

# FONDAMENTI DI STATISTICA <sup>1</sup>

Luigi Mussio <sup>(2)</sup> – Vincenza Tornatore

<sup>(1)</sup> Politecnico di Milano – DICA – Piazza Leonardo da Vinci, 32 – 20133 Milano  
Tel. 02-2399-6501 – Fax 02-2399-6602 – e-mail luigi.mussio@polimi.it

<sup>(2)</sup> Politecnico di Milano – DICA – Piazza Leonardo da Vinci, 32 – 20133 Milano  
Tel. 02-2399-6502 – Fax 02-2399-6530 – e-mail vincenza.tornatore@polimi.it

## RIASSUNTO

L'inferenza statistica (in buona parte basata su distribuzioni campionarie derivate da quella normale, oltreché sulla distribuzione normale stessa) e la teoria della stima (principalmente con il criterio dei minimi quadrati, rinviando ad altri lavori lo studio delle sue proprietà e di altri suoi metodi) costituiscono la parte centrale della statistica e permettono insieme la validazione dei dati e dei modelli e l'elaborazione dei dati di osservazione. In entrambi i casi, notevole è l'analisi multivariata con l'analisi di varianza e la regressione multipla.

## PARTE I – DISTRIBUZIONI CAMPIONARIE DERIVATE DA QUELLA NORMALE

### 1.1. Introduzione

Le distribuzioni di statistiche campionarie non dipendono dalla distribuzione dell'universo da cui i campioni sono estratti e sono asintoticamente normali, se i loro campioni molto numerosi. Invece se i campioni hanno piccole dimensioni, la loro numerosità  $n$  gioca un ruolo importante, nel determinare l'equazione e la forma della distribuzione delle varie statistiche campionarie, distribuzione che non può più essere approssimata con quella normale e si discosta da essa, tanto più, quanto più il campione è piccolo. Esiste tutta una teoria di campionamento esatto, dove si tiene conto del valore di  $n$ , la quale, se da un lato fornisce informazioni ugualmente accurate per tutti i valori di  $n$ , dall'altro è meno generale di quella per i grandi campioni, richiedendo sempre una o più ipotesi limitative. L'ipotesi limitativa fondamentale, alla base delle distribuzioni statistiche di piccoli campioni, è che i campioni stessi siano estratti da un universo normalmente distribuito. In questo caso, si può dimostrare che le medie e le varianze campionarie sono variabili casuali indipendenti.

### 1.2. Distribuzione *chi quadrato*

Date  $n$  variabili casuali indipendenti:  $x_1, x_2, \dots, x_n$ , normalmente distribuite, con  $M = 0$  e  $\sigma = 1$ , la somma dei loro quadrati è una variabile casuale, chiamata  $\chi^2$ , la cui densità di probabilità è:

$$f(\chi^2) = f_0 (\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}} \quad (2.1)$$

dove:  $\chi^2 = \sum_{i=1}^n x_i^2$ , e  $f_0$  è un fattore di normalizzazione, tale che:  $\int_0^{+\infty} f(\chi^2) d\chi^2 = 1$ .

In questo caso,  $\nu$  è uguale ad  $n$  (numero di variabili casuali indipendenti, presenti nel calcolo di  $\chi^2$ ) e prende il nome di *gradi di libertà*.

---

<sup>1</sup> Questo lavoro riporta, pressoché integralmente, nello stile degli autori, quanto esposto nei capitoli 4, 5, 6, 7 e 9 del libro: Fondamenti di statistica, di Giovanna Togliatti (Hoepli, Milano, 1976), dove le note, scritte dagli autori del presente lavoro, servono a colmare i quasi quaranta anni passati dall'epoca di edizione del libro suddetto, così da rendere questi cinque capitoli ancora pienamente attuali.

Rifacendosi a quanto ben noto sulle distribuzioni campionarie, si può dire che se  $x$  è una variabile casuale normale standardizzata, la somma dei quadrati di  $n$  estrazioni con ripetizioni, da essa, ha una distribuzione  $\chi^2$  con  $n$  gradi di libertà, definita fra 0 e  $+\infty$ . (che seconda del valore di  $\nu$ , assume una delle forme rappresentate in fig. 1.2.1).

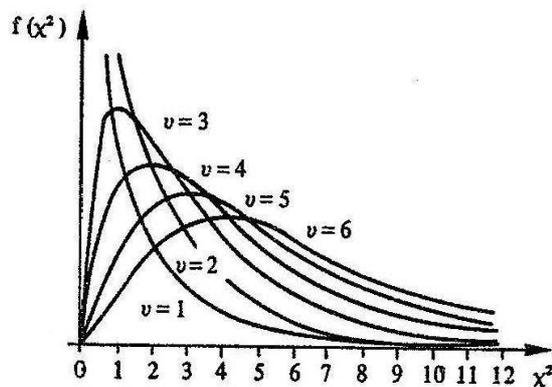


Fig. 1.2.1 – Densità di probabilità della variabile casuale  $\chi^2$  per diversi gradi di libertà

Ad eccezione delle curve corrispondenti a  $\nu = 1$  e  $2$  che sono anomale, la moda è:  $\chi^2 = \nu - 2$ , la media:  $M(\chi^2) = \nu$  e lo sqm:  $\sigma = \sqrt{2\nu}$ . All'aumentare di  $\nu$ , la distribuzione tende a diventare normale e, per  $\nu > 30$ , la variabile casuale  $\sqrt{2\chi^2}$  è normalmente distribuita con:  $M = \sqrt{2\nu - 1}$  e  $\sigma = 1$ .

I valori di  $F(\chi^2)$  sono solitamente tabulati per i valori di  $\nu$ . A riguardo, la variabile casuale  $\sqrt{2\chi^2}$  è usata per trovare i valori di  $\chi^2$ , per  $\nu > 30$ , in unità standardizzate:  $z = \sqrt{2\chi^2} - \sqrt{2\nu - 1}$ .

La variabile casuale  $\chi^2$  gode della cosiddetta proprietà di sommabilità e, se  $\chi_1^2$  e  $\chi_2^2$  hanno distribuzioni (2.1) indipendenti, rispettivamente con  $\nu_1$  e  $\nu_2$  gradi di libertà, anche la variabile casuale  $\chi_1^2 + \chi_2^2$  ha distribuzione  $\chi^2$ , con  $\nu_1 + \nu_2$  gradi di libertà.

Questo è particolarmente utile per ricavare la distribuzione delle varianze campionarie:

$$s^2 = \frac{1}{n} \left( (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \right)$$

Infatti se  $\sigma^2$  è la varianza dell'universo da cui proviene il campione, il secondo membro dell'espressione:

$$\frac{ns^2}{\sigma^2} = \frac{(x_1 - \bar{x})^2}{\sigma^2} + \frac{(x_2 - \bar{x})^2}{\sigma^2} + \dots + \frac{(x_n - \bar{x})^2}{\sigma^2} \quad (2.2)$$

è la somma dei quadrati di  $n$  variabili casuali normali standardizzate, del tipo:  $v_i = x_i - \bar{x}$ , legate dalla relazione:  $\sum (x_i - \bar{x}) = \sum v_i = 0$ .

Di conseguenza, esistono solo  $\nu = n - 1$  variabili casuali indipendenti tra loro, cosicché la variabile casuale  $ns^2/\sigma^2$  ha distribuzione  $\chi^2$ , con  $n - 1$  gradi di libertà (in generale, ogni qualvolta i dati del campione sono usati per stimare un parametro, in questo caso  $\bar{x}$ , il numero di gradi di libertà diminuisce di 1), e la variabile casuale delle varianze campionarie  $s^2$  ha una distribuzione  $(\sigma^2/n)\chi^2$ .

### 1.3. Distribuzione $t$ di Student

Date due variabili casuali indipendenti  $u$  e  $v^2$ , con  $u$  normalmente distribuita, con  $M = 0$  e  $\sigma = 1$  e  $v^2$  con distribuzione  $\chi^2$ , con  $\nu$  gradi di libertà, la variabile casuale  $t$  è derivata da queste, tramite la relazione sotto-riportata, e ha questa densità di probabilità:

$$t = \frac{u\sqrt{\nu}}{v} \qquad f(t) = f_0 \left( 1 + \frac{t^2}{\nu} \right)^{-\frac{\nu+1}{2}} \qquad (3.1) \text{ e } (3.2)$$

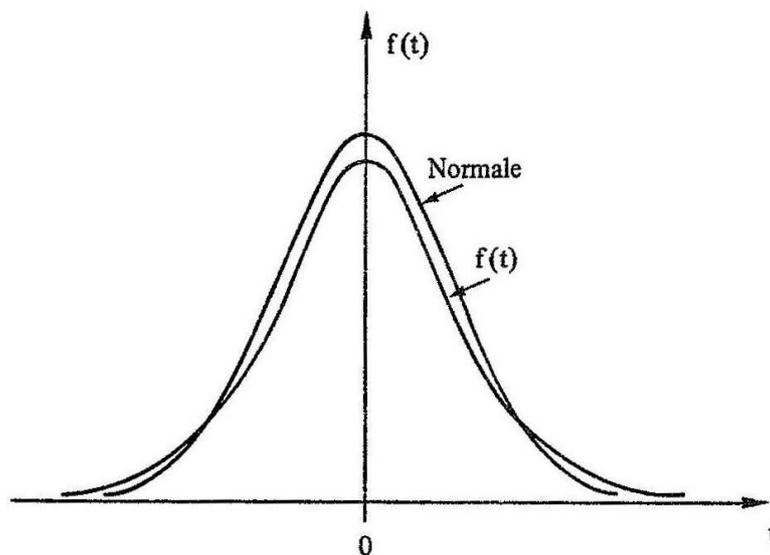


Fig. 1.3.1 – Confronto fra la densità di probabilità della variabile casuale  $t$  con  $\nu = 4$  e quella normale

La curva corrispondente è simmetrica e un po' più bassa della curva normale (Fig. 1.3.1). Per  $\nu \rightarrow \infty$  la (3.2) tende alla curva normale e se ne discosta tanto maggiormente quanto più  $\nu$  è piccolo.

Estratto da un universo normale:  $N(M, \sigma)$ , un campione di  $n$  elementi, la variabile casuale:  $(\bar{x} - M)/\sigma_{\bar{x}}$ , degli scarti standardizzati delle medie campionarie, rispetto alla media teorica, soddisfa le condizioni poste per  $u$ , e la variabile casuale  $ns^2/\sigma^2$ , dove  $s^2$  è la varianza campionaria, soddisfa le condizioni poste per  $v^2$ , con  $\nu = n - 1$  gradi di libertà. Dato che  $\bar{x}$  ed  $s^2$  sono variabili casuali indipendenti, perché l'universo è normale, si ha che il rapporto segue la distribuzione  $t$  con  $n - 1$  gradi di libertà:

$$t = \frac{\bar{x} - M}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{\sqrt{ns^2}} \sqrt{n-1} = \frac{\bar{x} - M}{s/\sqrt{n-1}} \qquad (3.3)$$

Nella variabile casuale. (3.3) compare lo sqm campionario  $s$  (e non quello dell'universo, come nella distribuzione delle medie di campioni numerosi), cosicché il calcolo della sua distribuzione non richiede alcuna stima di  $\sigma^2$  e questa variabile casuale rappresenta, in modo esatto, la popolazione delle medie di piccoli campioni.

Anche i valori di  $F(t)$  sono tabulati per i vari valori di  $v$  e, a pari probabilità, sono sempre maggiori di quelli normali. La (3.3), ricavata rigorosamente senza stimare la varianza dell'universo, può essere esplicitata in modo che risulti più chiaramente la struttura di variabile standardizzata del  $t$ :

$$t = \frac{\bar{x} - M}{\sqrt{\sum v_i^2 / (n(n-1))}} \quad \text{con:} \quad v_i = x_i - \bar{x}. \quad (3.4)$$

Infatti il denominatore della (3.4) è lo sqm della variabile casuale  $\bar{X}$  delle medie campionarie  $\bar{x}$ , ove si sia stimata la varianza dell'universo:  $\sigma_x^2 = \sum v_i^2 / (n-1)$ .

Le applicazioni della distribuzione del  $t$  di *Student* (pseudonimo dell'inglese Gosset suo scopritore) alle medie di piccoli campioni sono identiche a quelle per la distribuzione normale per le medie di campioni numerosi. Inoltre la distribuzione  $t$  può essere usata ogni qual volta il calcolo porti alla stima di statistiche che, per le loro proprietà caratteristiche, si possono considerare medie campionarie. In questo caso, il denominatore del  $t$  non è più uguale a quello della (3.3), valido solo nel caso in cui la  $\bar{x}$  sia calcolata come media aritmetica di un campione, ma deve essere  $\sigma_{\bar{x}}$  specifica del procedimento di calcolo usato.

Un'altra variabile casuale di notevole importanza cui si applica la distribuzione  $t$  è costituita dalle differenze di medie di piccoli campioni, purché questi siano estratti dallo stesso universo, oppure da universi aventi la stessa varianza (così oltre alla distribuzione normale per l'universo da cui si eseguono le estrazioni, si ha un'altra ipotesi restrittiva, relativa all'uguaglianza delle varianze).

Dati due universi normalmente distribuiti:  $X(M_x, \sigma)$  e  $Y(M_y, \sigma)$ , ed estratti due campioni indipendenti, di numerosità:  $n_x$  e  $n_y$ , medie campionarie:  $\bar{x}$  e  $\bar{y}$  e varianze campionarie:  $s_x^2$ ,  $s_y^2$ , le variabili casuali:

$$u = \frac{(\bar{x} - \bar{y}) - (M_x - M_y)}{\sigma_{\bar{x}-\bar{y}}} = \frac{(\bar{x} - \bar{y}) - (M_x - M_y)}{\sigma \sqrt{1/n_x + 1/n_y}} \quad \text{e} \quad v^2 = \frac{n_x s_x^2}{\sigma^2} + \frac{n_y s_y^2}{\sigma^2}$$

sono rispettivamente una variabile casuale normale standardizzata e, per la sommabilità di  $\chi^2$ , una variabile casuale  $\chi^2$  con  $(n_x - 1) + (n_y - 1) = n_x + n_y - 2$  gradi di libertà. Di conseguenza, la variabile casuale:

$$t = \frac{(\bar{x} - \bar{y}) - (M_x - M_y)}{\sqrt{n_x s_x^2 + n_y s_y^2}} \sqrt{\frac{n_x n_y (n_x + n_y - 2)}{n_x + n_y}} \quad \text{con} \quad v = n_x + n_y - 2 \quad (3.5)$$

ha una distribuzione  $t$  di Student con  $n_x + n_y - 2$  gradi di libertà. Anche nella (3.5) compaiono solo le varianze campionarie (e non delle stime di ) e, con qualche passaggio algebrico, si può evidenziare come la (3.5) sia una variabile casuale standardizzata, dove la varianza di  $\bar{X} - \bar{Y}$  è calcolata usando come stima della varianza comune di  $X$  e  $Y$  l'espressione:

$$\sigma^2 = \frac{\sum v_x^2 + \sum v_y^2}{n_x + n_y - 2} = \frac{n_x s_x^2 + n_y s_y^2}{n_x + n_y - 2}$$

Le applicazioni della distribuzione  $t$  di Student, alle differenze di medie di piccoli campioni, sono analoghe a quelle della distribuzione normale, per le differenze di medie di campioni numerosi.

#### 1.4. Distribuzione $F$ di Fisher

Date due variabili casuali Indipendenti  $v^2$  e  $w^2$ , entrambe con distribuzione  $\chi^2$ , rispettivamente con  $\nu_1$  e  $\nu_2$  gradi di libertà, la variabile casuale  $F$  (definita fra 0 e  $+\infty$ ) è derivata da queste, tramite la relazione sotto-riportata, e ha questa densità di probabilità:

$$F = \frac{v^2 / \nu_1}{w^2 / \nu_2} \qquad f(F) = f_0 \left( \nu_2 F^{\nu_1/2-1} + \nu_1 F^{-\nu_2/2-1} \right) \qquad (4.1)$$

La curva corrispondente dipende dai due parametri  $\nu_1$  e  $\nu_2$  e la tabulazione della funzione di distribuzione  $F(F)$  richiede tre dimensioni; tuttavia è uso tabulare la distribuzione, in funzione dei suoi gradi di libertà  $\nu_1$  e  $\nu_2$ , solo per i valori  $F$  tali che:

$$\int_0^F f(\xi) d\xi = 0.95 \qquad \text{e} \qquad \int_0^F f(\xi) d\xi = 0.99$$

Date due varianze campionarie  $s_x^2$  e  $s_y^2$ , poiché  $n_x s_x^2 / \sigma^2$  e  $n_y s_y^2 / \sigma^2$  sono variabili casuali indipendenti, entrambe con distribuzione  $\chi^2$  e gradi di libertà rispettivamente  $n_x - 1$  e  $n_y - 1$ , le variabili casuali:

$$\frac{v^2}{\nu_1} = \frac{n_x s_x^2}{(n_x - 1) \sigma^2} \qquad \text{e} \qquad \frac{w^2}{\nu_2} = \frac{n_y s_y^2}{(n_y - 1) \sigma^2}$$

soddisfano i requisiti imposti, perché la variabile casuale rapporto abbia la densità di probabilità della (4.1):

$$F = \frac{n_x s_x^2 / (n_x - 1)}{n_y s_y^2 / (n_y - 1)} \qquad (4.2)$$

Il numeratore (ed analogamente il denominatore della (4.2)) può essere così riscritto:

$$\frac{n_x s_x^2}{n_x - 1} = \frac{\sum v_x^2}{n_x - 1}$$

cosicché la variabile casuale  $F$  sia interpretabile come il rapporto fra la stima, non deviata, delle varianze delle due popolazioni  $X$  e  $Y$ , da cui sono estratti i campioni. Ad esempio, deve sempre essere  $F = 1$ , perché sia applicabile la distribuzione  $t$  di *Student* alla variabile casuale delle differenze di medie campionarie<sup>2</sup>.

In generale, nel calcolo di  $F$  campionario, il rapporto è eseguito ponendo al numeratore la maggiore delle due varianze. Infatti le tavole sono calcolate secondo il criterio di avere  $F \geq 1$ , cosicché il campo di definizione di  $F$  va da 1 a  $+\infty$ .

## PARTE II – INFERENZA STATISTICA

### 2.1. Controllo di ipotesi

L'inferenza statistica riguarda quei metodi con cui si cerca di dedurre informazioni su di una variabile casuale, per mezzo di informazioni ricavabili da campioni, estratti da questa. In questo modo, avendo a disposizione un campione di  $n$  elementi, estratti da una variabile casuale, si vuole sapere, se questa variabile casuale segue una determinata distribuzione di probabilità, caratterizzata da certi parametri. In alcuni casi, l'ipotetica distribuzione è completamente specificata: ad esempio, un campione potrebbe essere estratto da una data variabile casuale, distribuita normalmente, con media e varianza assegnate. Più frequentemente, si conosce solo il tipo di distribuzione e si cerca di determinare i suoi parametri, per definirne una particolare di quel tipo. Allora sulla base di dati sperimentali, si cerca di costruire il modello matematico più adatto a rappresentare il fenomeno, nel suo insieme, e di predire i risultati di altre future esperienze analoghe. Pertanto le inferenze statistiche riguardano, di solito, le funzioni di distribuzioni di variabili casuali, sotto il duplice aspetto del tipo di funzione, oppure dei momenti che la caratterizzano.

In generale, si chiama *ipotesi statistica* una supposizione sulla funzione di distribuzione di una o più variabili casuali. Tuttavia la distribuzione di un campione reale non coincide mai esattamente con la distribuzione ipotetica e così occorre valutare, se le deviazioni dal modello matematico, riscontrate nel campione, siano

---

<sup>2</sup> Welch e Tukey rimuovono l'ipotesi restrittiva:  $F = 1$ , e forniscono una distribuzione approssimata, per il confronto di medie di campioni normali ed indipendenti, anche di diversa varianza: In questo modo, il valore atteso della differenza standardizzata (identico a quello per campioni numerosi) segue ancora approssimativamente la distribuzione  $t$  di *Student*:  $\Delta \approx t_v$ , purché i suoi gradi di libertà  $v$  siano calcolati, tenendo conto non solo delle numerosità, dei due campioni estratti, ma anche delle loro varianze campionarie:

$$\Delta = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \quad \text{con} \quad v = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\left(\frac{s_x^2}{n_x}\right)^2 \frac{1}{n_x + 1} + \left(\frac{s_y^2}{n_y}\right)^2 \frac{1}{n_y + 1}} - 2$$

A riguardo, se il numero dei gradi di libertà  $v$  non è un numero intero, esso è arrotondato al numero intero più vicino, così da poter accedere alle usuali tabelle di distribuzione di probabilità.

dovute a fluttuazioni casuali, inevitabili in qualunque esperimento, oppure se invece denuncino un'effettiva differenza fra la distribuzione incognita della popolazione (da cui si è estratto il campione) e quella ipotetica. Il *controllo di un'ipotesi statistica* (in inglese: *test*) è un criterio per decidere, se accettare o respingere l'ipotesi statistica fatta. Allo statista è lasciata la completa libertà, nello stabilire la regola di decisione; tuttavia questi, nel progettare, è ovviamente guidato da quelle proprietà che fanno al caso suo. Tutti i test di significatività si basano sul concetto di valutare, se le deviazioni fra la distribuzione campionaria (od i suoi momenti) e la distribuzione ipotizzata per la variabile casuale da cui il campione è estratto (od i suoi momenti) si possono attribuire a fluttuazioni casuali, oppure se invece sono significative, tanto grandi cioè che l'ipotesi di partenza debba essere respinta. In ogni caso, l'inferenza statistica è un tipo di decisione basato sulla probabilità, in quanto la significatività (o meno) dei risultati osservati non può mai essere stabilita con un criterio di validità assoluta.

Formulata un'ipotesi fondamentale  $H_0$  ed una alternativa  $H_1$ , per prima cosa, si stabilisce la *regione critica* del test, cioè un sottoinsieme dei possibili valori argomentali della variabile casuale, in esame, dove l'ipotesi fondamentale  $H_0$  è respinta. Infatti se vale  $H_0$ , quei valori compresi nella regione critica sono così poco probabili che un loro presentarsi in un campione basta per poter concludere: non corrispondente alla realtà l'ipotesi  $H_0$  formulata. Per questo, si verifica se il valore argomentale, ricavabile dal campione, cade o no nella regione critica. Se no,  $H_0$  è accettata; se sì,  $H_0$  è respinta e si accetta l'ipotesi alternativa  $H_1$ . In quest'ultimo caso, è anche possibile sospendere la decisione, ovvero non accettare  $H_1$ , pur avendo rigettato  $H_0$ , in attesa di ulteriori informazioni. Tuttavia in generale, scartare un'ipotesi conduce prima o poi ad accettarne una alternativa. Nella decisione presa sono sempre possibili due tipi di errori.

- ❑ Il tipo – *respingere un'ipotesi giusta*: cioè dire che i risultati sperimentali sono significativi, ovvero che cadono nella regione critica, quando invece l'ipotesi statistica  $H_0$  è corretta (questo accade con tanta maggiore probabilità, quanto più ampia è la regione critica stabilita).
- ❑ Il tipo – *accettare un'ipotesi sbagliata*: cioè dire che i risultati non sono significativi, ovvero che le deviazioni fra dati sperimentali ed ipotesi fatta sono dovuti soltanto al caso, ed accettare l'ipotesi  $H_0$  che invece è falsa (questo accade con maggiore probabilità, quanto più piccola è la regione critica).

Pertanto è evidente che il problema di stabilire un test per un'ipotesi si riconduce a quello di fissare il tipo e l'ampiezza della regione critica del test o, in alternativa, il rischio che si è disposti a correre di commettere un errore di primo o di secondo tipo.

Si chiama *livello di significatività* di un test la massima probabilità di commettere un errore di primo tipo. Questa probabilità, generalmente indicata con  $\alpha$ , è fissata prima di estrarre il campione, per evitare che gli elementi contenuti influenzino la decisione. In pratica, sono molto comuni i valori 5% e 1%. Ad esempio, se si sceglie un livello di significatività del test di  $\alpha = 5\%$ , significa che solo in circa 5 casi su 100 si respinge l'ipotesi  $H_0$  (che dovrebbe invece essere accettata) ed allora la decisione presa è corretta al 95%. Il livello di significatività 5% è più restrittivo di quello 1%, perché può capitare di respingere  $H_0$  al

5% , mentre la si accetta all'1% (infatti la regione non critica al 5% è più ristretta di quella all'1% ).

Perché un test di ipotesi sia buono, deve essere progettato in modo da ridurre al minimo gli errori di decisione. Questo non è semplice perché, dato un certo campione, ogni tentativo di ridurre gli errori del primo tipo conduce ad aumentare quelli del secondo tipo. Nei casi concreti, si tratta di decidere quale di essi è più nocivo e regolarsi di conseguenza, dato che il solo modo per ridurli entrambi è aumentare la numerosità del campione.

Oltre al livello di significatività  $\alpha$  , definibile come la probabilità che un valore argomentale campionario della variabile casuale, con ipotesi statistica  $H_0$  , cada nella regione critica, quando  $H_0$  è vero, si definisce la probabilità di commettere un errore del secondo tipo, generalmente indicato con  $\beta$  , ovvero la probabilità di estrarre, a caso, un campione in possesso di un valore argomentale, compreso nella regione non critica, quando l'ipotesi corretta è invece  $H_1$  .

Nella Fig. 2.1.1, la curva di sinistra rappresenta la distribuzione ipotizzata con  $H_0$  la cui regione critica ha due code di area  $\alpha/2$  ciascuna (e questo significa che  $H_0$  è accettata, se il valore campionario è compreso fra  $-a$  e  $+a$  ). Nella stessa figura, la curva di destra rappresenta la distribuzione ipotizzata con  $H_1$  (ed in questo caso, essendo vera  $H_1$  , la probabilità di ottenere valori campionari compresi fra  $-a$  e  $+a$  è data dall'area  $\beta$  ). Dato che la regola di decisione è la stessa, in corrispondenza di questi valori, si accetta  $H_0$  , nonostante valga  $H_1$  , in realtà, ovvero si commette un errore di secondo tipo, con probabilità  $\beta$  .

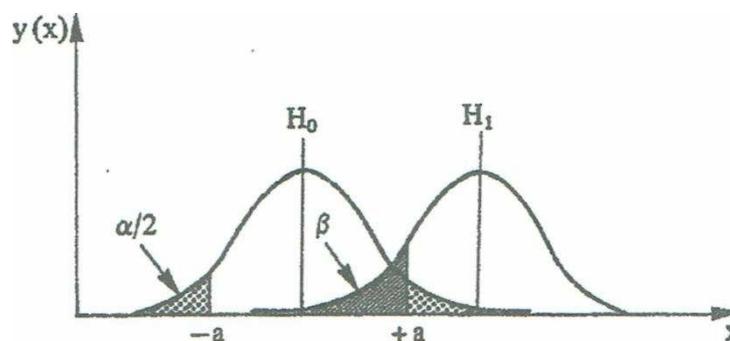


Fig. 2.1.1 – Rappresentazione grafica del significato degli errori di I e II

Un test di significatività può essere stabilito, a seconda dei casi, in tre modi.

- Assegnare la regione critica per cui sono considerati significativi i valori campionari esterni ad un fissato intervallo  $[A, B]$  . Eventualmente  $A$  o  $B$  possono essere gli estremi del campo di definizione della funzione di distribuzione e, in tal caso si dice che il test è fatto su una sola coda. Fissati  $A$  e  $B$  si determina a quale livello di significatività corrispondono, calcolando così  $\alpha$  (come nell'Esempio 2.2.1).
- Assegnare  $\alpha$  , specificando se il test è fatto su una o due code e calcolare conseguentemente i limiti  $A$  e  $B$  della regione critica (come nell'Esempio 2.3.1).
- Assegnare  $\alpha$  e  $\beta$  , in base ai quali si calcolano la numerosità del campione e la regione critica (come nell'Esempio 2.3.2).

Tuttavia la maggior parte dei problemi comporta più di una singola alternativa, in quanto lo sperimentatore ha spesso ragioni, teoriche o pratiche, per conoscere quale ipotesi fondamentale  $H_0$  provare, ma raramente sa quale ipotesi alternativa  $H_1$  adottare, se  $H_0$  si dimostra falsa.

Per queste classi, più generali, di alternative, l'entità dell'errore di secondo tipo dipende dalla particolare ipotesi alternativa  $H_1$ , presa in considerazione, in quanto  $H_1$  non è più costante, ma dipende da più entità suscettibili di assumere valori diversi, genericamente indicati con  $\theta$ . Allora per determinare l'efficacia del test scelto ed eventualmente confrontarlo con un altro, bisogna valutare l'entità di  $\beta(\theta)$ , cioè l'errore di secondo tipo, per tutte le possibili ipotesi alternative  $H_1(\theta)$ .

Anche in questo caso,  $\beta(\theta)$  è la probabilità per un valore campionario di cadere nella regione non critica, quando  $H_1(\theta)$  è l'ipotesi corretta. Dato che si preferisce evidenziare la regione critica, di solito, si calcola  $1 - \beta(\theta)$ , ovvero la probabilità per un valore campionario di cadere nella regione critica, quando  $H_1(\theta)$  è l'ipotesi corretta.

L'espressione  $P(\theta) = 1 - \beta(\theta)$  è detta *potenza del test* ed i diagrammi  $\theta, P(\theta)$  si chiamano *curve di potenza*. Invece i diagrammi  $\theta, \beta(\theta)$  sono detti *curve delle caratteristiche operative* del test (curve O.C.), cosicché usare l'uno o l'altro dei due tipi di curve è solo una questione di consuetudine, dato che la quantità di informazione contenuta è uguale.

Nel confronto fra due tipi di test, è da preferire quello la cui curva di potenza sia più alta, cioè quello per cui si ha, a parità di  $\theta$ , un valore di  $P(\theta)$  maggiore. Inoltre dallo studio delle curve di potenza, lo sperimentatore può determinare la probabilità di accettare (o meno) ipotesi alternative possibili e valutare, se l'esperimento è sufficientemente vasto da dare la fiducia, per qualunque decisione da prendere in base al test.

Il metodo di inferenza, per verificate ipotesi statistiche, può apparire artificioso, in quanto spesso non si ha un'ipotesi precisa:  $H_0 = H(\theta_0)$  da valutare, ma solo un'ipotesi approssimativa, basata sull'esperienza. Se questa ipotesi approssimativa è trattata come ipotesi precisa, da sottoporre a test, e se il test l'accetta, non significa che questa diventi improvvisamente del tutto vera, ma piuttosto che la realtà non è troppo discosta da  $H(\theta_0)$  e che, agli effetti pratici, si può considerare  $H(\theta_0)$  conforme a questa realtà.

Un procedimento più generale è estrarre non un campione di numerosità prefissata, ma un individuo alla volta, decidendo così, ad ogni passo, se accettare l'ipotesi o respingerla, oppure continuare ad aumentare il numero di individui del campione. Questo metodo, detto di campionamento sequenziale, permette spesso di raggiungere una decisione, con la stessa entità di errori di primo e secondo tipo, più velocemente e pertanto più economicamente di quello con campioni di numerosità prefissata.

## 2.2. Test relativi alla distribuzione binomiale

Ogni qual volta si abbia una variabile casuale relativa al numero od alla percentuale di eventi favorevoli su  $n$  prove, con  $n$  piccolo, si deve fare riferimento alla distribuzione binomiale.

Se il valore di  $n$  è fissato, l'unico parametro da sottoporre a test, è la probabilità  $p$  che definisce compiutamente la distribuzione.

### Esempio 2.2.1

Ad un esame, è distribuita una lista con 15 domande alle quali si deve rispondere: *sì* o *no*. Per valutare l'ipotesi che uno studente stia rispondendo a caso, dato che la probabilità di dare una risposta giusta è  $p = 0.5$ , è stabilita una regione critica, corrispondente alla seguente regola di decisione: se 10 o più risposte sono corrette lo studente non sta rispondendo a caso. Determinare il livello di significatività del test. L'ipotesi da sottoporre a test è:  $H_0: p = 0.5$  e la probabilità di dare 10 o più risposte giuste, se  $H_0$  è vera:

$$P = \binom{15}{10}(0.5)^{10}(0.5)^5 + \binom{15}{11}(0.5)^{11}(0.5)^4 + \binom{15}{12}(0.5)^{12}(0.5)^3 + \binom{15}{13}(0.5)^{13}(0.5)^2 + \binom{15}{14}(0.5)^{14}(0.5)^1 + \binom{15}{15}(0.5)^{15}(0.5)^0 = 0.1509$$

ottenendo così  $\alpha = 0.1509$ , cioè una probabilità del 15% circa di respingere l'ipotesi vera  $H_0: p = 0.5$ , ovvero di promuovere lo studente, quando sta rispondendo, a caso.

### **2.3. Test relativi alla distribuzione normale**

La distribuzione normale può essere usata, ogni qual volta si ha a disposizione un campione di numerosità  $n$  abbastanza grande, da poter ritenere sufficiente l'approssimazione in base alla quale queste variabili casuali possono essere considerate asintoticamente normali.

#### **2.3.1. Distribuzione binomiale con $n$ grande**

Valgono le stesse considerazioni fatte nel paragrafo precedente.

### Esempio 2.3.1

Fissare una regola di decisione, per controllare l'ipotesi che una moneta non sia truccata, avendo stabilito in precedenza di fare 49 lanci e di usare un livello di significatività del 5%. Se  $p$  è la probabilità di ottenere testa, in un lancio della moneta, si ha:

$H_0: p = 0.5$  con la moneta non truccata

$H_1: p \neq 0.5$  con la moneta truccata

Il test è fatto su due code perché è indifferente avere  $p < 0.5$  o  $p > 0.5$ , affinché si verifichi l'ipotesi  $H_1$ . Dato che  $\alpha = 0.05$ , ciascuna delle due aree tratteggiate della Fig. 2.3.1 è pari a 0.025 dell'area totale, sotto la curva normale standardizzata. I valori  $z_1$  e  $z_2$ , limiti della regione non critica, valgono  $-1.96$  e  $1.96$ . Nella ipotesi  $H_0$ , la media e lo sqm della distribuzione sono:

$$M = np = 49(0.5) = 24.5 \quad \text{e} \quad \sigma = \sqrt{npq} = \sqrt{49(0.5)(0.5)} = 3.50$$

e valori  $x$ , corrispondenti a  $z = \pm 1.96$ :

$$(x - np)/\sigma = (x - 24.5)/3.50 = \pm 1.96 \quad \text{da cui} \quad x_1 = 17.64 \quad \text{e} \quad x_2 = 31.36$$

Pertanto la regione critica comprende un numero di teste, su 49 lanci, compreso, fra 0 e 18 oppure fra 31 e 49, e così si respinge l'ipotesi  $H_0$  e si conclude che la moneta è truccata, se si ottiene un numero di teste compreso in questa zona.

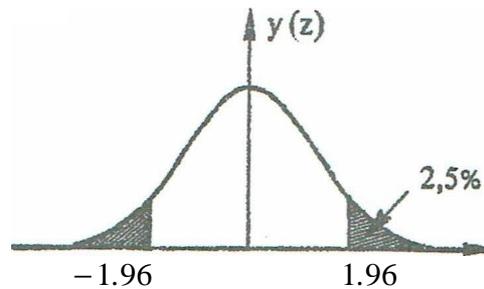


Fig. 2.3.1 – Zona critica per test su due code con  $\alpha = 5\%$  (punteggiato)

### Esempio 2.3.2

Determinare la zona critica e il minimo numero di lanci necessari per stabilire, se un dado è truccato o no, con rischi assegnati:  $\alpha = 0.025$  e  $\beta = 0.05$ . L'ipotesi  $H_0$  è che la probabilità, ad esempio, della faccia 1 sia uguale a  $1/6 = 0.16666$ . Allora si ritiene il dado truccato e  $H_0$  da respingere, qualora la probabilità della faccia 1 supera  $0.1\bar{6}$  di  $0.03$ , risultando  $H_1: P_1 = 0.1\bar{6} + 0.03 = 0.19\bar{6}$ .

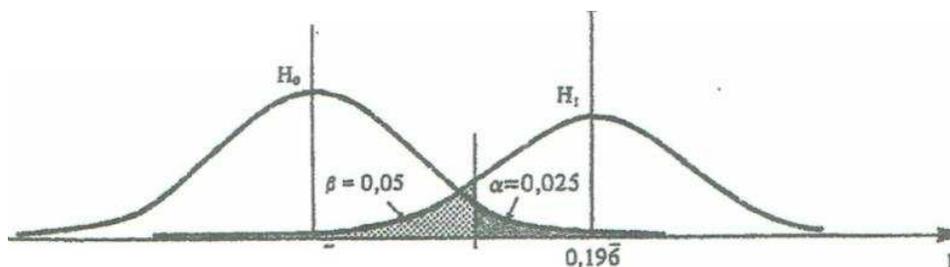


Fig. 2.3.2 – Rappresentazione grafica di un test su una coda, con  $\alpha$  e  $\beta$  assegnati

Nella figura 2.3.2, compaiono  $\alpha$  e  $\beta$  assegnati, mentre sono incogniti la numerosità  $n$  del campione ed il limite inferiore della zona critica  $p^*$ , a loro volta, legati con  $\alpha$  e  $\beta$ , da due relazioni:

- l'area alla destra di  $p^*$ , valutata nella curva normale standardizzata relativa ad  $H_0$ , vale:  $\alpha = 0.025$ ;
- l'area alla sinistra di  $p^*$ , valutata nella curva normale standardizzata relativa ad  $H_1$ , vale:  $\beta = 0.05$ ,

requisiti che si traducono nelle equazioni:

$$\frac{p^* - 0.1\bar{6}}{\sqrt{(0.1\bar{6})(0.83)/n}} = 1.96 \quad \frac{p^* - 0.19\bar{6}}{\sqrt{(0.19\bar{6})(0.83)/n}} = -1.645$$

cosicché:

$$p^* = 0.1\bar{6} + 0.730\sqrt{n} \quad n = 2129$$

$$p^* = 0.19\bar{6} - 0.654\sqrt{n} \quad p^* = 0.182$$

In questo modo, il dado deve essere lanciato, al minimo, 2129 volte e la zona critica è quella oltre 0.182. Se in 2129 lanci, ad esempio, la faccia 1 compare con frequenza inferiore a 0.182, si accetta l'ipotesi  $H_0: P = 0.1\bar{6}$ , ovvero l'ipotesi che il dado non sia truccato.

