delle famiglie e quindi sapere se quando andiamo ad allocare le famiglie nei vari mesi esse sono già uscite dalla popolazione o no.

*Terza parte*

# *Analisi statistica multivariata*

**INTRODUZIONE**

L'argomento che andiamo ad introdurre è l'uso di variabili ausiliarie in fase di stima, quindi non a monte durante l'estrazione del campione ma alla fine, dopo che sono tornati i dati, quindi non utilizzeremo lo stimatore di Hurvitz-Thompson ma qualcosa di un po' più complicato che tenga conto dell'introduzione di informazioni ausiliarie. Per fare questo però c'è bisogno di spiegare un po' come funziona una cosa che probabilmente non abbiamo fatto, la regressione multipla, noi finora la regressione lineare la sappiamo fare con una sola *X*, ossia,

$$Y = a + b \cdot x + \varepsilon$$

a statistica di base ci dovrebbero aver spiegato come si svolge la regressione quando vi è una sola variabile esplicativa e quindi come si trovano i coefficienti *a* e *b.* Il problema nasce quando le *x* sono più di una, quando abbiamo più variabili esplicative, e quindi, ad esempio,

$$Y = a + b \cdot x_1 + c \cdot x_2 + d \cdot x_3 + e \cdot x_4 + f \cdot x_5 + g \cdot x_6 + \varepsilon$$

Come si fa? Con il calcolo matriciale ormai siamo diventati degli esperti, ne abbiamo già parlato nelle pagine precedenti, ed è proprio quest'ultimo che ci permetterà di capire gli argomenti che tratteremo.

## 14. La regressione multipla

Innanzitutto l'espressione precedente la riscriviamo in termini matriciali, e quindi il vettore-colonna delle *Y*, dove abbiamo tutti i dati della *Y*,

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

un vettore-colonna $\varepsilon$ con tutti gli errori,

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

ed una matrice *X* dove abbiamo tutte le variabili esplicative, ovvero tutte le nostre *x*.

$$X = \left(\begin{array}{c|cccc} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & \ddots & & x_{k2} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & x_{1n} & \cdots & \cdots & x_{kn} \end{array}\right)$$

Questa *matrice-X* è costruita in questo modo: ha una prima colonna costituita di tutti valori unitari, una seconda colonna dove abbiamo la prima variabile esplicativa per la prima unità osservata *($x_{11}$)*, la prima variabile esplicativa per la seconda unità osservata*($x_{12}$)*, fino alla *k-esima* variabile esplicativa nell'ultima colonna per la prima unità osservata*($x_{k1}$)*, per la seconda unità*($x_{k2}$)* osservata e così via.
Poi definiamo $\beta$ un vettore di parametri di regressione e che dobbiamo stimare,

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

dove $\beta_0$ rappresenta quella che solitamente indichiamo con *a***,** ovvero l'intercetta.

Stando le cose in questo modo invece di scrivere tutte le somme in termini matriciali si può scrivere, che il vettore delle *y*

$$Y = X \cdot \beta + \varepsilon$$

è uguale alla matrice *X* per il vettore $\beta$ + il vettore $\varepsilon$ infatti se proviamo a fare il prodotto matriciale, scrivere in quel modo significa fare $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ..... + \beta_k x_k$ etc., abbiamo semplicemente riscritto la sommatoria in termini matriciali ovvero l'espressione equivale a dire che per ogni unità *i* è la stessa cosa fare,

$$y_i = \beta_0 + \sum_J \beta_J \cdot x_J + \varepsilon_i$$

ovvero le due forme di scrittura sono identiche, non è altro, in definitiva, che la definizione di prodotto matriciale. Ora il nostro obiettivo è riuscire a trovare una stima per il vettore $\beta$, ovvero dobbiamo trovare una stima dei parametri di regressione, che indicheremo con $\hat{\beta}_{\mathrm{LS}}$ dove *"LS"* sta per "minimi quadrati" in inglese.

