori	**SQ**	**GDL**	S^2	**F**	**P**
Totale	18742	44	---	---	---
A – Operatori	4440	2	2220,0	3,398	0,085
B – Campioni	8190	4	2047,5	3,134	0,079
Operatori x Campioni	5226	8	653,3	22,119	0,000
Errore	886	30	29,5	**---**	**---**

evidenzia soprattutto l'**altissima significatività della interazione**.
Come mostra la successiva tabella delle medie,

Campione	Medie $\overline{X}_{ij}$			
	Oper. A	Oper. B	Oper. C	$\overline{X}_{i.}$
1	*1018*	*1029*	*989*	**1012**
2	*983*	*1006*	*960*	**983**
3	*999*	*1018*	*1016*	**1011**
4	*987*	*1020*	*1023*	**1010**
5	*973*	*997*	*982*	**984**
$\overline{X}_{.j}$	**992**	**1014**	**994**	**1000**

tale significatività è determinata dal fatto che

- l'operatore A fornisce una stima sistematicamente maggiore della media dei tre operatori quando il valore è alto e minore quando in valore medio è basso, mentre l'operatore C fornisce risultati opposti.

Esiste poi variabilità tra campioni e tra operatori.

Ma mentre è logico che esista una differenza tra i campioni, per la ripetibilità del metodo di analisi è incongruo che la differenza tra operatori sia così grande. L'operatore B fornisce risultati sistematicamente maggiori di quelli degli altri due. E' conveniente individuarne la causa, per ottenere risultati corretti.

Nel calcolo **delle componenti della varianza,** utilizzando i dati della tabella precedente dove

- $n = 3$ è il numero di repliche effettuate dallo stesso operatore sullo stesso campione,
- $a = 3$ è il numero di operatori,
- $b = 5$ è il numero di campioni,

la successione di passaggi logici e dei calcoli è descritta nei quattro punti seguenti.

1 - La **prima componente** è la **varianza d'errore** o **errore puro** (***pure error***) σ^2, identificata nella **ripetibilità entro operatore** (***within-operator repeatability***) S_e^2.

Con i dati dell'esempio, è

$$\sigma^2 = S_e^2 = 29{,}5$$

2 - La **seconda componente** da stimare è il termine della **varianza d'interazione AB** σ_{AB}^2.

Con i dati dell'esempio, è

$$\sigma_{AB}^2 = \frac{S_{AB}^2 - S_e^2}{n} = \frac{653{,}3 - 29{,}5}{3} = 207{,}9$$

Questa **varianza attesa di interazione** σ_{AB}^2**,** è la varianza dell'**intera popolazione di operatori con l'intera popolazione di campioni.** Ovviamente, è calcolata a partire dai dati campionari S_{AB}^2 e S_e^2.

3 - La **terza componente** da stimare è il termine della **varianza di tutti gli operatori** σ_A^2.

Con i dati dell'esempio, è

$$\sigma_A^2 = \frac{S_A^2 - S_{AB}^2}{nb} = \frac{2220{,}0 - 653{,}3}{3 \cdot 5} = \frac{1566{,}7}{15} = 104{,}4$$

4 - La **quarta componente** da stimare è il termine della **varianza di tutti i campioni** σ_B^2.

Con i dati dell'esempio, è

$$\sigma_B^2 = \frac{S_B^2 - S_{AB}^2}{na} = \frac{2047{,}5 - 653{,}3}{3 \cdot 3} = \frac{1394{,}2}{9} = 154{,}9$$

Da queste quattro componenti , si ricava la **varianza stimata del metodo o strumento** σ_{gage}^2, la ***gage variance***.

In realtà essa riguarda solo tre fattori:

$$\sigma_{gage}^2 = \sigma^2 + \sigma_A^2 + \sigma_{AB}^2 = 29{,}5 + 104{,}4 + 207{,}9 = 341{,}8$$

poiché quella tra campioni non riguarda il metodo o lo strumento sottoposto a verifica.