### 2.3.2. Distribuzioni di medie campionarie di campioni numerosi

Per il teorema centrale, le medie campionarie sono distribuite in modo asintoticamente normale, con media  $M$  e sqm  $\sigma/\sqrt{n}$ , dove  $M$  e  $\sigma$  si riferiscono all'universo da cui è estratto il campione di numerosità  $n$ . In questo caso, le ipotesi da controllare sono del seguente tipo. Calcolata uguale ad  $A$  la media di una certa caratteristica d'interesse, nel campione, è accettabile l'ipotesi di un campione estratto, a caso, da un universo in cui questa caratteristica vale  $B$  (oppure il valore medio dell'universo non è  $B$ , ma un altro valore qualsiasi).

#### Esempio 2.3.3

Le funi, prodotte da una ditta, hanno carico di rottura medio e sqm di 130 kg e 10 kg. La ditta sostiene che per mezzo di nuove tecniche il carico di rottura medio è aumentato, mentre lo sqm è rimasto invariato. Su di un campione di 64 funi, si è valutato un carico di rottura medio di 134 kg. Per accettare l'affermazione della ditta ad un livello di significatività di 0.01, si definiscono le ipotesi  $H_0$  e  $H_1$ :

$$H_0: M = 130 \text{ la produzione è sempre la stessa}$$

$$H_1: M > 130 \text{ la produzione è migliorata}$$

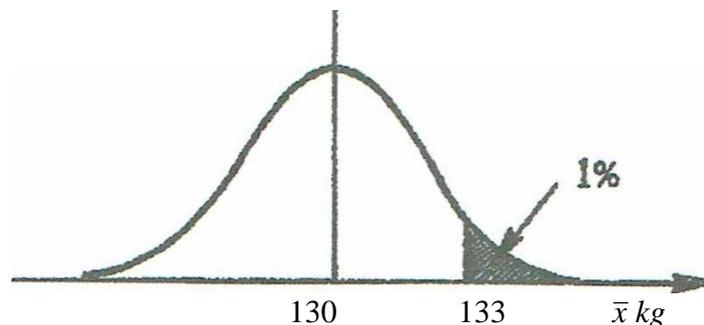


Fig. 2.3.3 – Zona critica corrispondente ad  $\alpha = 1\%$ , nell'Esempio 2.3.3

Il test sulla sola coda di destra, con l'area tratteggiata (di Fig. 2.3.3) pari all'1% dell'area totale, ha valore limite:  $z = 2.33$ . Nell'ipotesi  $H_0$ , la distribuzione delle medie campionarie ha rispettivamente media e sqm:

$$M_x = 130 \text{ kg} \quad \sigma_{\bar{x}} = 10/\sqrt{64} = 1.25 \text{ kg}$$

$$z = \frac{\bar{x} - M_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - 130}{1.25} = 2.33$$

$$\bar{x} = 1.25(2.33) + 130 = 132.91 \cong 133$$

La regola di decisione stabilisce che si respinga  $H_0$ , se la media di un campione supera  $133 \text{ kg}$ , mentre si accetta, in caso contrario. Dato che il campione esaminato ha  $\bar{x} = 134 \text{ kg}$ , l'ipotesi  $H_0$  è respinta e si deve accettare l'affermazione della ditta che la produzione è migliorata.

#### Esempio 2.3.4

Data la regola di decisione assunta nell'esempio precedente, occorre poi calcolare la probabilità di accettare  $H_0$ , quando il nuovo procedimento porta, in realtà, il carico di rottura medio dell'intera produzione a  $134 \text{ kg}$  (cioè calcolare la probabilità  $\beta$  di commettere un errore di secondo tipo, se  $H_1: M = 134 \text{ kg}$  è l'ipotesi corrispondente al vero).

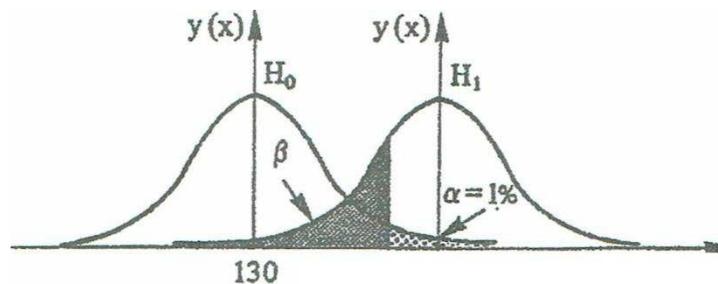


Fig. 2.3.4 – Rappresentazione grafica di un test su una coda con  $\alpha$  e regione critica assegnati

Le due curve in Fig. 2.3.4 rappresentano le distribuzioni delle medie dei campioni estratti da due universi le cui medie sono rispettivamente  $130$  e  $134 \text{ kg}$ .

Secondo la regola di decisione stabilita si accetta  $H_0$  per valori campionari  $\bar{x} \leq 133 \text{ kg}$ . Tuttavia se  $134 \text{ kg}$  è la vera media della produzione, si hanno medie campionarie inferiori od uguali a  $133 \text{ kg}$  con una probabilità uguale all'area  $\beta$  (tratteggiata in figura). Per determinare l'entità di  $\beta$  bisogna calcolare l'equivalente di  $133 \text{ kg}$  in unità standardizzate, data una distribuzione normale, con media  $134 \text{ kg}$  e  $\text{sqm}$  di  $1.25 \text{ kg}$ :  $z = (133 - 134)/1.25 = -0.80$ .

L'area, sotto la curva normale, alla sinistra di  $z = -0.80$  è  $0.2119$  e la probabilità  $\beta$  di non accettare  $H_1$ , cioè l'affermazione della ditta che la produzione è migliorata, quando è effettivamente migliorata ed il suo carico medio di rottura è diventato  $134 \text{ kg}$ , è il  $21\%$  circa.

Allora il compratore stabilisce, nella misura dell'  $1\%$ , il proprio rischio  $\alpha$  di ammettere un miglioramento della produzione (ed un aumento di prezzo), quando non esiste effettivamente. In funzione di  $\alpha$  è determinata la regione critica e la regola di decisione: se il valore medio campionario risulta inferiore a  $133 \text{ kg}$ , l'affermazione del produttore circa il miglioramento del prodotto non è accettata.

Dato che la numerosità del campione è preventivamente fissata in  $64$  pezzi, questa regola di decisione si

traduce in un rischio  $\beta$  del venditore uguale a circa il 21% (esiste cioè un 21% di probabilità che, pur essendo aumentato il carico medio di rottura, passando da 130 a 134 kg, il miglioramento non è riconosciuto).

E' ovvio che ben difficilmente il venditore, convinto della sua affermazione, accetta una così sfavorevole situazione, pertanto potrebbe chiedere, ad esempio, che il proprio rischio sia portato al valore  $\beta = 5\%$ . Se il compratore vuole mantenere il suo  $\alpha = 1\%$ , si tratta allora di determinare la numerosità del campione e la nuova regione critica che permettono insieme:  $\alpha = 1\%$  e  $\beta = 5\%$ . Il problema, analogo a quello dell'esempio 2.3.2, ponendo:  $H_0: M = 130$  e  $H_1: M = 134$ , si risolve trovando le incognite  $n$  e  $\bar{x}^*$  (limite inferiore della regione critica del test), cosicché:

$$z^* = \frac{\bar{x}^* - 130}{10/\sqrt{n}} = 2.33 \qquad \bar{x}^* = 130 + (2.33)10/\sqrt{n}$$

$$z^* = \frac{\bar{x}^* - 134}{10/\sqrt{n}} = -1.64 \qquad \bar{x}^* = 130 - (1.64)10/\sqrt{n}$$

da cui si ottiene:

$$n = 98 \qquad \bar{x}^* = 132.35$$

Per assicurare sia al compratore che al venditore il livello di rischio da essi desiderato, è necessario aumentare considerevolmente la numerosità del campione (da 64 a 98), mentre la zona critica (nella quale si accetta l'affermazione  $H_1$  del venditore) è ampliata, iniziando a 132.35 kg, invece che a 133 kg.

Al variare di  $H_1(\theta)$ , cioè per i vari nuovi valori medi, la curva di destra si sposta con continuità, facendo variare l'entità di  $\beta$ . Facendo assumere a  $\theta$  i valori 126, 128, ecc., fino a 138 kg, si può costruire per punti la curva O.C. o la curva di potenza:

$M =$	126	128	130	132	134	136	138
$\beta =$	1.0000	1.0000	0.9900	0.7881	0.2119	0.0082	0.0000

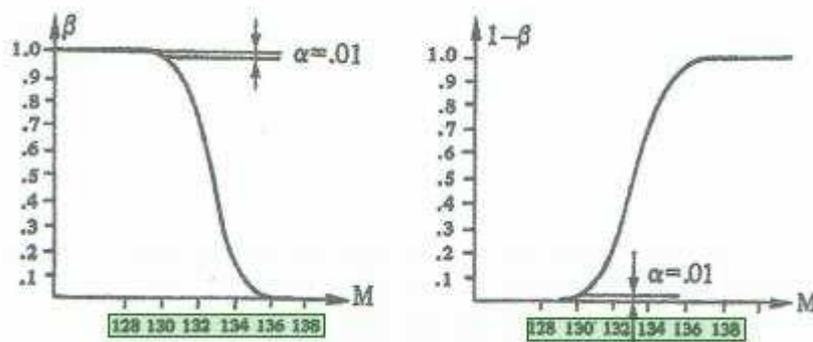


Fig. 2.3.5 – Curva O.C. e curva di potenza per il test a una coda dell'Esempio 2.3.4

Per  $M = 130 \text{ kg}$ , si ha:  $\beta = 1 - \alpha = 0.9900$ . Dalla curva O.C., con la regola di decisione adottata, la probabilità di accettare  $H_0$  (la produzione non è migliorata), quando la produzione media è inferiore a  $130 \text{ kg}$  è praticamente uguale ad 1. Dopo il valore  $130$ , la curva va rapidamente a zero, cosicché non si ha quasi alcun rischio di accettare  $H_0$ , quando il carico di rottura medio della produzione arriva a  $136 \text{ kg}$ .

### 2.3.3. Distribuzioni di differenze di medie per campioni numerosi

Sempre per il teorema centrale la differenza di medie campionarie per due campioni di numerosità  $n_1$  e  $n_2$ , estratti rispettivamente da due universi  $(M_1, \sigma_1)$  e  $(M_2, \sigma_2)$ , è distribuita in modo asintoticamente normale con media e sqm:

$$M(\bar{X}_1 - \bar{X}_2) = M_{\bar{x}_1} - M_{\bar{x}_2} = M_1 - M_2 \qquad \sigma(\bar{X}_1 - \bar{X}_2) = \sqrt{\sigma_{\bar{x}_1}^2 - \sigma_{\bar{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}}$$

Date le medie di una caratteristica campionaria, pari ad  $A$ , in un campione, ed a  $B$ , in un altro, si ricorre a questa distribuzione per controllare, se può ritenersi corretta l'ipotesi che  $A$  sia diverso da  $B$ , solo per motivi accidentali, e che i due campioni siano estratti dallo stesso universo o da universi con la stessa media (in alternativa, si deve invece concludere che i due campioni appartengono ad universi con medie differenti).

#### Esempio 2.3.5

Dati i voti medi di laurea, in due sessioni, con 40 e 50 laureati, rispettivamente pari a 74, con sqm 8, e 78, con sqm 7, si deve decidere, se la differenza, fra i risultati delle due sessioni, è significativa all'1%.

$H_0$ :  $M_1 = M_2$  la differenza è dovuta solo al caso

$H_1$ :  $M_1 \neq M_2$  il livello medio degli studenti è cambiato

Nella ipotesi  $H_0$ , entrambi i campioni provengono dalla stessa popolazione, con la media e lo sqm della variabile casuale costituita dalle differenze di medie campionarie:

$$M(\bar{X}_1 - \bar{X}_2) = 0 \qquad \sigma(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{8^2}{40} - \frac{7^2}{50}} = 1.606$$

avendo usato gli sqm non devianti, valutati sui campioni, come stima di quelli della popolazione.

L'equivalente, in unità standardizzate, della differenza  $74 - 78$ , appartenente ad una distribuzione normale con media nulla e sqm 1.606, è:  $z = (74 - 78)/1.606 = -2.49$ .

In un test su due code sono significativi, al livello di significatività dell'1% i valori di  $z$  esterni all'intervallo  $[-2.58, +2.58]$ , ed a questo livello la differenza non è significativa, cioè non si ha alcun cambiamento nei

voti medi di laurea. Si può altresì notare che  $z = -2,49$  sarebbe stato significativo al 5% e che, per una migliore discriminazione fra le due ipotesi, sarebbe opportuna qualche ulteriore indagine.

## 2.4. Test relativi alla distribuzione $t$ di Student

### 2.4.1. Distribuzioni di medie di piccoli campioni

Gli stessi criteri, già descritti in 2.3.2. per i campioni numerosi, sono usati, con l'avvertenza di servirsi della distribuzione del  $t$  di Student, con gradi di libertà:  $\nu = n - 1$ , invece della distribuzione normale. Esiste tuttavia la condizione limitativa che il campione deve essere estratto da un universo normale e, per una corretta applicazione del test  $t$ , andrebbe fatta un'indagine preliminare, in tal senso (qualora invece si tratti di errori di misura, si ritiene che la condizione di normalità sia sempre approssimativamente soddisfatta).

### 2.4.2. Distribuzioni di differenze di medie di piccoli campioni

Le ipotesi da sottoporre a controllo sono le stesse di quelle già descritte in 2.3.3, per i grandi campioni, con le condizioni limitative di universi (da cui sono estratti i due campioni) almeno approssimativamente normali e di uguale varianza. Trattandosi di controllo delle ipotesi, la condizione:  $F = 1$  (posta nel paragrafo 1.4), va interpretata come  $F$  significativamente uguale a 1 (il controllo preliminare di questa ipotesi è trattato nel paragrafo 2.5).

#### Esempio 2.4.1

Un soggetto, costituzionalmente con bassa pressione arteriosa, fa una cura, per cercare di aumentarla, ed esegue 10 misure  $x$ , in giorni consecutivi (prima di iniziare la cura), ed altrettante  $y$ , dopo un periodo di cura. Si vuol valutare l'ipotesi che il trattamento abbia aumentato la pressione media del soggetto, dati:

$x$ (prima della cura)	92	98	95	105	92	95	95	93	96	102
$y$ (dopo la cura)	96	98	103	98	105	97	99	110	100	96

Si suppone che le variazioni registrate, in una stessa situazione, siano normalmente distribuite e che  $\sigma_x = \sigma_y$ , cosicché le ipotesi da sottoporre a test sono:

$$H_0: M_x = M_y \qquad H_1: M_x < M_y$$

Con qualche calcolo, si ha:

$$\begin{aligned} \bar{x} &= 96.30 & n_x s_x^2 &= \sum v_x^2 = 164.10 & \sigma_x &= 4.27 \\ \bar{y} &= 100.20 & n_y s_y^2 &= \sum v_y^2 = 183.60 & \sigma_y &= 4.52 \end{aligned}$$

$$t = \frac{(100.20 - 96.30) - 0}{\sqrt{162.10 - 183.60}} \sqrt{\frac{100(18)}{20}} = 1.98$$

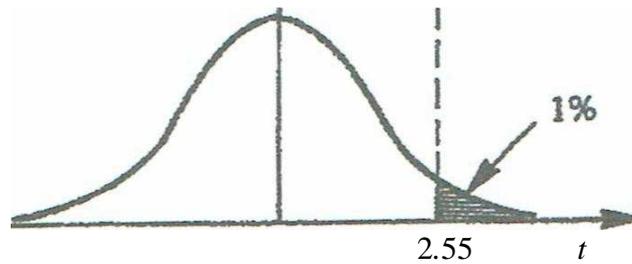


Fig. 2.4.1 – Zona critica per test t su una coda con  $\alpha = 1\%$

Il test è fatto su una sola coda, ad esempio, stabilendo i valori di  $t > t_{0,99}$ , come regione critica. L'area tratteggiata in Fig. 2.4.1 è l'1% dell'area totale sotto la curva  $y = f(t)$ . Con  $\nu = 18$ , si ha:  $t_{0,99} = 2.55$  per cui la differenza standardizzata, fra le pressioni medie, non è significativa e  $H_0$  va accettata<sup>3</sup>.

### 2.4.3. Distribuzione dei coefficienti di correlazione lineare

Una particolare applicazione della distribuzione normale e della distribuzione  $t$  di Student è relativa ai test su coefficienti di correlazione lineare. Infatti questi posseggono una loro distribuzione, più complessa (trovata da David, nel 1954), dove  $f(r)$ , rappresentante la densità di probabilità, contiene non solo la numerosità  $n$  del campione, ma anche il valore del coefficiente di correlazione  $\rho$  dell'universo al quale il campione appartiene. Si ha cioè:  $f(r) = f(r/n, \rho)$ , cosicché si hanno, a parità di  $n$ , infinite possibili distribuzioni per  $r$ , a seconda del valore assegnato a  $\rho$ . L'andamento di  $f(r)$  è notevolmente diverso, nei vari casi: simmetrico rispetto all'asse  $r = 0$ , per  $\rho = 0$ , e marcatamente asimmetrico per  $\rho \rightarrow \pm 1$ .

Tuttavia esiste la possibilità di eseguire test su coefficienti di correlazione, utilizzando il fatto che, se  $\rho = 0$ , la variabile casuale:

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

segue la distribuzione  $t$  di Student con  $\nu = n - 2$ ; se  $\rho \neq 0$ , si deve invece usare la cosiddetta trasformazione  $Z$  di Fischer, secondo la quale è distribuita in modo approssimativamente normale (con media e sqm sotto-indicati) la variabile casuale:

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

$$M(Z) = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

$$\sigma(Z) = \frac{1}{\sqrt{n-3}}$$

<sup>3</sup> Come detto in nota, al termine della Parte I, Welch e Tukey rimuovono l'ipotesi restrittiva sull'uguaglianza delle varianze e presentano un test approssimato, per il confronto di medie di campioni normali ed indipendenti, anche di diversa varianza. In questo modo, il valore atteso della differenza standardizzata (identico a quello per campioni numerosi) segue ancora approssimativamente la distribuzione  $t$  di Student, purché i suoi gradi di libertà siano calcolati, tenendo conto opportunamente non solo delle numerosità, dei due campioni estratti, ma anche delle loro varianze campionarie.

### Esempio 2.4.2

Un coefficiente di correlazione basato, su un campione di 20 coppie di elementi, risulta di 0.35 e, con un livello di significatività:  $\alpha = 0.05$ , occorre verificare, se sia possibile accettare l'ipotesi di un coefficiente di correlazione (dell'universo al quale il campione appartiene) nullo:

$$H_0: \rho = 0 \qquad H_1: \rho > 0$$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.35\sqrt{20-2}}{\sqrt{1-(0.35)^2}} = 1.59$$

Il test è fatto su una sola coda ed il limite della regione critica della distribuzione  $t$  di Student, con  $\nu = 18$ , è  $t_{0.95} = 1.73$ , cosicché  $H_0$  è accettato, con livello di significatività 5% .

### Esempio 2.4.3

Da due campioni di  $n_1 = 30$  e  $n_2 = 37$  coppie di elementi, si sono calcolati rispettivamente i coefficienti di correlazione  $r_1 = 0.50$  e  $r_2 = 0.20$ , ed occorre verificare, se esiste una differenza fra i due valori, al livello di significatività 5%. Applicando la trasformazione  $Z$  di Fisher, ai due valori sperimentali, si ottiene:

$$Z_1 = \frac{1}{2} \ln \frac{1+0.50}{1-0.50} = 0.5493 \qquad Z_2 = \frac{1}{2} \ln \frac{1+0.20}{1-0.20} = 0.2027$$

con 
$$\sigma(Z_1 - Z_2) = \sqrt{\sigma_{Z_1}^2 + \sigma_{Z_2}^2} = \sqrt{\frac{1}{30-3} + \frac{1}{37-3}} = 0.2578$$

Date le ipotesi, fondamentale ed alternativa:

$$H_0: M(Z_1) = M(Z_2) \qquad H_1: M(Z_1) \neq M(Z_2)$$

nell'ipotesi  $H_0$ , si ha:

$$z = \frac{(Z_1 - Z_2) - (M(Z_1) - M(Z_2))}{\sigma(Z_1 - Z_2)} = \frac{(0.5493 - 0.2027) - 0}{0.2578} = 1.34$$

I limiti della regione critica, per un test su due code ed  $\alpha = 0.05$ , sono  $z_{0.975} = \pm 1.96$ , cosicché  $H_0$  è accettato, non esistendo differenza fra  $r_1$  e  $r_2$ .

### Esempio 2.4.4

Un coefficiente di correlazione, su un campione di 28 coppie di elementi, è 0.80 e, al livello di significatività

$\alpha = 5\%$ , occorre verificare se si può respingere l'ipotesi che il coefficiente di correlazione della popolazione sia:  $\rho = 0.60$ . Date le ipotesi, fondamentale ed alternativa:

$$H_0: \rho = 0.60 \qquad H_1: \rho > 0.60$$

si ha:

$$Z = \frac{1}{2} \ln \frac{1+0.80}{1-0.80} = 1.0986 \qquad M(Z) = \frac{1}{2} \ln \frac{1+0.60}{1-0.60} = 0.6931 \qquad \sigma(z) = \sqrt{\frac{1}{28-3}} = 0.2$$

da cui:

$$z = \frac{1.0986 - 0.6931}{0.2} = 2.03$$

Pertanto con  $\alpha = 0.05$  ed il test su una sola coda, il limite della regione critica è:  $z_{0.95} = 1.64$ , ed allora  $H_0$  deve essere respinta.

## 2.5. Tests relativi alla distribuzione $F$ di Fisher

La distribuzione  $F$  di Fisher riguarda il rapporto fra le stime delle varianze di due universi normali. Essa può applicarsi all'Esempio 2.4.1 in cui, prima di procedere all'uso del test  $t$  di Student, si deve controllare che  $\sigma_x$  e  $\sigma_y$  non siano significativamente diverse. Date le ipotesi, fondamentale ed alternativa:

$$H_0: \sigma_x^2 = \sigma_y^2 \qquad H_1: \sigma_x^2 \neq \sigma_y^2$$

con:  $\nu_1 = \nu_2 = 9$ , si ha:

$$\sigma_x^2 = \frac{n_x s_x^2}{n_x - 1} = 18.23 \qquad \sigma_y^2 = \frac{n_y s_y^2}{n_y - 1} = 20.40 \qquad \text{e} \qquad F = 1.12$$

Volendo valutare  $H_0$  al livello di significatività  $5\%$ , si dovrebbe fare un test su due code (per tener conto di entrambe le possibilità  $\sigma_x \leq \sigma_y$  e  $\sigma_x \geq \sigma_y$ ), cercando sulle tavole, in corrispondenza di:  $\nu_1 = \nu_2 = 9$ , il valore di  $F_{0.975}$ , tale che sia  $0.025$  la probabilità di valori:  $F > F_{0.975}$ . Tuttavia poiché spesso sono date tavole che forniscono solo i valori  $F_{0.95}$  e  $F_{0.99}$ , calcolate nell'ipotesi che la maggiore delle due varianze sia sempre posta al numeratore, ottenendo così:  $F > 1$ , il test è eseguito sulla sola coda di destra. Dato che  $F_{0.95} = 3.18$ , il valore osservato:  $F = 1.12$ , non è significativo e si può accettare  $H_0$ .

Il test  $F$  di Fisher sull'ipotesi di uguaglianza delle varianze va applicato, contrariamente a quanto fatto in questo caso, prima di applicare il test  $t$  di *Student* (valutando la significatività della differenza di due medie). Tuttavia le principali applicazioni della variabile casuale  $F$  di Fisher riguardano l'analisi di varianza e la regressione multipla.

## 2.6. Tests relativi alla distribuzione *chi quadrato*

Nei problemi di inferenza statistica, la distribuzione  $\chi^2$  (*chi quadrato*) permette di effettuare test sulle varianze, sul buon adattamento di frequenze (alle corrispondenti probabilità di una variabile casuale) e di indipendenza.

- La distribuzione  $\chi^2$ , riferita alla variabile casuale:  $ns^2/\sigma^2$ , è usata quando si deve decidere, se una varianza campionaria è significativamente diversa da quella ipotizzata per la varianza dell'universo di provenienza del campione. In questi casi, l'ipotesi alternativa può essere la dispersione dell'universo intorno al suo valore medio aumentata o diminuita (cosa di estrema importanza, in tutti i processi produttivi, dove si tende a mantenere la dispersione entro limiti prefissati, il più possibile ristretti).

### Esempio 2.6.1

Una macchina dovrebbe riempire sacchi di materiale con uno sqm di  $0.10\text{ kg}$  ed invece, su un campione casuale di 22 sacchi, si è calcolato uno sqm di  $0.15\text{ kg}$ . Allora occorre verificare, se l'apparente aumento della dispersione è significativo ai livelli di probabilità: 0.05 e 0.01:

$$H_0: \quad \sigma = 0.10\text{ kg} \qquad H_1: \quad \sigma > 0.10\text{ kg}$$

Il valore  $\chi^2$  per il campione è:

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{22(0.15)^2}{(0.10)^2} = 49.5$$

Eseguendo il test su una sola coda, con  $\nu = 22 - 1 = 21$ , si ha:  $\chi_{0.95}^2 = 32.7$  e  $\chi_{0.99}^2 = 38.9$ . Il valore sperimentale  $\chi^2$  è maggiore di entrambi per cui  $H_0$  è respinta e la macchina deve essere sottoposta a revisione.

- Date le frequenze totali dei valori argomentali di una variabile casuale, nell'ipotesi che la stessa obbedisca ad una particolare funzione di distribuzione, spesso queste frequenze teoriche, così calcolate, sono in discreto accordo con quelle sperimentali osservate.

La distribuzione  $\chi^2$  si applica nei casi in cui si devono confrontare fra loro due serie di frequenze totali, la prima ricavata da un esperimento, la seconda calcolata ipotizzando che la variabile casuale, riferita a

questo esperimento, segua una certa distribuzione di probabilità. I parametri, caratterizzanti la densità di probabilità, possono essere ipotizzati o, più frequentemente, dedotti dai dati osservati.

Se un'ipotesi è corretta, la probabilità di un valore argomentale qualsiasi può essere espressa da:

$p = f_t/n$ , con  $f_t$  frequenza teorica corrispondente, calcolata in base all'ipotesi stessa. La frequenza osservata  $f_0$  di quel valore argomentale è una variabile casuale che può assumere i valori:  $0, 1, 2, \dots, n$ , essendo  $n$  il numero di individui esaminati.

Trattandosi di una variabile casuale, dove valori argomentali sono il numero delle volte che un evento di probabilità  $p$  si presenta, in  $n$  prove, la probabilità del generico valore  $k$  è data dalla legge binomiale, con media della distribuzione:  $M = np = f_t$ . Se poi  $n$  è sufficientemente grande, la variabile casuale:  $f_0 - f_t$ , si può considerare normale (con media:  $M = 0$ ) e la varianza della distribuzione può essere calcolata, per  $n \rightarrow \infty$ , come:

$$\sigma^2 = npq = n \frac{f_t}{n} \left(1 - \frac{f_t}{n}\right) = f_t - \frac{f_t^2}{n} \quad \text{per cui} \quad \lim_{n \rightarrow \infty} \sigma^2 = f_t$$

Pertanto la variabile casuale:  $(f_0 - f_t)/\sqrt{f_t}$ , può considerarsi, almeno approssimativamente, normale e standardizzata. Se poi, per ognuna delle  $m$  coppie di frequenze osservate e teoriche corrispondenti, si costruisce il rapporto:  $(f_0 - f_t)/\sqrt{f_t}$ , per le proprietà della variabile casuale  $\chi^2$ , la somma:

$$\frac{(f_0^{(1)} - f_t^{(1)})^2}{f_t^{(1)}} + \frac{(f_0^{(2)} - f_t^{(2)})^2}{f_t^{(2)}} + \dots + \frac{(f_0^{(m)} - f_t^{(m)})^2}{f_t^{(m)}} \quad (6.1)$$

segue una distribuzione  $\chi^2$ . Il numero dei gradi di libertà è:  $\nu = m - 1 - k$ , dove  $k$  è il numero dei parametri stimati, in base ai dati osservati. Se nessun parametro è stimato, si ha:  $\nu = m - 1$ . perché le frequenze teoriche sono legate a quelle osservate dalla relazione:  $\sum f_0 = \sum f_t = n$ . Il valore  $\chi^2$ , ricavato dalla (6.1) e detto sperimentale, è nullo, se i dati sperimentali sono identici a quelli attesi, in base all'ipotesi secondo cui si ricavano le frequenze teoriche. Invece più grande è il valore  $\chi^2$  sperimentale e più teoria ed osservazioni sono in disaccordo. Anche qui si tratta di valutare, se le deviazioni fra  $f_0$  e  $f_t$  possono essere dovute al caso o se pure sono così grandi, da far respingere l'ipotesi fatta, sulla distribuzione della variabile casuale, riferita all'esperimento.

A questo scopo, la distribuzione  $\chi^2$  serve egregiamente, potendosi fissare, come limite della regione critica, il valore  $\chi_\alpha^2$ , tale che si ha:  $P(\chi^2 > \chi_\alpha^2) = \alpha$ , in base alla distribuzione  $\chi^2$ , con  $\alpha$  il livello di significatività del test.

L'ipotesi che l'esperimento segua una certa distribuzione è respinta, se  $\chi^2$  sperimentale è maggiore di

$\chi^2_\alpha$ , perché il test è fatto su di una sola coda. Tuttavia è bene controllare anche che il valore  $\chi^2$  sperimentale non sia troppo prossimo allo zero, ad esempio, inferiore ad un  $\chi^2_b$  tale che:  $P(0 < \chi^2 < \chi^2_b) = 0.05$ . Infatti valori così bassi di  $\chi^2$  sperimentale si possono ottenere, per motivi puramente casuali, solo in 5 casi su 100, cosicché ottenerne uno può far pensare che l'ipotesi fatta sia troppo buona, cioè che, in qualche modo, si utilizzino i dati sperimentali, per costruire una teoria con la quale poi confrontarli.

Nella applicazione di  $\chi^2$  ai controlli di ipotesi di questo tipo, detti di *buon adattamento*, bisogna aver presenti alcune avvertenze.

- Di solito, si ha un'approssimazione sufficiente, se si ha:  $f_i \geq 5$  e  $\nu \geq 5$ , mentre è opportuno avere:  $f_i > 5$ , se invece:  $\nu < 5$ .
- Se le frequenze teoriche sono molto piccole, gli addendi in cui esse compaiono al denominatore hanno un peso preponderante, nella formazione di  $\chi^2$  sperimentale, e possono alterarne il valore macroscopicamente. In questo caso, è bene riunire, in una stessa classe, più valori argomentali contigui, sommando fra loro le rispettive frequenze osservate e teoriche, così da aumentare il valore di quest'ultima nell'unico addendo  $(f_0 - f_t)^2 / f_t$ , proveniente dal conglobamento fatto.
- Se si hanno a disposizione  $s$  serie di esperimenti dello stesso tipo, ma indipendenti fra loro, con i dati dei quali si potrebbero calcolare  $s$  valori  $\chi^2$  sperimentale con:  $\nu_1, \nu_2, \dots, \nu_s$ , gradi di libertà, è opportuno sfruttare la proprietà di sommabilità di  $\chi^2$ , determinando un  $\chi^2$  sperimentale uguale a:  $\chi^2_1 + \chi^2_2 + \dots + \chi^2_s$ , e poi confrontarlo con quello teorico, corrispondente al livello di significatività prefissato, della distribuzione  $\chi^2$ , con:  $\nu_1 + \nu_2 + \dots + \nu_s$ , gradi di libertà.

### Esempio 2.6.2

Avendo misurato, con precisione, lo spessore di 100 fili, con livello di significatività 5%, valutare l'adattamento delle frequenze sotto-riportate:

<i>sperimentali</i>	6	20	40	27	7
<i>teoriche</i>	4.93	22.05	38.49	26.31	7.03

$H_0$ : la distribuzione normale con:  $M = 49.27 \mu m$  e  $\sigma = 2.97 \mu m$ , ben si adatta bene a quella dello spessore dei fili prodotti

$H_1$ : questa distribuzione normale non si adatta ai dati sperimentali

In questo caso, la (6.1) fornisce il valore sperimentale:

$$\chi^2 = \frac{(6 - 4.93)^2}{4.93} + \frac{(20 - 22.05)^2}{22.05} + \frac{(40 - 38.49)^2}{38.49} + \frac{(27 - 26.31)^2}{26.31} + \frac{(7 - 7.03)^2}{7.03} = 0.50$$

Dato che il numero  $k$  di parametri della distribuzione teorica stimati dai dati osservati è uguale a 2 (avendo stimato:  $M$  e  $\sigma$ ), si ha:  $\nu = 5 - 1 - 2 = 2$ , e così:  $\chi^2_{0.95} = 5.99$ , per cui  $H_0$  è accettata e l'adattamento è molto buono (inoltre essendo:  $\chi^2_{0.05} = 0.10$ , l'adattamento non è troppo buono).

- Il test  $\chi^2$  è usato, in base agli stessi principi esposti nel punto precedente, per il controllo di ipotesi dette di *indipendenza*. In questo caso, un campione di  $n$  individui è esaminato sotto il punto di vista di due caratteristiche diverse, con lo scopo di stabilire, se l'ipotesi sull'indipendenza (una dall'altra), delle due caratteristiche, può essere accettata, oppure no. Le frequenze teoriche, da confrontare con quelle sperimentali, si derivano proprio in base alla ipotesi  $H_0$  di indipendenza fra le due caratteristiche.

### Esempio 2.6.3.

Nella prima delle due tabelle sotto-riportate, sono indicati il numero di individui, su 1000 intervistati, dichiaratisi rispettivamente favorevoli, contrari od indecisi, riguardo una questione sulla quale è in corso il dibattito parlamentare.

I mille individui sono estratti, a caso, in due comunità considerate potenzialmente diverse, i residenti in comuni con popolazione inferiore a 10.000 abitanti (Gruppo A) ed i residenti in comuni con popolazione superiore a 10.000 abitanti (Gruppo B). Le ipotesi, fondamentale e alternativa, sono:

$H_0$ : la risposta è indipendente dall'ambiente di residenza

$H_1$ : la risposta non è indipendente dall'ambiente di residenza

Se  $H_0$  è vera, le percentuali di favorevoli, contrari ed indecisi, presenti nei due gruppi, devono essere le stesse di quelle riscontrate nel totale degli individui esaminati ossia, nel gruppo A, dovrebbero essere rispettivamente favorevoli, contrari ed indecisi il 56.5%, 27.7% e 15.8% di 451 individui, mentre nel gruppo B si dovrebbero ritrovare le stesse percentuali su 549 individui. Allora nella seconda delle due tabelle sotto-riportate, sono indicate le frequenze teoriche, in base all'ipotesi  $H_0$ .

$f_0$	gruppo A	gruppo B	totale	$f_t$	gruppo A	gruppo B	totale
<i>favorevoli</i>	233	332	565	<i>favorevoli</i>	254.82	310.19	565
<i>contrari</i>	135	142	277	<i>contrari</i>	124.93	152.07	277
<i>indecisi</i>	83	75	158	<i>indecisi</i>	71.26	86.74	158
<i>totale</i>	451	549	1000	<i>totale</i>	451	549	1000

Dato che i totali per righe e per colonne devono essere gli stessi nelle due tabelle, le frequenze teoriche non sono tutte indipendenti fra loro. Nell'esempio, solo due e non nella stessa riga, potrebbero essere messe liberamente, mentre le altre si ricavano dai totali. Il numero di frequenze teoriche indipendenti costituisce il numero di gradi di libertà del problema e della distribuzione  $\chi^2$ , utilizzata per valutare la significatività della somma:

$$\chi^2 = \sum \frac{(f_0^{(1)} - f_t^{(1)})^2}{f_t^{(1)}} = 8.41$$

In generale, con tabelle di dimensioni:  $m \times n$ , si ha:  $\nu = (m-1)(n-1)$ , ed in questo caso:  $\nu = 2$ .

Dato che, con  $\nu = 2$ ,  $\chi_{0.95}^2 = 5.99$ , le frequenze delle due tabelle sono significativamente diverse fra loro e, con una probabilità di errore di primo tipo del 5%, si respinge l'ipotesi che l'opinione dei cittadini non risenta dell'influenza della località in cui vivono.

Contrariamente ai controlli di buon adattamento (dove ogni valore:  $f_t/n$ , rappresenta effettivamente la probabilità del valore argomentale corrispondente, in base alla distribuzione ipotizzata), nei controlli di indipendenza  $f_t/n$  tende alla probabilità, in base alla legge empirica del caso, ed affinché il test dia risultati attendibili, si richiede  $n$  elevato (nei test di indipendenza valgono poi le stesse avvertenze, fatte per i test di buon adattamento) <sup>4</sup>.

<sup>4</sup> L'aggettivo non parametrico (in inglese: distribution-free, anche se i due termini non sono sinonimi) qualifica un particolare gruppo di test statistici, sotto certe condizioni, sostitutivo dei test statistici classici. Infatti i test non parametrici, rispetto ai test classici, presentano i seguenti vantaggi:

- la loro comprensione è immediata ed elementare;
- le condizioni di validità sono meno forti (più ampie);
- i calcoli necessari non presentano, in generale, difficoltà computazionali.

D'altra parte, i test non parametrici presentano alcuni svantaggi: molta informazione è sprecata e la potenza del test è bassa, cosicché test poco potenti tendono ad essere troppo conservativi, cioè l'ipotesi fondamentale (o nulla) è accettata, anche quando dovrebbe valere l'ipotesi alternativa. Pertanto i test statistici classici sono preferibili, quando le condizioni di validità sono soddisfatte. Di seguito, sono presentati due test di rango (sui valori centrali e sulle dispersioni), per campioni indipendenti, e due test del segno (di Thompson, ancora sui valori centrali e sulle dispersioni), per campioni qualsiasi, oltre al test sul coefficiente di correlazione sui ranghi.