$$Y = X \cdot \beta + \varepsilon$$
$$\uparrow$$
$$\hat{\beta}_{\mathrm{LS}}$$

Ma cosa significa "minimi quadrati"? Significa semplicemente che se il *vettore dei residui* altro non è che

$$\varepsilon = y - x \cdot \beta$$

e visto che la somma dei quadrati dei residui altro non è che

$$\varepsilon^T \cdot \varepsilon$$

allora dobbiamo trovare quel $\beta$ che minimizza

$$\underset{\beta}{MIN}\left(\varepsilon^T \cdot \varepsilon\right) = \underset{\beta}{MIN}\left[\left(y - x \cdot \beta\right)^T \cdot \left(y - x \cdot \beta\right)\right]$$

dove $\left(\varepsilon^T \cdot \varepsilon\right)$ che è un vettore trasposto per se stesso, non è altro che

$$\left(\varepsilon^T \cdot \varepsilon\right) = \begin{bmatrix}\varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n\end{bmatrix} \cdot \begin{bmatrix}\varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n\end{bmatrix} = \varepsilon_1 \cdot \varepsilon_1 + \varepsilon_2 \cdot \varepsilon_2 + ..... + \varepsilon_n \cdot \varepsilon_n == \varepsilon_1^{\,2} + \varepsilon_2^{\,2} + ..... + \varepsilon_n^{\,2}$$

ovvero la somma dei quadrati dei residui. Ora per trovare il minimo, come facevamo a statistica di base per una sola variabile, dobbiamo calcolare la derivata rispetto a $\beta$ dell'espressione che esprime il problema di minimo ed uguagliare a zero,

$$\underset{\beta}{MIN}\left(\varepsilon^T \cdot \varepsilon\right) = \underset{\beta}{MIN}\left[\left(y - x \cdot \beta\right)^T \cdot \left(y - x \cdot \beta\right)\right] = 0$$

e quello che otteniamo uguagliando a zero è uno stimatore dei minimi quadrati.

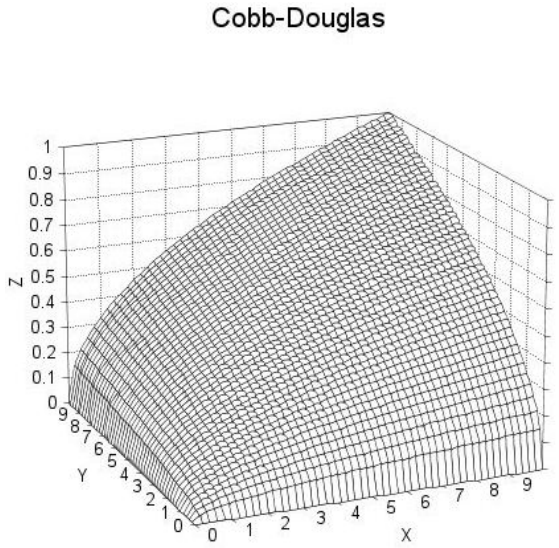
$$\hat{\beta}_{LS} = \left(x^T \cdot x\right)^{-1} \cdot x^T \cdot y$$

Come possiamo notare le formule riscritte in termini matriciali sono molto più semplici, se ci pensiamo già con la regressione lineare la formula è molto più lunga senza utilizzare le matrici, ora la formula è molto più semplice a prescindere dal numero di variabili esplicative che abbiamo, ossia per regredire su una *Y* possiamo mettere anche *10* variabili esplicative ma la formula non cambia, aumenta solo il numero di colonne da scrivere nella matrice *X*. La regressione multipla con i nostri stimatori centra perché se dobbiamo utilizzare delle variabili ausiliare, utilizzare delle variabili ausiliarie significa che se abbiamo una variabile *Y* che deve essere rilevata avremo delle variabili ausiliare

nell'archivio, ed il modo migliore per utilizzare le variabili ausiliare è farci una regressione sopra che sarà *lo stimatore di regressione*. Ma prima di arrivare a capire come si utilizza lo stimatore di regressione bisogna capire cosa è la regressione, un po' più complessa della regressione lineare semplice che si fa a statistica di base.

## 14.1 Ma cos'è la regressione lineare?

Tra le altre cose la regressione lineare multipla è qualcosa che come laureati o laureandi in economia dovremmo conoscere; supponiamo di dover andare a cercare quali sono le esplicative in una variabile economica, ovvero - qualcuno si ricorda cosa sono le funzioni di produzione? Stiamo parlando di economia, non di statistica - ricordate cos'è la *Cobb-Douglas*? La *Coubb-Douglas* ci da la produzione per determinate quantità di capitale e lavoro,