Per un confronto e per meglio valutare le cause della variabilità nei risultati, è vantaggioso **trasformare le singole componenti dal valore assoluto alla percentuale di quella totale** del **gage**

$$\sigma_{gage}^2 = \sigma^2 + \sigma_A^2 + \sigma_{AB}^2 = 8{,}6\ \% + 30{,}6\ \% + 60{,}8\ \% = 100{,}0\ \%$$

Emerge con maggior chiarezza che, in questo caso,

1 - la variabilità dovuta alle singole ripetizioni è piccola e quindi **ogni operatore tende a dare sempre la stessa misura del medesimo campione**,

2 - **esiste una differenza tra operatori**, per cui alcuni tendono a fornire un valore maggiore e altri un valore minore della media generale, che rappresenta la stima migliore del valore vero, ma essa non è casuale o sistematica;

3 - infatti **esiste interazione**: le differenze tra operatori non sono costanti, ma cambiano con il valore della misura, se alto o basso. Questa ultima componente risulta quella maggiore. E' quindi importante descriverne le caratteristiche per individuarne le cause.

In questo caso, come in parte osservato nella tabella delle medie, l'operatore A

- nel campione 5 che ha una media generale bassa (984), fornisce una sua media (973) che è minore di quella degli altri due operatori,

- mentre nel campione 1 che ha una media generale alta (1012), fornisce una sua media (1018) che è maggiore di quella degli altri due operatori.

E' importante sottolineare che, dalla stima delle componenti della varianza, è esclusa quella dovuta alle differenze tra campioni. Il motivo è che **devono essere prese in considerazione solo le varianze che riguardano gli operatori**. Esse dovrebbero essere ridotte al minimo, per un **corretto funzionamento del *gage***.

In modo più dettagliato, ritornando all'analisi delle componenti,

- la **prima** (σ^2) riguarda la variabilità **di (entro) ogni operatore**, quando ripete le misure sullo stesso campione; in termini tecnici internazionali, è detta ***within-operator repeatability;***

- le **altre due** (σ_A^2 e σ_{AB}^2) riguardano le differenze **tra gli operatori**; insieme, formano la ***operator-to-operator reproducibility.***

ESEMPIO 2. (MODEL III, A EFFETTI MISTI CON GLI STESSI DATI)

Servendosi dello stesso disegno sperimentale e degli stessi dati, si supponga ora che i possibili utilizzatori dello strumento siano solamente i tre tecnici assunti dall'azienda**.** Essi quindi sono **un fattore fisso** e non più un campione **random** di tutti i possibili utilizzatori.

Diventa possibile applicare il **modello III**, supponendo che **i campioni analizzati siano random**.

Benché gli operatori siano un fattore fisso, l'**interazione operatori-campioni** rimane un fattore **random,** del quale pertanto può essere stimata la componente,

con la formula già riportata

$$\sigma_B^2 = \frac{S_B^2 - S_{AB}^2}{na} = \frac{2047{,}5 - 653{,}3}{3 \cdot 3} = \frac{1394{,}2}{9} = 154{,}9$$

Ovviamente, la variazione più interessante (quella dovuta agli operatori) non può essere stimata come varianza random. E' un limite grave di questa impostazione.

Il risultato dell'ANOVA diventa

Fattori	**SQ**	**GDL**	S^2	**F**	**P**
Totale	18742	44	---	---	---
A – Operatori	4440	2	2220,0	3,398	0,085
B – Campioni	8190	4	2047,5	69,328	0,000
Operatori x Campioni	5226	8	653,3	22,119	0,000
Errore	886	30	29,5	---	---

dove, come nell'esempio precedente,

- il valore di **F tra Operatori** è stato calcolato ponendo al denominatore la varianza d'interazione

$$(F = S^2_{Operatrori} \big/ S^2_{AB})$$

e, a differenza dell'esempio precedente,

- il valore di **F tra Campioni** è stato calcolato ponendo al denominatore la varianza d'errore

$$F = S^2_{Campioni} \big/ S^2_e \ .$$

Dalla lettura dei dati riportati nell'ultima tabella, il **test F tra Operatori** (P = 0,085) non risulta significativo, in quanto maggiore di α = 0.05 anche se non molto distante da esso.