#### Test di Mann-Whitney

L'ipotesi  $H_0: \mu_x = \mu_y$ , porta al confronto dei valori centrali di due campioni  $X$  e  $Y$  indipendenti. A riguardo, i dati dei campioni sono sostituiti dai corrispondenti ranghi i cui valori vanno da 1, per il dato di valore argomentale minimo, a  $(N_x + N_y)$ , per il dato di valore argomentale massimo. Detta  $\hat{R}_x$  la somma dei ranghi del campione  $X$ , si ha:

$$\frac{\hat{R}_x - \frac{N_x(N_x + N_y + 1)}{2}}{\sqrt{\frac{N_x N_y (N_x + N_y + 1)}{12}}} \approx z \quad \text{con:} \quad z = N(0,1)$$

#### Test di Siegel-Tuckey

L'ipotesi  $H_0: \sigma_x^2 = \sigma_y^2$ , porta al confronto dei valori di dispersione di due campioni  $X$  e  $Y$  indipendenti. A riguardo, i dati dei campioni sono sostituiti dai corrispondenti ranghi i cui valori vanno da 1, per il dato il cui scarto in valore assoluto rispetto alla mediana è minimo, a  $(N_x + N_y)$ , per il dato il cui scarto in valore assoluto rispetto alla mediana è massimo. Detta  $\hat{R}_x$  la somma dei ranghi del campione  $X$ , si ha la stessa espressione, asintoticamente normale, del test di Mann-Whitney.

#### Test del segno (per i valori centrali)

L'ipotesi  $H_0: \mu_x = \mu_y$ , porta al confronto dei valori centrali di due campioni  $X$  e  $Y$  qualsiasi. Infatti nel caso dei cosiddetti studi "prima e dopo", cioè quando si misura due volte lo stesso campione, si ottengono due campioni  $X$  (valore misurato "prima") e  $Y$  (valore misurato "dopo") non indipendenti. Per ogni coppia di valori argomentali, si determina il segno (*più* o *meno*, scartando le differenze nulle), secondo la convenzione:

valore " prima"		valore " dopo"	segno
$X$	>	$Y$	-
$X$	<	$Y$	+
$X$	=	$Y$	nessuno

## 2.7. Tests sequenziali

Qualora si impongano valori abbastanza piccoli di  $\alpha$  e  $\beta$ , come nell'Esempio 2.3.2, può essere necessario esaminare un campione molto numeroso, prima di decidere se accettare o scartare l'ipotesi fondamentale. In questo caso, se le osservazioni di un esperimento sono fatte in serie, nel senso che il risultato  $x$  di ogni singola prova è noto prima di effettuare la successiva, si può seguire una procedura diversa che, nella maggioranza dei casi, riduce molto il numero di esperimenti necessari, per poter prendere una decisione, risultando così molto più economica.

Pertanto in questi tipi di test, detti sequenziali, non è fissata a priori la numerosità del campione, in esame, ed il test è fatto dopo ogni osservazione sull'insieme dei dati, accumulati fino a quel momento, ripetendo l'esperimento, fino a quando non è possibile decidere quale delle due ipotesi alternative accettare con il prestabilito livello di significatività. I test sequenziali richiedono un grafico sul quale sono riportate:

- in ascissa, la numerosità del campione, fino a quel momento;
- in ordinata, una particolare funzione  $f(x)$  dei valori ottenuti, a seconda del particolare tipo di test.

Sul grafico, nel caso più semplice, si tracciano anche due linee di confine la cui posizione dipende dall'entità dei rischi  $\alpha$  e  $\beta$ , dall'entità della differenza dei valori del parametro  $\theta$ , nelle due ipotesi  $H_0$  e  $H_1$ , ecc. che delimitano ed individuano tre zone:

- l'accettazione dell'ipotesi fondamentale  $H_0$ ;
- l'accettazione dell'ipotesi alternativa  $H_1$ ;

Detti:  $N_p = n$ . di segni "+",  $N_m = n$ . di segni "-" e  $N_{tot} = N_p + N_m$  e calcolata la frazione dei segni "più", sul totale dei segni:

$\hat{f} = N_p / N_{tot}$ , si ha:

$$\frac{\hat{f} - 0.5}{0.5 / \sqrt{N_{tot}}} \approx z \quad \text{con:} \quad z = N(0,1)$$

### Test del segno (per i valori di dispersione)

L'ipotesi  $H_0: \sigma_x^2 = \sigma_y^2$ , porta al confronto dei valori di dispersione di due campioni  $X$  e  $Y$  qualsiasi. Infatti anche in questo caso, si ottengono due campioni non indipendenti. Per ogni coppia di scarti in valore assoluto rispetto alla mediana, si determina il segno (*più* o *meno*, scartando sempre le differenze nulle), secondo la convenzione:

valore "prima"	valore "dopo"	segno	
$ X - \text{mediana}(X) $	$>$	$ Y - \text{mediana}(Y) $	-
$ X - \text{mediana}(X) $	$<$	$ Y - \text{mediana}(Y) $	+
$ X - \text{mediana}(X) $	$=$	$ Y - \text{mediana}(Y) $	nessuno

Dopodiché si esegue la stessa procedura, con la stessa espressione, asintoticamente normale, del test del segno (per i valori centrali).

### Test di Spearman

La procedura per il calcolo del coefficiente di correlazione sui ranghi, fra due campioni qualsiasi, si attua nei seguenti passi:

- ordinare i dati per ciascuna componente  $X$  e  $Y$ ;
- assegnare i ranghi, separatamente, a ciascuna componente, nell'ordine crescente dei valori argomentali;
- calcolare, elemento ad elemento, le differenze  $\Delta_i$  fra i ranghi delle due componenti;
- calcolare il coefficiente di correlazione sui ranghi (di Spearman):  $\hat{r}_{xy} = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n \Delta_i^2$ ;
- effettuare il test  $t_{N-2}$  di Student, nel caso di incorrelazione:  $\rho = 0$ , ed test con la trasformata  $Z$  di Fisher, in tutti gli altri casi.

□ l'impossibilità di prendere una decisione.

Se il diagramma di  $f(x)$  si mantiene nella terza zona, l'esperimento deve continuare ad essere ripetuto, mentre non appena il diagramma entra nelle prima o seconda zona si accetta rispettivamente l'ipotesi  $H_0$  o l'ipotesi  $H_1$ , interrompendo la sperimentazione. E' evidente che la numerosità del campione necessaria, per poter prendere una decisione, varia a seconda dei valori ottenuti, contrariamente a quanto accade nei test non sequenziali.

Il tipo di linee confine e la forma della funzione  $f(x)$  sono qui ricavate solo per il caso, molto semplice, in cui si ha un solo parametro da sottoporre a test, ad esempio, la media, supponendo che gli altri parametri siano noti e che si tratti di un test su una sola coda.

A riguardo, si supponga che questo parametro possa assumere solo due valori  $\theta_0$  e  $\theta_1 (> \theta_0)$ . Lo scopo del test è decidere quale delle due alternative sia quella vera e, a questo scopo, si introduce il *rapporto di verosimiglianza*:

$$\lambda = \frac{\text{probabilità di un certo campione quando } \theta = \theta_0}{\text{probabilità dello stesso campione quando } \theta = \theta_1} \quad (7.1)$$

dove, se la distribuzione è continua, il rapporto si esegue fra le densità di probabilità.

L'ipotesi  $H_0: \theta = \theta_0$  è accettabile quando  $\lambda$  è grande e l'ipotesi  $H_1: \theta = \theta_1$  è accettabile quando  $\lambda$  è piccolo. Di conseguenza, al fine dell'esecuzione del test sequenziale, si stabiliscono due limiti  $\lambda_0$  e  $\lambda_1$ , e si calcola il valore di  $\lambda$ , dopo ogni osservazione, cosicché:

- $H_0$  è accettato non appena  $\lambda \geq \lambda_0$ ;
- $H_1$  è accettato non appena  $\lambda \leq \lambda_1$ ;
- un altro elemento va aggiunto al campione, se  $\lambda_1 < \lambda < \lambda_0$ .

Il valore  $\lambda_0$  può essere calcolato, introducendo nella (7.1), invece del generico valore campionario, l'insieme dei valori che portano all'accettazione di  $H_0$ . La probabilità di questi valori è  $1 - \alpha$ , se  $\theta = \theta_0$ , e  $\beta$ , se  $\theta = \theta_1$ , per cui:

$$\lambda_0 = \frac{1 - \alpha}{\beta}$$

Analogamente il valore  $\lambda_1$  è il rapporto fra la probabilità dei valori campionari che portano all'accettazione di  $H_1$ , essendo  $\theta = \theta_0$ , e quella degli stessi valori essendo  $\theta = \theta_1$ , da cui:

$$\lambda_1 = \frac{\alpha}{1 - \beta}$$

Dato che le entità di  $\alpha$  e  $\beta$  sono fissate a priori, si possono facilmente determinare i due valori limite con i quali confrontare il valore  $\lambda$ , calcolato ad ogni successivo esperimento. Tuttavia in pratica, è più opportuno usare qualche semplice funzione dei valori campionari  $x_i$ , a seconda del tipo di test da fare.

Ad esempio, per un test su medie, su una sola coda, essendo la popolazione distribuita normalmente con varianza nota, si ha  $H_0: M = \mu_0$  e  $H_1: M = \mu_1 > \mu_0$  con  $\mu_0$ ,  $\mu_1$  e  $\sigma$  noti,  $\alpha$  e  $\beta$  assegnati. Se sono fatte  $n$  osservazioni, la densità di probabilità di un certo gruppo di valori  $x_i$ , se è vera  $H_0$ , è:

$$L(x_1, x_2, \dots, x_n / \theta_0) = \frac{1}{\sqrt{(2\pi)^n \sigma^n}} e^{-\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2}}$$

mentre la densità di probabilità dello stesso gruppo di valori, se è vera  $H_1$ , risulta:

$$L(x_1, x_2, \dots, x_n / \theta_1) = \frac{1}{\sqrt{(2\pi)^n \sigma^n}} e^{-\frac{\sum_{i=1}^n (x_i - \mu_1)^2}{2\sigma^2}}$$

Il rapporto di verosimiglianza è:

$$\lambda = \frac{L(x_1, x_2, \dots, x_n / \theta_0)}{L(x_1, x_2, \dots, x_n / \theta_1)} = \frac{e^{-\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2}}}{e^{-\frac{\sum_{i=1}^n (x_i - \mu_1)^2}{2\sigma^2}}} \quad \text{e} \quad \ln \lambda = \frac{\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2} - \frac{\sum_{i=1}^n (x_i - \mu_1)^2}{2\sigma^2}$$

da cui, con qualche passaggio, si ottiene:

$$\sum_i^n x_i = T_n = \frac{-\sigma^2 \ln \lambda}{\mu_1 - \mu_0} + \frac{n}{2}(\mu_1 + \mu_0) \quad (7.2)$$

In questo modo, invece di calcolare  $\lambda$ , dopo ogni osservazione, e confrontarlo con i valori limite  $\lambda_0$  e  $\lambda_1$ , si può calcolare la somma dei valori osservati:

$$T_n = \sum_{i=1}^n x_i$$

e confrontarla con i valori  $T_0$  e  $T_1$ , ottenuti ponendo nella (7.2) rispettivamente  $\lambda = \lambda_0$  e  $\lambda = \lambda_1$ , cosicché:

$$T_0 = h_0 + ns \quad T_1 = h_1 + ns \quad (7.3)$$

dove:

$$h_0 = -\frac{b\sigma^2}{\delta} \qquad h_1 = -\frac{a\sigma^2}{\delta} \qquad s = \frac{1}{2}(\mu_1 + \mu_0) = \mu_0 + \frac{1}{2}\delta$$

$$\delta = \mu_1 - \mu_0 \qquad a = \ln \frac{1-\beta}{\alpha} \qquad b = \ln \frac{1-\alpha}{\beta}$$

Le due rette (7.3) rappresentano le linee di confine, nel grafico (mostrato in Fig. 2.7.1) di un test sequenziale di tipo lineare. La loro distanza, in direzione parallela a  $T_n$ , cioè l'ampiezza del corridoio dove si ha nessuna decisione, è direttamente proporzionale alla varianza della popolazione ed inversamente proporzionale alla differenza fra i parametri  $\mu_1$  e  $\mu_0$  (cosicché la discriminazione fra due alternative molto vicine richiede un maggior numero di osservazioni).

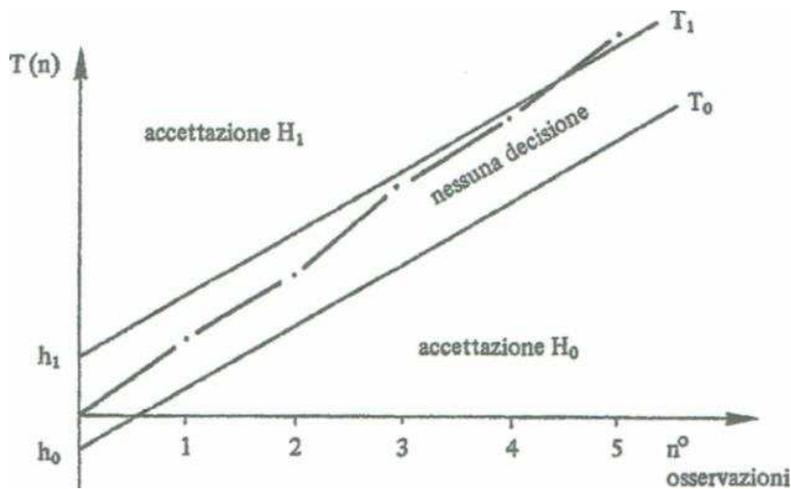


Fig. 2.7.1 – Schema di test sequenziale ad una sola coda con  $\theta_1 > \theta_0$

In questo tipo di test sequenziale ed in tutti gli altri casi di test lineari, è possibile ricavare l'equazione di una curva rappresentante il numero medio di osservazioni, da fare per raggiungere una decisione in funzione del valore effettivo del parametro  $\theta$ . Questa curva che ha un andamento simile a quello di Fig. 2.7.2., come facilmente prevedibile, ha il suo massimo fra  $\theta_0$  e  $\theta_1$ , e ha ordinate molto inferiori al valore richiesto da un test non sequenziale, per la maggioranza dei valori di  $\theta$ .

Il metodo, presentato nel grafico di Fig. 2.7.1, è applicabile per decidere, se un valore medio è più grande significativamente di un valore  $\mu_0$  assegnato, essendo nota la varianza. Un procedimento analogo si può seguire per decidere, se il valore medio è significativamente inferiore a  $\mu_0$ . Se  $\mu_0 = 0$ , il grafico risultante risulta simmetrico, rispetto all'asse  $n$  (delle ascisse), di quello di Fig. 2.7.1. Invece se l'ipotesi alternativa è

$H_1: M \neq \mu_0$ , cioè quando si ha un test su due code, il grafico del test consiste in una combinazione della Fig. 2.7.1 e della sua simmetrica, presentando così uno schema analogo a quello di Fig.2.7.3 nella quale si hanno quattro zone, rappresentanti rispettivamente le seguenti decisioni:

- accettazione di  $H_1: M < \mu_0$ ;
- accettazione di  $H_1: M > \mu_0$ ;
- accettazione di  $H_0: M = \mu_0$ ;
- nessuna decisione.

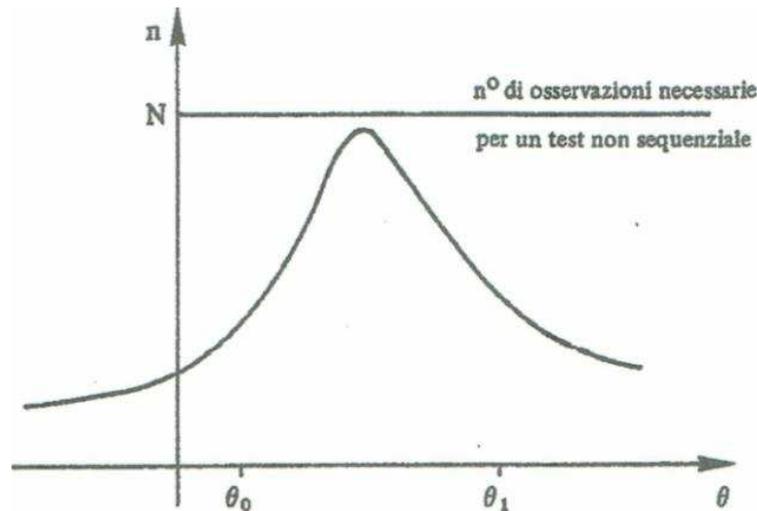


Fig. 2.7.2 – Andamento del numero medio di osservazioni richieste da un test sequenziale

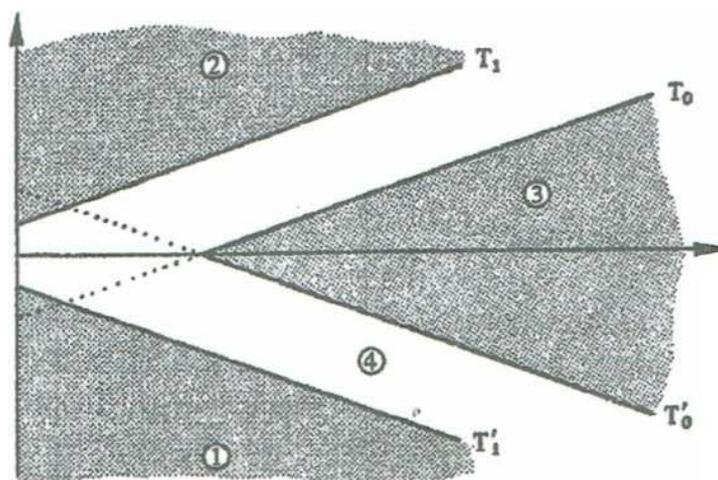


Fig. 2.7.3 – Schema di test sequenziale a due code

### Esempio 2.7.1

In un'indagine sui fattori, influenzanti la resistenza alla rottura di una fibra tessile, si introduce una modifica nella fabbricazione della stessa, preparando due serie di esemplari, una secondo la procedura tradizionale, l'altra secondo la procedura modificata. Il risultato di ogni coppia di osservazioni è già noto prima della fabbricazione della successiva coppia, cosicché è consigliabile l'applicazione di un test sequenziale, per decidere, se la modifica apportata ha realmente variato il valore medio del carico di rottura. A riguardo, i valori delle differenze fra le tensioni di rottura delle fibre, preparate secondo le due diverse modalità, sono:

$$\Delta = x_1 - x_2 = 7, 5, 8, -11, 10, 8, -9, 6, -7$$

Da precedenti esperimenti, è appurato che lo sqm delle tensioni di rottura è  $\sigma = 7.07$  unità ed il test sequenziale è progettato, in modo da correre un rischio  $\alpha = 0.05$ , di affermare la presenza di cambiamenti non esistenti, e contemporaneamente di evidenziare, con il 90% di probabilità, una variazione pari a  $\pm 10$  unità. Allora dati:

$$\sigma_{\Delta} = \sigma_x \sqrt{2} = 10$$

$$\alpha = 0.05$$

$$\beta = 1 - 0.90 = 0.10$$

$$H_0: \mu_0 = \mu_1$$

$$H_1: \mu_0 = \mu_1 \pm 10$$

$$\text{essendo: } \delta = 0.10$$

le equazioni delle rette limiti sono:

$$\begin{cases} T_0 = h_0 + ns = -22.8 + 5n \\ T_1 = h_1 + ns = 35.8 + 5n \end{cases}$$

$$\begin{cases} T'_0 = h'_0 - ns' = 22.8 - 5n \\ T'_1 = h'_1 - ns' = -35.8 - 5n \end{cases}$$

dove

$$h_0 = -b\sigma^2/\delta = -h'_0$$

$$h_1 = -a\sigma^2/\delta = -h'_1$$

$$s = \delta/2 = -s'$$

$$a = \ln \frac{1-\beta}{\alpha/2}$$

$$b = \ln \frac{1-\alpha/2}{\beta}$$

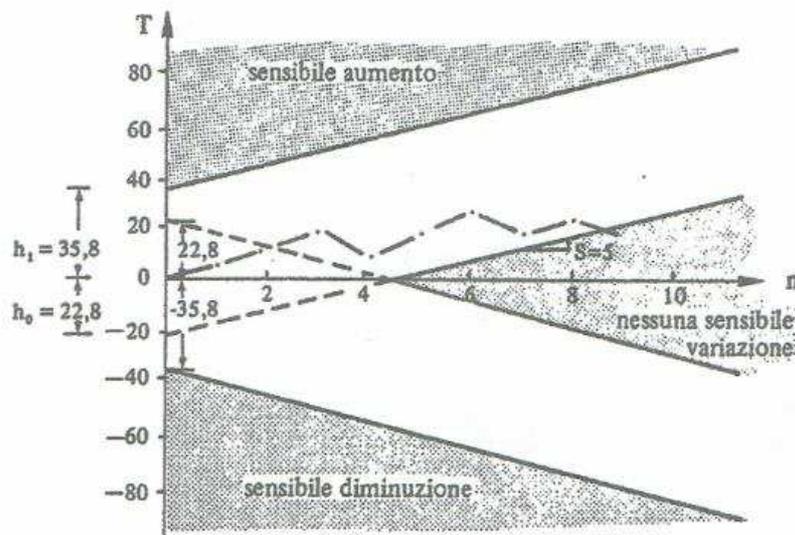


Fig. 2.7.4 – Schema di test sequenziale a due code, relativo all'Esempio 2.7.1

La Fig. 2.7.4 rappresenta la posizione delle rette limite e delle varie zone di decisione. I valori  $\Delta$ , ottenuti dai successivi esperimenti, portano a valori cumulati (riportati nel diagramma):

$$T(n) = 7, 12, 20, 9, 19, 27, 18, 24, 17$$

L'ultimo valore  $T_9$  risulta fuori dalla zona di nessuna decisione e si trova invece in quella di accettazione di  $H_0$ . Pertanto si conclude che il diverso procedimento di fabbricazione non porta sensibili variazioni nella resistenza alla rottura della fibra in esame.

Qualora la varianza della popolazione, invece di essere nota, è stimata dai valori osservati, si applica il test sequenziale non lineare di Barnard, equivalente al test  $t$  non sequenziale. La funzione dei valori osservati, utilizzata per delimitare le diverse zone del grafico, è così:

$$U(n) = \frac{\sum_{i=1}^n (x_i - \mu_0)}{\sqrt{\sum_{i=1}^n (x_i - \mu_0)^2}}$$

I valori  $U_0$  ed  $U_1$  sono forniti dalle tavole Davies, in funzione di un parametro  $D$ , rappresentante la differenza fra le medie, considerata sufficiente per concludere accettando:  $H_1 : M = \mu_1 > \mu_0$ , oppure  $H_1 : M = \mu_1 < \mu_0$ , espressa in termini di scarto quadratico medio:

$$D = \frac{\mu_1 - \mu_0}{\sigma}$$

L'andamento del grafico del test sequenziale è analogo a quello di Fig. 2.7.5 e le quattro zone hanno lo stesso significato di quelle della precedente Fig. 2.7.3.

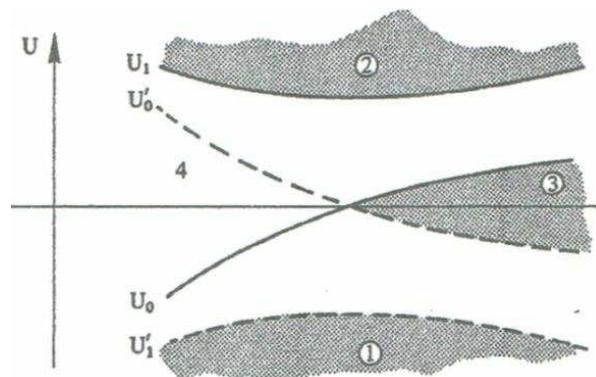


Fig. 2.7.5 – Schema di test sequenziale, non lineare di Barnard, a due code

I test sequenziali per percentuali e differenze di percentuali, ovvero i test relativi alla distribuzione binomiale, sono nuovamente di tipo lineare, con rette limite di equazione:

$$X_0 = h_0 + ns \qquad X_1 = h_1 + ns$$

dove:

$$h_0 = -b/(P+Q) \qquad h_1 = a/(P+Q) \qquad s = Q/(P+Q)$$

$$P = \ln(p_1 - p_0) \qquad Q = \ln(1 - p_0)/(1 - p_1)$$

$$a = \ln(1 - \beta)/\alpha \qquad b = \ln(1 - \alpha)/\beta$$

In questo caso, la funzione  $f(x)$  usata è semplicemente il numero totale di casi in cui si è verificato l'evento di interesse ed un analogo test sequenziale di tipo lineare è utilizzato per test su varianze.

## 2.8. Limiti fiduciarî per statistiche campionarie

Data una statistica campionaria è possibile determinare un intervallo  $[A, B]$ , tale che, se la statistica teorica della popolazione è compresa in questo intervallo, la statistica campionaria possa considerarsi estratta da questa popolazione, con un prefissato livello di probabilità o *fiducia*. In termini più intuitivi, anche se meno corretti: data una statistica campionaria, si possono anche stabilire i limiti  $A$  e  $B$  entro i quali è compresa la corrispondente (incognita) statistica dell'universo.

L'intervallo, detto *fiduciario*, per la statistica in questione, è tanto più ampio quanto più piccola è la probabilità di errore di prima specie  $\alpha$  commesso. Infatti stabilire  $\alpha$  piccolo implica che l'informazione ottenuta, a parità di numerosità  $n$  del campione, sulla statistica dell'universo è più scarsa che con  $\alpha$  più elevato, aumentando così il rischio di errore. A parità di  $\alpha$ , il solo modo per restringere l'intervallo entro cui è compresa la statistica dell'universo, è aumentare la numerosità del campione.

L'attendibilità di un intervallo fiduciario è di solito indicata dal valore  $1 - \alpha$ : se si stabilisce  $\alpha = 0.05$ , con la probabilità del 95%, la statistica dell'universo in esame è effettivamente compresa entro i limiti  $A$  e  $B$  trovati, in conseguenza di  $\alpha = 0.05$ . Di conseguenza, 95% è una misura della fiducia nella correttezza dell'intervallo stabilito.

Ad esempio, dato lo sqm calcolato su un campione di 180 misure angolari risulta di 15", si possono trovare i limiti fiduciarî al 95% per lo sqm della popolazione alla quale quel campione appartiene. Infatti se  $\chi_1^2$  e  $\chi_2^2$  sono due valori tali che:

$$P(0 < \chi^2 < \chi_1^2) + P(0 < \chi^2 < \chi_2^2) = \alpha$$

si sa che con probabilità  $1 - \alpha$  si ha:

$$\chi_1^2 \leq \frac{ns^2}{\sigma^2} \leq \chi_2^2$$

dove  $s^2$  è la varianza campionaria nota e  $\sigma^2$  la varianza incognita della popolazione, cosicchè:

$$\frac{ns^2}{\chi_2^2} \leq \sigma^2 \leq \frac{ns^2}{\chi_1^2}$$

Di conseguenza, i due numeri  $ns^2/\chi_2^2$  e  $ns^2/\chi_1^2$  forniscono i due estremi  $A^2$  e  $B^2$  di un intervallo in cui  $\sigma^2$  è compreso, con il livello di fiducia  $1 - \alpha$ .

Nell'esempio, dati  $\chi_1^2 = \chi_{0.025}^2$  e  $\chi_2^2 = \chi_{0.975}^2$ , ed il campione numeroso ( $n = 180$ ), per la determinazione di questi due valori, si usa la variabile casuale normale standardizzata:  $z = \sqrt{2\chi^2} - \sqrt{2\nu - 1}$ :

$$\begin{aligned} z_{0.025} &= \sqrt{2\chi_{0.025}^2} - \sqrt{2(179) - 1} = -1.96 & \chi_{0.025}^2 &= 143.4 \\ z_{0.975} &= \sqrt{2\chi_{0.975}^2} - \sqrt{2(179) - 1} = 1.96 & \chi_{0.975}^2 &= 217.5 \end{aligned}$$

cosicché i due limiti fiduciarî al 95% per lo sqm della popolazione sono:

$$A = \frac{15\sqrt{180}}{\sqrt{217.5}} = 13.65'' \quad \text{e} \quad B = \frac{15\sqrt{180}}{\sqrt{143.4}} = 16.81''$$

Inoltre essendo il campione numeroso, lo stesso problema può essere risolto, utilizzando la distribuzione degli sqm. campionari, ottenendo rispettivamente:  $A = 13.59''$  e  $B = 16.72''$ .

In modo perfettamente analogo, si possono trovare i limiti fiduciarî per la media di una popolazione, note la media e la varianza campionarie, di campioni numerosi e di piccoli campioni. Nel primo caso, si utilizza la distribuzione normale e, nel secondo, la distribuzione  $t$  di *Student*. Gli estremi dell'intervallo fiduciario si ricavano risolvendo due disuguaglianze, nell'incognita  $M$ :

$$z_{\alpha/2} \leq \frac{\bar{x} - M}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2} \quad \text{per campioni numerosi} \quad (8.1)$$

$$t_{\alpha/2} \leq \frac{\bar{x} - M}{\sigma/\sqrt{n-1}} \leq t_{1-\alpha/2} \quad \text{per piccoli campioni numerosi} \quad (8.2)$$

La  $\sigma$  della (8.1) può essere conosciuta a priori, oppure stimata tramite la varianza campionaria  $s^2$ , con la nota relazione:

$$\sigma^2 = \frac{n}{n-1} s^2 = \frac{1}{n-1} \sum v^2$$

Ancora se un controllo, sulla significatività della differenza di due medie campionarie, porta a respingere l'ipotesi  $H_0: M_x = M_y$ , può essere interessante stabilire un intervallo fiduciario, per la differenza fra le due medie:  $M_x - M_y$ . La distribuzione della variabile casuale:  $\bar{X} - \bar{Y}$ , è normale, se le stime  $\bar{x}$  e  $\bar{y}$  derivano da campioni numerosi, e quella  $t$  di *Student*, con  $\nu = n_x + n_y - 2$ , se i campioni sono piccoli.

Infine si possono trovare i limiti fiduciarî per percentuali o differenze di percentuali, noti i valori rilevati sui campioni e la numerosità degli stessi, dove la distribuzione utilizzata è quella normale, nella forma:

$$z = \frac{x - p}{\sqrt{p(1-p)/n}}$$

### Esempio 2.8.2

In un seggio con 250 elettori, si ha una percentuale del 18% , a favore di un certo candidato. In un secondo seggio di 300 elettori, scelto a caso, in un'altra zona, diversa come composizione sociale, si riscontra una percentuale di voti favorevoli pari al 10% . Volendo conoscere i limiti fiduciarci al 95% , per la differenza di percentuali di voti favorevoli, nella popolazione delle due zone, stabilito che la differenza riscontrata, fra i due campioni, è dell'8% , occorre trovare lo sqm di tale differenza:

$$\sigma_{P_1-P_2}^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} = \frac{0.18 \cdot 0.82}{250} + \frac{0.10 \cdot 0.90}{300} = (0.0298)^2$$

cosicché:

$$z_{0.025} \leq \frac{0.08 - (P_1 - P_2)}{0.0298} \leq z_{0.975}$$

e poi, essendo:  $-z_{0.025} = z_{0.975} = 1.96$  , si ha:

$$0.02 = 0.08 - 0.06 = 0.08 - 1.96 \cdot 0.0298 \leq P_1 - P_2 \leq 0.08 + 1.96 \cdot 0.0298 = 0.08 + 0.06 = 0.14$$

da cui si deduce che la differenza  $P_1 - P_2$  è compresa, con il 95% di probabilità, fra 0.02 e 0.14 . Di conseguenza, la notevole ampiezza dell'intervallo, permette solo una stima molto modesta di  $P_1 - P_2$  e d è dovuta ai valori delle numerosità campionarie che non permettono una valutazione accurata di  $P_1$  e di  $P_2$  .

## **2.9. Analisi di varianza**

### **2.9.1. Schema a casualizzazione completa**

Un particolare procedimento di analisi statistica trova una delle sue applicazioni là dove si vuole indagare sulle conseguenze di vari *iter* sperimentali, seguiti da materiali (od individui), inizialmente omogenei, fino alla determinazione del valore numerico caratteristico. Originariamente l'analisi di varianza è introdotta da Fisher, come strumento di indagine nella sperimentazione agricola. Gli individui sono i vari appezzamenti di terreno, fertilizzati con diversi tipi di concimi, e le rilevazioni numeriche, da assoggettarsi ad analisi di varianza, sono i raccolti per ettaro, ottenuti dai singoli campi.

In prima ipotesi, è necessario supporre l'uniforme fertilità naturale dei campi, l'uniforme qualità della semente adoperata, le uniformi condizioni di irrigazione, ecc. Qualora una (od alcune) di queste ipotesi vengano sensibilmente a mancare, lo schema dell'analisi di varianza si complica, passando da uno *schema a casualizzazione completa* ad uno *schema a blocchi casualizzati*, a *quadrati latini*, e su, su ad altri schemi più complessi, per tener conto del progressivo allontanarsi del materiale, sotto esperimento, dall'iniziale ipotesi di omogeneità. E' evidente che, anche al di fuori del campo della sperimentazione agricola, sono numerosi gli esperimenti nei quali pezzi di materiale, da ritenersi all'origine omogenei o differenziabili, per cause già

individuate, si diversificano via, via, a causa del *trattamento* che subiscono, prima di raggiungere lo stadio finale nel quale è misurato il valore dell'argomento *a* presente in ciascun pezzo.

Il caso più semplice è l'esperimento a *casualizzazione completa* che ipotizza materiale originariamente uniforme, attribuzione a caso a ciascun *pezzo* del trattamento modificante ed un certo numero di repliche di ogni trattamento.

L'insieme dei valori sperimentali può essere raccolto in una tabella nella quale sono anche indicate le medie e le varianze dei gruppi di valori  $a_{ij}$ , provenienti da tutti quegli elementi che subiscono trattamento  $t_j$ .

		Trattamenti				
<b>Replicazioni</b>		$a_{11}$	$a_{12}$	...	$a_{1m}$	
		$a_{21}$	$a_{22}$	...	$a_{2m}$	
		...	...	...	...	
		$a_{i1}$	$a_{i2}$	...	$a_{im}$	
		...	...	...	...	
		$a_{n1}$	$a_{n2}$	...	$a_{nm}$	
<b>Medie</b>		$\bar{a}_1$	$\bar{a}_2$	...	$\bar{a}_m$	$\bar{a}$
<b>Varianze</b>		$\sigma_1^2$	$\sigma_2^2$	...	$\sigma_m^2$	$\sigma^2$

Il valore numerico finale  $a_{ij}$ , presente in ogni pezzo, può essere scomposto in due componenti:

- ❑ una sistematica dovuta al trattamento  $t_j$ , subito dall'individuo;
- ❑ una casuale  $u_{ij}$ , dovuta all'insieme di un certo numero di cause modificanti, non identificabili, legate alle macchine, agli operatori, all'ambiente, ecc.

Nei limiti di validità del teorema centrale, si possono supporre normalmente distribuiti gli effetti perturbanti di tali cause, potendosi scrivere:

$$a_{ij} = t_j + u_{ij} \tag{9.1}$$

dove  $t_j$  è uguale per tutte le repliche, di uno stesso campione, ed  $u_{ij}$  sono normalmente distribuiti, con media nulla e varianza assegnata. L'analisi di varianza ha lo scopo di stabilire, se i valori medi  $\bar{a}_j$ , di ogni trattamento, sono significativamente uguali tra loro (il che porta a concludere che i diversi trattamenti non hanno efficacia differenziante). Dato che, in generale, il numero delle repliche è piuttosto modesto, anche per questo confronto tra medie di piccoli campioni, si richiedono le stesse ipotesi di 2.4.2, cioè la

distribuzione normale dei valori  $a_{ij}$  (giustificabile in base alla presunta normalità di  $u_{ij}$ ) e l'uguale varianza per gli universi dai quali i campioni sono estratti. Questo significa che le componenti accidentali  $u_{ij}$  devono avere tutte la stessa varianza (cioè che ogni  $\sigma_j^2$  deve essere uguale a tutte le altre). Pertanto si indica con  $\sigma^2$  l'unica varianza dei singoli universi normali, legati ad ogni trattamento, i quali possono differire nelle loro medie  $T_j$ , se i trattamenti sono significativi.

Riassumendo si pone, con un prefissato livello di significatività:

$$H_0: T_1 = T_2 = \dots = T_m$$

Tuttavia la decisione relativa all'accettazione (o meno) di  $H_0$  è qui impostata in modo del tutto diverso di quanto fatto nel paragrafo 1.2, in quanto si fa ora uso della stima di quell'unica  $\sigma^2$ , già ipotizzata, attraverso strade diverse, per mettere in risalto, oppure nascondere l'eventuale effetto, prodotto dai trattamenti. La singola media per colonna  $\bar{a}_j$ , analogamente alla (9.1), può scriversi:

$$\bar{a}_j = t_j + \bar{u}_j \quad (9.2)$$

ed evidentemente risente dell'effetto  $t_j$ , se questo esiste. Invece la stima di  $\sigma^2$ , valutata attraverso i dati del  $j$ -esimo universo, non risente delle eventuali conseguenze del trattamento:

$$\sigma^{2(j)} = \frac{\sum_{i=1}^n (a_{ij} - \bar{a}_j)^2}{n-1} = \frac{\sum_{i=1}^n ((t_j + u_{ij}) - (t_j + \bar{u}_j))^2}{n-1} = \frac{\sum_{i=1}^n (u_{ij} - \bar{u}_j)^2}{n-1}$$

Allora posto che, da ciascuna colonna, si può avere un'analogha stima di  $\sigma^2$ , si assume, utilizzando i dati di tutte le colonne

$$\sigma_R^2 = \frac{\sum_{j=1}^m \sum_{i=1}^n (a_{ij} - \bar{a}_j)^2}{m(n-1)} \quad (9.3)$$

il simbolo  $\sigma_R^2$  indica che, nella stima di  $\sigma^2$ , effettuata con la (9.3), influiscono solo i *residui*, ovvero le componenti accidentali, presenti in  $a_{ij}$ , e spariscono le componenti sistematiche dovute ai trattamenti.