$$Y = \alpha \cdot K^{\beta} \cdot L^{\gamma}$$

*Fig. 14.1 – La funzione di produzione*

Ma cosa rappresentano $\beta$ e $\gamma$? Non sono altro che le elasticità di sostituzione del capitale e del lavoro, ovvero di quanto aumenta la produzione se aumentiamo di un'unità la quantità di capitale o la quantità di lavoro. Quindi se abbiamo i dati sul capitale e sul

lavoro dell'economia italiana come facciamo a determinare i parametri $\alpha$, $\beta$ e $\gamma$ ? Oppure in altro modo, abbiamo *100* imprese, a ciascuna chiediamo quant'è la produzione, il capitale ed il lavoro, e vogliamo sapere quant'è l'elasticità di queste imprese, l'elasticità della produzione al capitale ed al lavoro, o ancora, tradotto in italiano, l'amministratore delegato potrebbe chiederci: se io aumento un'unità di lavoro di quanto mi aumenta la produzione? Come facciamo a trovare quei valori? $\alpha$, $\beta$ e $\gamma$? Sempre restando in tema passiamo ai logaritmi, la funzione di produzione in ambo i membri la *linearizziamo* e diventa

$$\ln Y = \ln \alpha + \beta \cdot \ln K + \gamma \cdot \ln L$$

che a prima vista sembra proprio una regressione lineare in cui abbiamo tre parametri, $\alpha$, $\beta$ e $\gamma$, e tre variabili, una dipendente e due indipendenti ovvero *ln(Y)*, *ln(K)* e *ln(L)*, e quindi basta che riscriviamo il vettore delle *Y*,

$$Y = \begin{pmatrix} \ln y_1 \\ \ln y_2 \\ \vdots \\ \ln y_n \end{pmatrix}$$

la matrice delle *x* diventa

$$X = \left(\begin{array}{c|cc} 1 & \ln k_1 & \ln L_1 \\ 1 & \ln k_2 & \ln L_2 \\ \vdots & \vdots & \vdots \\ 1 & \ln k_n & \ln L_n \end{array}\right)$$

dove la prima colonna serve per stimare il parametro $\alpha$, nella seconda colonna il logaritmo del capitale della prima impresa, della seconda fino all'*n*-esima e stessa cosa per la terza colonna per il lavoro. Successivamente applichiamo la nostra formula,

$$\hat{\beta}_{LS} = \left(x^T \cdot x\right)^{-1} \cdot x^T \cdot y$$

ed *otteniamo i parametri che minimizzano i residui* della nostra funzione *Cobb-Douglas.*

Chiarendo ulteriormente il concetto della minimizzazione dei residui, graficamente significa (in sole due dimensioni ma ci stiamo occupando di più dimensioni) ricordando l'andamento della funzione *Coubb-Douglas*, vuol dire trovare tra le tante possibili curve, tra le tante possibili *Coubb-Douglas* che si possono ottenere al variare di $\alpha$, $\beta$ e $\gamma$, bisogna trovare quella per cui sono minimi gli scarti dei dati osservati dalla curva teorica, che nel nostro caso è la *Coubb-Douglas*, mentre i segmenti rappresentano gli scarti, i residui di cui abbiamo parlato finora, ogni segmento è un ε; detto in altro modo minimizzare i residui significa trovare tra le tante possibili curve quella che più si avvicina ai dati osservati, cosa abbastanza ragionevole.