Tuttavia, in quanto sono un fattore fisso, è ugualmente ragionevole **misurare le differenze sistematiche o bias tra operatori**. Per valutare la significatività della **differenza** tra i **singoli operatori**, il metodo più potente è un confronto a priori , benché sia possibile ricorrere ai confronti a posteriori (sia per i test a priori sia per quelli a posteriori, si rinvia al capitolo relativo).

Nel calcolo **delle componenti della varianza,** utilizzando i dati ANOVA della tabella precedente e sempre con

- n = 3 è il numero di repliche effettuate dallo stesso operatore sullo stesso campione,
- a = 3 è il numero di operatori,
- b = 5 è il numero di campioni,

si ottengono le tre stime seguenti.

1 - La **prima componente** è σ^2, che misura la **ripetibilità entro operatore** (***within-operator repeatability***) S^2_e.

Con i dati dell'esempio, è

$$\sigma^2 = S^2_e = 29{,}5$$

2 - La **seconda componente** σ^2_{AB}**,** il termine della **varianza d'interazione AB**.

Con i dati dell'esempio, è

$$\sigma^2_{AB} = \frac{S^2_{AB} - S^2_e}{n} = \frac{653{,}3 - 29{,}5}{3} = 207{,}9$$

3 - La **terza componente** è il termine della **varianza di tutti i campioni** σ^2_B.

Con i dati dell'esempio e **in modo differente dall'esempio precedente**,
è

$$\sigma_B^2 = \frac{S_B^2 - S_e^2}{na} = \frac{2047{,}5 - 29{,}5}{3 \cdot 3} = \frac{2018}{9} = 224{,}2$$

La varianza per il fattore operatori non può essere stimata, in quanto **non è una variabile random.**

Da queste tre, si ricava la **varianza stimata del metodo o strumento** σ_{gage}^2, la ***gage variance***.

In realtà essa è determinata solamente da
due fattori:

$$\sigma_{gage}^2 = \sigma^2 + \sigma_{AB}^2 = 29{,}5 + 207{,}9 = 237{,}4$$

Si può osservare che, rispetto alla stima Model II in cui la **gage variance** era σ_{gage}^2 **= 341,8**

- questa **gage variance** σ_{gage}^2 **= 237,3** è minore del 31%:

In modo implicito, nella discussione di questi due esempi emerge un problema importante, che è bene evidenziare chiaramente, per una corretta comprensione degli studi ***gage R&R***.
I campioni sono un fattore random oppure fisso?

Sotto l'aspetto teorico, quasi sempre i campioni utilizzati per gli studi ***gage R&R*** sono solamente una parte delle misure che saranno successivamente analizzate con lo strumento. E' anche il caso dell'esempio utilizzato, nel quale viene misurata la glicemia in campioni di sangue. Quindi, tecnicamente sono un campione random. I campioni da analizzare dovrebbero sempre essere scelti in random.
Tuttavia lo studio ***gage R&R*** **come scopo specifico misura la variabilità tra operatori** e pertanto **si disinteressa di quella tra i campioni.**
Come visto nell'esempio 1, è importante anche la variabilità Operatori per Campioni. Essa è necessaria, per meglio caratterizzare il comportamento degli operatori, quando si passa da campioni con valori piccoli a campioni con valori grandi.
Da queste considerazioni, si può dedurre che negli studi ***gage R&R*** non si dovrebbe mai applicare l'ANOVA model I o a fattori fissi, ma solamente la model II oppure la model III.
Tra scelta tra questi ultimi due modelli dipende dalle considerazioni fatte sugli operatori.

Gli esempi discussi impiegano analisi che non sono distruttive del campione.

Spesso gli studi ***gage R&R*** sono distruttivi, come avviene quando si devono analizzare i tempi o l'intensità di una esplosione, la resistenza di un materiale alla rottura, il tempo in cui un farmaco degrada alle varie condizioni ambientali (temperatura, umidità, esposizione alla luce, ecc.).
Se le prove **non sono distruttive** oppure **sono distruttive**, il disegno sperimentale deve essere impostato in modo diverso.