Un'altra via per stimare  $\sigma^2$  si ha con la determinazione preliminare della varianza riscontrabile tra le medie  $\bar{a}_j$  di ogni trattamento. Infatti osservando la (9.2), in ogni  $\bar{a}_j$ , è integralmente presente la parte sistematica  $t_j$ , mentre gli effetti accidentali sono mediati ed il valore  $\bar{u}_j$  converge, in probabilità, a zero. Dalla varianza tra le medie di trattamento, si può poi risalire alla stima di  $\sigma^2$  della popolazione, dati  $n$  trattamenti:

$$\sigma_{\bar{a}_j}^2 = \frac{\sum_{j=1}^n (\bar{a}_j - \bar{a})^2}{m-1} \quad \Rightarrow \quad \sigma_T^2 = \frac{n \sum_{j=1}^n (\bar{a}_j - \bar{a})^2}{m-1} \quad (9.4)$$

La stima di  $\sigma^2$ , fatta attraverso la (9.4), esalta gli effetti dei trattamenti, se essi esistono, occultando invece la variabilità accidentale. Infine è altresì possibile stimare  $\sigma^2$ , con una formula:

$$\sigma_G^2 = \frac{\sum_{j=1}^n \sum_{i=1}^m (a_{ij} - \bar{a})^2}{nm-1} \quad (9.5)$$

che utilizza *globalmente*  $nm$  risultati dell'esperimento. e dove sono presenti, in modo inscindibile, sia la parte sistematica che quella accidentale della variabilità, non potendo così essere d'aiuto nella valutazione di quanto la prima parte prevalga sulla seconda.

La genesi delle tre diverse stime:  $\sigma_R^2$ ,  $\sigma_T^2$  e  $\sigma_G^2$ , della stessa  $\sigma^2$ , qui chiarita estensivamente solo per le implicazioni metodologiche contenute, è basata sull'applicazione del teorema di decomposizione ortogonale della varianza. Infatti con qualche banale cambiamento di indici, sussiste la relazione:

$$\sum_{j=1}^m \sum_{i=1}^n (a_{ij} - \bar{a})^2 = n \sum_{j=1}^m (\bar{a}_j - \bar{a})^2 + \sum_{j=1}^m \sum_{i=1}^n (a_{ij} - \bar{a}_j)^2$$

indicabile sommariamente come:

$$S_G^2 = S_T^2 + S_R^2 \quad (9.6)$$

dove  $S_G^2$  (generale) rappresenta la somma dei quadrati di tutti gli scarti fra gli  $a_{ij}$  e la media (generale)  $\bar{a}$  che può essere scomposta nella somma dei quadrati degli scarti fra le medie per trattamento ed  $\bar{a}$  ( $S_T^2$ ) ed in una parte residua  $S_R^2$ , comprendente la somma dei quadrati delle componenti accidentali. Da  $S_T^2$  e  $S_R^2$ , si possono ottenere le due diverse stime di  $\sigma^2$ , una ( $\sigma_T^2$ ) rispecchiante l'effetto dei trattamenti e l'altra ( $\sigma_R^2$ ) depurata da esso, dividendo rispettivamente per i loro gradi di libertà:  $m-1$  e  $m(n-1)$ . Una terza stima di  $\sigma^2$  si ottiene da  $S_G^2$ , dividendo per  $nm-1$ , numero di gradi di libertà della varianza generale  $\sigma_G^2$ . Tra i gradi di libertà delle tre stime sussiste una relazione di sommabilità come tra  $S_R^2$ ,  $S_T^2$  e  $S_G^2$ :

$$nm-1 = (m-1) + m(n-1) \quad (9.7)$$

che non vale invece tra le varianze corrispondenti.

Per il calcolo di  $S_G^2$  e  $S_T^2$  sono utili le espressioni, dedotte dalla consueta relazione:  $\sigma^2 = M_2 - M^2$ .

$$S_G^2 = \sum_{ij} (a_{ij} - \bar{a})^2 = \sum_{ij} a_{ij}^2 - nm \left( \frac{\sum_{ij} a_{ij}}{nm} \right)^2 = \sum_{ij} a_{ij}^2 - \frac{\left( \sum_{ij} a_{ij} \right)^2}{nm}$$

$$S_T^2 = n \sum_j (\bar{a}_j - \bar{a})^2 = n \left( \sum_j \bar{a}_j^2 - \frac{\left( \sum_j \bar{a}_j \right)^2}{m} \right) = n \left( \sum_j \left( \frac{\sum_i a_{ij}}{n} \right)^2 - \frac{\left( \sum_j \sum_i \frac{a_{ij}}{n} \right)^2}{m} \right) = \frac{\sum_j \left( \sum_i a_{ij} \right)^2}{n} - \frac{\left( \sum_{ij} a_{ij} \right)^2}{nm}$$

cosicché  $S_R^2$  è poi calcolato per differenza (notando che nel calcolo non intervengono medie parziali, né varianze parziali, ma solo le somme dei risultati sperimentali o le somme dei loro quadrati).

Introducendo l'ipotesi fondamentale  $H_0$  secondo cui gli effetti dei trattamenti non siano diversi fra loro, si può supporre che anche  $\sigma_T^2$  risenta solo della variabilità accidentale. Allora se  $\sigma_R^2$  corrisponde al vero il rapporto:

$$F = \sigma_T^2 / \sigma_R^2 \tag{9.8}$$

dovrebbe essere approssimativamente uguale ad 1.

Invece quanto più  $H_0$  non corrisponde alla realtà, cioè quanto più un effetto dei trattamenti induce una variabilità nettamente più sensibile di quella accidentale, tanto più si ottengono valori di  $F$  maggiori dell'unità. Pertanto in base al livello di significatività assegnato ed ai valori teorici, forniti dalle tavole, si può decidere, se respingere  $H_0$  (o meno).

Gli elementi calcolati per l'analisi di varianza possono essere riassunti nella seguente tabella.

Componenti	Somma dei quadrati degli scarti	Gradi di libertà	Varianze	Indice F
<b>Trattamenti</b>	$S_T^2$	$m - 1$	$\sigma_T^2$	$F = \sigma_T^2 / \sigma_R^2$
<b>Errori (residui)</b>	$S_R^2$	$m(n - 1)$	$\sigma_R^2$	
<b>Totale</b>	$S_G^2$	$nm - 1$	$\sigma_G^2$	

nella quale la  $S_G^2$  è calcolata, per comodità, per poterne dedurre  $S_R^2$ , per differenza, in quanto poi  $\sigma_G^2$  non è utilizzata in alcun modo.

Se è accertata la significatività dei trattamenti, si tratta tuttavia sempre di un responso globale, coinvolgente l'insieme dei trattamenti stessi, per cui bisogna procedere all'analisi delle singole medie  $\bar{a}_j$ , determinando per ciascuna i limiti fiduciarî, con un prefissato livello di fiducia. Il calcolo segue lo schema del paragrafo 2.8:

$$A_j = \bar{a}_j \pm t_{\alpha/2} \sigma / \sqrt{n} \tag{9.9}$$

Tra le tre stime disponibili di  $\sigma$  si utilizza  $\sigma_R$ , cioè quella che misura la variabilità accidentale usando tutti i dati raccolti nell'esperimento e non solo quelli del campione  $j$ -esimo, cosicché i gradi di libertà della variabile casuale  $t$  di Student sono, di conseguenza,  $m(n-1)$ .

### Esempio 2.9.1

Si ricercano gli effetti di 4 diversi regimi alimentari (trattamenti), in un gruppo di pulcini della stessa covata, avendo eseguito, per ciascun trattamento, attribuito a caso, 5 replicazioni. I dati della tabella rappresentano l'aumento percentuale di peso, riscontrato dopo l'applicazione dei trattamenti.

Replicazioni	Trattamenti				
	A	B	C	D	
1	55	61	42	169	
2	49	112	97	137	
3	42	30	81	169	
4	21	89	95	85	
5	52	63	92	154	
<b>Somme</b>	219	355	407	714	1695
<b>Medie</b>	43.8	71.0	81.4	142.8	84.75

Ai fini dell'analisi della varianza, i valori stimati possono essere così riassunti:

$$S_G^2 = 55^2 + 49^2 + \dots + 154^2 - 1695^2/20 = 37793.75$$

$$S_T^2 = (219^2 + 355^2 + 407^2 + 714^2)/5 - 1695^2/20 = 26234.95$$

$$S_R^2 = S_G^2 - S_T^2 = 37793.75 - 26234.95 = 11558.80$$

Componenti	Somma dei quadrati degli scarti	Gradi di libertà	Varianze	Indice F
Trattamenti	26234.95	3	8744.98	12.1
Errori (residui)	11558.80	16	722.42	
Totale	37793.75	19		

Per 3 e 16 gradi di libertà, la tavola della distribuzione di Fisher-Snedecor dà il valore critico:  $F_{0.01} = 5.29$ .

Di conseguenza, il valore osservato  $F = 12.1$  cade nella regione critica e la significatività del test  $F$  denota che l'ipotesi  $H_0$ , nel caso in esame, corrispondente ad ammettere nessuna differenza esistente tra i diversi regimi alimentari, deve essere respinta.

### 2.9.2. Schema a blocchi casualizzati

La significatività dei trattamenti deriva dal confronto fra  $\sigma_T^2$  e  $\sigma_R^2$ ; se quest'ultima è molto elevata, può arrivare a mascherare l'effetto dei trattamenti, e questo può accadere, se sono erroneamente considerate cause accidentali altre cause di variabilità che, ad una più attenta analisi, si rivelerebbero invece di natura sistematica. Si tratta cioè di valutare se  $S_R^2$  della (9.6) non possa essere ulteriormente scomposto, dando luogo a un  $S_R^2$ , nuovo e più piccolo, in una relazione del tipo:

$$S_G^2 = S_T^2 + S_B^2 + S_R^2 \quad (9.10)$$

dove  $S_B^2$  rappresenta la porzione di  $S_G^2$ , ad esempio, dovuta a differenze nel materiale originario sottoposto a sperimentazione. Si passa così dallo schema a casualizzazione completa a quello a blocchi casualizzati, essendo i blocchi composti, nel loro interno, da materiale omogeneo, mentre differiscono dall'uno all'altro per qualche causa, identificabile o supposta. Supponendo  $n$  blocchi, ciascuno costituito di  $m$  elementi, lo schema riassuntivo dei risultati sperimentali è identico a quello precedente, con la differenza che ora le varie righe corrispondono ai diversi blocchi (per cui i valori non possono più essere spostati, nell'ambito della stessa colonna, come quando ogni campione rappresentava pure replicazioni). Rimane ancora affidato al caso l'accoppiamento fra gli elementi di ogni blocco ed i trattamenti da sperimentare.

Alla scomposizione (9.10), corrisponde un'analogia scomposizione fra i gradi di libertà:

$$nm - 1 = (m - 1) + (n - 1) + (m - 1)(n - 1) \quad (9.11)$$

e dividendo ciascun  $S^2$  per i suoi gradi di libertà, si hanno quattro varianze:  $\sigma_R^2$ ,  $\sigma_T^2$ ,  $\sigma_B^2$  e  $\sigma_G^2$ , e si può calcolare due distinti  $F$  sperimentali:

$$F_T = \sigma_T^2 / \sigma_R^2 \quad \text{e} \quad F_B = \sigma_B^2 / \sigma_R^2$$

Per  $\sigma_B^2$  si possono fare ragionamenti analoghi a quelli del disegno a casualizzazione completa, essendo strutturata per evidenziare l'apporto, nella variabilità dei risultati, dell'ipotizzata suddivisione in blocchi. Può anche accadere che la suddivisione eseguita non sia significativa e valga l'ipotesi fondamentale aggiuntiva:

$$H_0^*: B_1 = B_2 = \dots = B_m$$

La decisione sulla significatività dei blocchi spetta al valore del rapporto  $F_B$ . Infatti nel disegno a blocchi casualizzati è possibile valutare sia la significatività dei trattamenti che quella dei blocchi, indipendentemente una dall'altra. E' chiaro che, proseguendo con questo ragionamento, ovvero scomponendo  $S_G^2$  (ed i suoi gradi di libertà), si può via, via individuare e valutare varie cause apportatrici di variabilità, fino a quando  $S_R^2$  rappresenti soltanto la porzione della variabilità totale, effettivamente dovuta al caso. Questo dà luogo a

schemi di sperimentazione un po' più complessi dal punto di vista organizzativo, ma molto logici per l'analisi delle cause sistematiche, concorrenti a modificare gli oggetti della sperimentazione.

### Esempio 2.9.2

L'esperimento riguarda 4 varietà *A*, *B*, *C* e *D* di grano (trattamenti), seminate in 5 appezzamenti (blocchi), ritenuti diversi tra loro come composizione chimica del terreno. I valori corrispondenti alle varie produzioni possono essere ordinati, ai fini dell'analisi della varianza, nella tabella sottostante e, sulla base dei suoi dati, si possono calcolare le somme dei quadrati degli scarti già precedentemente specificate:

Blocchi	Trattamenti				Somme	Medie
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>		
1	32.3	33.3	30.8	29.3	125.7	31.4
2	34.0	33.0	34.3	26.0	127.3	31.8
3	34.3	36.3	35.3	29.8	135.7	33.9
4	35.0	36.8	32.3	28.0	132.1	33.0
5	36.5	34.5	35.8	28.8	135.6	33.9
<b>Somme</b>	172.1	173.9	168.5	141.9	656.4	
<b>Medie</b>	34.4	34.8	33.7	28.4		32.8

Ai fini dell'analisi della varianza, i valori stimati possono essere così riassunti:

$$S_G^2 = 32.3^2 + 34.0^2 + \dots + 28.8^2 - 656.4^2/20 = 182.17$$

$$S_T^2 = (172.1^2 + 173.9^2 + 168.5^2 + 141.9^2)/5 - 656.4^2/20 = 134.45$$

$$S_B^2 = (125.7^2 + 127.3^2 + 135.7^2 + 132.1^2 + 135.6^2)/4 - 656.4^2/20 = 21.46$$

$$S_R^2 = S_G^2 - S_T^2 - S_B^2 = 26.26$$

Componenti	Somma dei quadrati degli scarti	Gradi di libertà	Varianze	Indice F
<b>Trattamenti</b>	134.45	3	44.82	20.47
<b>Blocchi</b>	21.46	4	5.37	2.45
<b>Errori (residui)</b>	26.26	12	2.19	
<b>Totale</b>	182.17	19	9.59	

Il valore limite, con 3 e 12 gradi di libertà, è  $F_{0.01} = 5.95$ , cosicché il valore sperimentale  $F = 20.47$  si trova nella regione critica, evidenziando la notevole significatività dei trattamenti. Invece il valore limite, con 4 e 12 gradi di libertà, è  $F_{0.05} = 3.26$ , cosicché la suddivisione in blocchi, con un valore sperimentale  $F = 2.45$ , non causa una differenziazione sensibile tra gli elementi, dove si applicano i trattamenti.

## PARTE III – IL PROBLEMA DELLA STIMA

### 3.1 Proprietà degli stimatori

Il problema della stima dei parametri di una popolazione si basa sull'informazione ottenibile da un campione estratto dalla stessa. Inizialmente questo problema è deliberatamente accantonato, limitandosi a ritenere, su basi un po' intuitive, ad esempio, che la media di un campione sia una soddisfacente stima della media della popolazione. Invece bisogna ora esaminare quali condizioni deve soddisfare una *buona stima* e se esiste la *migliore* stima nel senso corrente del termine. Ovviamente il problema si pone solo quando il campione è casuale, perché nulla può inferirsi, sulla popolazione originale, se nella formazione del campione interviene qualche distorsione del meccanismo dell'estrazione a caso. Ad esempio, tutto quanto è detto sul problema della stima può riferirsi alle misure ripetute di una stessa grandezza, solo se gli errori da cui sono affette sono di natura casuale, o *accidentale*, termine che è generalmente contrapposto a *sistematico* (laddove l'analisi della formazione degli errori accidentali o sistematici esula dalla presente trattazione). Innanzitutto è utile distinguere fra *stima* e *stimatore*, intendendosi con la prima il valore che si calcola di una certa statistica campionaria, assunto a rappresentare la corrispondente statistica dell'universo, e con il secondo quella particolare funzione dei valori campionari  $t$  che dà luogo alla stima. In questo contesto, il problema non è calcolare stime, ma piuttosto investigare le caratteristiche dei metodi di stima, ovvero degli stimatori.

### 3.2 Consistenza

Dato il comportamento della media e varianza della distribuzione delle medie campionarie, in generale, facendo ricorso all'uso delle funzioni caratteristiche, si può dimostrare che, se l'universo da cui sono estratti i campioni, ha la seguente distribuzione normale  $N(\theta,1)$ .

$$dF(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} dx$$

e se come stimatore di  $\theta$  si assume  $t = \bar{x} = \sum x/n$ , la distribuzione della variabile casuale delle medie di campioni di  $n$  elementi è data da:

$$dF(\bar{x}) = \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-\frac{1}{2}(\bar{x}-\theta)^2} d\bar{x} \quad (2.1)$$

il che sta ad indicare che  $\bar{x}$  è distribuita normalmente intorno a  $\theta$  con varianza  $1/n$ . Due cose essenziali sono da rilevare nella distribuzione (2.1):

- la sua media è uguale a  $\theta$  (come pure la sua mediana e la sua moda);
- all'aumentare di  $n$ , la dispersione dei possibili valori  $\bar{x}$  intorno a  $\theta$  diventa sempre più piccola, ovvero che *l'attendibilità di  $\bar{x}$  aumenta con  $n$* .

Quest'ultima proprietà è comune a molti stimatori, ma non a tutti, e dove essa valga, cioè quando lo stimatore converge in probabilità alla statistica della popolazione  $\theta$ , lo stimatore stesso è detto *consistente*.

### 3.3 Assenza di deviazioni (*Unbiased estimators*)

La consistenza è solo una proprietà asintotica, essendo relativa al comportamento dello stimatore per  $n$  tendente all'infinito, e non pone requisiti allo stimatore stesso per  $n$  finito. Inoltre se esiste uno stimatore consistente  $t_n$ , se ne possono costruire infiniti altri (ad esempio, anche:  $(n-a)/(n-b) \cdot t_n$ , con  $a$  e  $b$  costanti arbitrarie, è uno stimatore consistente). Allora è necessario stabilire un criterio adottare per scegliere uno stimatore piuttosto che un altro e, a tal fine, uno stimatore è definito *corretto* o *non deviato*, se il suo valor medio coincide con  $\theta$ , qualunque sia  $n$ . La scelta della relazione:  $M(t) = \theta$ , come caratteristica peculiare di uno stimatore non deviato, è arbitraria, nel senso che solo motivi di facilità di calcolo fanno preferire la media ad altri valori centrali.

A riguardo, è noto che la varianza campionaria sia uno stimatore deviato della varianza dell'universo:

$$M\left(\sum (x_i - \bar{x})^2 / n\right) = (n-1)/n \cdot \sigma^2$$

e pertanto è sostituita dallo stimatore non deviato:

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

Questo esempio mostra come la consistenza non implichi necessariamente l'assenza di deviazioni e si può dimostrare che vale anche il viceversa. Inoltre in generale, esiste più di uno stimatore consistente tra gli stimatori non devianti.

Ad esempio, la mediana campionaria è uno stimatore non deviato della mediana (e della media) della popolazione, per evidenti ragioni di simmetria, ed è consistente, in quanto la sua varianza è uguale a:  $\pi\sigma^2/2n$  (per  $n$  elevato), e tende a zero per  $n$  tendente a  $\infty$ . Dopodiché fra due stimatori, entrambi consistenti e non devianti, è logico scegliere quello con varianza minore che, in generale, è distribuito in un intorno più ristretto di  $\theta$ . Nel caso della media e della mediana campionarie si ha:

$$\sigma_{\bar{x}}^2 = \sigma^2 / n \quad \text{e} \quad \sigma_{\text{mediana}}^2 = \pi\sigma^2 / (2n)$$

Dato che  $\pi/2 \cong 1.57$ , la media è meno dispersa della mediana, rispetto a  $\theta$ , ed è da preferirsi, come stimatore del valore centrale della distribuzione.

### 3.4 Minima varianza

La valutazione della varianza di uno stimatore, come criterio per la sua accettabilità, risale a tempi lontani. Tuttavia poi è dimostrato che la varianza di uno stimatore è inferiormente limitata. Uno stimatore la cui varianza raggiunge l'estremo inferiore è detto limite di minima varianza (Lim. Min. Var.).

Per stabilire la relazione, soddisfatta dalla densità di probabilità del campione e dalla funzione  $\tau(\theta)$  da stimare, è necessario premettere la definizione della funzione di verosimiglianza  $L$  di un campione di  $n$  elementi indipendenti, ognuno dei quali ha densità di probabilità  $f(x/\theta)$ :

$$L(x_1, x_2, \dots, x_n / \theta) = f(x_1 / \theta) \cdot f(x_2 / \theta) \cdot \dots \cdot f(x_n / \theta) \quad (4.1)$$

Essendo  $L$  la densità di probabilità di una variabile casuale a  $n$  dimensioni di cui  $(x_1, x_2, \dots, x_n)$  è una estrazione a caso, si ha:  $\int \dots \int L dx_1 dx_2 \dots dx_n = 1$ . A riguardo, si può dimostrare che, volendo stimare una funzione  $\tau(\theta)$ , la varianza di un suo stimatore  $t$ :

$$\theta_i^2 = M(t - \tau(\theta))^2$$

sul quale non è fatta alcuna ipotesi, soddisfa la relazione:

$$\theta_i^2 \geq (\tau'(\theta))^2 / M\left(\frac{\partial \ln L}{\partial \theta}\right)^2 \quad (4.2)$$

dove  $\tau'(\theta)$  è la derivata della funzione  $\tau(\theta)$ . Con qualche semplice passaggio, si dimostra che si raggiunge il Lim. Min. Var., ovvero vale il segno di uguale nella (4.2), se e solo se:

$$\frac{\partial \ln L}{\partial \theta} = A(\theta)(t - \tau(\theta)) \quad (4.3)$$

cioè se  $\partial \ln L / \partial \theta$  è esprimibile come una costante (eventualmente funzione di  $\theta$ ), moltiplicata per lo scarto fra lo stimatore e la funzione da stimare. In questo caso:

$$\sigma_i^2 = (\tau')^2 / M(A^2(t - \tau(\theta))^2) = (\tau')^2 / A^2 M(t - \tau(\theta))^2 = \tau' / (A^2 \sigma_i^2)$$

da cui

$$\sigma_i^2 = \tau'(\theta) / A(\theta) \quad (4.4)$$

Se vale la (4.3), si può valutare, contemporaneamente allo stimatore Lim. Min. Var. di  $\tau(\theta)$ , anche la sua varianza. Questo accade per una vasta classe di distribuzioni la cui densità di probabilità è esprimibile nella forma generale che comprende la maggioranza delle variabili casuali di uso corrente:

$$f(x / \theta) = e^{K(\theta)B(x)+C(x)+D(\theta)} \quad (4.5)$$

L'equazione (4.3) stabilisce la condizione cui deve soddisfare la funzione di verosimiglianza (e la densità di probabilità), perché esista uno stimatore Lim. Min. Var. di una certa funzione di  $\theta$ :  $\tau(\theta)$ . Se questa non è soddisfatta, può ancora esistere uno stimatore di  $\tau(\theta)$  che, pur senza raggiungere il Lim. Min. Var., abbia varianza minore di qualunque altro stimatore. In questo caso, è detto stimatore di minima varianza (Min. Var.) e, con procedimento piuttosto elaborato, si può dimostrare che, se esiste, è anche unico.

#### Esempio 3.4.1

Si vuole stimare la media  $\lambda$  di una distribuzione Poissoniana, sulla base di un campione  $x_1, x_2, \dots, x_n$ .

Ponendo  $\lambda = \theta$  la funzione di verosimiglianza ed il suo logaritmo sono:

$$L = \frac{\theta^{x_1}}{x_1!} e^{-\theta} \frac{\theta^{x_2}}{x_2!} e^{-\theta} \dots \frac{\theta^{x_n}}{x_n!} e^{-\theta} = \frac{\theta^{\sum x_i}}{\prod x_i!} e^{-n\theta} \quad \ln L = -\ln \prod x_i! + \sum x_i \ln \theta - n\theta$$

da cui

$$\frac{\partial \ln L}{\partial \theta} = \frac{\sum x_i}{\theta} - n = \frac{n}{\theta} \left( \frac{\sum x_i}{n} - \theta \right) \quad (4.6)$$

La (4.6) risulta della forma:  $A(\theta)(t - \tau(\theta))$ , dove:  $\tau(\theta) = \theta$  e  $t = \sum x_i/n$ . Di conseguenza, la media campionaria  $\bar{x} = \sum x_i/n$  è uno stimatore Lim. Min. Var. della media della popolazione  $\lambda = \theta$  ed inoltre:

$$\sigma_t^2 = \sigma_{\bar{x}}^2 = \frac{1}{n/\theta} = \frac{\lambda}{n}$$

per la (4.4), cosicché risulta, in accordo con quanto già dimostrato, che nella distribuzione Poissoniana:

- la varianza della variabile casuale è uguale alla sua media;
- fra la varianza della variabile casuale delle medie campionarie di  $n$  elementi e quella della variabile casuale da cui si estrae il campione vale la relazione:  $\sigma^2(\bar{x}) = \sigma^2/n$ , già precedentemente trovata.

Infine si può notare che la densità di probabilità della distribuzione Poissoniana è del tipo (4.5):

$$f(x/\theta) = \frac{\theta^x}{x!} e^{-\theta} = e^{x \lg \theta - \lg x! - \theta}$$

per il quale esiste uno stimatore Lim. Min. Var. di  $\tau(\theta)$  (in questo caso  $\tau(\theta) = \theta$ ).

### 3.5 Efficienza

La trattazione sugli stimatori di minima varianza non comporta alcun vincolo sulla numerosità  $n$  del campione. Tuttavia anche se non esistono stimatori Min. Var. per piccoli campioni, quasi sempre ne esiste uno se  $n$  è elevato. Infatti la maggioranza degli stimatori di uso corrente ha, in base al Teorema Centrale, distribuzione asintoticamente normale, dipendente solo dai due parametri  $M$  e  $\sigma$ . Allora dato che, se lo stimatore è consistente, di solito è asintoticamente non deviato, la sua varianza può essere usata come criterio di scelta fra stimatori equivalenti, sotto gli altri punti di vista esaminati precedentemente. Più in generale, si dicono *efficienti* quegli stimatori consistenti, asintoticamente normali, i quali, per  $n$  elevato, hanno Min. Var. Di un qualunque altro stimatore si potrà *misurare* l'efficienza  $E$  nei confronti dello stimatore efficiente, come rapporto inverso delle relative varianze. Pertanto riprendendo l'esempio del paragrafo 3.3, il confronto delle varianze della media e mediana campionarie, entrambe consistenti ed asintoticamente normali, porta al calcolo dell'efficienza della mediana:

$$E_{\text{mediana}} = 2/\pi = 0.637$$

inferiore di circa un terzo a quella della media che è lo stimatore efficiente.

### 3.6 Sufficienza

La trattazione dei criteri di stima di un parametro  $\theta$  si può ulteriormente approfondire con l'introduzione del concetto di *sufficienza*, una delle proprietà più pregevoli di una statistica. Considerando il caso in cui si deve stimare un solo parametro, in base a  $n \geq 2$  elementi di un campione, con questi elementi si può costruire un numero illimitato di possibili stimatori di  $\theta$ , tra i quali si deve poi scegliere. Siano essi  $t_1, t_2, \dots$ , essendo  $t$  quello da analizzare. Si dice che  $t$  è una statistica sufficiente di  $\theta$ , in base agli elementi  $(x_1, x_2, \dots, x_n)$  di un campione, se la probabilità composta di questi valori, condizionata da  $t$  (ovvero la probabilità che, dato  $t$ , si presentino insieme  $x_1, x_2, \dots, x_n$ ), non dipende da  $\theta$ . In questo caso, la probabilità composta del contemporaneo verificarsi di  $\theta$  e  $x_1, x_2, \dots, x_n$  può decomporre nel prodotto di due fattori di cui uno dipende solo da  $t$  e  $\theta$ , mentre l'altro solo da  $t$  e  $x_1, x_2, \dots, x_n$ . Una delle formulazioni più semplici di questo concetto è espressa da una condizione sulla funzione di verosimiglianza che deve essere così esprimibile:

$$L(x_1, x_2, \dots, x_n / \theta) = g(t / \theta) k(t, x_1, x_2, \dots, x_n) \quad (6.1)$$

dove  $g$  è funzione solo di  $t$  e  $\theta$ , mentre  $k$  è indipendente da  $\theta$ . In questo caso, tutta l'informazione che si può trarre dal campione, circa  $\theta$ , è data dalla conoscenza della statistica  $t$  e della sua distribuzione, mentre nessun'altra statistica  $t_i = t_i(x_1, x_2, \dots, x_n)$  può aggiungere ulteriori informazioni su  $\theta$ .

#### Esempio 3.6.1

La stima del valor medio  $\theta$  di una variabile casuale normale, fornita dalla media campionaria  $\bar{x}$ , è sufficiente: Infatti:

$$L(x_1, \dots, x_n / \theta) = \frac{1}{\sqrt{2\pi}^n \sigma^n} e^{-\frac{1}{2\sigma^2} \sum (x_i - \theta)^2} \quad \text{con} \quad \sum (x_i - \theta)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2$$

da cui:

$$L = e^{-\frac{n}{2\sigma^2} (\bar{x} - \theta)^2} \cdot \frac{1}{\sqrt{2\pi}^n \sigma^n} e^{-\frac{1}{2\sigma^2} \sum (x_i - \bar{x})^2}$$

Allora la funzione di verosimiglianza  $L$  risulta spezzata in due fattori di cui uno è proporzionale alla densità di probabilità di  $\bar{x}$ , dato  $\theta$ , e l'altro a quella composta di  $(x_1, x_2, \dots, x_n)$ , dato  $\bar{x}$ . Pertanto  $\bar{x}$  è una stima sufficiente. Inoltre se vale la (6.1), si ha anche:

$$\frac{\partial \ln L}{\partial \theta} = \frac{\partial \ln g(t/\theta)}{\partial \theta} \quad (6.2)$$

la quale mostra come la sufficienza sia una condizione meno restrittiva della condizione per l'esistenza di

uno stimatore Lim. Min. Var. di  $\tau(\theta)$ . Infatti la (4.3) è un caso particolare della (6.2) che è sicuramente verificata, mentre può ugualmente esistere una statistica sufficiente, anche se non è valida la (4.3) (ovvero se esiste uno stimatore Lim. Min. Var., è anche una statistica sufficiente).

Più in generale, è dimostrato che lo stimatore non deviato di Min. Var. di  $\tau(\theta)$  è sempre una funzione di una statistica sufficiente. Inoltre si può dimostrare che, per tutte quelle variabili casuali la cui densità di probabilità è esprimibile sotto la forma (4.5):

$$f(x/\theta) = e^{K(\theta)B(x)+C(x)+D(\theta)} \quad (6.3)$$

vale una condizione per l'esistenza di uno stimatore Lim. Min. Var., per qualche funzione  $\tau(\theta)$ , il cui intervallo di definizione sia indipendente da  $\theta$ , in base alla quale il metodo di stima, detto di massima verosimiglianza (di cui al paragrafo 3.7), fornisce una statistica sufficiente per  $\theta$ . Infatti sotto condizioni di regolarità molto ampie, si ha una corrispondenza biunivoca fra l'esistenza di una statistica sufficiente di  $\theta$  e l'esistenza di uno stimatore Lim. Min. Var. per alcune funzioni  $\theta$ .

Allora se vale la (6.3), esiste una statistica sufficiente per  $\theta$  ed esiste una sola funzione  $t$  di questa statistica che soddisfa la (4.3) e pertanto è lo stimatore Lim. Min. Var. di qualche funzione  $\tau(\theta)$ . Inoltre nei campioni numerosi, qualunque funzione della statistica sufficiente è uno stimatore Lim. Min. Var. del corrispondente valore dell'universo, mentre per  $n$  arbitrario, qualunque funzione della statistica sufficiente stima il rispettivo valore teorico con la minima varianza raggiungibile.

### 3.7 Criteri di stima: massima verosimiglianza

Il principio di Massima Verosimiglianza (dall'inglese *Maximum Likelihood*) è molto frequentemente applicato, come metodo di stima, anche se spesso sono omesse le dimostrazioni delle sue caratteristiche, sotto i punti di vista precedentemente esposti. Secondo questo principio, la stima  $\hat{\theta}$  del parametro  $\theta$  è quel valore che rende massima la funzione di verosimiglianza (4.1):  $L(x_1, x_2, \dots, x_n / \hat{\theta}) \geq L(x_1, x_2, \dots, x_n / \theta)$ . Se poi questa funzione ammette derivate prima e seconda, in tutto il suo campo di definizione, la stima di  $\theta$  è data dalla maggiore fra le radici dell'equazione:

$$\frac{\partial L(x/\theta)}{\partial \theta} = 0 \quad (7.1)$$

con la condizione  $L''(x/\hat{\theta}) < 0$ . Nella pratica, per facilitare i calcoli, nell'equazione (7.1) si sostituisce  $\ln L$  a  $L$ , dato che, essendo  $L > 0$ , i massimi di  $L$  coincidono con quelli di  $\ln L$ , cercando così le soluzioni dell'equazione:

$$\frac{\partial \ln L(x/\theta)}{\partial \theta} = 0$$

per le quali:  $(\ln L)' < 0$  e, dove ne esista più di una, si assume la maggiore, come stima di  $\theta$ .

Il principio di massima verosimiglianza è evidentemente arbitrario, perché non si presenta sempre, all'atto di una prova, l'evento con la massima probabilità. Tuttavia la sua accettazione è giustificata proprio per le caratteristiche delle stime che si ottengono per mezzo di esso. Innanzitutto si dimostra che, se esiste una statistica sufficiente di  $\theta$ , il suo stimatore di massima verosimiglianza deve essere una funzione di questa. Infatti l'esistenza di una statistica sufficiente implica la fattorizzazione della funzione di verosimiglianza in due termini di cui il secondo indipendente da  $\theta$ :  $L(x/\theta) = g(t/\theta)h(x,t)$ , per cui la ricerca di  $\theta$  che renda massima  $L(x/\theta)$  equivale alla ricerca di  $\theta$  che massimizzi  $g(t/\theta)$  e che è funzione solo di  $t$ .

Inoltre il paragrafo 6.6 mostra che, in una vasta classe di casi, se esiste una statistica sufficiente, è possibile trovare uno stimatore Lim. Min. Var.  $t$  per  $\tau(\theta)$ . Quest'ultimo, se esiste la soluzione  $\hat{\theta}$  dell'equazione di massima verosimiglianza, è  $t = \tau(\hat{\theta})$ , in quanto, dove esiste uno stimatore Lim. Min. Var., esso è dato dal metodo di massima verosimiglianza.

Ad esempio, la media:  $t = \bar{x} = \sum x_i/n$ , stimata nell'Esempio 3.4.1, è una stima di massima verosimiglianza, dato che la funzione di verosimiglianza, con qualche artificio, può così porsi:

$$L = \frac{\theta^{\sum x_i}}{\prod x_i!} e^{-n\theta} = \frac{e^{-n\theta} \cdot (n\theta)^{n\bar{x}}}{(n\bar{x})!} \frac{(n\bar{x})!}{n^{n\bar{x}} \cdot \prod (x_i)!} = g(\bar{x}/\theta)h(\bar{x}, x_1, x_2, \dots, x_n)$$

Inoltre questa stima della media è una statistica sufficiente, cosicché è inutile cercarne una migliore, in quanto è anche stimatore Lim. Min. Var.

Le proprietà ottimali delle statistiche sufficienti sono trasferite agli stimatori di massima verosimiglianza i quali, se possibile, sono stimatori Lim. Min. Var. e, nel caso più generale, sono gli stimatori cui compete la minima varianza raggiungibile. Inoltre gli stimatori di massima verosimiglianza, con condizioni molto poco restrittive sulle densità di probabilità, sono consistenti, efficienti e asintoticamente normali. Tuttavia va precisato che, al di fuori del campo delle statistiche sufficienti, le proprietà ottimali degli stimatori di massima verosimiglianza sono solo asintotiche.

Per contro, gli stimatori di massima verosimiglianza possono essere devianti. Infatti in generale:  $M(\tau(\hat{\theta})) \neq \tau(M(\hat{\theta}))$ , in quanto, anche se  $\hat{\theta}$  è uno stimatore non deviato di  $\theta$ , non sempre  $\tau(\hat{\theta})$  è uno stimatore non deviato di  $\tau(\theta)$ , cosicché bisogna apportare allo stimatore stesso, in questo caso, una correzione che ne annulli il *bias*.

### 3.8 Criteri di stima: minimi quadrati

Lo stimatore di massima verosimiglianza della media  $M = \hat{\theta}$  di una distribuzione normale, in base ad un campione di  $n$  elementi  $x_1, x_2, \dots, x_n$ , è ottenuto cercando il massimo della funzione di verosimiglianza:

$$\ln L(x/\theta) = -\frac{1}{2} n \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \theta)^2 \quad \Rightarrow \quad \sum_{j=1}^n (x_j - \theta)^2 = \min$$

In questo caso, il principio di massima verosimiglianza equivale al, più noto e più antico, principio dei *minimi quadrati*. In generale, supponendo che la media della popolazione sia funzione lineare di alcuni parametri:

$$M = \sum_{j=1}^k a_j \theta_j$$

dove  $a_j$  sono costanti note. La stima di  $\theta_j$  si ha imponendo:

$$\sum_{j=1}^n \left( x_j - \sum_{i=1}^k a_{ji} \theta_i \right)^2 = \min$$

Se poi le  $n$  osservazioni non provengono dalla stessa popolazione normale, ma da  $n$  popolazioni normali con diversa media  $M_j$ , essendo sempre funzioni di un certo numero di parametri  $\theta$ :

$$M_j = \sum_{i=1}^k a_{ji} \theta_i \quad j = 1, 2, \dots, n \quad (8.1)$$

si ottengono i parametri  $\theta_i$  imponendo:

$$\sum_{j=1}^n \left( x_j - \sum_{i=1}^k a_{ji} \theta_i \right)^2 = \min$$

Come per ogni altro principio di stima, l'adozione del principio dei minimi quadrati dipende dalla proprietà degli stimatori ottenuti. Tuttavia a differenza del metodo di massima verosimiglianza, quello dei minimi quadrati non possiede, in generale, proprietà ottimali, neppure asintotiche. Per contro, in un'importantissima classe di applicazioni (costituita dai cosiddetti *modelli lineari* di cui la (8.1) è un esempio), anche per piccoli campioni, dà stimatori non devianti, lineari nei valori osservati cui compete la minima varianza raggiungibile. Questo accade quando le osservazioni sono funzioni lineari di parametri incogniti e, in questo caso, le proprietà ottimali del metodo non richiedono l'ipotesi di normalità delle osservazioni. Tuttavia va notato che quest'ipotesi ridiventa necessaria, qualora si vogliano sottoporre a test di significatività i parametri stimati.