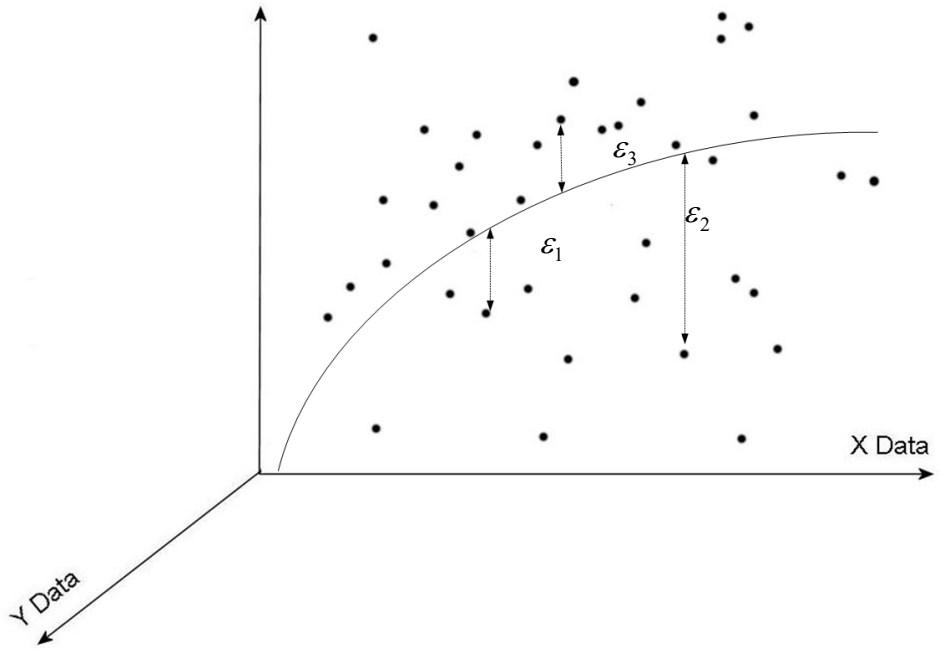



*Fig. 14.2 – I residui nella regressione multipla*

Ora, rimanendo sulla nostra funzione di produzione, i punti che hanno dei residui positivi, abbiamo detto che abbiamo *1000* imprese - i punti sopra la curva che informazione danno? Semplicemente significa che a parità di input la produttività delle imprese indicata dai punti sopra la curva è maggiore ovvero queste imprese sono più

efficienti a parità di fattori produttivi rispetto a quelle che si trovano sopra o sotto la curva. Nella figura abbiamo indicato ambedue i fattori ma bisogna ragionare solo con il lavoro, ribadiamo che siamo in due dimensioni, quindi supponiamo di avere solo il lavoro e che il capitale sia uguale per tutte le imprese, vuol dire che all'aumentare del lavoro aumenta la produzione secondo la nostra *Cobb-Douglas*. Avere un residuo positivo significa che le imprese che abbiamo considerato, se la nostra *Cobb-Douglas* indica un andamento medio, allora la curva indica quanto dovrebbero produrre mediamente con le rispettive quantità di lavoro aumentando o diminuendo la quantità di input; le imprese che stanno sopra la curva quindi sono quelle che si comportano meglio tra le *1000*, sono le imprese più efficienti, producono di più con la stessa quantità di lavoro, viceversa le imprese che hanno un residuo negativo. Quindi, la stima di una funzione di produzione è uno dei modi classici e standard per calcolare efficienze o inefficienze di imprese o addirittura di settori, nulla ci vieta infatti, ad esempio, di considerare *1000* sportelli bancari dove la produzione, l'output, è rappresentato dai depositi bancari e l'input solo dal lavoro in quanto il capitale conta poco in questo settore.

Quello che abbiamo detto finora è uno dei tanti esempi di utilizzo della regressione multipla, infatti andando a spulciare un qualsiasi testo di economia, di esempi di relazioni lineari o linearizzabili, (come nell'esempio che abbiamo fatto dove la relazione era apparentemente non-lineare ma facilmente linearizzabile passando per i logaritmi) se ne possono trovare molti, spesso si trovano dei parametri che non si sa mai come calcolarli, poi lo si fa con la regressione lineare semplice.

## 14.2 Gauss in più dimensioni, multicollinearità ed eteroschedasticità?

Torniamo agli aspetti statistici e tralasciamo per ora gli aspetti applicativi della regressione multipla; la relazione che abbiamo enunciato,

$$\hat{\beta}_{LS} = \left(x^T \cdot x\right)^{-1} \cdot x^T \cdot y$$

quand'è che salta? Quand'è che non funziona? Sicuramente quando l'espressione $\left(x^T \cdot x\right)$ *non è quadrata*, ma sappiamo che lo è sempre, quindi non è un problema che dobbiamo porci; oppure quando il *rango non è pieno ossia quando una colonna o una riga è combinazione lineare delle altre,*