Nel **caso di prove non distruttive**, come quelle impostate nei due esempi precedenti, i due fattori presi in considerazione (Operatori e Campioni) **sono crossed**.
Tutti gli Operatori analizzano ripetutamente (2 o 3 volte) lo stesso Campione.
Nel **caso di prove distruttive**, i due fattori presi in considerazione (Operatori e Campioni) **sono nested** o meglio
- i Campioni sono nested entro Operatori.
Ne consegue che non è più possibile analizzare l'interazione Operatori per Campioni, che rappresenta sempre un aspetto importante, come evidenziato nei due esempi svolti.

ESEMPIO 3. (ANOVA MODEL II. Tratto, dal ***website online*** di L. M. **Bland** (May 2004), che riporta l'articolo di Doug **Altman** e di Martin **Bland** ***How do I analyse observer variation studies?***).
Mediante apparecchiatura ad ultrasuoni, 4 medici hanno valutato la circonferenza addominale del feto in 3 donne in gravidanza. Ogni medico ha ripetuto la misura 3 volte in modo indipendente, con i risultati riportati nella tabella seguente:

Circonferenza addominale (cm) con ultrasuoni									
Medico	Paziente I			Paziente II			Paziente III		
	Prov 1	Prov 2	Prov 3	Prov 1	Prov 2	Prov 3	Prov 1	Prov 2	Prov 3
A	13,6	13,3	12,9	14,7	14,8	14,7	17,1	17,1	18,3
B	13,8	14,2	13,2	14,9	14,1	14,5	17,2	17,5	17,6
C	13,2	13,1	13,1	14,5	14,2	13,8	16,3	15,2	16,1
D	13,7	13,7	13,4	14,4	14,3	13,6	16,8	16,8	17,5

Stimare le componenti della varianza dovuta ai medici.
E' un disegno sperimentale **Model II o a effetti random**, in quanto ovviamente le pazienti sono solamente un campione di tutte quelle possibili e l'interesse della ricerca è rivolta esplicitamente a valutare se tra gli utenti dell'apparecchiatura (i medici) esiste in generale (nella popolazione dei medici) una varianza grande.

E' un'ANOVA a due criteri, dove

- i medici sono $m = 4$,
- le pazienti sono $p = 3$,
- le repliche di ogni medico sulla stessa paziente sono $r = 3$,

L'output del computer è

Fattori	**SQ**	**GDL**	S^2	**F**	**P**
Totale	90,4222	35	2,5835	---	---
Pazienti	79,9439	2	39,9719	250,26	<0,0001
Medici	3,9089	3	1,3030	8,16	0,0006
Interazione	2,7361	6	0,4560	2,86	0,0300
Errore	3,8333	24	0,1597	**---**	**---**

(Come facile dedurre dai valori, i tre test F sono stati calcolati ponendo al denominatore la varianza d'errore S_e^2)

Le varianze attese sono

Fonte	**GDL**	**Varianze**
Totale	$mpr - 1 = 35$	----
Pazienti	$p - 1 = 2$	$mr\sigma_P^2 + r\sigma_I^2 + \sigma_e^2$
Medici	$m - 1 = 3$	$pr\sigma_M^2 + r\sigma_I^2 + \sigma_e^2$
Interazione	$(p-1)\cdot(m-1) = 6$	$r\sigma_I^2 + \sigma_e^2$
Errore	$(m-1)pr = 24$	σ_e^2