## PARTE IV – ELABORAZIONE DEI DATI DI OSSERVAZIONE

### 4.1 Errori accidentali e sistematici

Il principio di massima verosimiglianza (mediante la massimizzazione di una probabilità composta tra tutte le osservazioni, in funzione delle stime attese) fornisce il valore da assumere, come misura di una grandezza, dove le osservazioni eseguite della stessa sono in numero esuberante. Per buona parte, si tratta di cose già note che conviene tuttavia ricomporre, in un tutto organico riferito a quell'operazione di campionamento, data dai risultati di misure ripetute, invece che a generiche operazioni di campionamento.

A questo proposito, occorre sottolineare che le misure ripetute di una stessa grandezza possono essere affette da errori sia accidentali che sistematici, ma che il trattamento statistico delle misure stesse è possibile solo se gli errori sono di tipo accidentale, cioè distribuiti in modo completamente casuale intorno a valori medi nulli. La modellazione degli errori sistematici è molto più ardua: in generale, hanno la caratteristica di mantenersi invariati od almeno di segno costante, nella ripetizione delle misure. Si usa dire che essi, in quanto dovuti a cause ben determinate e individuali, possono essere eliminati con particolari accorgimenti. In

realtà, questo è vero solo in parte e, in particolare, non per misure di precisione molto elevata.

In alcuni casi, si possono eliminare gli effetti di piccoli errori sistematici strumentali, eseguendo le misure in condizioni di simmetria, cosicché la loro influenza è annullata. Tuttavia quest'ultima può essere molto ridotta nelle cosiddette misure relative, cioè quelle in cui interessa solo la differenza di due grandezze in luoghi o tempi diversi; ponendo così la massima cura nell'effettuare le misure in condizioni quanto possibile identiche, affinché tutti gli errori sistematici abbiano la stessa influenza ed i risultati ne siano esenti. In ogni caso, la riduzione degli errori sistematici comporta un'attenta analisi delle modalità strumentali ed ambientali con le quali le misure sono eseguite.

Invece le osservazioni affette da errori puramente accidentali sono trattate con i consueti procedimenti della statistica, atti a dedurre, dall'insieme delle osservazioni stesse, alcune stime delle grandezze da misurarsi. In generale, è accettata l'ipotesi che le misure ripetute di una stessa grandezza affette solo da errori accidentali abbiano distribuzione normale. Tuttavia negli ultimi tempi sono comparsi parecchi studi che ipotizzano altre distribuzioni, simili a quella normale, ma più rispondenti al comportamento delle osservazioni ripetute, come constatato molto spesso dagli sperimentatori. Per contro, tutti i metodi di stima sottoesposti non mutano che formalmente qualora si applichino questi diversi tipi di distribuzione di errori.

#### 4.2 Osservazioni dirette di uguale precisione

Dato un campione  $x_1, x_2, \dots, x_n$  di misure ripetute della stessa grandezza, eseguite con modalità strumentali, ambientali e personali identiche, si pone il problema della determinazione dei due parametri  $M$  e  $\sigma$ , per definirne completamente la distribuzione normale.

Supponendo noto  $\sigma$  si può determinare  $M = \theta$  applicando il metodo di massima verosimiglianza. A riguardo, la funzione di verosimiglianza ed il suo logaritmo sono rispettivamente:

$$L(x/\theta, \sigma) = \frac{1}{\sqrt{(2\pi)^n \sigma^n}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2} \quad \ln L = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2} \quad (2,1) \text{ e } (2,2)$$

da cui:

$$\frac{\partial \ln L}{\partial \theta} = \frac{\sum_{i=1}^n (x_i - \theta)}{\sigma^2} = \frac{n}{\sigma^2} \left( \sum_{i=1}^n x_i / n - \theta \right) \quad (2,3)$$

Il secondo membro della (2.3) risulta della forma  $A(t - \tau(\theta))$ , dove  $\tau(\theta) = \theta$  e  $t = \sum_{i=1}^n x_i / n$ , cosicché la

media campionaria:  $t = \bar{x} = \sum_{i=1}^n x_i / n$ , ottenuta con il metodo di Massima Verosimiglianza, è uno stimatore

Limite Minima Varianza della media della popolazione  $\theta$ ; la cui varianza è:  $\sigma_t^2 = \sigma_x^2 = \frac{1}{n/\sigma^2} = \frac{\sigma^2}{n}$ .

Si ritrova così, tramite un ragionamento molto più generale la relazione che lega la varianza della variabile casuale delle medie campionarie di  $n$  elementi, alla varianza della variabile casuale da cui si estraggono i campioni. Invece se si suppone la media  $M$  nota e si vuole stimare la varianza, la (2.2) assume la forma:

$$\ln L = -\frac{n}{2} \ln 2\pi - n \ln \theta - \sum_{i=1}^n (x_i - M)^2 / 2\theta^2$$

$$\frac{\partial \ln L}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum_{i=1}^n (x_i - M)}{\theta^3} = \frac{n}{\theta^3} \left( \sum_{i=1}^n (x_i - M)^2 / n - \theta^2 \right) \quad (2.4)$$

Lo stimatore (varianza della popolazione):

$$t = \sum_{i=1}^n (x_i - M)^2 / n$$

della funzione:  $\tau(\theta) = \theta^2$ , è del tipo Limite Minima Varianza, la cui varianza associata è:

$$\sigma_t^2 = \sigma_{\sigma^2}^2 = 2\theta \frac{\theta^3}{n} = \frac{2\theta^4}{n} = \frac{2\sigma^4}{n}$$

In questo caso, occorre altresì osservare:

- l'unicità della funzione  $\tau(\theta)$  di cui esiste uno stimatore Limite Minima Varianza (infatti questa proprietà esiste solo per  $\tau = \sigma^2$ , ma non per  $\tau = \sigma$  o per altre funzioni di  $\sigma$ );
- la valutazione corretta dello stimatore, ottenuta calcolando teoricamente gli scarti:  $x_i - M$ , rispetto alla media dell'universo (cosicché lo stimatore risulta deviato con fattore di *bias*:  $(n-1)/n$ , dovendo usare, nella pratica, gli scarti:  $x_i - \bar{x}$ , rispetto alla media campionaria, per cui si ritorna alla nota formula:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n v_i^2}{n-1}. \quad (2.5)$$

### 4.3 Osservazioni dirette di diversa precisione

Se le  $n$  osservazioni di una stessa grandezza sono eseguite con diversa precisione, esse possono considerarsi come estratte da  $n$  diverse popolazioni normali, tutte con la stessa media, ma con varianze diverse. Volendo stimare la media comune, le (2.1), (2.2), (2.3) diventano:

$$L = \frac{1}{\sqrt{(2\pi)^n \sigma_1 \cdots \sigma_n}} e^{-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \theta)^2}{\sigma_i^2}} \quad \ln L = -\frac{n}{2} \ln 2\pi - \sum_{i=1}^n \ln \sigma_i - \sum_{i=1}^n \frac{(x_i - \theta)^2}{2\sigma_i^2} \quad (3.1)$$

$$\frac{\partial \ln L}{\partial \theta} = \sum_{i=1}^n \frac{(x_i - \theta)}{\sigma_i^2} = \left( \sum_{i=1}^n \frac{1}{\sigma_i^2} \right) \left( \frac{\sum_{i=1}^n x_i / \sigma_i^2}{\sum_{i=1}^n 1 / \sigma_i^2} - \theta \right)$$

Introducendo quantità inversamente proporzionali alle varianze dette *pesi*, secondo la relazione:

$$p_i = \sigma_0^2 / \sigma_i^2 \quad (3.2)$$

dove  $\sigma_0^2$  è una costante di proporzionalità arbitraria (anche se convenientemente opportuna), si ottiene:

$$\frac{\partial \lg L}{\partial \theta} = \frac{\sum_{i=1}^n p_i}{\sigma_0^2} \left( \frac{\sum_{i=1}^n p_i x_i}{\sum_{i=1}^n p_i} - \theta \right).$$

In questo caso, lo stimatore Limite Minima Varianza di  $\theta$  è detto *media ponderata*:

$$t = \bar{x}_p = \frac{\sum_{i=1}^n p_i x_i}{\sum_{i=1}^n p_i} \quad (3.3)$$

e la sua varianza è:

$$\sigma_t^2 = \sigma_{\bar{x}_p}^2 = \frac{1}{\sum p_i / \sigma_0^2} = \frac{\sigma_0^2}{\sum p_i} \quad (3.4)$$

da cui si può notare che il peso della media ponderata è la somma dei pesi delle singole osservazioni.

Resta ancora da stimare  $\sigma_0$ , detto *errore medio dell'unità di peso* (perché infatti, se  $p_i = 1$ ,  $\sigma_i = \sigma_0$ ). Esso è assunto arbitrariamente e dovrebbe essere noto a priori, se si conoscessero con esattezza gli sqm  $\sigma_i$  delle osservazioni. Tuttavia dato che questi sono noti, in generale, solo in modo molto grossolano,  $\sigma_0$  deve essere stimato, basandosi sugli scarti tra i valori osservati e la stima della media ponderata. Introducendo nella (3.1), al posto di  $\sigma_i$ , i valori:  $\sigma_0^2 / \sqrt{p_i} = \theta / \sqrt{p_i} = \sigma_i$ , si ha:

$$\ln L = -\frac{n}{2} \ln 2\pi - \sum_{i=1}^n \ln \frac{\theta}{\sqrt{p_i}} - \frac{\sum_{i=1}^n p_i (x_i - M)^2}{2\theta^2} = -\frac{n}{2} \ln 2\pi - n \ln \theta + \frac{1}{2} \sum_{i=1}^n \ln p_i - \frac{1}{2\theta^2} \sum_{i=1}^n p_i (x_i - M)^2$$

da cui si ricava l'equivalente della (2.4), valevole nel caso di diversa precisione delle osservazioni:

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta^3} \left( \sum_{i=1}^n p_i (x_i - M)^2 / n - \theta^2 \right)$$

Come già detto per la stima della media, lo stimatore di  $\sigma_0^2$ :

$$t = \sum_{i=1}^n p_i (x_i - M)^2 / n$$

è deviato, se si usa  $\bar{x}_p$ , al posto di  $M$ ; e così, dopo correzione del *bias*, si ha:

$$\sigma_0^2 = \frac{\sum_{i=1}^n p_i (x_i - \bar{x}_p)^2}{n-1} = \frac{\sum_{i=1}^n p_i v_i^2}{n-1} \quad (3.5)$$

Il problema delicato della media ponderata è la scelta dei pesi o, se si vuole, l'assegnazione delle varianze alle varie osservazioni. I loro veri valori sono sempre evidentemente incogniti e si possono conoscere stime abbastanza valide, solo se ciascuna osservazione sia, a sua volta, la media campionaria di un campione abbastanza numeroso. Ad esempio, questo si verifica in certe determinazioni di alta precisione, come quelle assolute di gravità, effettuate nello stesso luogo con diversi procedimenti, poiché ogni misura risulta dalla media di un grande numero di prove (tuttavia in generale, questo caso è raro). In molte altre occasioni, si procede ad una assegnazione dello sqm in base alla conoscenza delle precisioni degli strumenti usati, come pure ad una certa valutazione delle condizioni ambientali. In altri casi ancora, non si conoscono gli sqm o si preferisce non calcolarli, perché poco sicuri, ma è noto che sono proporzionali a determinate grandezze che intervengono indirettamente nella misura.

Allora una domanda pertinente riguarda l'attendibilità della media ponderata, dato che i pesi possono essere affetti da errori sensibili. A riguardo, occorre innanzitutto osservare che non ha senso mediare valori ottenuti con precisione molto diverse, poiché è preferibile scartare misure poco precise e mantenere quelle migliori. In secondo luogo, anche se i pesi sono determinati in modo piuttosto grossolano, i valori della media hanno variazioni che sono ampiamente contenute entro il suo errore medio e pertanto non si considerano rilevanti. In alcuni casi di incertezza, si eseguono due o più calcoli, con diverse distribuzioni di pesi, e si decide poi quale adottare, ad esempio, scegliendo quella che fornisce il minor valore dello sqm della media.

#### Esempio 4.3.1

Di uno stesso angolo, eseguite tre serie di misure *A*, *B* e *C*, con diverse modalità e diversa precisione. occorre calcolare, per ciascuna serie, il valore medio, lo sqm delle misure, lo sqm della media. Usando poi i valori medi  $\bar{x}_A$ ,  $\bar{x}_B$  e  $\bar{x}_C$ , come osservazioni di peso diverso, si deve ricavare la loro media ponderata, lo sqm della unità di peso e lo sqm della media ponderata.

<i>A</i>	<i>B</i>	<i>C</i>
87°12'23"	87°12'21"	87°12'23"
87°12'27"	87°12'25"	87°12'30"
87°12'25"	87°12'20"	87°12'22"
87°12'27"	87°12'26"	87°12'26"
87°12'23"	87°12'28"	87°12'27"
87°12'24"	87°12'23"	
87°12'24"	87°12'22"	
87°12'22"	87°12'24"	
	87°12'23"	
	87°12'27"	

Applicando le formule relative alle osservazioni dirette di uguale precisione, si ha:

$$\bar{x}_A = \frac{\sum x_A}{n_A} = 87^\circ 12' 24''.37$$

$$\bar{x}_B = 87^\circ 12' 23''.90$$

$$\bar{x}_C = 87^\circ 12' 25''.60$$

$$\sigma_{x_A} = \sqrt{\frac{\sum (x_A - \bar{x}_A)^2}{n_A - 1}} = 1''.85$$

$$\sigma_{x_B} = 2''.60$$

$$\sigma_{x_C} = 3''.21$$

$$\sigma_{x_A}^- = \frac{\sigma_{x_A}}{n_A} = 0''.65$$

$$\sigma_{x_B}^- = 0''.82$$

$$\sigma_{x_C}^- = 1''.44$$

Per calcolare la media ponderata  $\bar{x}_p$  dei tre valori  $\bar{x}_A$ ,  $\bar{x}_B$  e  $\bar{x}_C$ , si introducono come pesi gli inversi delle loro varianze:

$$p_A = \frac{1}{0.65^2} = 2.37$$

$$p_B = \frac{1}{0.82^2} = 1.49$$

$$p_C = \frac{1}{1.44^2} = 0.48$$

cosicché si ha poi, con la (3.3):

$$\bar{x}_p = 87^\circ 12' 20'' + \frac{4.37 \cdot 2.37 + 3.90 \cdot 1.49 + 5.60 \cdot 0.48}{4.34} = 87^\circ 12' 24''.34 \quad \text{essendo: } \sum_i p_i = 4.34$$

Il calcolo di  $\sigma_{x_p}^-$  richiede la stima di  $\sigma_0^2$  a posteriori, secondo la (3.5), e precedentemente il calcolo di  $v_i$ :

$$v_A = 0.03$$

$$v_B = -0.44$$

$$v_C = 1.26$$

$$\sigma_0 = \sqrt{\frac{\sum_i p_i v_i^2}{m-1}} = 0.725$$

essendo:  $m = 3$ , da cui infine:

$$\sigma_{\bar{x}_p} = \frac{\sigma_0}{\sqrt{\sum p}} = 0''.35$$

Come controllo dei calcoli deve poi essere:  $\sum_i p_i v_i = 0$  e, di fatto:  $\sum p v \cong 0$

#### Esempio 4.3.2

La quota del vertice  $X$  è determinata, con lo stesso strumento e le stesse modalità operative, partendo da quattro vertici, di quota nota,  $A_1$ ,  $A_2$ ,  $A_3$  ed  $A_4$ , di cui sono anche note le distanze da  $X$  (come mostra la Fig. 4.3.1), ed aggiungendo a tali quote i quattro dislivelli misurati, lungo questi lati. Il metodo di misura adottato permette di stabilire che i pesi dei dislivelli misurati (e delle quote) sono inversamente proporzionali ai quadrati di tali distanze e di calcolare così il valore più probabile della quota di  $X$  ed il suo errore medio.

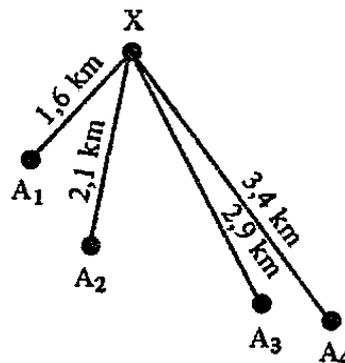


Fig. 4.3.1 – Schema delle distanze fra i punti utilizzati per le quattro determinazioni della quota di  $X$

$$\begin{array}{llll}
q_1 = 251.18 m & d_1 = 1.6 km & q_3 = 251.23 m & d_3 = 2.9 km \\
q_2 = 251.15 m & d_2 = 2.1 km & q_4 = 251.23 m & d_4 = 3.4 km
\end{array}$$

Per l'ipotesi fatta sulle modalità delle misure, i pesi risultano:

$$\begin{array}{ll}
p_1 = \frac{k}{(1.6)^2} = \frac{k}{2.56} & p_2 = \frac{k}{(2.1)^2} = \frac{k}{4.41} \\
p_3 = \frac{k}{(2.9)^2} = \frac{k}{8.41} & p_4 = \frac{k}{(3.40)^2} = \frac{k}{11.56}
\end{array}$$

Per avere pesi maggiori di 1, per comodità, si pone  $k = 11,56$ , cosicché:

$$p_1 = 4.52 \quad p_2 = 2.62 \quad p_3 = 1.37 \quad p_4 = 1 \quad \sum_i p_i = 9.51$$

ed il valore più probabile della quota incognita di  $X$ , con il suo errore medio, risulta:

$$\bar{q}_X = \frac{\sum p_i q_i}{\sum p_i} = \frac{251.18 \cdot 4.52 + 251.15 \cdot 2.62 + \dots}{9.51} = 251.19 m \quad \sigma_{\bar{q}_X}^2 = \frac{\sum p_i (q_i - \bar{q}_X)^2}{(n-1) \sum p_i}$$

A tal fine, si devono prima calcolare gli scarti delle osservazioni, rispetto al valore più probabile:

$$\begin{array}{lll}
v_1 = q_1 - \bar{q}_X = -0.01 & v_2 = q_2 - \bar{q}_X = -0.04 & \\
v_3 = q_3 - \bar{q}_X = 0.04 & v_4 = q_4 - \bar{q}_X = 0.09 & \sum_i p_i v_i^2 \cong 0.015 m^2
\end{array}$$

da cui:

$$\sigma_0^2 = \frac{\sum_i p_i v_i^2}{n-1} = \frac{0.015}{3} = 0.005 m^2 \quad \sigma_0 = \pm 0.07 m$$

ed infine:

$$\sigma_{\bar{q}_X}^2 = \frac{0.015}{3 \cdot 9.51} = 5.26 \cdot 10^{-4} m^2 \quad \sigma_{\bar{q}_X} = \pm 2.3 cm$$

Il valore  $\sigma_0$  rappresenta lo sqm della quota alla quale è stato dato peso unitario, cioè  $q_4$ . Dato che la distanza  $A_4 X$  è  $3.4 km$  e che gli sqm crescono in ragione delle distanze, si può dire che lo sqm relativo a quote misurate da un kilometro di distanza è di  $0.07/3.4 = 0.02 m$ .

In questo caso, non ha senso confrontare, con un opportuno test statistico, la stima di  $\sigma_0^2$  a posteriori, con  $k = 11.56$ , assunto inizialmente, in quanto qui  $k$  funge solo da costante di proporzionalità. Questo perché, a differenza dell'esempio 3.3.1., dove si possono stimare gli sqm delle varie misure, introdotte nella media ponderata, non si conoscono ora gli sqm delle quote, utilizzate per la stima del valore più probabile della quota incognita di  $X$ , ma si sa solo in che rapporto questi sqm stanno tra loro.

### Esempio 4.3.3

In una stessa località, sono effettuate tre misure assolute di gravità  $g_i$  ( $i = 1, 2, 3$ ), con tre apparati diversi. Poiché ciascuna misura è ripetuta parecchie volte, è possibile stimare i loro scarti quadratici medi  $\sigma_i$  e poi calcolare la media ponderata, lo scarto quadratico medio della unità di peso e quello della media. I valori ottenuti sono i seguenti:

$g$	980.35841	980.35838	980.35836	$gal = cm \cdot sec^{-2}$
$\sigma$	$3.5 \cdot 10^{-5}$	$2.3 \cdot 10^{-5}$	$1.2 \cdot 10^{-5}$	

Per completezza, occorre poi ripetere il calcolo con gli stessi valori di  $g$ , ma ponendo lo sqm della prima osservazione pari a  $1.5 \cdot 10^{-5} gal$ .

Poiché i valori  $g_i$  sono uguali fino ai millesimi di  $gal$ , la media e gli scarti quadratici medi possono essere calcolati utilizzando solo le ultime due cifre; risultando così espressi in centesimi di  $mgal$  oppure, il che è lo stesso, in decine di  $\mu gal$ .

Il calcolo dei pesi deriva dalla formula:  $p = \sigma_0^2 / \sigma^2$ , ove  $\sigma_0$  è una costante arbitraria. Scegliendo  $\sigma_0 = 3.5 \cdot 10^{-5} gal$ , si ha:  $p_1 = 1$ ,  $p_2 = 2.3$ ,  $p_3 = 8.5$  e, poiché i pesi sono molto diversi, potrebbe essere dubbio l'utilizzo della prima misura. Il valore medio è:

$$\bar{g}_p = \frac{41 \cdot 1.0 + 38 \cdot 2.3 + 36 \cdot 8.5}{11,8} = 36.8 \quad \text{da cui:} \quad \bar{g}_p = 980.358368 gal.$$

Gli scarti hanno valori: 4.2, 1.2, -0,8 e si ha:  $\sum pv \approx 0$ , verificando così il calcolo della media.

Dopodiché il calcolo dell'errore medio dell'unità di peso è:

$$\sigma_0^2 = \frac{\sum p_i v_i^2}{n-1} = \frac{26.4}{2} = 13.2 \quad \sigma_0 = 3.6.$$

e lo sqm della media ponderata è:

$$\sigma_{\bar{g}} = \frac{\sigma_0}{\sqrt{\sum p}} = \frac{3.6}{\sqrt{11.8}} \cong 1.1 \cdot 10^{-5} gal;$$

Il valore di  $\sigma_0$  è poco diverso da quello assunto inizialmente e questo comportamento può essere indice di un calcolo corretto degli sqm delle osservazioni e, in particolare, di assenza di errori sistematici.

Invece lo sqm della media ponderata è praticamente uguale a quello della misura più precisa per cui l'utilizzo di misure meno precise non dà un contributo sensibile alla conoscenza della gravità nel luogo considerato.

La differenza tra  $\sigma_{\bar{g}}$  e lo sqm dell'ultima misura è inferiore all'errore medio della media ed insignificante. In particolare, togliendo la prima misura si ha  $\bar{g} = 36.4$ , cosicché il contributo della prima misura è, del tutto, irrilevante e probabilmente più dannoso che utile, essendo la variazione di  $\bar{g}$  dell'ordine di  $4 \cdot 10^{-6} gal$ , mentre l'errore medio è  $11 \cdot 10^{-6} gal$ .

Si supponga ora che la prima misura abbia lo stesso valore, ma uno sqm di 1.5 e, ripetendo il calcolo con  $\sigma_0 = 2.3$ , i pesi siano: 2.3, 1 e 3.6, da cui:

$$\bar{g}_p = \frac{41 \cdot 2.3 + 38 \cdot 1 + 36 \cdot 3.6}{6.9} = 38.0 \quad \sigma_0 = 4.1 \quad \sigma_{\bar{g}} = 1.6$$

Questo risultato merita un breve commento: il valore  $\sigma_{\bar{g}}$  è nettamente superiore agli sqm della prima e dell'ultima misura. Questo significa che uno di essi non è calcolato correttamente, in quanto la misura o le misure sono affette da errori sistematici di cui non si tiene conto. Purtroppo ciò accade non troppo di rado, poiché l'eliminazione degli errori sistematici è assai ardua, in alcune misure e, tra esse, in quelle di gravità assoluta, con apparecchiature moderne, dove gli sqm delle misure sono ridotti a pochissimi  $\mu gal$ .

La presenza di probabili errori sistematici è rivelata anche dall'ispezione delle misure. Infatti lo scarto effettivo tra la prima e l'ultima è  $5 \cdot 10^{-5} gal$ . Lo sqm della differenza delle due misure dovrebbe essere:

$\sigma_d = \sqrt{1.5^2 + 1.2^2} = 1.9$ , mentre la differenza trovata è poco meno del triplo di  $\sigma_d$  (ciò è possibile, ma con probabilità molto piccola, per cui la presenza di effetti sistematici è altamente probabile). In questo caso, la media ponderata delle tre misure è del tutto giustificata, in quanto probabilmente riduce sensibilmente gli effetti sistematici presenti nei singoli risultati, errori i quali sono certamente diversi, da misura a misura, e conseguentemente, nel gruppo delle tre misure, si comportano come accidentali.

#### 4.4 Funzioni di quantità osservate direttamente

Sia  $x$  una quantità dedotta mediante la misura diretta di  $n$  grandezze  $x_1, x_2, \dots, x_n$ , legate ad essa da una relazione analitica:

$$x = f(x_1, \dots, x_n). \quad (4.1)$$

Se i valori osservati delle  $x_i$  sono  $O_1, O_2, \dots, O_n$ , l'espressione:  $O = f(O_1, O_2, \dots, O_n)$  è una stima non deviata di  $x$ , nel senso che il suo valor medio coincide con  $x$ , almeno in prima approssimazione. Infatti, posto:  $O_i - x_i = v_i$ , nell'ipotesi che gli errori  $v_i$  siano *accidentali*, cioè *a media nulla ed abbastanza piccoli da poterne trascurare i quadrati e le potenze superiori*, si ha, sviluppando in serie la (4.1) nell'intorno del punto  $(O_1, O_2, \dots, O_n)$ :

$$x = f(O_i) + \sum \left( \frac{\partial f}{\partial x_i} \right)_{O_i} (x_i - O_i) = O - \sum a_i v_i \quad \text{con:} \quad a_i = \left( \frac{\partial f}{\partial x_i} \right)_{O_i} \quad (4.2)$$

Dalla (4.2) si ottiene:  $M_O = M_x + \sum_{i=1}^n a_i M_{v_i} = M_x$ .

#### 4.4.1 Varianza di una funzione di quantità osservate

Per il calcolo della varianza di  $x$  basta osservare che la (4.2), ponendo  $O - x = v_x$ , assume la forma:

$$v_x = a_1 v_1 + a_2 v_2 + \dots + a_n v_n \quad (4.3)$$

dove  $v_x$  e  $v_i$  sono variabili casuali. Pertanto alla (4.3) si può applicare la legge di propagazione degli scarti, a seconda che le osservazioni  $O_i$  siano indipendenti, o meno, fra loro:

$$\sigma_x^2 = \left( \frac{\partial f}{\partial x_1} \right)_{O_i}^2 \sigma_{x_1}^2 + \left( \frac{\partial f}{\partial x_2} \right)_{O_i}^2 \sigma_{x_2}^2 + \dots + \left( \frac{\partial f}{\partial x_n} \right)_{O_i}^2 \sigma_{x_n}^2 + \dots + 2 \left( \frac{\partial f}{\partial x_k} \right)_{O_i} \left( \frac{\partial f}{\partial x_l} \right)_{O_i} r_{x_k x_l} \sigma_{x_k} \sigma_{x_l} + \dots \quad (4.4)$$

dove  $r_{x_k x_j}$  è il coefficiente di correlazione lineare fra  $x_k$  e  $x_j$  (la (4.4) è importante e di frequentissima applicazione).

#### 4.4.2 Coefficiente di correlazione funzionale

Un'analisi delle funzioni di quantità osservate porta al concetto, ed al calcolo, del coefficiente di correlazione lineare di tipo funzionale. A riguardo, si supponga di avere due funzioni  $x_i$  e  $x_j$  di quantità osservate direttamente  $y_1, y_2, \dots, y_n$ , in tutto od in gran parte, uguali fra loro (ed eseguite in maniera indipendente o, in generale, correlate fra loro):

$$\begin{cases} x_i = f_i(y_1, y_2, \dots, y_n) \\ x_j = f_j(y_1, y_2, \dots, y_n) \end{cases} \quad (4.5)$$

Poiché per il calcolo di  $x_i$  e  $x_j$  si introducono nelle (4.5) gli stessi valori misurati delle grandezze  $y$ , gli errori di queste ultime fanno sì che, attraverso  $f_i$  e  $f_j$ , si stabiliscano, fra  $x_i$  e  $x_j$ , correlazioni per cui esse non si possono considerare indipendenti. In tal caso, dalla (4.3), il coefficiente di correlazione lineare  $r_{x_k x_j}$  è così calcolato:

$$\begin{cases} v_{x_i} = a_{i1} v_{y_1} + a_{i2} v_{y_2} + \dots + a_{in} v_{y_n} \\ v_{x_j} = a_{j1} v_{y_1} + a_{j2} v_{y_2} + \dots + a_{jn} v_{y_n} \end{cases} \quad (4.6)$$

dove  $a_{ik}$  e  $a_{jl}$  hanno ancora il significato di derivate parziali delle funzioni  $f_i$  e  $f_j$ , calcolate per i valori misurati direttamente di  $y$ . Sostituendo le (4.6) nell'espressione del coefficiente lineare ed applicando le proprietà della media di variabile casuale, si ottiene:

$$r_{x_k x_j} = \frac{M(v_{x_i} v_{x_j})}{\sigma_{x_i} \sigma_{x_j}} = \frac{M((a_{i1} v_{y_1} + a_{i2} v_{y_2} + \dots + a_{in} v_{y_n})(a_{j1} v_{y_1} + a_{j2} v_{y_2} + \dots + a_{jn} v_{y_n}))}{\sigma_{x_i} \sigma_{x_j}} =$$

$$= \frac{a_{i1} a_{j1} \sigma_{y_1}^2 + a_{i2} a_{j2} \sigma_{y_2}^2 + \dots + a_{in} a_{jn} \sigma_{y_n}^2 + \dots + (a_{ik} a_{jl} + a_{il} a_{jk}) \sigma_{y_k} \sigma_{y_l} + \dots}{\sigma_{x_i} \sigma_{x_j}} \quad (4.7)$$

Se le misure dirette  $y_1, y_2, \dots, y_n$  sono eseguite in maniera indipendente, allora la media dei prodotti misti degli scarti si annulla ed il coefficiente di correlazione lineare di tipo funzionale fra le misure indirette  $x_i$  e  $x_j$ , diventa:

$$r_{x_k x_j} = \frac{a_{i1} a_{j1} \sigma_{y_1}^2 + a_{i2} a_{j2} \sigma_{y_2}^2 + \dots + a_{in} a_{jn} \sigma_{y_n}^2}{\sigma_{x_i} \sigma_{x_j}} \quad (4.8)$$

Il denominatore della (4.7) e della (4.8) è, a sua volta, calcolato applicando l'espressione (4.4) alle (4.6).

#### Esempio 4.4.1

La lunghezza di una sbarra metallica alle diverse temperature è data dalla nota relazione:  $L_t = L_0(1 + \alpha t)$ , dove  $L_0$  ed  $\alpha$  sono parametri, noti in precedenza, per ricavare il corrispondente valore  $L_t$ , ad ogni temperatura  $t$ . In generale,  $L_0$  ed  $\alpha$  si ricavano contemporaneamente, mediante opportune misure dirette di  $t$  e  $L_t$ . A tal fine, data una sbarra di metallo della lunghezza di circa  $1 m$ , si misura la lunghezza della sbarra, a due temperature diverse:  $t_1 = 5.5^\circ C$  e  $t_2 = 28.7^\circ C$ , ottenendo:

$$L_1 = 1.000312 m \quad L_2 = 1.000553 m$$

Le misure della  $t$  e della lunghezza sono fra loro indipendenti e sono eseguite rispettivamente con  $\sigma_t = \pm 0.5^\circ C$  e  $\sigma_L = \pm 4 \mu m$ ; di conseguenza, si ricavano dapprima le espressioni per calcolare  $\alpha$  e  $L_0$ , mediante le misure di  $L$  e  $t$  a due diverse temperature:

$$L_0 = \frac{L_1 t_2 - L_2 t_1}{t_2 - t_1} = 1.0000254 m \quad \alpha = \frac{L_2 - L_1}{L_1 t_2 - L_2 t_1} = 10,4 \cdot 10^{-6} (^\circ C)^{-1}$$

Le due espressioni sono di tipo (4.5), essendo entrambe funzioni delle stesse misure, e da esse si ricavano, mediante l'applicazione della (4.4), le espressioni delle varianze delle due misure indirette di  $L_0$  e  $\alpha$ :

$$\sigma_{L_0}^2 = \frac{t_2^2 + t_1^2}{(t_2 - t_1)^2} (\sigma_L^2 + L_0^2 \alpha^2 \sigma_t^2) = (8.2 \mu m)^2 \quad \sigma_\alpha^2 = \frac{L_1^2 + L_2^2}{L_0^2 (L_1 t_2 - L_2 t_1)^2} (\sigma_L^2 + L_0^2 \alpha^2 \sigma_t^2) = (0.4 \cdot 10^{-6} ^\circ C^{-1})^2$$

A sua volta, il coefficiente di correlazione lineare di tipo funzionale fra le misure indirette di  $\alpha$  e  $L_0$  si ricava applicando direttamente la (4.8) ed eseguendo facili passaggi analitici:

$$r_{L_0\alpha} = -\frac{L_2 t_2 + L_1 t_1}{\sqrt{(t_2^2 + t_1^2)(L_2^2 + L_1^2)}} = -0.83$$

Come evidente, la correlazione tra  $L_0$  ed  $\alpha$  è assai alta e non può essere trascurata, quando si vuole ricavare la varianza della misura indiretta  $L_t$  in funzione, non solo della varianza della temperatura  $t$ , ma anche di quelle dei parametri  $L_0$  ed  $\alpha$ , presenti nella relazione:

$$\sigma_{L_t}^2 = L_0^2 \alpha^2 \sigma_t^2 + \left( \frac{L_t^2}{L_0^2} \sigma_{L_0}^2 + L_0^2 t^2 \sigma_\alpha^2 + 2r_{L_0\alpha} L_t t \sigma_{L_0} \sigma_\alpha \right)$$

Nell'ultima espressione, compare solo il termine che tiene conto della correlazione fra  $L_0$  ed  $\alpha$  perché, per evidenti ragioni, non si ha correlazione fra il parametro  $\alpha$  e la generica misura  $t$  e neppure tra questa e  $L_0$ .

#### 4.5 Osservazioni indirette con modello lineare

Dato il caso di parametri da stimare  $\theta_1, \theta_2, \dots, \theta_g$ , legati da  $n$  relazioni lineari con un gruppo di quantità osservate indipendentemente o, nel caso più generale, di funzioni di quantità osservate:

$$x_i - (a_i \theta_1 + b_i \theta_2 + \dots + g_i \theta_g) = 0 \quad i = 1, 2, \dots, n \quad (5.1)$$

dove  $x_1, x_2, \dots, x_n$  sono le quantità osservate o funzioni di queste:

- se  $n = g$  ed il determinante del sistema è diverso da zero, il calcolo di  $\theta$  si riduce alla soluzione di un sistema di equazioni lineari;
- se  $n > g$ , cioè se il numero delle osservazioni è superiore a quello delle incognite, il sistema, se è ben impostato fisicamente, dovrebbe avere determinante nullo, in quanto le soluzioni, ottenute da un gruppo qualunque di  $g$  equazioni, dovrebbero essere valide anche per le rimanenti  $n - g$ , combinazioni lineari delle precedenti (tuttavia nella pratica, questa ipotesi non è mai soddisfatta, perché è impossibile trovare un gruppo di soluzioni valide per tutte le (5.1), a causa degli errori di osservazione da cui sono affette le quantità osservate o funzioni di queste  $x_i$ ).

Per la soluzione di questo nuovo problema, occorre scegliere un metodo di calcolo che permetta di ottenere un insieme di stime dei parametri  $\theta_1, \theta_2, \dots, \theta_g$  le quali soddisfano a tutte, o solo ad alcune, proprietà degli stimatori. Questi valori stimati delle incognite  $\theta$  non soddisfano ovviamente le equazioni (5.1) ed i secondi membri di queste hanno valori:  $v_i \neq 0$ , per cui le (5.1) si possono scrivere nella forma, detta anche equazione agli errori.

$$x_i - (a_i\theta_1 + b_i\theta_2 + \dots + g_i\theta_g) = v_i \quad (5.2)$$

Il criterio dei minimi quadrati fornisce stime non deviate e di minima varianza, cioè la determinazione dei valori  $\theta$ , ottenuti imponendo:

$$\sum_{i=1}^n v_i^2 = \sum_{i=1}^n (x_i - (a_i\theta_1 + b_i\theta_2 + \dots + g_i\theta_g))^2 = \min \quad (5.3)$$

Poiché  $x_i$  sono osservazioni indipendenti, nelle formulazioni (5.2) e (5.3), si suppongono di uguale varianza  $\sigma_0^2$ . Nel caso più generale, ad ogni  $x_i$ , compete una sua varianza  $\sigma_i^2$  e queste ultime, per le (5.2), sono anche le varianze dei residui  $v_i$ . A riguardo, ammettere diverse varianze, per le osservazioni, significa che non tutte le equazioni (5.1) hanno lo stesso peso, ovvero non contribuiscono tutte in modo ugualmente attendibile alla determinazione di  $\theta$ . Allora introducendo la consueta espressione del peso:  $p_i = \sigma_0^2 / \sigma_i^2$ , la condizione di minimo (5.3) diventa:

$$\sum_{i=1}^n p_i v_i^2 = \sum_{i=1}^n \frac{\sigma_0^2}{\sigma_i^2} (x_i - (a_i\theta_1 + b_i\theta_2 + \dots + g_i\theta_g))^2 = \min$$

e pertanto, ciascuna delle (5.2) deve essere divisa per  $\sigma_i$  di competenza, ovvero moltiplicata per la radice quadrata del proprio peso:  $\sqrt{p_i}$ .