$$DET\left(x^T \cdot x\right) = 0$$

*una matrice non è invertibile* quando una *X* tra le varie colonne è combinazione di un'altra *X* o più X, ovvero esiste una riga-*i* o una colonna-*j* per cui vale la relazione precedente, detto in altre parole una variabile ausiliaria è calcolabile in modo deterministico sulla base di un'altra, allora in questo caso la matrice non è invertibile.
Questo problema di chiama <u>*Multicollinearità*</u> *delle variabili ausiliarie* ovvero le ausiliarie sono troppo correlate l'una con l'altra o ancora ce n'è qualcuna di troppo; ad esempio supponiamo di avere la nostra *Y = spesa* in ricerca e sviluppo *(R&S)* e la mettiamo in regressione con gli addetti e il fatturato. Siccome abbiamo detto che le variabili ausiliarie addetti e fatturato sono molto correlate l'una con l'altra ci troviamo in una situazione di *Multicollinearità*, la quale indica che stiamo facendo una cosa, tra l'altro illogica, che non serve a nulla, perché utilizziamo come esplicative due variabili praticamente identiche, una delle due la dobbiamo togliere, quindi la *multicollinearità* si risolve togliendo una o più variabili tra quelle che troppo correlate l'una con l'altra, ripetendo che è illogico regredire su delle ausiliare che sono troppo correlate l'una con l'altra perché ridicono la stessa cosa. Ora ipotizziamo che il vettore $\varepsilon$ si distribuisce come una gaussiana con media zero e matrice di varianze e covarianze

$$\varepsilon \sim N\left(0, \sigma^2 I\right)$$

Probabilmente finora abbiamo sentito parlare di gaussiana ad una dimensione

$$X_1 \sim N\left(\mu, \sigma^2\right)$$

ovvero una normale con media $\mu$ e varianza $\sigma^2$,

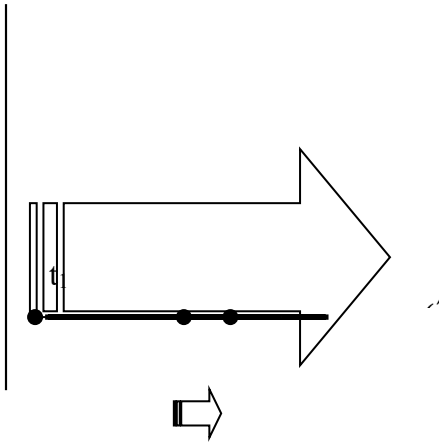


*Fig. 14.3 – Gaussiana in due dimensioni*

con due variabili aleatorie invece graficamente abbiamo una forma a tre dimensioni con una gobba centrale, (una sorta di lenzuolo preso al centro e tirato verso l'alto).

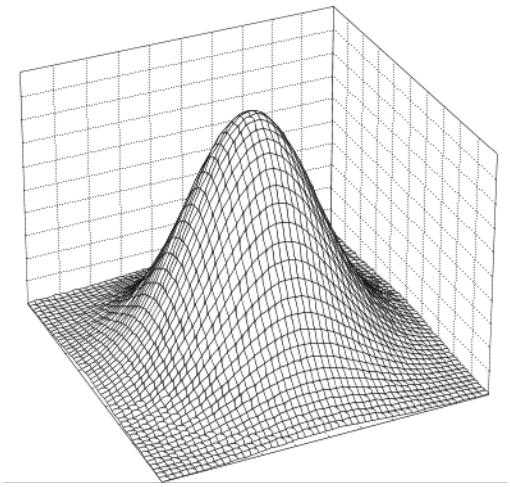


*Fig. 14.4(a) – Gaussiana in tre dimensioni*

che se sezionata sia rispetto ad una variabile sia rispetto all'altra genera sempre delle gaussiane in due dimensioni. Infatti come si vede dalla figura seguente, sezionando con

un piano (parallelo) rispetto ad $X_2$ otteniamo sempre una gaussiana, stessa cosa rispetto ad $X_1$.