Con i dati dell'esempio, le componenti della varianza sono le seguenti

1 – Per l'errore: σ_e^2 **= 0,1597**

2 – Per l'interazione: $r\sigma_I^2 + \sigma_e^2 = 3 \cdot \sigma_I^2 + \sigma_e^2 = 0{,}4560$

da cui

$$\sigma_I^2 = \frac{0{,}4560 - 0{,}1597}{3} = 0{,}0988$$

3 – Per i medici : $pr\sigma_M^2 + r\sigma_I^2 + \sigma_e^2 = 3 \cdot 3 \cdot \sigma_M^2 + 3 \cdot \sigma_I^2 + \sigma_e^2 = 1{,}3030$

da cui

$$\sigma_M^2 = \frac{1{,}3030 - 0{,}4560}{3 \cdot 3} = 0{,}0941$$

4 – Per le pazienti: $mr\sigma_P^2 + r\sigma_I^2 + \sigma_e^2 = 4 \cdot 3 \cdot \sigma_P^2 + 3 \cdot \sigma_I^2 + \sigma_e^2 = 39{,}9719$

da cui

$$\sigma_P^2 = \frac{39{,}9719 - 0{,}4560}{4 \cdot 3} = 3{,}2930$$

Da questi, è ora possibile calcolare la σ_{gage}^2, vale a dire la **varianza delle osservazioni sullo stesso paziente da parte di due medici,** con

$$\sigma_{gage}^2 = \sigma_e^2 + \sigma_I^2 + \sigma_M^2 = 0{,}1597 + 0{,}0988 + 0{,}0941 = 0{,}3526$$

poiché quella tra le pazienti non deve essere presa in considerazione.

Diventa possibile osservare che la **varianza strumentale** ($\sigma_{gage}^2 = 0{,}3526$), considerando anche le differenze tra medici e la loro interazione con le caratteristiche delle pazienti, **è più del doppio di quella dovuta a un singolo medico** o entro medici ($\sigma_e^2 = 0{,}1597$).

Sempre secondo Bland, con questi dati è possibile stimare anche

- la **differenza massima probabile** (***maximum difference likely***) tra le misure dello **stesso medico** o **operatore**, che corrisponde al concetto di **ripetibilità** (***repeatability***),

con

$$Repeatability = 2{,}83 \cdot \sigma_e = 2{,}83 \cdot \sqrt{0{,}1597} = 1{,}13$$

- la **differenza massima probabile** tra le misure di più **medici** o **operatore**,

con

$$2{,}83 \cdot \sigma_{gage} = 2{,}83 \cdot \sqrt{0{,}3526} = 1{,}68$$

- la varianza totale σ_T^2 delle misure di differenti medici su differenti soggetti

con

$$\sigma_T^2 = \sigma_P^2 + \sigma_M^2 + \sigma_I^2 + \sigma_e^2 = 3{,}2930 + 0{,}3526 = 3{,}6456$$

- il **coefficiente di correlazione intra-classe** (***intra-class correlation coefficient***) o **ICC** per le misure di differenti medici o operatori

con

$$ICC = \frac{\sigma_P^2}{\sigma_T^2} = \frac{3{,}2930}{3{,}6456} = 0{,}90$$

- il **coefficiente di correlazione intra-classe entro medici**, cioè per le misure dello stesso medico,

con

$$intra-observer \quad ICC = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_e^2} = \frac{3{,}2930}{3{,}2930 + 0{,}1597} 00{,}95$$

Questa ultima ICC (***intra-class correlation coefficient***) risulta maggiore della precedente, in quanto l'utilizzo di medici differenti aumenta la variazione delle misure.

Come già evidenziato in precedenza, una difficoltà che può presentarsi in questo metodo di analisi è che una varianza potrebbe risultare negativa. E' logicamente impossibile, trattandosi di un quadrato, anche se effettivamente avviene per variazione casuale dei dati. La soluzione del problema consiste nell'attribuire zero al valore negativo della varianza.

Le condizioni di validità di questa metodologia fondata sull'ANOVA sono due:

- la deviazione standard entro soggetti è indipendente dalla media,
- gli errori entro soggetti sono distribuiti in modo normale, almeno approssimativamente.

Ma non sempre queste condizioni sono rispettate. Per giungere a una soluzione anche a questi casi e per ulteriori approfondimenti della metodologia, si rinvia all'articolo di **Bland** qui riportato e a quello di C. G. **Moertel** e J. A. **Hanley** del 1976 ***The effect of measuring error in the results of therapeutic trials in advanced cancer*** (sulla rivista **Cancer** Vol. 38 pp.: 388 – 394).