Supponendo di avere fatto questa operazione, le (5.2) (di cui per semplicità di scrittura si mantiene invariata l'espressione) sono tutte ridotte allo stesso peso e le osservazioni hanno tutte una stessa varianza  $\sigma_0^2$ , costante ed arbitraria. Introducendo ora la notazione matriciale:

$$A = \begin{pmatrix} a_1 & b_1 & \dots & g_1 \\ a_2 & b_2 & & g_2 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_n & b_n & \dots & g_n \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \cdot \\ \cdot \\ \theta_g \end{pmatrix} \quad v = \begin{pmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{pmatrix}$$

Il sistema delle equazioni agli errori (5.2) si può scrivere brevemente nella forma:

$$x - A\theta = v \quad (5.4)$$

mentre la norma (5.3) diventa:

$$v^T v = \min \quad (5.5)$$

dove  $v^T$  è il vettore trasposto  $[v_1 v_2 \cdots v_n]$  del vettore  $v$ .

L'operazione  $v^T v$ , per le note regole sul prodotto di matrici, dà origine allo scalare  $\sum_{i=1}^n v_i^2$ . Pertanto la (5.5) è soddisfatta, se è nullo il differenziale del primo membro:

$$dv^T v + v^T dv = 0. \quad (5.6)$$

Allora essendo i termini  $dv^T v$  e  $v^T dv$  due scalari uguali, la (5.6) si può scrivere:

$$2dv^T v = 0 \quad \Rightarrow \quad dv^T v = 0. \quad (5.7)$$

Dalla (5.4) si ha, ricordando che la trasposta di un prodotto di matrici è uguale al prodotto delle trasposte in ordine invertito (la stessa regola vale per l'operazione di inversione):

$$v^T = x^T - \theta^T A^T \quad \Rightarrow \quad dv^T = -d\theta^T A^T \quad (5.8) \text{ e } (5.9)$$

e sostituendo le (5.9) e (5.4) nella (5.7):

$$dv^T (A^T x - A^T A \theta) = 0. \quad (5.10)$$

Questa relazione lega tra loro linearmente i differenziali delle grandezze indipendenti  $\theta_i$  e, affinché essa sia identicamente soddisfatta, la matrice dei coefficienti deve annullarsi:

$$A^T x - A^T A \theta = 0. \quad (5.11)$$

La matrice  $D = A^T A$  è quadrata, di dimensioni  $(g \times g)$ , ed  $A^T x$  è un vettore colonna, di dimensioni  $(g \times 1)$ . Il sistema lineare (5.11), nelle incognite  $\theta_1, \theta_2, \dots, \theta_g$ , è univocamente risolubile, perché  $D$  non può essere degenere essendo indipendenti tra loro tutte le (5.2). La matrice  $D$ , così costituita:

$$D = \begin{vmatrix} \sum a^2 & \sum ab & \cdots & \sum ag \\ \sum ab & \sum b^2 & & \sum bg \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \sum ag & \cdot & \cdots & \sum g^2 \end{vmatrix} \quad (5.12)$$

detta matrice normale, è simmetrica e si può anche dimostrare che, per  $D$  e per ogni matrice normale, i termini sulla diagonale principale sono preponderanti, rispetto a quelli fuori diagonale, ovvero è valida la relazione:

$$-1 \leq \frac{a_{rs}}{\sqrt{a_{rr}a_{ss}}} \leq +1 \quad (5.13)$$

dove  $a_{ij}$  è un generico elemento della  $D$ . La matrice inversa  $D^{-1}$  (dove un qualsiasi termine generico:  $\alpha_{ij} = (-1)^{i+j} \partial(D_{ij})/\partial(D)$ , è dato dal rapporto fra il determinante del minore complementare  $D_{ij}$  e il determinante della  $D$ ) è essenziale per la soluzione del sistema (5.4):

$$\theta = D^{-1} A^T x \quad \begin{pmatrix} \theta_1 \\ \theta_2 \\ \cdot \\ \cdot \\ \theta_g \end{pmatrix} = \begin{pmatrix} \sum a^2 & \sum ab & \cdots & \sum ag \\ \sum ab & \sum b^2 & & \sum bg \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \sum ag & \cdot & \cdots & \sum g^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum ax \\ \sum bx \\ \cdot \\ \cdot \\ \sum gx \end{pmatrix} \quad (5.14)$$

La (5.14) fornisce l'insieme delle stime dei  $g$  parametri:  $\theta_1, \theta_2, \dots, \theta_g$ , i cui valori, di entità:  $z_1, z_2, \dots, z_g$ , introdotti nella (5.4), permettono di calcolare, senza alcuna difficoltà, le altre  $n$  incognite del problema, cioè  $n$  scarti, le cui stime sono:  $\lambda_1, \lambda_2, \dots, \lambda_n$ . Introducendo i due vettori colonna  $z$  e  $\lambda$ , costituiti dalle stime dei parametri e degli scarti, la relazione:

$$x - Az = \lambda \quad (5.15)$$

è fra entità numeriche note (a riguardo, si indicano qui le soluzioni che sono numeri, con simboli diversi, dalle variabili casuali:  $\theta_1, \theta_2, \dots, \theta_g$  e  $v_1, v_2, \dots, v_n$ , per non creare confusioni concettuali). Dopodiché come controllo dei calcoli, si dimostra la validità delle seguenti relazioni:

$$A^T \lambda = 0 \quad \lambda^T A = 0. \quad (5.16)$$

Infatti moltiplicando per  $A^T$  la (5.15), si ottiene:  $A^T x - A^T A z = A^T \lambda = 0$ , che dimostra entrambe le (5.16). Inoltre moltiplicando la (5.15) per  $\lambda^T$ , si ha:  $\lambda^T x - \lambda^T A z = \lambda^T \lambda = \lambda^T \lambda$ , altra relazione di controllo.

#### 4.6. Varianze delle grandezze determinate indirettamente e coefficienti di correlazione

La formula risolutiva:

$$z = D^{-1} A^T x \quad (6.1)$$

fornisce le stime  $z_i$  delle incognite  $\theta_i$ , attraverso la matrice inversa  $D^{-1}$  della matrice normale e, da questa

matrice, si possono ottenere le varianze di  $z$ , in funzione della varianza  $\sigma_0^2$  degli scarti  $v$  delle equazioni agli errori, uguale a quella dei termini noti  $x$ . Indicate simbolicamente con  $u$  e  $v$ :

$$u = \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ u_g \end{pmatrix} \quad \text{e} \quad v = \begin{pmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ v_n \end{pmatrix}$$

le variabili casuali rappresentano gli errori delle incognite e dei termini noti.

Dalla (6.1), analoga alla (4.3), si ha poi:

$$u = D^{-1} A^T v \tag{6.2}$$

potendo così trovare il valore medio del prodotto  $u_i u_s^T$  di due qualsiasi delle variabili casuali  $u$ . A tal fine, si considera la matrice:

$$uu^T = \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ u_g \end{pmatrix} \begin{pmatrix} u_1 & u_2 & \cdots & u_g \end{pmatrix} = \begin{pmatrix} u_1^2 & u_1 u_2 & \cdots & u_1 u_g \\ u_2 u_1 & u_2^2 & \cdots & u_2 u_g \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ u_g u_1 & u_g u_2 & \cdots & u_g^2 \end{pmatrix} \tag{6.3}$$

Dalle (6.2), si ha:  $u^T = v^T A (D^{-1})^T$  e  $uu^T = D^{-1} A^T v v^T A (D^{-1})^T$ , dove  $vv^T$  è una matrice  $(n \times n)$ , ricavata come la (6.3). Applicando l'operazione di media, si ottiene:

$$M(uu^T) = D^{-1} A^T M(vv^T) A (D^{-1})^T. \tag{6.4}$$

e, dato che le  $v_i$  sono errori di osservazioni indipendenti, già ridotte allo stesso peso, si ha:

$$M_{v_i^2} = \sigma_0^2 \quad \text{e} \quad M_{v_i v_j} = M_{v_i} M_{v_j} = 0 \quad \forall j \neq i$$

dove  $\sigma_0^2$  è l'errore medio dell'unità di peso, cioè la varianza comune a tutte le osservazioni. La matrice  $M(vv^T)$  ha così la struttura:

$$M(v \cdot v^T) = \begin{vmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_0^2 & \cdots & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & \cdots & \sigma_0^2 \end{vmatrix} = I\sigma_0^2$$

essendo  $I$  la matrice identità di dimensioni  $(n \times n)$ . Sostituendo nella (6.4) e tenendo presente che  $(D^{-1})^T = D^{-1}$ , perché  $D^{-1}$  è simmetrica, si ottiene:

$$M(uu^T) = D^{-1}A^T\sigma_0^2AD^{-1} = D^{-1}A^TAD^{-1}\sigma_0^2 = D^{-1}DD^{-1}\sigma_0^2 = D^{-1}\sigma_0^2. \quad (6.5)$$

cosicché gli elementi della matrice inversa  $D^{-1}$  risultano proporzionali alle medie dei prodotti degli errori delle incognite, prese a due a due. Esplicitamente indicando con  $\alpha_{rs}$  il generico elemento di  $D^{-1}$ , si ha:

$$M_{u_r u_s} = \alpha_{rs}\sigma_0^2. \quad (6.6)$$

In particolare, la varianza di una incognita  $z_i$  è:

$$\sigma_{z_i}^2 = M_{u_i^2} = \alpha_{ij}\sigma_0^2 \quad (6.7)$$

da cui  $1/\alpha_{ij}$  è il peso di  $z_i$ . Infine il coefficiente di correlazione delle incognite  $z_i$  e  $z_j$  è:

$$r_{ij} = \frac{M_{u_i u_j}}{\sqrt{M_{u_i^2} M_{u_j^2}}} = \frac{\alpha_{ij}}{\sqrt{\alpha_{ii} \alpha_{jj}}} \quad (6.8)$$

il quale, per le proprietà delle matrici normali, soddisfa la condizione:  $-1 \leq r \leq +1$ .

Per mezzo dei termini della matrice  $D^{-1}$  si ottengono non solo le varianze di tutte le incognite, ma anche i coefficienti di correlazione fra due qualunque di esse i quali non sono nulli, in generale, perché le incognite sono ottenute indirettamente, attraverso la soluzione del sistema normale e, proprio per questa origine comune, non sono indipendenti fra loro. Dall'espressione di  $r_{ij}$  risulta, in modo chiaro, che la dipendenza fra le incognite non proviene da errori di misura, riflessi contemporaneamente su alcune di esse. Infatti nella formula (6.8) non compare  $\sigma_0^2$  il quale dipende dagli errori delle osservazioni, ma solo  $\alpha_{rs}$  la cui struttura è determinata da  $a_{rs}$ , elementi di  $D$ , a loro volta, dipendenti dal tipo di relazioni analitiche (5.1), esistenti fra le grandezze misurate direttamente e quelle calcolate indirettamente.

Come già nel caso della media ponderata, il valore di  $\sigma_0^2$  non può essere ricavato dalla nota relazione che lo lega al peso delle equazioni agli errori (5.2), ma deve essere ottenuto a posteriori, in base agli scarti, funzioni delle incognite calcolate.

Se le grandezze osservate fossero esenti da errori, si potrebbero ottenere i valori  $\theta'$  e termini noti (in questo caso, indicati con  $x'$ ) fra cui varrebbe la relazione:

$$x' - A\theta' = 0 \tag{6.9}$$

Sottraendo quest'ultima dalla (5.15), si ottiene:

$$v - Au = \lambda \quad \Rightarrow \quad v = \lambda + Au \tag{6.10}$$

dove  $u$  sono ancora variabili casuali, rappresentanti gli errori delle incognite,  $v$  gli errori dei termini noti, ovvero delle equazioni agli errori, ridotte allo stesso peso, e  $\lambda$  entità numeriche. Un valore di  $\sigma_0^2$  che tenga conto di tutte le possibili osservazioni è dato da:

$$\sigma_0^2 = M\left(\frac{\sum v_i^2}{n}\right) = \frac{M(v^T v)}{n} \tag{6.11}$$

dove  $\sum v^2$  è una variabile casuale, costituita dai valori campionari ottenuti calcolando la somma dei quadrati di  $n$  scarti estratti, a caso, dalla variabile casuale  $v$ . Sviluppando la (6.11) e ricordando le (6.6) e (6.10), si ha:

$$\begin{aligned} M(vv^T) &= M((\lambda^T + u^T A^T)(\lambda + Au)) = M(\lambda^T \lambda + \lambda^T Au + u^T A^T \lambda + u^T A^T Au) = \\ &= M(\lambda^T \lambda + u^T Du) = \sum \lambda_i^2 + M(u^T Du) = \\ &= \sum \lambda_i^2 + M\left(\sum a^2 u_1^2 + \sum ab u_2 u_1 + \dots + \sum ag u_g u_1 + \right. \\ &+ \sum ab u_1 u_2 + \sum b^2 u_2^2 + \dots + \sum bg u_g u_2 + \\ &+ \dots + \\ &+ \left. \sum ag u_1 u_g + \sum bg u_2 u_g + \dots + \sum g^2 u_g^2\right) = \\ &= \sum \lambda_i^2 + \left(\sum a^2 \alpha_{11} + \sum ab \alpha_{21} + \dots + \sum ag \alpha_{g1} + \right. \\ &+ \sum ab \alpha_{12} + \sum b^2 \alpha_{22} + \dots + \sum bg \alpha_{g2} + \\ &+ \dots + \\ &+ \left. \sum ag \alpha_{1g} + \sum bg \alpha_{2g} + \dots + \sum g^2 \alpha_{gg}\right) \sigma_0^2 \end{aligned} \tag{6.12}$$

Ciascuna delle  $g$  righe fra parentesi della (6.12) è il prodotto della  $j$ -esima riga della matrice normale  $D$  per la  $i$ -esima colonna della sua inversa  $D^{-1}$ . Dato che  $DD^{-1} = I$ , questi prodotti forniscono i termini diagonali della matrice identità e sono tutti uguali ad 1. Dalla (6.12), si ottiene così:

$$n \sigma_0^2 = \sum \lambda_i^2 + g \sigma_0^2 \quad \Rightarrow \quad \sigma_0^2 = \frac{\sum \lambda_i^2}{n - g} \quad (6.13)$$

in perfetta analogia con la (2.5) nella quale da  $n$  misure occorre stimare una sola incognita. Il denominatore di  $\sigma_0^2$  rappresenta il numero di gradi di libertà del problema, ovvero il numero delle misure esuberanti. Noto  $\sigma_0^2$ , sono determinabili le varianze delle incognite (6.7):

$$\sigma_{z_i}^2 = \alpha_{ij} \frac{\sum \lambda_i^2}{n - g} \quad (6.14)$$

Anche in questo caso, è opportuno notare che una differenza significativa fra i valori di  $\sigma_0^2$ , introdotto a priori, per la riduzione di tutte le equazioni allo stesso peso, e calcolato con la (6.13), è indice di probabile presenza di errori sistematici (valendo ancora le note fatte negli esempi 3.3.1, 3.3.2 e 3.3.3).

Nella (6.14),  $n$  e  $g$ , rispettivamente numero di equazioni e di incognite, sono determinate dal problema, gli  $\alpha_{ii}$  sono costanti, dipendenti solo dalla forma delle relazioni (5.1), e  $\sum \lambda_i^2$  è ricavato, applicando il criterio di stima (5.3), e calcolato in modo da risultare minimo., di conseguenza, le stime dei parametri  $\theta$ , ottenute con il metodo dei minimi quadrati, sono non deviate e di minima varianza, come si potrebbe facilmente dimostrare.

Ad esse, sono applicabili le metodologie di inferenza statistica, valide per le *medie campionarie*, con l'avvertenza che, se il numero di gradi di libertà è elevato, nessuna ipotesi aggiuntiva è necessaria, mentre, per bassi gradi di libertà, dovendosi utilizzare la distribuzione  $t$  di *Student*, occorre che le osservazioni fatte si possano considerare appartenenti alla distribuzione normale. I parametri stimati, funzioni lineari delle osservazioni, possono essere considerati come medie campionarie di piccoli campioni, appartenenti a una variabile casuale  $t$  di *Student*, con  $n - g$  gradi di libertà. Lo sqm di  $z_i$ , posto al denominatore del  $t$  di *Student* sperimentale, in questo caso, è  $\sigma_0 \sqrt{\alpha_{ii}}$ , in conseguenza della (6.7).

Sempre con la consueta impostazione, possono eseguirsi test di significatività per le differenze fra due valori stimati con il procedimento di minimi quadrati, provenienti anche da due diversi sistemi normali. Ad esempio, se  $z_i$  e  $z_j^*$  sono due parametri, stimati in base a due diversi gruppi di equazioni agli errori, ciascuno caratterizzato dal proprio sqm dell'unità di peso ( $\sigma_0$  e  $\sigma_0^*$ ) e da diversi valori di  $n$  e  $g$ , si può utilizzare la distribuzione  $t$  di *Student*, per valutare la significatività della differenza fra essi ponendo:

$$H_0: \theta_i = \theta_j^*$$

$$H_1: \theta_i \neq \theta_j^*$$

$$t = \frac{z_i - z_j^*}{\sqrt{\alpha_{ii} \sigma_0^2 + \alpha_{jj}^* \sigma_0^{*2}}} \quad \text{con} \quad v = (n - g) + (n^* - g^*)$$

#### 4.7. Osservazioni indirette con modello non lineare

Nel caso più generale, le relazioni fra le quantità osservate e le incognite non sono lineari, ma possono avere forma qualsiasi, indicata brevemente con:

$$f_i(\theta/x) = 0 \quad i = 1, \dots, n \quad (7.1)$$

ove  $x$  sono certe quantità osservate che possono essere diverse nelle varie equazioni e  $\theta$  il vettore delle  $g$  incognite. In questo caso, si ritorna al modello lineare, determinando valori approssimati  $\theta^0$  di  $\theta$ , cosa solitamente non difficile, dato che, in quasi tutti i problemi di natura fisica, si conoscono, a priori, i valori approssimati delle incognite. Si pone pertanto:

$$\theta_i = \theta_i^0 + v_i$$

e le incognite non sono più  $\theta$ , ma le correzioni  $v_i$ , da apportare ai valori approssimati. L'approssimazione deve essere tale, da poter trascurare i quadrati di  $v_i$  e le potenze superiori. Sviluppando le (7.1), in serie di Taylor, arrestate ai termini lineari, si ha:

$$f_i(\theta/x) = f_i(\theta^0/x) + \left( \frac{\partial f_i}{\partial \theta_1} \right)_{x, \theta^0} v_1 + \dots + \left( \frac{\partial f_i}{\partial \theta_g} \right)_{x, \theta^0} v_g = 0 \quad i = 1, \dots, n \quad (7.2)$$

I termini noti  $f_i(\theta^0/x)$  contengono le quantità osservate  $x$  e le costanti  $\theta^0$ , e sono l'equivalente di  $x_i$  nelle (5.1). Inoltre applicando la (4.4), si possono calcolare le varianze, in funzione di quelle delle quantità osservate, e valutare i pesi delle singole equazioni. In questo caso, i coefficienti delle incognite  $v_i$  non sono costanti, come nelle (5.1), in quanto contengono, anch'essi, le grandezze osservate e non potrebbero, a rigore, essere considerati indipendenti da queste e dai loro errori. In realtà, dato che questi coefficienti moltiplicano quantità piccole  $v_i$ , in teoria, dello stesso ordine di grandezza degli errori accidentali di misura, si può ritenere trascurabile l'effetto che la presenza di errori di misura, nei coefficienti delle incognite, ha nella determinazione delle stesse. Ponendo allora:

$$f_i(\theta^0/x) = x_i \quad \text{e} \quad \left( \frac{\partial f_i}{\partial \theta_1} \right)_{x, \theta^0} = a_i, \quad \left( \frac{\partial f_i}{\partial \theta_2} \right)_{x, \theta^0} = a_i, \quad \dots, \quad \left( \frac{\partial f_i}{\partial \theta_g} \right)_{x, \theta^0} = a_i$$

la (7.2) prende la forma (5.1) e, per essa, vale tutto quanto detto per il metodo dei minimi quadrati, applicato ai modelli lineari.

In quasi tutti i problemi che danno luogo a un modello non lineare, una volta ricavate le correzioni  $v_i$  e le incognite:  $\theta_i^{(1)} = \theta_i^0 + v_i$ , si utilizzano queste ultime, per una nuova linearizzazione, in cui esse fungono da

nuovi valori approssimati. Si ottiene così un nuovo sistema lineare del tipo (7.2) dal quale è possibile ricavare nuovi valori  $\vartheta_i$  (che si possono indicare come  $\vartheta_i^{(1)}$ , contrapponendoli ai precedenti, ora chiamati  $\vartheta_i^{(0)}$ ). Procedendo in questo modo, si ottiene una sequenza di parametri:

$$\begin{aligned} \theta_i^{(1)} &= \theta_i^{(0)} + \vartheta_i^{(0)} \\ \theta_i^{(2)} &= \theta_i^{(1)} + \vartheta_i^{(1)} \\ &\dots\dots\dots \\ \theta_i^{(k+1)} &= \theta_i^{(k)} + \vartheta_i^{(k)} \end{aligned}$$

Le iterazioni proseguono, fino a quando i successivi valori  $\theta_i$  non subiscono più variazioni sensibili, da una iterazione all'altra, cioè finché:

$$|\theta_i^{(k)} - \theta_i^{(k-1)}| \leq \varepsilon_i \quad i = 1, 2, \dots, g$$

dove  $\varepsilon_i$  è una costante prefissata (ad esempio, 1% del valore di  $\theta_i^{(k-1)}$ ).

L'introduzione di valori approssimati delle incognite è eseguita, quasi sempre, anche se le equazioni agli errori sono già di tipo lineare, principalmente per i seguenti motivi.

- Le incognite  $\theta_i$  possono essere di entità molto diverse, mentre è più opportuno, dal punto di vista della soluzione numerica del sistema normale, che siano dello stesso ordine di grandezza. Con l'introduzione dei valori approssimati  $\theta_i^{(0)}$  si ritorna sempre in questa situazione, in quanto le correzioni  $\vartheta_i$ , apportate ad essi (costituenti le nuove incognite del problema), sono tutte necessariamente di piccola entità.
- Il problema può essere di tipo lineare, ma ugualmente contenere valori misurati delle incognite, oltreché ovviamente nel termine noto. In questo caso, non è più possibile calcolare il peso dell'equazione come peso del solo termine noto e l'intero problema della riduzione delle equazioni allo stesso peso presenta aspetti di difficile e spesso dubbia soluzione. Invece l'introduzione dei valori approssimati fa ricadere nel caso in cui tutte le nuove incognite  $\vartheta_i$  sono piccole e, in ciascuno dei termini:  $a_i\vartheta_1, b_i\vartheta_2$ , ecc., la componente errata (dovuta agli errori di misura, presenti nei coefficienti:  $a_i, b_i$ , ecc.) può essere considerata del 2° ordine, rispetto agli errori di misura presenti nel termine noto, e così trascurabile.

**Esempio 4.7.1**

Le differenze di quota (dislivelli) tra quattro punti (caposaldi), nel centro di Milano, sono determinate secondo lo schema indicato in figura 4.7.1. Il procedimento di misura adottato è tale per cui gli scarti quadratici medi dei dislivelli sono proporzionali alle radici quadrate delle distanze, percorse per andare da un caposaldo all'altro (i dislivelli e le distanze sono riportati nella tabella sottostante).

Le incognite sono le quote dei caposaldi; le quantità osservate i dislivelli e, dato che si cercano le quote *relative*, occorre conoscere il valore della quota di un caposaldo qualunque od assegnarla arbitrariamente (in questo caso, si è assunta la quota convenzionale:  $Q_1 = 120.000 \text{ m}$ , per il caposaldo di Brera).

Le equazioni agli errori, in numero di sei, hanno la forma semplicissima e risultano lineari, nelle tre incognite  $Q_2$ ,  $Q_3$  e  $Q_4$ :  $q_{ij} - (Q_i - Q_j) = v_{ij}$  (essendo:  $Q_i$  la quota del caposaldo i-esimo).

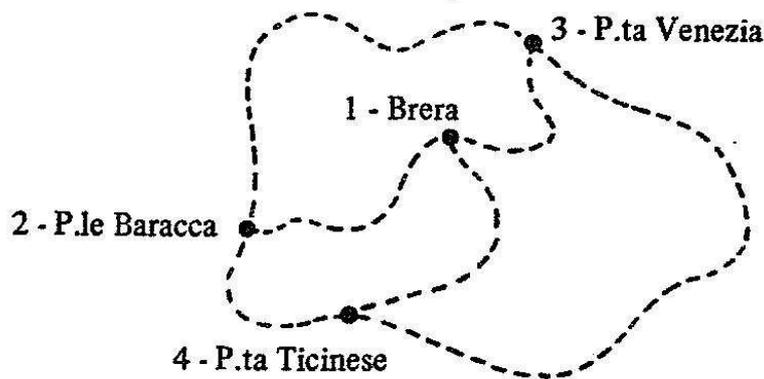


Fig. 4.7.1 – Schema di distanze e dislivelli misurati

	<i>Dislivelli misurati</i>	<i>Distanze</i>
<i>Brera-P.ta Venezia</i>	$q_{13} = +177.4 \text{ mm}$	$d_{13} = 1.74 \text{ km}$
<i>P.ta Venezia-P.ta Ticinese</i>	$q_{34} = +5584.8 \text{ mm}$	$d_{34} = 4.40 \text{ km}$
<i>P.ta Ticinese-Brera</i>	$q_{41} = -5763.3 \text{ mm}$	$d_{41} = 3.25 \text{ km}$
<i>P.ta Ticinese-P.le Baracca</i>	$q_{42} = -4953.5 \text{ mm}$	$d_{42} = 2.43 \text{ km}$
<i>P.le Baracca-Brera</i>	$q_{21} = -809.4 \text{ mm}$	$d_{21} = 2.49 \text{ km}$
<i>P.le Baracca- P.ta Venezia</i>	$q_{23} = -634.4 \text{ mm}$	$d_{23} = 4.65 \text{ km}$

Per i sei dislivelli misurati si ha il seguente sistema, esprimendo le quote ed i dislivelli in mm:

$$\begin{cases} + 177.4 - (120.000 - Q_3) = v_{13} \\ + 5584.8 - (Q_3 - Q_4) = v_{34} \\ - 5763.3 - (Q_4 - 120.000) = v_{41} \\ - 4953.5 - (Q_4 - Q_2) = v_{42} \\ - 809.4 - (Q_2 - 120.000) = v_{21} \\ - 634.4 - (Q_2 - Q_3) = v_{23} \end{cases}$$

Anche se non necessario, anche in questo caso, è opportuno assumere valori approssimati delle quote, affinché le incognite, cioè le correzioni da apportare alle quote approssimate, risultino piccole ed i calcoli possano essere effettuati con poche cifre significative. Le quote approssimate sono ricavate aggiungendo alla quota di Brera tre dislivelli:  $q_{21}$ ,  $q_{31}$ ,  $q_{41}$ , arrotondati al mm:  $Q_2^0 = 119.191 \text{ m}$ ;  $Q_3^0 = 119.823 \text{ m}$  e  $Q_4^0 = 114.237 \text{ m}$ . Dopodiché indicando con  $v_i$  le tre correzioni da apportare a  $Q_i^0$  (poiché  $v_1 = 0$ ), cioè ponendo:  $Q_i = Q_i^0 + v_i$ , il sistema delle equazioni agli errori diventa:

$$\left\{ \begin{array}{l} +0.4 - ( \quad -v_3 \quad ) = v_{13} \\ -1.2 - ( \quad v_3 \quad -v_4 ) = v_{34} \\ -0.3 - ( \quad \quad v_4 \quad ) = v_{41} \\ +0.5 - ( -v_2 \quad v_4 \quad ) = v_{42} \\ -0.4 - ( v_2 \quad \quad ) = v_{21} \\ -2.4 - ( v_2 \quad -v_3 \quad ) = v_{23} \end{array} \right.$$

Le equazioni precedenti hanno diverso peso, poiché, come detto, gli scarti quadratici medi dei termini noti, cioè dei dislivelli, dipendono dalla distanza. Detta  $\sigma_0^2$  la varianza per la distanza di 1 km, si ha:  $\sigma_{ij}^2 = \sigma_0^2 d_{ij}$  e poiché:  $p_{ij} = \sigma_0^2 / \sigma_{ij}^2$  risulta:  $p_{ij}^2 = 1/d_{ij}$ . Nello specchio seguente, sono indicati i pesi e le loro radici, cioè le quantità per cui devono essere moltiplicate le equazioni agli errori per ridurle allo stesso peso:

$$\begin{array}{l} d_{ij} = \quad 1.74 \quad 4.40 \quad 3.25 \quad 2.43 \quad 2.49 \quad 4.65 \\ p_{ij} = \quad 0.57 \quad 0.23 \quad 0.31 \quad 0.41 \quad 0.40 \quad 0.22 \\ \sqrt{p_{ij}} = \quad 0.76 \quad 0.48 \quad 0.55 \quad 0.64 \quad 0.63 \quad 0.46 \end{array}$$

Di seguito, sono riportate le equazioni agli errori, ridotte allo stesso peso, la matrice dei coefficienti delle incognite ed i termini noti:

$$\left\{ \begin{array}{l} +0.30 - ( \quad -0.76v_3 \quad ) = v_{13} \\ -0.58 - ( \quad +0.48v_3 \quad -0.48v_4 ) = v_{34} \\ -0.17 - ( \quad \quad +0.55v_4 \quad ) = v_{41} \\ +0.32 - ( -0.64v_2 \quad \quad +0.64v_4 ) = v_{42} \\ -0.25 - ( +0.63v_2 \quad \quad ) = v_{21} \\ -1.10 - ( +0.46v_2 \quad -0.46v_3 ) = v_{23} \end{array} \right.$$

$$A = \begin{vmatrix} 0 & -0.76 & 0 \\ 0 & +0.48 & -0.48 \\ 0 & 0 & +0.55 \\ -0.64 & 0 & +0.64 \\ +0.63 & 0 & 0 \\ +0.46 & -0.46 & 0 \end{vmatrix} \quad x = \begin{vmatrix} +0.30 \\ -0.58 \\ -0.17 \\ +0.32 \\ -0.25 \\ -1.10 \end{vmatrix}$$

Le matrici  $D$ ,  $D^{-1}$ ,  $A^T x$  ed il sistema normale risultano:

$$D = \begin{vmatrix} 1.02 & -0.21 & -0.41 \\ -0.21 & 1.02 & -0.23 \\ -0.41 & -0.23 & +0.94 \end{vmatrix} \quad D^{-1} = \begin{vmatrix} 1.348 & 0.434 & 0.694 \\ 0.434 & 1.177 & 0.477 \\ 0.694 & 0.477 & 1.483 \end{vmatrix} \quad A^T x = \begin{vmatrix} -0.87 \\ 0.00 \\ +0.39 \end{vmatrix}$$

$$\begin{cases} 1.02v_2 - 0.21v_3 - 0.41v_4 = -0.87 \\ -0.21v_2 + 1.02v_3 - 0.23v_4 = 0.00 \\ -0.41v_2 - 0.23v_3 + 0.94v_4 = 0.39 \end{cases} \quad \text{da cui: } \begin{cases} v_2 = -1.348 \cdot 0.87 + 0.694 \cdot 0.39 = -0.90 \\ v_3 = -0.434 \cdot 0.87 + 0.477 \cdot 0.39 = -0.19 \\ v_4 = -0.694 \cdot 0.87 + 1.483 \cdot 0.39 = -0.03 \end{cases}$$

I valori stimati delle quote, approssimati a 0.1 mm, sono:

$$\begin{aligned} Q_1 &= 120.0000 & Q_2 &= 119.191 - 0.0009 = 119.1901 \\ Q_3 &= 119.823 - 0.0002 = 119.8228 & Q_4 &= 114.237 - 0.0000 = 114.2370 \end{aligned}$$

Per ottenere le varianze delle incognite, si applica la (6.7):

$$\alpha_{22} = 1.348 \qquad \alpha_{33} = 1.177 \qquad \alpha_{44} = 1.483$$

Il valore di  $\sigma_0^2$  è dato dalla formula (6.13), con  $n = 6$ ,  $g = 3$ ;  $\lambda$  sono i residui delle equazioni agli errori, ridotte allo stesso peso, ottenuti introducendo in essi i valori calcolati delle incognite.

$$\begin{aligned} \lambda_1 &= +0.30 \quad -0.14 & & = +0.16 \\ \lambda_2 &= -0.58 \quad +0.09 \quad -0.01 & & = -0.50 \\ \lambda_3 &= -0.17 \quad +0.02 & & = -0.15 \\ \lambda_4 &= +0.32 \quad -0.58 \quad +0.02 & & = -0.24 \\ \lambda_5 &= -0.25 \quad +0.57 & & = +0.32 \\ \lambda_6 &= -1.10 \quad +0.41 \quad -0.09 & & = -0.78 \end{aligned}$$

Si ha pertanto:

$$\sigma_0^2 = \frac{1.07}{3} = 0.36 \qquad \sigma_0 = 0.60 \text{ mm (sqm chilometrico)}$$

ed allora:

$$\sigma_2 = \sqrt{1,348} \cdot 0,60 = 0,70 \quad \sigma_3 = \sqrt{1,177} \cdot 0,60 = 0,65 \quad \sigma_4 = \sqrt{1,483} \cdot 0,60 = 0,73 .$$

#### Esempio 4.7.2

Una precedente serie di misure, eseguite con lo stesso riferimento a Brera, dà per il caposaldo di P.ta Venezia una quota  $Q_3^* = 119.8251 \text{ m}$  (con sqm  $\sigma_3^* = 0.40 \text{ mm}$ ). Se si vuole sapere se l'abbassamento relativo di P.ta Venezia rispetto a Brera è significativo, dato che gli errori di misura si possono considerare normalmente distribuiti ed indipendenti, si può applicare il test di *Student*, con un livello di significatività, ad esempio, del 5%. Esprimendo quote e sqm in mm, si ha:

$$t = \frac{(119825.1 - 119822.8) - 0}{\sqrt{0.40^2 + 0.65^2}} = \frac{2.3}{0.76} = 3.026$$

con un numero dei gradi di libertà pari alla somma di quelli parziali:  $\nu = (n - g) + (n^* - g^*) = 3 + 3 = 6$ .  
Dalle tavole, si ha:  $t(\nu = 6, \alpha = 5\%) = 1.94$ , e l'abbassamento risulta significativo.

#### 4.8. Varianza di una funzione di quantità osservate indirettamente

Data una qualsiasi relazione analitica:  $z = F(z_1, z_2, \dots, z_g)$ , per ricavare una grandezza  $z$ , in funzione di un gruppo di altre grandezze:  $z_1, z_2, \dots, z_g$ , se le stime di  $z_i$  sono ottenute con un procedimento come quello descritto nei paragrafi 3.6 o 3.7, le misure:  $z_1, z_2, \dots, z_g$ , non sono fra loro indipendenti. Pertanto ricordando la (4.4), la varianza di  $z$  è:

$$\sigma_z^2 = h_1^2 \sigma_{z_1}^2 + h_2^2 \sigma_{z_2}^2 + \dots + h_g^2 \sigma_{z_g}^2 + \dots + 2h_i h_j \sigma_{z_i} \sigma_{z_j} + \dots \quad (8.1)$$

e sostituendo nella (8.1) le (6.7) e le (6.8), si ottiene:

$$\sigma_z^2 = (h_1^2 \alpha_{11} + h_2^2 \alpha_{22} + \dots + h_g^2 \alpha_{gg} + \dots + 2h_i h_j \alpha_{ij} + \dots) \sigma_0^2 \quad (8.2)$$

ove  $\alpha_{ij}$  sono i termini della matrice inversa della matrice normale con cui si ricavano  $z_i$ ,  $\sigma_0^2$  è l'errore medio della unità di peso delle equazioni agli errori e  $h$  sono le derivate rispetto a;  $z_1, z_2, \dots, z_g$  della funzione  $F$ .

##### Esempio 4.8.1

Volendo calcolare lo sqm del dislivello fra P.ta Venezia e P.ta Ticinese, in base ai dati dell'esempio 3.7.1, il dislivello in questione può essere indicato come:  $\Delta_{34} = Q_4 - Q_3 = 114.2370 - 119.8228 = -5.5858 \text{ m}$ . Le quote  $Q_3$  e  $Q_4$  non sono indipendenti per cui, in primo luogo, si può cercare il coefficiente di correlazione fra esse  $r_{34}$  che, in base alla (6.8), risulta:

$$r_{34} = \frac{\alpha_{34}}{\sqrt{\alpha_{33} \cdot \alpha_{44}}} = \frac{0.477}{\sqrt{1.177 \cdot 1.483}} = 0.361$$

Tuttavia il calcolo di  $\sigma_{\Delta}^2$  non richiede il valore di  $r_{34}$ , in quanto nella (8.2) compaiono solo  $\alpha_{ij}$  e le derivate della funzione  $\Delta_{34}$ , rispetto a  $Q_3$  e  $Q_4$ , le quali sono rispettivamente uguali a:  $+1$  e  $-1$ . Si ha allora:

$$\sigma_{\Delta_{34}}^2 = (\alpha_{33} + \alpha_{44} - 2\alpha_{34}) \sigma_0^2 = (1.177 + 1.483 - 2(0.477)) \sigma_0^2 = 0.614 \quad \text{e:} \quad \sigma_{\Delta_{34}} = 0.78 \text{ mm}$$

## PARTE V – REGRESSIONE LINEARE MULTIPLA

### 5.1 Regressioni e relazioni funzionali

Nel problema della regressione lineare, problema fondamentale in quasi tutti i campi di applicazione delle teorie statistiche, la generica equazione:

$$y = b_0 + b_1 x \quad (1.1)$$

esprime la dipendenza fra una variabile indipendente  $x$  e la variabile dipendente  $y$ , potendosi riferire, nella stessa forma, a due problemi sostanzialmente diversi, una regressione (propriamente detta) e relazioni funzionali, in base tipo di universo dal quale sono estratte  $x$  e  $y$ .