24.16. VISIONE GENERALE DELLE STIME RICHIESTE NELL'ANALISI DI PROCESSO

Quanto presentato nei paragrafi precedenti è la parte introduttiva all'**analisi di processo** (***measurements process***), con i suoi test fondamentali. Per approfondimenti ulteriori, si rimanda ai testi scritti appositamente sul controllo di qualità. Nei programmi informatici, i punti fondamentali sono

1 - l'analisi della capability,
2 - la ripetibilità e la riproducibilità, R&R
3 - l'analisi di Weibull,
4 - il piano di campionamento,
5 - la scomposizione della varianza per gli effetti random.

1 - **L'analisi della Capability** è fondata su indici specifici, che possono essere applicati sia con una distribuzione di dati raggruppati, sia con dati singoli. In letteratura si trovano gli indici

- **C_p, C_r, C_{pk}, C_{pl}, C_{pu}**, K, **C_{pm}, P_p, P_r, P_{pk}, P_{pl}, P_{pu},**
- e i limiti di tolleranza normali e non normali, con le indicazioni della ***capability*** corrispondente.

Gli indici precedenti sono fondati sulla distribuzione Normale. Ma alcune stime, come **C_{pk}, C_{pl}, C_{pu},** possono essere fondati sul metodo dei percentili, in accordo con distribuzioni non normali, sia discrete come le curve di Johnson e Pearson per i momenti, sia continue come le leggi Beta, Esponenziale, Valore Estremo (Tipo I , Gumbel), Gamma, Log-Normale, Rayleigh e Weibull.

E' possibile stimare i parametri di queste distribuzioni con il metodo della massima verosimiglianza e verificare l'accordo della distribuzione osservata con la distribuzione teorica, mediante il test di Kolmogorov-Smirnov, l'uso degli istogrammi, il metodo Probabilità-Probabilità (P-P) e il metodo Quantile-Quantile (Q-Q).

Con i programmi informatici è possibile verificare congiuntamente tutte queste ipotesi, per scegliere quella che maggiormente si avvicina alla distribuzione campionaria dei dati. Sotto l'aspetto della logica statistica non è una procedura corretta; ma lo sviluppo dell'informatica permette queste verifiche, in modo estremamente rapido e semplice.

2 - L'analisi della **ripetabilità,** della **riproducibilità** e della **R&R** è quasi sempre effettuata mediante l'ANOVA, in quanto più completo dell'altro metodo (***Range & Average***), che fondamentalmente rappresenta un procedimento abbreviato per calcoli manuali ed è approssimato. Il metodo dell'ANOVA è quello raccomandato dalla Società Americana di Controllo di Qualità e da associazioni di produttori, come quella delle aziende automobilistiche.

3 – Oltre alla distribuzione normale, è possibile utilizzare le **probabilità Weibull** e stimare **i parametri** di questa distribuzione, con i loro **intervalli di confidenza**.

4 – Possono essere prodotti **piani di campionamento** fissi e sequenziali, per medie distribuite in modo normale oppure in accordo con la binomiale oppure la poissoniana. Si ricavano pure le dimensioni del campione richiesto e la stima della potenza del test.

5 - Con l'ANOVA è possibile stimare le componenti della varianza, come la ripetibilità o variazione del gruppo di operatori, la variazione della stima o di un singolo operatore, la variazione dei campioni, la variazione dell'interazione operatore per campione.

Vari programmi informatici forniscono anche gli intervalli di confidenza delle componenti della varianza e la percentuale di tolleranza. Allegano pure tutte le statistiche descrittive, in relazione a ogni singolo operatore, ai singoli campioni, ai campioni analizzati da ogni singolo operatore.

24.17. STORIA DEL SEI-SIGMA; UN SECOLO DI EVOLUZIONE DEI METODI STATISTICI, PER IL CONTROLLO DI QUALITA'

Nell'anno 1980, la **Motorola** ha inventato e applicata una metodologia, che si è diffusa con il nome di ***Six-Sigma***, proposto dal suo responsabile della pubblicità. Fondata su test e concetti statistici, è esplicitamente finalizzata al miglioramento della produzione. Il suo successo è stato rapido tra le aziende. A metà degli anni '90, era adottata da altre grandi industrie americane di settori diversi, quali Honeywell, Johnson & Johnson, General Electric, DuPont, American Express e Ford, per ricordare alcune tra le maggiori. Attualmente è sinonimo di qualità totale e di perfezione del prodotto.

La metodologia statistica impiegata rappresenta lo sviluppo di un secolo di evoluzione dei test, come descrive Ronald D. **Snee** nell'articolo del 2004 ***Six-Sigma: the evolution of 100 years of business improvement methodology*** (pubblicato all'inizio del primo numero della rivista **International Journal of Six Sigma and Competitive Advantage** (IJSSCA) Vol. 1, No 1, pp.: 4-20).

La parte di organizzazione del lavoro e della gestione industriale ha visto i promotori scientifici

- in Adam **Smith** con l'opera del 1776 **The Wealth of Nations**
- e in Frederick W. **Taylor** con il volume del 1911 ***The Principles of Scientific Management*** (W. W. Norton and Company, New York).

Ponendo attenzione solo alla parte statistica, il primo lavoro importante per l'industria e le applicazioni alla produzione è l'articolo di W. S. **Gossett** del 1908, scritto con lo pseudonimo di **Student** mentre lavorava in una birreria inglese:

- ***The probable error of a mean*** (su **Biometrics** Vol. VI, pp.: 1-25).

Tra le numerose applicazioni scientifiche, ha anche quella di verificare la significatività statistica di un processo di miglioramento produttivo.

Lo sviluppo immediatamente successivo è la nascita della statistica moderna. Negli anni '20, dirigendo il ***Rothamsted Agricultural Experimental Station*** di Londra, R. A. **Fisher** crea l'**Analisi della Varianza,** come generalizzazione del t di Student, e l'approccio statistico del **Disegno Sperimentale** (***Design of Experiments***).

Nella divulgazione scientifica assumono un ruolo fondamentale

- il testo del 1925 **Statistical Methods for Research Workers** (1st ed. Oliver and Boyd, Edinburg , Scotland, 239 pp. +6 tables) che ha la 13ª edizione nel 1958 (12th ed. Hafner, New York 356 pp.)
- e quello pubblicato nel 1935 **The Design of Experiments** (Oliver and Boyd, London).

Un'altra pietra miliare, collocata storicamente all'inizio degli anni '30, è la nascita del ***control chart***, sviluppato da Walter A. **Shewhart** lavorando nei ***Bell Laboratories*** durante gli anni '30

- con il volume del 1931 ***Economic Control of Quality of Manufactured Product*** (Van Nostrand, New York). In esso illustra le modalità tecniche per il controllo e il miglioramento della produzione, che anche attualmente risultano più utilizzate. Sono grafici che permettono di evidenziare nel tempo le caratteristiche del prodotto e quindi di valutare quando il processo sta cambiando.

Avendo lavorato anch'essi nei ***Bell Laboratories*** durante gli anni '20 e '30**,** Harold F. **Dodge** e Harry G. **Romig** nel 1944

- con il volume ***Sampling Inspection Tables*** (John Wiley and Sons, New York) sviluppano i metodi per decidere quando, sulla base dell'analisi di un campione, tutta la produzione del lotto è accettabile. A partire dalla seconda guerra mondiale, sono la procedura adottate dall'esercito, dalla marina e dall'aviazione degli Stati Uniti per decidere se la qualità di una commessa rientra nelle specifiche del contratto oppure se il lotto deve essere rifiutato. Sono metodologia importante anche per il materiale in cui la prova di funzionamento diventa necessariamente distruttiva, come la verifica per la qualità del materiale esplosivo.