Un esempio è dato dalla differenza di una relazione del tipo (1.1), fra l'altezza ed il peso degli individui di una certa regione, e con la stessa relazione, usata come taratura di un termometro a platino, in cui la  $y$  è la resistenza e  $x$  la temperatura. Infatti nel primo caso esiste una effettiva variabilità strutturale fra le due variabili, in quanto il comportamento della variabile doppia  $(x, y)$  può essere rappresentato compiutamente solo in modo stocastico, con la distribuzione congiunta dell'altezza e del peso.

Se questa distribuzione è di tipo normale, le equazioni delle due curve di regressione, luogo geometrico delle medie di  $y$  condizionata da  $x$  e viceversa, sono rette e la regressione è di tipo lineare, rappresentabile con un'equazione del tipo (1.1) o dall'analoga:

$$x = b'_0 + b'_1 y \quad (1.2)$$

Pertanto le (1.1) ed 1.2) possono essere usate per predire il valor medio di  $y$ , in corrispondenza di un prefissato valore  $x_k$  di  $x$ , oppure il valor medio di  $x$ , in corrispondenza di un prefissato valore  $y_k$  di  $y$ , purchè i valori osservati in base ai quali sono stimati i parametri:  $b_0$ ,  $b_1$ ,  $b'_0$  e  $b'_1$ , siano un campione preso a caso dall'universo.

Se una sola delle due relazioni interessa, ad esempio:  $y = b_0 + b_1 x$ , si può limitare la casualità dell'estrazione solo a  $y$  in corrispondenza a prefissati (e non casuali) valori di  $x$ , cosa a volte molto conveniente negli esperimenti pianificati. Ovviamente l'insieme dei valori  $x$  e  $y$ , così raccolti non permette di stimare contemporaneamente i parametri  $b'_0$  e  $b'_1$ .

Nel secondo caso, la distribuzione congiunta è priva di significato, perché il legame fra temperatura e resistenza, in assenza di errori di misura, dovrebbe essere rappresentato rigorosamente da una relazione funzionale, invertibile e valida per predire l'effettivo valore  $y_k$ , corrispondente ad un  $x_k$ , o viceversa.

Infatti una relazione funzionale lineare esiste laddove le variazioni, intorno ad un'opportuna retta, possono essere attribuite solo ad errori di misura. Fra questi ultimi, devono essere prevalenti quelli della variabile dipendente, o meglio, è assunta, come variabile indipendente, quella che può essere, più facilmente, fissata su certi predeterminati valori, praticamente senza errore, valori in corrispondenza dei quali sono effettuate le

misure della variabile dipendente. Se così non fosse, l'equazione della regressione (1.1) darebbe una stima deviata della relazione funzionale, dove l'entità del bias dipende dagli errori della variabile indipendente e dall'intervallo su cui si estendono le misure di questa.

Allora laddove si vogliono stimare i parametri di una relazione funzionale, è opportuno predisporre un esperimento controllato che permetta di fissare i valori della variabile indipendente, con un alto grado di precisione e sull'intervallo più esteso possibile.

Per discriminare fra questi due possibili significati dell'equazione (1.1), si conviene d'indicare con le lettere maiuscole i valori previsti, cioè i valori medi della variabile dipendente, in corrispondenza di un certo valore presissato della variabile indipendente, cosicchè la (1.1), intesa come regressione (e non come relazione funzionale) risulta:

$$Y = b_0 + b_1 x$$

mentre le lettere minuscole continuano a rappresentare le osservazioni o le variabili casuali da cui sono estratte.

## 5.2 Stima dei parametri e scomposizione degli scarti

Per stimare i parametri  $b_0$  e  $b_1$  dell'equazione:

$$Y = b_0 + b_1 x \tag{2.1}$$

se  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  sono le osservazioni, a disposizione, ed  $\bar{x}$  e  $\bar{y}$  le loro medie, si può ridurre la stessa equazione nella forma:

$$Y = \bar{y} + b_1(x - \bar{x}) \tag{2.2}$$

dove:

$$b_1 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \tag{2.3}$$

Volendo riesaminare quanto sopra alla luce della teoria delle osservazioni indirette, svolta nella Parte III, si hanno una serie di  $n$  equazioni agli errori del tipo:

$$(y_i - \bar{y}) - b_1(x_i - \bar{x}) = v_i \quad i = 1, 2, \dots, n \quad \text{dove} \quad v_i = y_i - Y_i$$

con un'unica incognita  $b_1$ . Pertanto la matrice normale e la sua inversa sono rispettivamente gli scalari:

$$D = \sum(x_i - \bar{x})^2 \quad \text{e} \quad D^{-1} = \frac{1}{\sum(x_i - \bar{x})^2} \tag{2.4}$$

mentre il vettore dei termini noti normalizzati è:

$$A^T (y - \bar{y}) = \sum (x_i - \bar{x})(y_i - \bar{y}) \quad (2.5)$$

Dalle (2.4) e (2.5) si ricava subito la (2.3):

$$\vartheta = b_1 = D^{-1} A^T (y - \bar{y}) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \text{e} \quad b_0 = \bar{y} - b_1 \bar{x}$$

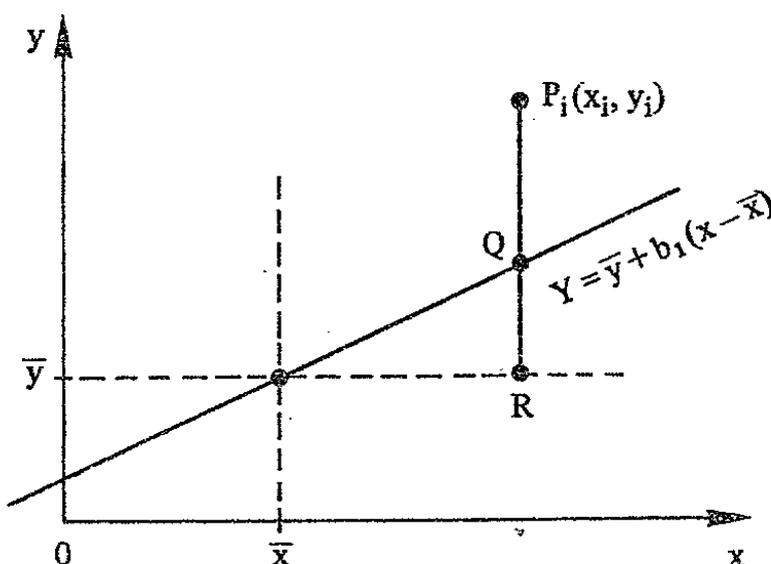
Sostituendo ora la stima ricavata per  $b_1$  nelle equazioni agli errori, si ottengono i valori di  $v_i$  con cui calcolare la stima della varianza dell'unità di peso  $\sigma_0^2$  (cui competono  $n - 2$  gradi di libertà, perché due sono incognite del problema e  $n$  le coppie di osservazioni, a disposizione). L'ausiliaria del peso  $\alpha_{11}$  è l'unico termine della matrice  $D^{-1}$  da cui:

$$\sigma_{b_1}^2 = \sigma_0^2 \frac{1}{\sum (x_i - \bar{x})^2} = \frac{\sum v_i^2}{n - 2} \cdot \frac{1}{\sum (x_i - \bar{x})^2} = \frac{\sum (y_i - Y_i)^2}{(n - 2) \sum (x_i - \bar{x})^2} \quad (2.6)$$

Il numeratore della (2.6) è costituito dalla somma dei quadrati degli scarti fra  $y_i$  osservato ed i valori della retta di regressione (2.1), in corrispondenza delle rispettive  $x_i$ . Questi scarti:  $y_i - Y_i$ , sono detti *scarti intorno alla regressione* (o *scarti dalla regressione*), per contrapporli a due altri tipi di scarti che si possono individuare per ciascun punto  $P_i(x_i, y_i)$ . Infatti nella figura 5.2.1 il segmento  $P_i R$  può essere scomposto in:

$P_i R = P_i Q + QR$ , cosicché:

$$(y_i - \bar{y}) = (y_i - Y_i) + (Y_i - \bar{y}) \quad (2.7)$$



$y_i - \bar{y} = \text{scarto dalla media}$   
 $y_i - Y_i = \text{scarto dalla regressione}$   
 $Y_i - \bar{y} = \text{scarto della regressione dalla media}$

Fig. 5.2.1 – Rappresentazione dei tre tipi di scarti

Dall'identità (2.7), si può ricavare:

$$\sum(y_i - \bar{y})^2 = \sum((y_i - Y_i) + (Y_i - \bar{y}))^2 = \sum(y_i - Y_i)^2 + \sum(Y_i - \bar{y})^2 + 2 \sum(y_i - Y_i)(Y_i - \bar{y})$$

dove l'ultimo termine è nullo, per il principio dei minimi quadrati:

$$\begin{aligned} \sum(y_i - Y_i)(Y_i - \bar{y}) &= \sum(y_i - Y_i)b_1(x_i - \bar{x}) = \sum((y_i - \bar{y}) - b_1(x_i - \bar{x}))b_1(x_i - \bar{x}) = \\ &= \sum\left((y_i - \bar{y}) - \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}(x_i - \bar{x})\right) \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}(x_i - \bar{x}) = \\ &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \sum(x_i - \bar{x})(y_i - \bar{y}) - \left(\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right)^2 \sum(x_i - \bar{x})^2 = \\ &= \frac{(\sum(x_i - \bar{x})(y_i - \bar{y}))^2}{\sum(x_i - \bar{x})^2} - \frac{(\sum(x_i - \bar{x})(y_i - \bar{y}))^2}{\sum(x_i - \bar{x})^2} = 0 \end{aligned}$$

avendo sostituito il coefficiente  $b_1$  e gli scarti:  $v_i = y_i - Y_i$ , con il loro valore, cosicché:

$$\sum(y_i - \bar{y})^2 = \sum(y_i - Y_i)^2 + \sum(Y_i - \bar{y})^2 \qquad S_G^2 = S_R^2 + S_S^2 \qquad (2.8)$$

La relazione (2.8) è fondamentale per l'analisi della regressione. Infatti significa che la somma dei quadrati degli scarti dalla media  $S_G^2$  (*scarti generali*) può essere scomposta nella somma dei quadrati *scarti residui* intorno alla regressione  $S_R^2$  più la somma dei quadrati degli scarti della regressione dalla media  $S_S^2$  (*scarti spiegati*, come spiegato nel seguito). Analogamente a quanto visto nel paragrafo sull'analisi di varianza, si possono anche scomporre i rispettivi gradi di libertà  $\nu$ :  $n-1 = (n-2) + 1$ , ed il rapporto fra ogni  $S^2$  ed il rispettivo  $\nu$  fornisce una diversa stima della varianza.

Di conseguenza, si può costruire una tabella, analoga a quelle per l'analisi di varianza, in cui compaiono il tipo degli scarti, la entità degli scarti  $S^2$  corrispondenti, i rispettivi gradi di libertà, e le varianze ricavate.

Scarti dovuti alla regressione	$\sum(Y_i - \bar{y})^2 = b_1^2 \sum(x_i - \bar{x})^2$	1	$\sigma_S^2$
Scarti intorno alla regressione	$\sum(y_i - Y_i)^2 = \sum(y_i - \bar{y})^2 - b_1^2 \sum(x_i - \bar{x})^2$	$n-2$	$\sigma_0^2 = \sigma_R^2$
Scarti totali dalla media	$\sum(y_i - \bar{y})^2$	$n-1$	$\sigma_G^2$

La notazione  $\sigma_S^2$  proviene da un'espressione, abbastanza adottata, che chiama scarti spiegati quelli dei punti della regressione rispetto alla media. Infatti se il modello (2.1) è corretto, la variabilità di  $y$  intorno ad  $\bar{y}$  è *spiegata*, almeno per una frazione, dalla retta di regressione, costituente il modello stesso.

Un'espressione equivalente di  $S_S^2$  chiarisce ulteriormente il concetto; infatti introducendo il coefficiente di correlazione lineare  $r$ ,  $S_S^2$  può essere così riscritto:

$$S_S^2 = b_1^2 \sum (x_i - \bar{x})^2 = \left( \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \right)^2 \sum (x_i - \bar{x})^2 = \left( \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \right)^2 \sum (y_i - \bar{y})^2 = r^2 \sum (y_i - \bar{y})^2 = r^2 S_G^2 \quad (2.9)$$

dove  $r^2$ , detto anche *indice di determinazione*, spiega la percentuale di  $S_G^2$  giustificata dalla regressione. Per contro, se l'equazione di regressione fosse nota con esattezza o stimata in base ad un numero molto elevato di punti, lo sqm  $\sigma_0$  intorno alla regressione rappresenterebbe l'errore con cui si potrebbe predire il valore da osservarsi per  $y$ , in corrispondenza di un predeterminato valore  $x$ . Per questo motivo,  $\sigma_0^2$  è detto varianza degli errori o varianza residua ed è sempre di fondamentale importanza in quanto la precisione con cui si arriva a determinare l'equazione di regressione, sulla base degli elementi di un campione, dipende dalla sua entità. Infatti dalla (2.6) lo sqm di  $b_1$  è dato da:

$$\sigma_{b_1} = \sigma_0 / \sqrt{\sum (x_i - \bar{x})^2} \quad (2.10)$$

ed i limiti fiduciali di  $b_1$ , all'  $(1 - 2\alpha)\%$ , sono:

$$\beta_1 = b_1 \pm t_\alpha \sigma_0 / \sqrt{\sum (x_i - \bar{x})^2}$$

dove  $t_\alpha$  è ricavato dalle tavole con  $\nu = n - 2$ . Anche gli scarti  $S_R^2$  possono essere introdotti in una espressione deducibile dalla (2.9), dato l'indice di determinazione  $r^2$ :

$$S_R^2 = \sum (y_i - Y_i)^2 = (1 - r^2) \sum (y_i - \bar{y})^2 \quad \text{da cui: } r^2 = 1 - \frac{\sum (x_i - Y_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{S_R^2}{S_G^2} \quad (2.11)$$

dove  $S_R^2/S_G^2$  è la percentuale della variabilità generale non spiegata dal modello di regressione. Pertanto  $r^2$  ha valori compresi fra 0 ed 1:

- $r^2 = 0$ , per  $S_R^2 = S_G^2$ , cioè quando la componente spiegata dalla regressione  $S_S^2$  è nulla e la retta di regressione è la parallela dell'asse  $x$ , passante per  $\bar{y}$ ;
- $r^2 = 1$ , se  $S_R^2 = 0$ , ovvero quando tutti i punti, rappresentanti le osservazioni, stanno sulla retta di regressione, essendo così:  $S_G^2 = S_S^2$ .

Tuttavia la significatività del modello è di solito stabilita in base al rapporto  $\sigma_S^2/\sigma_R^2$  che, se le osservazioni sono indipendenti e normalmente distribuite, segue distribuzione  $F$  di Fisher, con 1 e  $(n-2)$  gradi di libertà. Allora in base al livello di significabilità  $\alpha$  prefissato, si stabilisce se respingere (o meno) l'ipotesi di adeguatezza del modello (2.1), come rappresentativo del fenomeno.

### 5.3. Varianza dei vari elementi della regressione

La (2.10) fornisce lo sqm del coefficiente  $b_1$ , ma anche  $b_0$ , l'altro parametro da cui dipende la regressione, è soggetto ad errore che determina un possibile spostamento della retta, parallelamente a se stessa. Dato che si può dimostrare che  $\bar{y}$  e  $b_1$  sono indipendenti, la varianza  $\sigma_{b_0}^2$  può essere ricavata applicando la legge di propagazione degli scarti nel caso di variabili casuali indipendenti alla relazione:

$$\sigma_{b_0}^2 = \sigma_{\bar{y}}^2 + \bar{x}^2 \sigma_{b_1}^2 = \frac{\sigma_0^2}{n} + \bar{x}^2 \frac{\sigma_0^2}{\sum(x_i - \bar{x})^2} = \sigma_0^2 \frac{\sum(x_i - \bar{x})^2 + n\bar{x}^2}{n \sum(x_i - \bar{x})^2} = \sigma_0^2 \frac{\sum x_i^2}{n \sum(x_i - \bar{x})^2} \quad (3.1)$$

dove per il calcolo di  $\sigma_{\bar{y}}^2$  si è usata formula  $\sigma_0^2/n$  e, ad esempio, non  $\sigma_G^2/n$ , in quanto la  $\sigma_G^2$  non rappresenta solo la varianza di tipo stocastico da cui sono affette le osservazioni  $y$ , ma contiene, oltre a questa, tutta la variabilità sistematica, indotta in  $y$ , dal variare di  $x$ .

L'entità di  $\sigma_0$ , invece, misura la variabilità puramente stocastica delle  $y$ , ossia è un indice della loro dispersione *intorno* alla retta di regressione.

La (3.1) si sarebbe potuta ricavare anche direttamente con la (3.6.7), purché si fossero mantenute in evidenza, nel sistema normale, entrambe le incognite  $b_0$  e  $b_1$ , invece di eliminare la prima con la traslazione dell'origine nel punto  $(\bar{x}, \bar{y})$ .

L'indipendenza di  $\bar{y}$  e  $b_1$  permette anche di ricavare la varianza del valore  $Y_k$ , stimato tramite la regressione in corrispondenza di  $x_k$ . Infatti dato:  $Y_k = \bar{y} + b_1(x_k - \bar{x})$ , si ha:

$$\sigma_{Y_k}^2 = \frac{\sigma_0^2}{n} + (x_k - \bar{x})^2 \sigma_{b_1}^2 = \sigma_0^2 \left( \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right) \quad (3.2)$$

D'altra parte, la (3.2) rappresenta solo la varianza dovuta a variazioni casuali, nell'ambito del modello di un punto della regressione, cioè del suo valor medio  $Y_k$ , in funzione di  $x_k$  corrispondente, mentre un valore osservato  $y$  può ulteriormente variare, intorno alla regressione, con sqm  $\sigma_0$ . Queste due variazioni sono indipendenti, per cui, quando si vuole usare la (2.2) come stima di  $y$  che si osserva in corrispondenza a  $x_k$ , la varianza di questa stima è la somma di due, quella del valor medio e quella intorno al valor medio:

$$\sigma_{y_k}^2 = \sigma_0^2 \left( 1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right) \quad (3.3)$$

ed i limiti fiduciali di  $y_k$ , all'  $(1 - 2\alpha)\%$ , sono:

$$= Y_k \pm t_\alpha \sigma_0 \sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \quad (3.4)$$

La (3.4) mostra che l'ampiezza dei limiti fiduciali è una funzione di  $x$ , minima per  $x_k = \bar{x}$  e crescente, come rappresentato in Fig. 5.3.1, secondo l'equazione di un'iperbole.

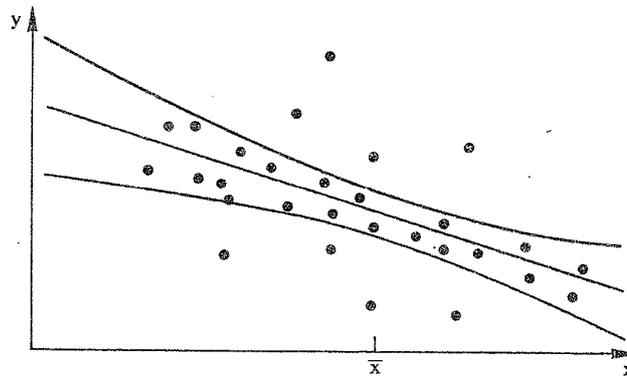


Fig. 5.3.1 – Limiti fiduciali intorno alla retta di regressione (entro questa zona devono cadere l'  $(1 - 2\alpha)\%$  dei punti della popolazione)

Qualora interessi la regressione di  $x$  su  $y$ , tutte le formule precedenti si possono invertire, scambiando  $x$  con  $y$  ed arrivando, in generale, ad un diverso valore  $b'_1$  del coefficiente di regressione, ovvero ad una diversa retta di regressione. Invece solo nel caso in cui si abbiano fondati motivi, per ritenere che la regressione rappresenti la relazione lineare funzionale fra le due variabili, si può dedurre la stima di  $x$ , corrispondente ad un dato  $y_k$ , come:

$$x_k = (y_k - \bar{y})/b_1 + \bar{x}$$

con una semplice inversione della (2.2), anche se questo modo di procedere ha senso solo nel caso in cui i valori  $x_k$  siano determinati con un errore trascurabile.

#### 5.4 Regressione lineare multipla

Se la percentuale degli scarti residui, rispetto a quelli generali, cioè  $S_R^2/S_G^2$ , è molto elevata, si può supporre che il modello ipotizzato non tenga conto di qualche fattore determinante, ovvero che, in realtà,  $y$  dipenda non solo da  $x$ , ma da una o più altre variabili. Ammettendo che la dipendenza sia di tipo lineare, la (2.1) può essere sostituita, ad esempio, con la relazione:

$$Y = b_0 + b_1x_1 + b_2x_2 \tag{4.1}$$

dove i valori dei parametri, presenti nella (4.1), non sono gli stessi ricavati con le due regressioni semplici di  $y$  su  $x_1$  o di  $y$  su  $x_2$ . Per evitare ogni confusione si dovrebbe usare una notazione più complessa:

$$Y = b_{y,12} + b_{y1,2} + b_{y2,1}x_2 \tag{4.2}$$

e per le due regressioni semplici:

$$Y = b_{y.1} + b_{y1}x_1 \quad \text{e} \quad Y = b_{y.2} + b_{y2}x_2 \quad (4.3)$$

Il posizionamento dei punti chiarisce il tipo di dipendenza; ad esempio,  $b_{y.12}$  significa che  $y$  è la variabile dipendente e  $x_1$  e  $x_2$  sono quelle indipendenti, mentre  $b_{y.1.2}$  rappresenta il legame fra  $y$  e  $x_1$ , dove  $x_2$  rappresenta invece la variabile extra, introdotta per giustificare matematicamente la variabilità (o parte di essa), rimasta fra i dati dopo l'assunzione, come modello, della prima delle (4.3). Analogo significato ha il coefficiente  $b_{y2.1}$ . Questa simbologia facilita anche la comprensione della differenza concettuale fra i coefficienti di regressione parziali (4.2) e totali (4.3).

Il coefficiente  $b_{y1.2}$  rappresenta l'effetto su  $y$  di un aumento unitario in  $x_1$ , quando  $x_2$  costante, costituendo così l'effetto netto di  $x_1$  su  $y$ . Analogamente  $b_{y2.1}$  misura l'incremento in  $y$ , dovuto ad un incremento unitario in  $x_2$ , con  $x_1$  costante, ovvero l'effetto netto di  $x_2$  su  $y$ . Per contro, nella (4.3),  $b_{y1}$  rappresenta l'effetto su  $y$  di un aumento unitario in  $x_1$ , quando  $x_2$  può variare senza restrizioni, e costituisce l'effetto totale di  $x_1$  su  $y$ , comprendente anche eventuali mutue influenze, esplicate attraverso  $x_2$ . Ad esempio, può capitare che gli effetti spaziali siano più grandi di quelli totali, a causa di un coefficiente di correlazione negativo fra  $x_1$  e  $x_2$  che determina, nell'ambito di un campione, la presenza di elevati valori di  $x_1$  a fronte di bassi valori di  $x_2$ . In questo modo, l'effetto totale di  $x_1$  su  $y$  è parzialmente annullato dall'effetto contrario di  $x_2$  su  $y$  che esiste, anche se nelle (4.3) non si dà adeguata formulazione matematica.

Chiarita l'importante distinzione fra coefficienti di regressione parziali e totali, nel caso più semplice, si può passare ad indicare la generica equazione in cui compaiono  $p$  variabili indipendenti:

$$Y = b_{y.12\dots p} + b_{y1.2\dots p}x_1 + b_{y2.1\dots p}x_2 + \dots + b_{yp.1\dots p-1}x_p$$

la quale, per non appesantire troppo le notazioni, è indicata nella forma analoga alla (4.1):

$$Y = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p \quad (4.4)$$

Le relazioni fra le due simbologie sono evidenti e quella più complessa è usata solo se indispensabile per evitare confusioni. Anche in questo caso, riferendo tutte le osservazioni ai loro valori medi, si può eliminare dalla (4.4) il termine costante:

$$Y = \bar{y} + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + \dots + b_p(x_p - \bar{x}_p) \quad (4.5)$$

Nel sistema normale che consente di ricavare i coefficienti  $b$ , compaiono tutti i termini del tipo:

$$C_{ii} = \sum_k (x_{ik} - \bar{x}_i)^2$$

$$C_{ij} = \sum_{k,v} (x_{ik} - \bar{x}_i)(x_{jv} - \bar{x}_j) \quad (4.6)$$

$$C_{yi} = \sum_k (y_i - \bar{y})(x_{ik} - \bar{x}_i)$$

dove le lettere  $C$  si usano per richiamare la natura di covarianze di questi termini. Per analogia, s'introduce:

$$C_{yy} = \sum (y_i - \bar{y})^2 = S_G^2 \quad (4.7)$$

Infine in questo caso, i termini della matrice  $D^{-1}$  sono indicati con  $C^{ij}$  (con gli indici in alto), invece che con il simbolo usuale  $\alpha_{ij}$ , per uniformarsi alla più diffusa letteratura di programmi per calcolatori.

I valori di  $b$  sono calcolati sempre con il procedimento consueto delle osservazioni indirette, tenendo conto che i termini noti sono espressi dal vettore  $(y - \bar{y})$ :

$$\theta = b = D^{-1}A^T(y - \bar{y}) \quad (4.8)$$

Per il calcolo di  $\sigma_0^2$  bisogna esprimere  $S_R^2$  come differenza fra  $S_G^2$  e  $S_S^2$ . Dalla (2.9):

$$S_S^2 = \sum (Y_i - \bar{y})^2 = b_1^2 \sum (x_i - \bar{x})^2 = b_1 \sum (x_i - \bar{x})(y_i - \bar{y}) = b_1 C_{y1}$$

ed analogamente, nel caso multidimensionale, si ha:

$$S_S^2 = b_1 C_{y1} + b_2 C_{y2} + \dots + b_p C_{yp} \quad (4.9)$$

$$S_R^2 = C_{yy} - b_1 C_{y1} - b_2 C_{y2} \dots - b_p C_{yp} \quad (4.10)$$

$$\sigma_0^2 = \frac{1}{n - p - 1} (C_{yy} - b_1 C_{y1} - b_2 C_{y2} - \dots - b_p C_{yp}) \quad (4.11)$$

essendo  $n$  i gruppi d'osservazione e  $p + 1$  le incognite  $b$ , potendo così ricavare le loro varianze:

$$\begin{aligned} \sigma_{b_i}^2 &= \sigma_0^2 C^{ii} & i = 1, 2, \dots, p \\ \sigma_{b_0}^2 &= \sigma_{\bar{y}}^2 + \bar{x}_1^2 \sigma_{b_1}^2 + \bar{x}_2^2 \sigma_{b_2}^2 \dots + \bar{x}_p^2 \sigma_{b_p}^2 + 2\bar{x}_1 \bar{x}_2 \text{cov}(b_1 b_2) + \dots = \\ &= \sigma_0^2 (1/n + \bar{x}_1^2 C^{11} + \bar{x}_2^2 C^{22} + \dots + \bar{x}_p^2 C^{pp} + 2C^{12} \bar{x}_1 \bar{x}_2 + \dots) \end{aligned} \quad (4.12)$$

La significatività di  $b$ , come al solito, è valutata dal rapporto con il loro sqm:

$$t = b_i / \sigma_0 \cdot \sqrt{C^{ii}}$$

confrontato con il  $t$  di Student prestabilito, con  $\nu = n - p - 1$  gradi di libertà. I limiti fiduciali, per ognuna  $b_i$ , presa isolatamente sono:  $\beta_i = b_i \pm t_\alpha \sigma_0 \sqrt{C^{ii}}$

Qualora si cerchi la regione fiduciaria all'  $(1 - \alpha)\%$ , per una coppia di  $\beta_i, \beta_j$ , da considerare congiuntamente, bisogna tener presente che, se  $b_i$  e  $b_j$  sono i valori campionari, ricavati dalla soluzione del sistema normale, la seguente espressione segue la distribuzione  $F$  di Fisher, con 2 e  $(n - p - 1)$  gradi di libertà:

$$\frac{C^{jj}(b_i - \beta_i)^2 - 2C^{ij}(b_i - \beta_i)(b_j - \beta_j) + C^{ii}(b_j - \beta_j)^2}{2\sigma_0^2(C^{ii}C^{jj} - (C^{ij})^2)} = F_{2, n-p-1} \quad (4.13)$$

La (4.13), ponendo al posto di  $F$  il suo valore numerico  $F_\alpha$  e di  $b_i, b_j$  i valori trovati, diventa l'equazione di un'ellisse, con il centro nel punto  $b_i, b_j$  (come mostra la figura 5.3.2). La regione del piano, racchiusa dall'ellisse, è quella entro la quale si può ritenere, con l'  $(1 - \alpha)$  di confidenza, incluso il punto rappresentante i coefficienti di regressione  $\beta_i, \beta_j$ .

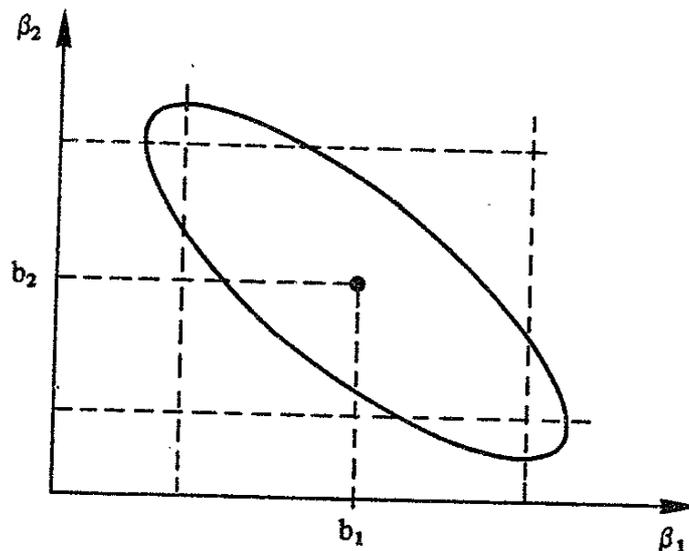


Fig. 5.3.2 – Limiti fiduciali per una coppia di coefficienti di regressione

Per contro, se si fosse voluta determinare questa regione operando separatamente su  $\beta_i$  e  $\beta_j$ , si sarebbe trovato, invece di un'ellisse, un rettangolo di area generalmente superiore a questa e molto meno utile per individuare i limiti fiduciali. Infatti i lati del rettangolo sono costruiti in modo che passino per i limiti fiduciali all'  $(1 - \alpha/2)$ , per ciascun coefficiente, tenendo conto della probabilità, per entrambi, di stare nei limiti calcolati all'  $(1 - \alpha/2)^2 \cong 1 - \alpha$

Per quanto riguarda poi la varianza di un valore previsto  $Y_k$  o di un valore osservato  $y_k$ , si possono estendere, senza difficoltà, le (3.2) e (3.3), così come la tabella relativa all'analisi di varianza.

SCARTI	$S^2$	$\nu$	$\sigma^2$
della regressione su tutte le $p$ variabili	$b_1 C_{y1} + b_2 C_{y2} + \dots + b_p C_{yp}$	$p$	$\sigma_s^2$
intorno alla regressione (residui)	$C_{yy} - \sum b_i C_{yi}$	$n - p - 1$	$\sigma_r^2 = \sigma_0^2$
totali	$C_{yy}$	$n - 1$	$\sigma_G^2$

Tuttavia in questo caso, l'analisi di varianza della regressione presenta aspetti più complicati di quelli nel caso della regressione semplice, perché più della significatività globale ottenuta, utilizzando tutte le  $p$  variabili, occorre analizzare il contributo alla riduzione di  $S_R^2$ , dovuto all'introduzione di ogni singola variabile. Pertanto la tipica analisi di varianza si presenta invece nella forma della seguente tabella.

SCARTI	$S^2$	$\nu$	$\sigma^2$
della regressione di $y$ su $x_1, x_2, \dots, x_{p-1}$	$S_S^2(1, 2, \dots, p-1)$	$p-1$	$\sigma_S^2(1, 2, \dots, p-1)$
aggiunta di $x_p$	$S_S^2(p)$	1	$\sigma_S^2(p)$
della regressione totale di $y$ su $x_1, x_2, \dots, x_p$	$S_S^2(1, 2, \dots, p)$	$p$	$\sigma_S^2(1, 2, \dots, p)$
intorno alla regressione (residui)	$S_R^2$	$n - p - 1$	$\sigma_r^2 = \sigma_0^2$
totali	$S_G^2$	$n - 1$	$\sigma_G^2$

Allora se si usano solo le variabili  $x_1, x_2, \dots, x_{p-1}$ , per predire  $y$ , si ottengono dal sistema normale le stime dei coefficienti di regressione  $b'_1, b'_2, \dots, b'_{p-1}$ , da utilizzare in una formula analoga alla (4.9) per calcolare  $S_S^2$ , dovuta alla regressione cui competono  $\nu = p - 1$  gradi di libertà. Invece se si usano  $p$  variabili, si hanno altri valori  $b$  ed un altro valore  $S_S^2$  con  $\nu = p$ . La differenza fra questi due valori dà la frazione  $S_S^2(p)$ , dovuta alla introduzione della  $p$ -esima variabile, con  $\nu = p - (p - 1) = 1$  gradi di libertà. La significatività del rapporto  $F$  di Fisher parziale, fra  $\sigma_S^2(p)$  e  $\sigma_0^2$ , permette di valutare la significatività del contributo della  $p$ -esima variabile agli effetti della riduzione di  $\sigma_0^2$ , cioè al perfezionamento dell'aderenza fra modello e fenomeno.

Anche per la regressione multipla si può, ad ogni passo, calcolare l'indice di determinazione multipla:

$$R_{y,1,2,\dots,p}^2 = 1 - \frac{\sum(\text{scarti residui})^2}{\sum(\text{scarti generali})^2} = 1 - \frac{C_{yy} - \sum b_i C_{yi}}{C_{yy}} \quad (4.14)$$

che tenderà ad avvicinarsi sempre più ad 1 via, via che tutte le variabili  $x$ , realmente influenzanti  $y$  sono introdotte nella regressione. Il valore di  $R^2$  si può calcolare, ad ogni passo, in modo ricorrente, a partire dai coefficienti di correlazione lineare di ordine zero ( $r_{y1}, r_{12}, \dots$ ), tramite i coefficienti di correlazione parziali:

$$1 - R_{y,12}^2 = (1 - r_{y1}^2)(1 - r_{y2,1}^2)$$

$$1 - R_{y,123}^2 = (1 - r_{y1}^2)(1 - r_{y2,1}^2)(1 - r_{y3,12}^2)$$

*ecc.*

(4.15)

Ad esempio, mentre  $r_{y1}$  è il coefficiente di correlazione fra  $y$  e  $x_1$ , indipendentemente dalle altre variabili,  $r_{y2,1}$  è il coefficiente di correlazione fra  $y$  e  $x_2$ , con  $x_1$  costante, e  $r_{y3,12}$  è il coefficiente di correlazione fra  $y$  e  $x_3$ , con  $x_1$  e  $x_2$  costanti. Anche questi ultimi possono poi essere facilmente messi in relazione fra loro:

$$r_{y1,2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}} \quad (4.16.a)$$

$$r_{y1,23} = \frac{r_{y1,3} - r_{12,3}r_{12,3}}{\sqrt{(1 - r_{y2,3}^2)(1 - r_{12,3}^2)}} = \frac{r_{y1,2} - r_{y3,2}r_{13,2}}{\sqrt{(1 - r_{y3,2}^2)(1 - r_{13,2}^2)}} \quad (4.16.b)$$

## 5.5 Ricerca della migliore equazione di regressione

La ricerca della migliore equazione è uno dei problemi più seri, al lato pratico delle cose, anche perché non esiste un criterio teorico assoluto, da guida, e molto è lasciato al buon senso dello sperimentatore. Fra i vari metodi in uso si accenna solo ai più noti, premettendo che possono non condurre allo stesso risultato, anche se questo non capita, nella maggioranza dei casi.

### 5.5.1 Procedimento di eliminazione all'indietro

Questo procedimento si avvia con una regressione su tutte le variabili che si pensa possono influire su  $y$ ; dopodiché calcolata questa regressione:

- ❑ con il procedimento della tabella precedente, si valutano i valori  $F$  di Fisher parziali, relativi ad ognuna variabile  $x$ , trattando quella sotto esame, come se essa sia l'ultima ad essere entrata nella regressione;
- ❑ tra tutti i valori  $F$  di Fisher parziali, si seleziona il più piccolo, ad esempio, si supponga  $F_k$ , e lo si confronta con un valore teorico  $F_\alpha$  prefissato, in corrispondenza al livello di significatività voluto ed ai gradi di libertà del problema:
  - ❑ se  $F_k < F_\alpha$ , la variabile  $x_k$  è eliminata e la regressione è calcolata ex-novo, senza  $x_k$ ;
  - ❑ se invece  $F_k > F_\alpha$  l'equazione di regressione originale è adottata senza variazioni.

### 5.5.2 Procedimento di selezione in avanti

Mentre il procedimento 5.5.1 usa la più ampia regressione possibile, fin dall'inizio ed eventualmente la riduce via, via, in questo caso si procede in senso inverso, aumentando il numero delle variabili fino a quando l'equazione risulta soddisfacente. L'ordine di ingresso in regressione delle variabili è determinato dall'entità dei coefficienti di correlazione parziali che misurano l'importanza delle variabili, non ancora presenti nella regressione. In questo modo:

- si identifica quella  $x$  con il più elevato coefficiente di correlazione lineare, di ordine zero, con  $y$  (ad esempio, si supponga  $x_1$ ) e si costruisce la regressione come  $Y = f(x_1)$ ;
- si trovano i coefficienti di correlazione parziali fra tutte le altre  $x$  e  $y$  ( $r_{yx_i.1}$ ), con formule analoghe alla (4.16.a) e si seleziona la variabile  $x$  con il coefficiente più elevato (ad esempio, si supponga  $x_2$ ), per farla entrare nella regressione, come seconda variabile.
- si costruiscono la nuova regressione  $Y = f(x_1, x_2)$  ed i nuovi coefficienti di correlazione parziali, con la seconda delle (4.16) ed analoghe, e così via,

ad ogni nuova variabile, entrata in regressione, si calcolano:

- l'indice di determinazione  $R^2$ ;
- Il valore  $F$  di Fisher parziale relativo alla variabile  $x$ , entrata per ultima, il quale permette di valutare, se questa variabile ha sostanzialmente diminuito l'entità di  $S_R^2$ , rispetto a quanto già fatto dalle precedenti variabili (non appena il valore sperimentale  $F$  di Fisher, relativo all'ultima variabile entrata, diventa non significativo, il procedimento termina).