Durante gli anni '40 e '50, nelle imprese ha inizio lo ***Statitical Quality Control***, che impiega gli strumenti statistici che in quel periodo si erano già consolidati nella cultura tecnica: il ***control chart*** di **Shewhart**, gli ***acceptance sampling methods*** di **Dodge** e **Romig**, il ***design of experiments*** di **Fisher.**

Dalla fine degli anni '40 e fino agli anni '60, Gorge E. **Box** con i suoi collaboratori dell'***Imperial Chemicals Industries*** dell'Inghilterra adatta il disegno sperimentale di Fisher alle esigenze del processo industriale. La parte logica e i metodi sono riportati

- nel volume di G. E. P. **Box** e K. B. **Wilson** del 1951 ***On the experimental attainment of optimum conditions*** (pubblicato da **Journal of the Royal Statistical Society**, Series B, Vol. 13, 1ff),
- e in quello di G. E. P. **Box,** W. G. **Hunter** e J. S. **Hunter** del 1978 ***Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building*** (John Wiley and Sons, Inc., New York).

Il motivo per cui questi testi hanno assunto tanta importanza per l'ingegneria e per il sistema produttivo è nella presentazione di questo ultimo: *Even more important than learning about statistical techniques is the development of what might be called a capability for* ***statistical thinking***.

Il punto conclusivo dello sviluppo dei metodi per il controllo del processo e del prodotto è raggiunto negli anni '80, con il testo di G. **Taguchi** del 1986 ***Introduction to Quality Engineering – Design***

Quality into Products and Process (Asian Productivity Organization, acquistabile da Krauss International Publications, White Plains, NY).

Durante lo stesso periodo (dagli anni ’50 agli anni ’70), nei ***Bell Laboratories*** sono introdotti i concetti e i metodi dell’analisi grafica dei dati, che negli anni successivi diventeranno le procedure semplici e rigorose che stanno alla base del processo di miglioramento della produzione. La metodologia è illustrata dal maestro di questi ricercatori, John W. **Tukey**, che nel 1977 pubblica il volume ***Exploratory Data Analysis*** (Addison-Wesley, Reading, MA).

In particolare nelle industrie giapponesi e americane dell’automobile, dalla fine degli anni ’60 con il miglioramento della tecnica **control charts** sono introdotti:
- i **process capability studies** (le statistiche **Cp** e **Cpk**) per lo **Statistical Process Control**,
- le **Pareto charts** per identificare le fonti dei difetti,
- gli studi **Gage Repeatabilty and Reproducibility** per valutare i sistemi di misurazione e gli altri strumenti impiegati nel controllo e nel miglioramento dei processi produttivi.

A partire dai primi anni ’80, la competizione con i paesi dell’Asia per i mercati mondiali spinge le aziende degli Stati Uniti a dare inizio al **Total Quality Management**. Da obiettivo e metodo della produzione, la qualità diventa la filosofia, lo scopo e il metodo fondamentali di gestione dell’azienda nel suo complesso. Questo nuovo pensiero è proposto da W. Edwards **Deming** nel 1982, con il volume ***Out of the Crisis*** (MIT, Center for Advanced Engineering Study, Cambridge, MA), grandemente influenzato dalle idee di **Shewhart**. Il concetto di base è statistico: ***ridurre la variabilità.***
- “If I had to reduce my message for management to just a few words, I would say it all had to do with reducing variation”.

Per raggiungere la **Total Quality Management**, la Mototola propone di passare dal programma 3-sigma al programma 6- sigma. Partendo da un processo industriale in cui l’obiettivo è che i difetti abbiano una frequenza di 66.807 pezzi per un milione di unità, pervenire a nuovi processi che abbiano solo 3,4 errori ogni milione di unità prodotte. E raggiungere l’obiettivo, anche quando la taratura del processo a lungo termine ha uno spostamento di 1,5 sigma, rispetto al valore ideale o programmato di ogni prodotto.
Sotto l’aspetto statistico,
- 66,807 parti per milione è il valore sottostante la curva normale, collocata oltre 1,5 (3 – 1,5) sigma,
- **3,4 parti per milione è il valore sottostante la curva normale, collocata oltre 4,5 (6 – 1,5) sigma**, ovviamente considerando entrambe le code della distribuzione.