Questo metodo è senz'altro migliore del procedimento 5.5.1, perché evita di lavorare con più variabili del necessario. Tuttavia la sua debolezza consiste nel fatto che nessuno sforzo è compiuto per controllare quale effetto può avere l'introduzione di una nuova variabile sul comportamento delle altre, già entrate.

### 5.5.3 Procedimento di regressione sequenziale

E' analogo al procedimento 5.5.2, salvo che ad ogni passo sono riesaminate tutte le variabili, entrate nel modello di regressione, in precedenza. Infatti una variabile che, ad un certo punto del procedimento, può essere la migliore, successivamente può anche diventare superflua, per effetto dei legami esistenti fra essa e le altre variabili, entrate dopo. Questo metodo è il più usato e dà generalmente ottimi risultati. Tuttavia, soprattutto nel caso di correlazioni molto elevate, fra le variabili indipendenti, è consigliabile porre livelli di accettazione o rigetto poco restrittivi, così da poter analizzare, nel modello, un elevato numero di variabili.

### 5.6 Ricerca delle trasformazioni sulle variabili

Un caso molto frequente tratta di una regressione, lineare nei parametri, ma necessitante invece di alcune semplici trasformazioni su  $x$  (o su  $y$ ), sotto forma  $x^2, 1/x, \sqrt{x}, \ln x, ecc..$  La determinazione della migliore funzione di trasformazione ha luogo empiricamente e, in mancanza d'informazioni preliminari, per tentativi, fino a raggiungere, per ogni  $x_i$ , quella particolare funzione  $f(x_i)$  che, insieme a  $f(y)$ , dà luogo al più elevato coefficiente di correlazione lineare parziale. Spesso poi può essere utile, per discriminare tra le molte possibilità, esistenti a priori, eseguire una regressione multipla sui dati bruti ed esaminare l'andamento del grafico dei residui di ogni coppia  $(y, x_i)$ , previa depurazione dell'influenza delle altre variabili.

Le Fig.5.5.1 a) e b) mostrano un tipico caso in cui la correlazione, fra  $y$  ed  $x$ , non è lineare e lo diventa, dopo un'opportuna trasformazione.

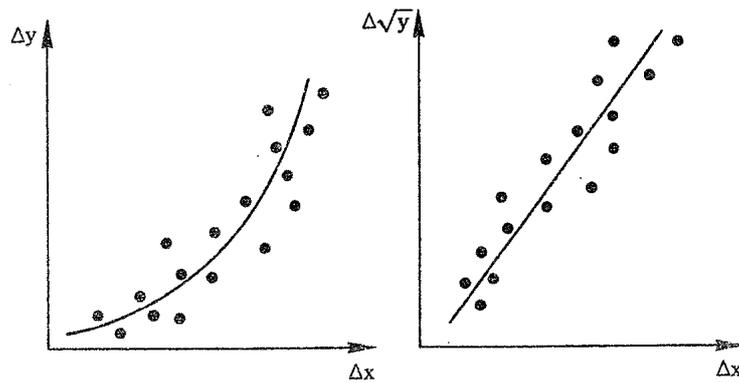


Fig. 5.5.1 a Fig. 5.5.1b  
Esempio di trasformazione sui dati originali per aumentare la correlazione lineare

I residui, riportati nei grafici, nel caso si voglia esaminare l'andamento della  $x_1$ , sono ad esempio:

$$\Delta y_i(x_i) = y_i - (b_0 + b_2 x_2 + b_3 x_3)$$

$$\Delta x_{1i} = x_{1i} - (d_0 + d_2 x_2 + d_3 x_3)$$

dove  $b$  e  $d$  sono rispettivamente i coefficienti di regressione multipla parziale fra  $y$  e tutte le variabili, non sotto esame, ed i coefficienti di regressione multipla parziale fra  $x_1$ , variabile in esame, e tutte le rimanenti: questi scarti sono quelli che, con il loro andamento, permettono di intuire la miglior forma di legame fra  $y$  e  $x_1$ , depurato dalle influenze delle altre variabili (infatti il coefficiente di correlazione lineare di grado zero, fra  $\Delta y$  e  $\Delta x_1$ , è un'altra formulazione del coefficiente di correlazione parziale  $r_{y1,2...p}$ , già in (4.16)).

Qualora nasca un'ambiguità nella scelta fra due diversi tipi di funzioni, approssimativamente con lo stesso andamento, si esegue la trasformazione dei dati originali con entrambe. Si calcolano poi i coefficienti di correlazione lineare fra i residui trasformati, con l'una e l'altra formula, scegliendo quella delle due cui compete  $r$  più elevato.

#### Esempio 5.5.1

Un programma di calcolo è applicato ai dati (di tabella 1), riferiti alla ricerca dell'equazione di regressione, fra i materiali, componenti una certa qualità di cemento, ed il calore sviluppato.

☐ Variabili indipendenti (misurate in percentuale del peso dell'inerte):

- ☐  $x_1 = 3CaO \cdot Al_2O_3$
- ☐  $x_2 = 3CaO \cdot SiO_2$
- ☐  $x_3 = 4CaO \cdot Al_2O_3 \cdot Fe_2O_3$
- ☐  $x_4 = 2CaO \cdot SiO_2$

☐ Variabile dipendente (calore sviluppato misurato in calorie/gr. di cemento):

- ☐  $x_5 = y$

### **Procedimento di eliminazione all'indietro**

Tutte le variabili sono introdotte nella regressione (come mostra la tabella 2) ed i valori sperimentali  $F$  di Fisher parziali misurano il contributo di ciascuna variabile alla riduzione di  $\sigma_0^2$ . A questo punto, si confronta il più piccolo tra questi:  $F_3 = 0.0182345$ , con il corrispondente valore critico, ad esempio, per  $\alpha = 0.10$ ,  $F(1.8, 0.90) = 3.46$  (dove i numeri fra parentesi sono i gradi di libertà ed  $(1-\alpha)$ ) e, dato che risulta:  $F_3 < F(1.8, 0.90)$ , si elimina la variabile  $x_3$ .

Successivamente si cerca la regressione con le sole variabili  $x_1, x_2, x_4$  (come mostra la tabella 3). Allora il valore  $F$  di Fisher globale è:  $F = 166.83 > F(3.9, 0.999) = 13.90$ , e pertanto la regressione è significativa, nel suo complesso. Tuttavia non è significativo il contributo della variabile  $x_4$  che è eliminato, perché si ha invece:  $F_4 = 1.86 < F(1.9, 0.90) = 3.36$ .

Infine si ricava (in tabella 4) la regressione:  $Y = f(x_1, x_2)$ , che è significativa, perché il valore  $F$  di Fisher globale è:  $F = 229.50 > F(2.10, 0.999) = 14.91$ . Di conseguenza, entrambe le variabili  $x_1$  e  $x_2$  danno un contributo significativo e l'equazione di regressione è:  $Y = 52.58 + 1.47 x_1 + 0.66 x_2$ .

### **Procedimento di selezione in avanti**

La variabile  $x_j$  con il più elevato coefficiente di correlazione con  $y = x_5$  è  $x_4$  (come mostra la tabella 1):  $r_{45} = -0.82130513$ . Pertanto  $x_4$  è la prima variabile ad entrare in regressione (ed i dati per l'equazione:  $Y = f(x_4)$ , sono riportati in tabella 5).

Dopodiché fra i quadrati dei coefficienti di correlazione parziali, con le variabili non ancora in regressione, il più elevato è  $r_{51,4}^2 = 0.91541$ . Allora si costruisce l'equazione:  $Y = f(x_4, x_1)$  (come mostra la tabella 6).

Quest'equazione ha una percentuale  $R^2$  di 0.972% ed è significativa, perché il valore  $F$  di Fisher globale è:  $F = 176.63 > F(2.10, 0.999) = 14.91$ . Infatti la variabile  $x_1$  fornisce una significativa diminuzione di  $\sigma_0^2$ , come provato dal relativo valore  $F$  di Fisher parziale:  $F_1 = 108.22 > F(1.10, 0.999) = 21.04$ .

A questo punto, il coefficiente  $r^2$  più elevato è:  $r_{52,14}^2 = 0.35833$ , e così la variabile  $x_2$  entra in regressione, con la nuova equazione:  $Y = f(x_4, x_1, x_2)$  (come mostra la tabella 7). Con quest'equazione  $R^2$  arriva a 0.92% e l'aggiunta della variabile  $x_2$ , alla regressione, è significativa, in quanto, se si assume  $\alpha = 0.10$ , il valore  $F$  di Fisher parziale è:  $F_2 = 5.03 > F(1.9, 0.90) = 3.36$ .

Dato che finora ogni variabile introdotta produce una riduzione di  $S_R^2$ , si procede ad introdurre anche l'ultima variabile  $x_3$  (come mostra la tabella 8). Tuttavia il valore  $F$  di Fisher parziale è:  $F_3 = 0.18$ , e non è significativo, cosicché la variabile  $x_3$  deve essere eliminata. Di conseguenza, l'analisi di varianza completa è riassumibile nella seguente tabella e l'equazione di regressione, da questa dedotta, in base al procedimento di selezione in avanti, risulta essere:  $Y = 71.65 - 0.24x_4 + 1.45x_1 + 0.42x_2$ .

TIPO DI SCARTI	$S^2$	$\nu$	$\sigma^2$
<i>Regressione</i>	2667.90	4	
$x_4$	1831.90	1	1831.90
$x_1/x_4$	809.10	1	809.10
$x_1/x_4, x_1$	26.79	1	26.79
$x_1/x_4, x_1, x_2$	0.11	1	0.11
<i>Residui</i>	47.86	8	5.98
<i>Totale</i>	2715.76	12	

### **Procedimento di regressione sequenziale**

Le tabelle 9 e 10 riportano i primi due passi di questo procedimento, identici a quelli del procedimento di selezione in avanti. Tuttavia giunti all'equazione:  $Y = f(x_4, x_1)$ , si esamina anche il contributo dato dalla variabile  $x_4$ , con la variabile  $x_1$ , entrata in regressione per prima (questo contributo è significativo, essendo il valore  $F$  di Fisher parziale:  $F_4 = 159.295$ ).

Procedendo come prima, si calcola:  $Y = f(x_4, x_1, x_2)$ , dove il contributo della variabile  $x_2$  è significativo (con  $\alpha = 0.10$ ), e si eseguono poi i test sulle variabili  $x_1$  e  $x_4$ , per decidere se devono essere mantenute od eliminate dalla regressione, trovando che la variabile  $x_4$  deve essere eliminata, dato che il valore  $F$  di Fisher parziale è:  $F_4 = 1.863 < F(1.9, 0.10) = 3.36$  (come mostra la tabella 11).

In tabella 12, è ricalcolata l'equazione di regressione (senza la variabile  $x_4$  e la migliore possibile); infatti l'unica variabile rimasta è  $x_3$ , ma è eliminata subito, cosicché il procedimento di regressione sequenziale termina, ottenendo un'equazione di regressione uguale a quella del procedimento di eliminazione all'indietro.

**Tabella 1**

#### **Dati originali o trasformati**

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
1	7.00000000	26.00000000	6.00000000	60.00000000	78.50000000
2	1.00000000	29.00000000	15.00000000	52.00000000	74.30000000
3	11.00000000	56.00000000	8.00000000	20.00000000	104.30000000
4	11.00000000	31.00000000	8.00000000	47.00000000	87.60000000
5	7.00000000	52.00000000	6.00000000	33.00000000	95.90000000
6	11.00000000	55.00000000	9.00000000	22.00000000	109.20000000
7	3.00000000	71.00000000	17.00000000	6.00000000	102.70000000
8	1.00000000	31.00000000	22.00000000	44.00000000	72.50000000
9	2.00000000	54.00000000	18.00000000	22.00000000	93.10000000
10	21.00000000	47.00000000	4.00000000	26.00000000	115.90000000
11	1.00000000	40.00000000	23.00000000	34.00000000	83.80000000
12	11.00000000	66.00000000	9.00000000	12.00000000	113.30000000
13	10.00000000	68.00000000	8.00000000	12.00000000	109.40000000

**Medie**

	7.46153830	48.15384500	11.76923000	29.99999900	95.42307500
--	------------	-------------	-------------	-------------	-------------

**Scarti quadratici medi**

	5.88239440	15.56087900	6.40512590	16.73817800	15.04372400
--	------------	-------------	------------	-------------	-------------

**Matrice di correlazione**

1	.99999991	.22857948	-.82413372	-.2454512	.73071745
2	.22857948	1.00000010	-.13924238	-.97295516	.81625268
3	-.82413372	-.13924238	.99999991	.02953701	-.53467065
4	-.24544512	-.97295516	.02953701	1.00000010	-.82413372
5	.73071745	.81625268	-.53467065	-.82413372	.99999999

<b>Numero di osservazioni</b>	13
<b>Variabile indipendente</b>	$y = x_5$
<b>Livello fiduciario per i coefficienti</b>	95%
<b>Valore <math>F</math> di Fisher per accettare o rigettare una variabile</b>	3.28

**PROCEDIMENTO 1****Tabella 2****Informazioni di controllo al 1° passo**

Indice di determinazione $R^2$	98.2375700
Sqm dei residui	2.4460044
Gradi di libertà	8

**Analisi di varianza**

<i>TIPO DI SCARTI</i>	<i>Gradi di libertà</i>	$S^2$	$\sigma^2$	<i>F totale</i>
<i>Totali</i>	12	2715.7635000		
<i>Re gressione</i>	4	2667.9000000	666.9750000	111.4795200
<i>Re sidui</i>	8	47.8634980	5.9829372	

**Coefficienti  $b$  e loro limiti fiduciari**

<i>N. var</i>	<i>Coefficienti <math>b</math></i>	<i>Limiti Superiore / Inferiore</i>	<i>sqm</i>	<i>F parziali</i>
4	-.1440588	1.4909970 -1.7791144	.7090441	.0412794
3	.1019111	1.8422494 -1.6384272	.7547001	.0182345
2	.5101700	2.1792063 -1.1588665	.7237799	.4968402
1	1.5511043	3.2685233 -.1663147	.7447611	4.3375858

**Termine costante nell'equazione di regressione**

62.4051530

**$r^2$  per le variabili non in regressione**

Variabili	$r^2$
5	1.0000

**Tabella 3****Informazioni di controllo al 2° passo**

Variabili non in regressione	3
Indice di determinazione $R^2$	98.2335600
Sqm dei residui	2.3087418
Gradi di libertà	9

**Analisi di varianza**

TIPO DI SCARTI	Gradi di libertà	$S^2$	$\sigma^2$	$F$ totale
<i>Totali</i>	12	2715.7635000		
<i>Re gressione</i>	3	2667.7911000	889.2637000	166.8321800
<i>Re sidui</i>	9	47.9725980	5.3302886	

**Coefficienti  $b$  e loro limiti fiduciar**

N. var	Coefficienti $b$	Limiti Superiore / Inferiore	$sqm$	$F$ parziali
2	.4161107	.8359611 -.0037398	.1856103	5.0258974
1	1.4519380	1.7165861 1.1872899	.1169974	154.0080400
4	-.2365395	.1554371 -.6285160	.1732876	1.8632548

**Termine costante dell'equazione di regressione**

71.6482410

 **$r^2$  per le variabili non in regressione**

Variabili	$r^2$
3	.00227
5	1.0000

**Tabella 4<sup>5</sup>****Informazioni di controllo al 3° passo**

Variabili non in regressione	3,4
Indice di determinazione $R^2$	97.8678500
Sqm	2.4063327
Gradi di libertà	10

<sup>5</sup> Questo passo conclude il Procedimento 1, con un risultato, in questo caso specifico, uguale a quello del Procedimento 3.

**Analisi di varianza**

<i>TIPO DI SCARTI</i>	<i>Gradi di libertà</i>	$S^2$	$\sigma^2$	<i>F totale</i>
<i>Totali</i>	12	2715.7635000		
<i>Re gressione</i>	2	2657.8593000	1328.9296000	229.5042100
<i>Re sidui</i>	10	57.9043680	5.7904368	

**Coefficienti *b* e loro limiti fiduciari**

<i>N. var</i>	<i>Coefficienti b</i>	<i>Limiti Superiore / Inferiore</i>	<i>sqm</i>	<i>F parziali</i>
2	.6622507	.7644149 .5600865	.045847	208.5823200
1	1.4683057	1.7385638 1.1980476	.1213008	146.5229400

Termine costante nell'equazione i regressione

52.5773400

 **$r^2$  per le variabili non in regressione**

<i>Variabili</i>	$r^2$
3	.16914
4	.17152
5	1.00000

**PROCEDIMENTO 2****Tabella 5****Informazione di controllo al 1° passo**

Variabili non in regressione	1, 2, 3
Variabile entrante	4
<i>F</i> parziale della variabile entrante	22.7985280
Indice di determinazione $R^2$	67.4542100
Sqm dei residui	8.9639014
Gradi di libertà	11

**Analisi di varianza**

<i>TIPO DI SCARTI</i>	<i>Gradi di libertà</i>	$S^2$	$\sigma^2$	<i>F totale</i>
<i>Totali</i>	12	2715.7635000		
<i>Re gressione</i>	1	1831.8968000	1831.8968000	22.7985300
<i>Re sidui</i>	11	883.8668200	80.3515290	

**Coefficienti *b* e loro limiti fiduciari**

<i>N. var</i>	<i>Coefficienti b</i>	<i>Limiti Superiore / Inferiore</i>	<i>sqm</i>	<i>F parziali</i>
4	-.7381619	-.3978962 -1.0784277	.1545960	22.7985270

Termine costante nell'equazione di regressione

117.5679300

$r^2$  per le variabili non ancora in regressione <sup>6</sup>

Variabili	$r^2$
1	.91541
2	.01696
3	.80117
5	1.00000

**Tabella 6**

**Informazione di controllo al 2° passo**

Variabili non in regressione	2,3
Variabili entrante	1
$F$ parziale della variabile entrante	108.2238900
Indice di determinazione $R^2$	97.2471100
Sqm dei residui	2.7342662
Gradi di libertà	10

**Analisi di varianza**

TIPO DI SCARTI	Gradi di libertà	$S^2$	$\sigma^2$	$F$ totale
<i>Totali</i>	12	2715.7635000		
<i>Re gressione</i>	2	2641.0015000	1320.5007000	176.6269800
<i>Re sidui</i>	10	74.7621170	7.4762117	

**Coefficienti  $b$  e loro limiti fiduciarci**

N. var	Coefficienti $b$	Limiti Superiore / Inferiore	$sqm$	$F$ parziali
1	1.4399582	1.7483504 1.1315660	.1384166	108.2238900
4	-.6139537	-.5055737 -.7223338	.0486446	159.2952400

Termine costante nell'equazione di regressione

103.0973800

$r^2$  per le variabili non ancora in regressione

Variabili	$r^2$
2	.35833
3	.32003
5	1.00000

<sup>6</sup> Il Procedimento 2, di selezione in avanti, inserisce questa variabile, superflua con gli altri procedimenti (come il Procedimento 1 di eliminazione all'indietro ed il Procedimento 3 di regressione sequenziale), senza più riuscire ad eliminarla. Resta poi da precisare, come fortuita (cioè legata a questo caso specifico), l'uguaglianza del risultato fra il Procedimento 1 ed il Procedimento 3, in quanto questo (ultimo) procedimento, proprio perché sequenziale, è capace di unire i pregi della selezione in avanti e dell'eliminazione all'indietro.

**Tabella 7****Informazioni di controllo al 3° passo**

Variabili non in regressione	3
Variabile entrante	2
$F$ parziale della variabile entrante	5.0258974
Indice di determinazione $R^2$	98.2335600
Sqm dei residui	2.3087418
Gradi di libertà	9

**Analisi di varianza**

<i>TIPO DI SCARTI</i>	<i>Gradi di libertà</i>	$S^2$	$\sigma^2$	$F$ totale
<i>Totali</i>	12	2715.7635000		
<i>Re gressione</i>	3	2667.7911000	889.2637000	166.8321800
<i>Re sidui</i>	9	47.9725980	5.3302886	

**Coefficienti  $b$  e loro limiti fiduciari**

<i>N. var</i>	<i>Coefficienti <math>b</math></i>	<i>Limiti Superiore / Inferiore</i>	<i>sqm</i>	<i>F parziali</i>
2	.416107	.8359611 -.0037398	.1856103	5.0258974
1	1.4519380	1.7165861 1.1872899	.1169974	154.0080400
4	-.2365395	.1554371 -.6285160	.1732876	1.8632548

**Termine costante nell'equazione di regressione**

71.6482410

 **$r^2$  per le variabili non ancora in regressione**

<i>Variabili</i>	$r^2$
3	.00227
5	1.00000

**Tabella 8****Informazioni di controllo al 4° passo**

Variabili non in regressione	<i>nessuna</i> <sup>7</sup>
Variabile entrante	3
$F$ parziale della variabile entrante	0.0182345
Indice di determinazione $R^2$	98.2375700
Sqm dei residui	2.4460044
Gradi di libertà	8

<sup>7</sup> In generale, anche il Procedimento 1 può dare un risultato diverso dal Procedimento 3, arrestando prima l'eliminazione all'indietro.

**Analisi di varianza**

<i>TIPO DI SCARTI</i>	<i>Gradi di libertà</i>	$S^2$	$\sigma^2$	<i>F totale</i>
<i>Totali</i>	12	2715.7635000		
<i>Re gressione</i>	4	2667.9000000	666.9750000	111.4795200
<i>Re sidui</i>	8	47.8634980	5.9829372	

**Coefficienti *b* e loro limiti fiduciari**

<i>N. var</i>	<i>Coefficienti b</i>	<i>Limiti Superiore / Inferiore</i>	<i>sqm</i>	<i>F parziali</i>
4	-.1440588	1.4909970 -1.7791144	.7090441	.0412794
3	.1019111	1.8422494 -1.6384272	.7547001	.0182345
2	.5101700	2.1792063 -1.1588665	.7237799	.4968402
1	1.5511043	3.2685233 -.1663147	.7447611	4.3375858

Termine costante nell'equazione di regressione

62.4051530

 **$r^2$  per le variabili non ancora in regressione**

<i>Variabili</i>	$r^2$
5	1.00000

**PROCEDIMENTO 3****Tabella 9****Informazioni di controllo al 1° passo**

Variabile entrante	4
<i>F</i> parziale della variabile entrante	22.7985280
Indice di determinazione $R^2$	67.452100
Sqm dei residui	8.9639014
Gradi di libertà	11

**Analisi di varianza**

<i>TIPO DI SCARTI</i>	<i>Gradi di libertà</i>	$S^2$	$\sigma^2$	<i>F totale</i>
<i>Totali</i>	12	2715.7635000		
<i>Re gressione</i>	1	1831.8968000	1831.8968000	22.7985300
<i>Re sidui</i>	11	883.8668200	80.3515290	

**Coefficienti *b* e loro limiti fiduciari**

<i>N. var</i>	<i>Coefficienti b</i>	<i>Limiti Superiore / Inferiore</i>	<i>sqm</i>	<i>F parziali</i>
4	-.7381620	-.3978962 -1.0784277	.1545960	22.7985270

Termine costante nell'equazione di regressione

117.5679300

$r^2$  per le variabili non ancora in regressione

Variabili	$r^2$
1	.91541
2	.01696
3	.80117
5	1.00000

Tabella 10

Informazioni di controllo al 2° passo

Variabile entrante	1
$F$ parziale della variabile entrante	108.2240500
Indice di determinazione $R^2$	97.2471100
Sqm dei residui	2.7342642
Gradi di libertà	10

Analisi di varianza

TIPO DI SCARTI	Gradi di libertà	$S^2$	$\sigma^2$	$F$ totale
<i>Totali</i>	12	2715.7635000		
<i>Re gressione</i>	2	2641.0015000	1320.5007000	176.6272400
<i>Re sidui</i>	10	74.7620080	7.4762008	

Coefficienti  $b$  e loro limiti fiduciarci

N. var	Coefficienti $b$	Limiti Superiore / Inferiore	$sqm$	$F$ parziali
4	-.6139538	-.5055738 -.7223338	.0486445	159.2954900
1	1.4399582	1.7483502 1.1315662	.1384165	108.2240500

Termine costante nell'equazione di regressione

103.0973800

$r^2$  per le variabili non ancora in regressione

Variabili	$r^2$
2	.35833
3	.32003
5	1.00000

Tabella 11

Informazioni di controllo al 3° passo

Variabile entrante	2
--------------------	---

$F$ parziale della variabile entrante	5.0258747
Indice di determinazione $R^2$	98.2335500
Sqm dei residui	2.3087426
Gradi di libertà	9

#### Analisi di varianza

TIPO DI SCARTI	Gradi di libertà	$S^2$	$\sigma^2$	$F$ totale
<i>Totali</i>	12	2715.7635000		
<i>Re gressione</i>	3	2667.7908000	889.2636000	166.830500
<i>Re sidui</i>	9	47.9726310	5.3302923	

#### Coefficienti $b$ e loro limiti fiduciar

$N.$ $var$	Coefficienti $b$	Limiti Superiore / Inferiore	$sqm$	$F$ parziali
4	-.2365401	.1554367 -.6285170	.1732877	1.8632619
1	1.4519379	1.7165861 1.1872897	.1169975	154.0079500
2	.4161100	.8359608 -.0037408	.1856104	5.0258730

#### Termine costante nell'equazione di regressione

71.6482910

#### $r^2$ per le variabili non ancora in regressione

Variabili	$r^2$
3	.00227
5	1.00000

**Tabella 12**

#### Informazioni di controllo al 4° passo

Variabile uscente	4
$F$ parziale della variabile uscente	1.8632611
Indice di determinazione $R^2$	97.8678500
Sqm dei residui	2.4063325
Gradi di libertà	10

#### Analisi di varianza

TIPO DI SCARTI	Gradi di libertà	$S^2$	$\sigma^2$	$F$ totale
<i>Totali</i>	12	2715.7635000		
<i>Re gressione</i>	2	2657.8593000	1328.9296000	229.5042500
<i>Re sidui</i>	10	57.9043570	5.79044357	

**Coefficienti  $b$  e loro limiti fiduciar**

$N.$ $var$	Coefficienti $b$	Limiti Superiore / Inferiore	$sqm$	$F$ parziali
1	1.4683057	1.7385638 1.1980476	.1213008	146.5229500
2	.6622507	.7644147 .5600864	.0458547	208.5821200

**Termine costante nell'equazione di regressione**

52.5773400

**$r^2$  per le variabili non ancora in regressione**

Variabili	$r^2$
3	.16914
4	.17152
5	1.00000

Generalizzazioni della regressione multipla e, più in generale, dell'analisi di varianza sono altresì possibili <sup>8</sup>.

<sup>8</sup> Ad esempio, alcuni test non-parametrici verificano sempre l'uguaglianza tra valori centrali, ma per campioni non necessariamente normali (generalizzando i test di rango), oppure anche non indipendenti (generalizzando i test di segno). Altri test della normalità e non-parametrici verificano l'uguaglianza tra valori di dispersione (ovvero studiano le componenti della varianza o di altri valori di dispersione) e la significatività della correlazione (studiando cioè la struttura di covarianza od altre modellazioni della dipendenza lineare).

**Test di Kruskal-Wallis per campioni indipendenti**

Dati d'ingresso: numero di campioni  $n$   
 numerosità di ciascun campione  $m_j$   
 numerosità totale  $N = \sum_{j=1}^n m_j$   
 ranghi, ovvero numeri ordinali, in corrispondenza all'unione di tutti i valori argomentali ordinati in modo crescente  $r_{ij}$   
 livello di significatività  $\alpha$

Ipotesi fondamentale:  $H_0$  : uguaglianza dei valori centrali

Valore atteso: 
$$\chi_e^2 = \frac{12}{N(N+1)} \sum_{j=1}^n \frac{R_j^2}{m_j} - 3(N+1)$$
 dove:  $R_j = \sum_{i=1}^{m_j} r_{ij} \quad \forall j$   

$$v = n - 1 \quad gdl$$

Confronto d'ipotesi:  $\chi_e^2 \leq \chi_t^2$

**Test di Friedman per campioni qualsiasi**

Dati d'ingresso: numero di campioni  $n$   
 numerosità di ciascun campione  $m$   
 ranghi, in corrispondenza alle unioni, un elemento alla volta per ogni campione, dei valori argomentali ordinati in modo crescente  $r_{ij}$   
 livello di significatività  $\alpha$

Ipotesi fondamentale:  $H_0$  : uguaglianza dei valori centrali

Valore atteso: 
$$\chi_e^2 = \frac{12}{mn(n+1)} \sum_{j=1}^n R_j^2 - 3m(n+1) = \frac{6 \sum_{j=1}^n (R_j - \bar{R})^2}{\sum_{j=1}^n R_j}$$
 dove:  $R_j = \sum_{i=1}^{m_j} r_{ij} \quad \forall j$   

$$v = n - 1 \quad gdl$$

Confronto d'ipotesi:  $\chi_e^2 \leq \chi_t^2$

**Test di Bartlett per campioni normali**

Dati d'ingresso: numero di campioni  $n$

numerosità di ciascun campione	$m_j$
numerosità totale	$N = \sum_{j=1}^n m_j$
componenti della varianza	$\sigma_j^2$
livello di significatività	$\alpha$

Ipotesi fondamentale:  $H_0$  : uguaglianza delle componenti della varianza

Valore atteso:  $\chi_e^2 = -2 \ln \Lambda$

$$\text{dove: } \Lambda = \frac{\prod_{j=1}^n (\sigma_j^2)^{m_j/2}}{\left( \sum_{j=1}^n m_j \sigma_j^2 / N \right)^{N/2}}$$

$v = n - 1$  gdl

Confronto d'ipotesi:  $\chi_e^2 \leq \chi_t^2$

□ Test di Kruskal-Wallis per campioni indipendenti

Dati d'ingresso:	numero di campioni	$n$
	numerosità di ciascun campione	$m_j$
	numerosità totale	$N = \sum_{j=1}^n m_j$
	ranghi, in corrispondenza all'unione dei moduli degli scarti (rispetto alle mediane parziali) ordinati in modo crescente	$r_{ij}$
	livello di significatività	$\alpha$

Ipotesi fondamentale:  $H_0$  : uguaglianza delle componenti della varianza

Valore atteso, gradi di libertà e confronto d'ipotesi come per il test di Kruskal-Wallis (per valori centrali).

□ Test di Friedman per campioni qualsiasi

Dati d'ingresso:	numero di campioni	$n$
	numerosità di ciascun campione	$m$
	ranghi, in corrispondenza alle unioni, un elemento alla volta per ogni campione, dei moduli degli scarti (rispetto alle mediane parziali)	
	ordinati in modo crescente	$r_{i,j}$
	livello di significatività	$\alpha$

Ipotesi fondamentale:  $H_0$  : uguaglianza delle componenti della varianza

Valore atteso, gradi di libertà e confronto d'ipotesi come per il test di Friedman (per valori centrali).

□ Test di Hotelling per campioni normali

Dati d'ingresso:	numero di componenti del campione (multidimensionale)	$n$	
	numerosità di ciascun componente	$m$	
	matrice di varianza-covarianza	$C_{xx}$	(dove: $\sigma_x^2$ varianze degli elementi)
	livello di significatività	$\alpha$	

Ipotesi fondamentale:  $H_0$  : incorrelazione tra le componenti

Valore atteso:  $\chi_e^2 = -2 \ln \Lambda$

$$\text{dove: } \Lambda = \frac{(\det C_{xx})^{m/2}}{\left( \prod_{j=1}^n \sigma_{xj}^2 \right)^{m/2}}$$

$v = n(n - 1) / 2$  gdl

Confronto d'ipotesi:  $\chi_e^2 \leq \chi_t^2$

□ Test Lawley per campioni normali

Dati d'ingresso:	numero di componenti del campione (multidimensionale)	$n$
	numerosità di ciascun componente	$m$
	coefficiente di correlazione fra le componenti	$r_{ij}$
	livello di significatività	$\alpha$

Ipotesi fondamentale:  $H_0$  : incorrelazione tra le componenti

Valore atteso:  $\chi_e^2 = \left(m - 1 - \frac{2n + 5}{6}\right) \sum_{i=1}^{n-1} \sum_{j=i+1}^n r_{ij}^2$  con  $m > \frac{2n + 11}{6}$   
 $\nu = n(n - 1) / 2 \quad gdl$

Confronto d'ipotesi:  $\chi_e^2 \leq \chi_t^2$

□ Test di Wilcoxon–Wilcox modificato secondo Lawley per campioni qualsiasi

Dati d'ingresso: numero di componenti del campione (multidimensionale)  $n$   
 numerosità di ciascun componente  $m$   
 coefficienti di correlazione sui ranghi di Spearman, disposti in una matrice di correlazione  $r_{ij}$   
 livello di significatività  $\alpha$

Ipotesi fondamentale:  $H_0$  : incorrelazione tra le componenti

Valore atteso:  $\chi_e^2 = \left(m - 1 - \frac{2n + 5}{6}\right) \sum_{i=1}^{n-1} \sum_{j=i+1}^n r_{ij}^2$   $\nu = n(n - 1) / 2 \quad gdl$

Confronto d'ipotesi:  $\chi_e^2 \leq \chi_t^2$

Concludendo quest'elencazione di test statistici, giova ribadire che l'inferenza statistica multivariata è forse la parte più ampia dell'analisi multivariata, cosicché innumerevoli sono i test multipli, noti in letteratura. A fianco ed oltre l'analisi di varianza, **controlli numerici**, altrettanto importanti ed utili, riguardano invece il **condizionamento del sistema** da risolvere e l'**affidabilità dello schema d'osservazione**.

Un discorso sul buon condizionamento del sistema da risolvere parte dall'ovvia considerazione che esso non deve essere, globalmente o localmente, singolare ed estende le precauzioni da prendere a tutte quelle situazioni prossime alla singolarità. In questo caso, si dice che il sistema da risolvere è, globalmente o localmente, mal-condizionato.

Pertanto una strategia d'allarme consiste nel calcolare un numero di condizione e nel valutarlo troppo prossimo a zero, ed una strategia di ricerca si effettua invece identificando tutti i valori singolari prossimi al valore singolare minimo. Un'opportuna aggiunta di osservazioni è capace di superare i problemi di condizionamento, formando un sistema ben condizionato.

In pratica, una strategia d'allarme calcola un numero di condizione, fondato su norme di matrici, da valutare se troppo prossime a zero:

$$\kappa_{\infty} = \frac{1}{\|C\|_{\infty} \cdot \|C^{-1}\|_{\infty}} \quad \text{dove:} \quad \|Q\|_{\infty} = \max_{i=1,n} \left( \sum_{j=1}^n |q_{ij}| \right)$$

essendo  $C$  la matrice normale,  $C^{-1}$  la sua matrice inversa e  $\| \cdot \|_{\infty}$  la norma dell'estremo superiore.

Dopodiché una strategia alternativa di ricerca si effettua identificando, nella matrice dei coefficienti di correlazione dei parametri, tutti quei coefficienti il cui valore assoluto è, relativamente, prossimo ad uno:  $R = (I * C^{-1})^{-1/2} C^{-1} (I * C^{-1})^{-1/2}$ , essendo  $C^{-1}$  la matrice inversa della matrice normale,  $I$  una matrice identità ed il simbolo  $*$  indica il prodotto di Hadamard.

Uno schema d'osservazione si dice affidabile, quando è in grado di identificare uno o più dati anomali nell'insieme delle osservazioni. Questo significa che la presenza di dati anomali, per quanto abbia sempre effetti distorcenti sulle stime, è grazie alla ridondanza globale e locale dello schema d'osservazione, comunque, evidenziata (cioè si sa che i dati anomali sono presenti) e localizzata (cioè si sa dove i dati anomali sono accaduti).

Come noto, per il teorema di decomposizione ortogonale della varianza, la ridondanza locale ha valore zero, quando un'osservazione è indispensabile, mentre ha valore limite uno, quando la stessa è del tutto superflua. Essendo ovviamente impossibile avere sempre il valore uno, valori superiori ad un quinto o un quarto dell'unità si considerano comunemente al di sopra di una ragionevole soglia di sicurezza, provvedendo ad un'opportuna aggiunta di osservazioni, ben mirata localmente, in caso di difetto. Allora un esame approfondito, della questione dell'affidabilità conduce alla definizione di affidabilità interna ed affidabilità esterna.

Si chiama affidabilità interna di una generica osservazione la quantità che rimane nel corrispondente scarto-residuo, a seguito dell'immissione di un effetto perturbativo. Questa è misurabile tramite l'espressione:

$$\nabla(\delta_i - y_{0i}) = \tau \sigma_{y_{0i}} / \sqrt{V_i}$$

essendo:  $\sigma_{y_{0i}}$  lo sqm dell' $i$ -esima osservazione,  $V_i$  la corrispondente ridondanza locale e  $\tau$  l'ascissa corrispondente al valore della curva di potenza (per la distribuzione di probabilità della variabile casuale di Thompson), scelto un livello di significatività ed una potenza del test.

Si chiama invece affidabilità esterna di una generica osservazione la quantità che fluisce nella stima di un certo parametro, distorcendo il suo valore atteso, a seguito dell'immissione dello stesso effetto perturbativo. Questa è misurabile tramite l'espressione (essendo:  $e_i$ , il versore unitario diretto secondo la componente  $i$ -esima del vettore delle osservazioni):

$$\nabla x_j = - \left( (A^T P A)_j^{-1} \right)^T A^T P e_i \nabla(\delta_i - y_{0i}) \quad \forall j$$