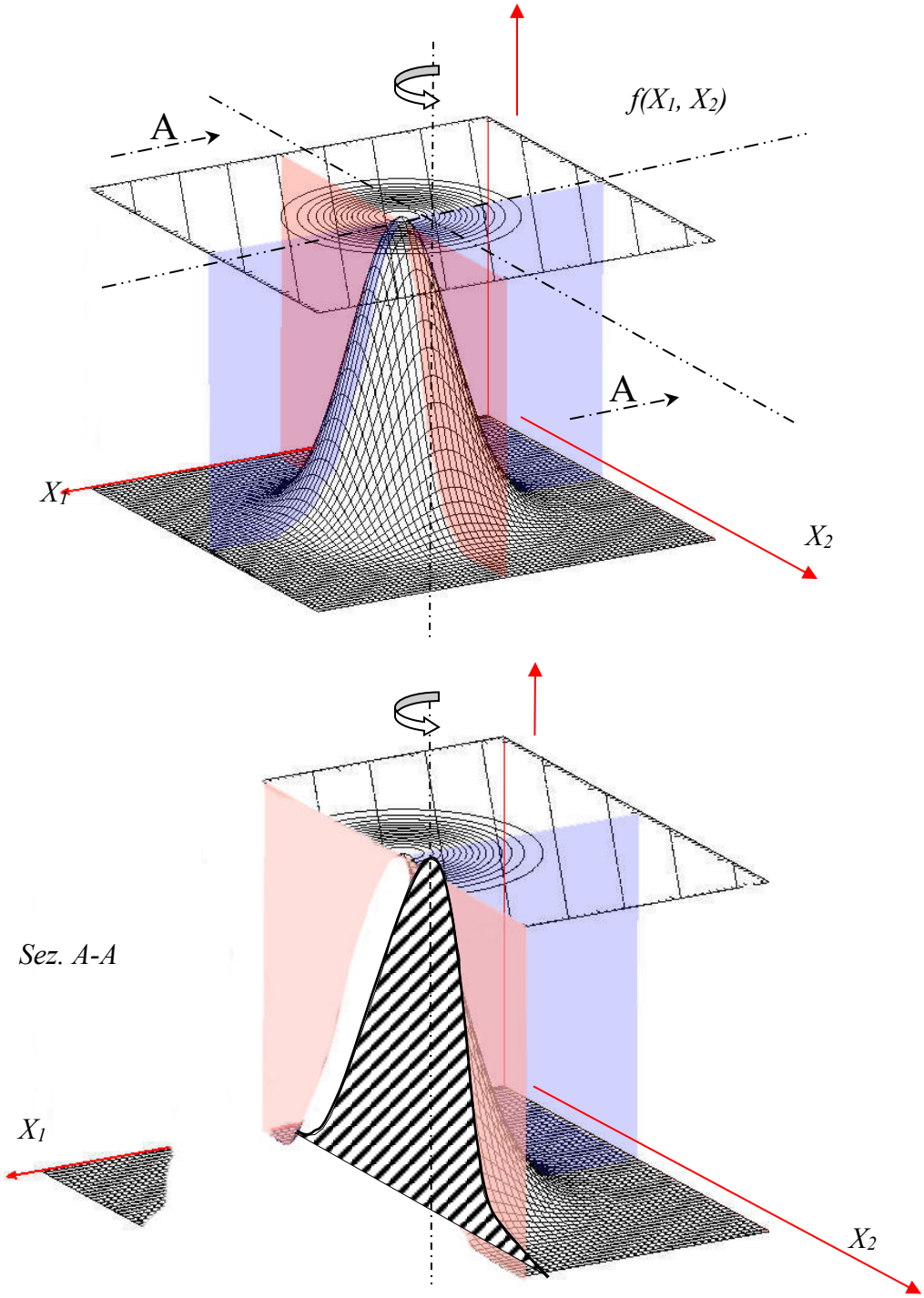
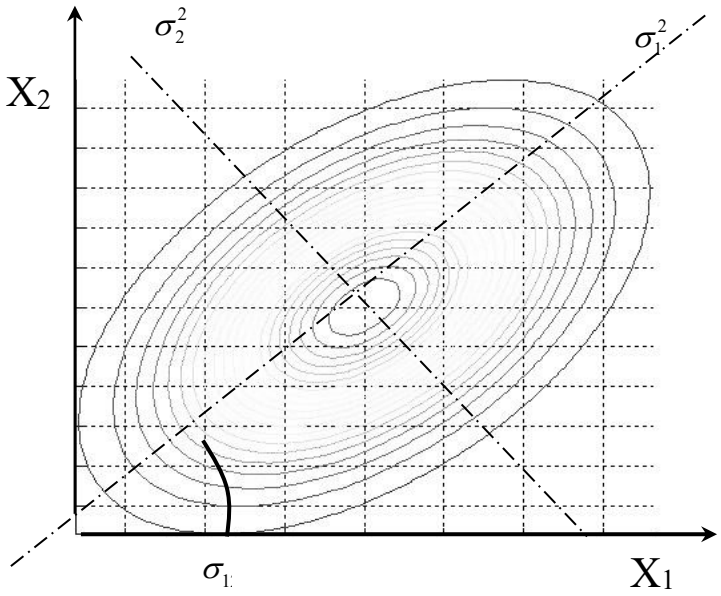
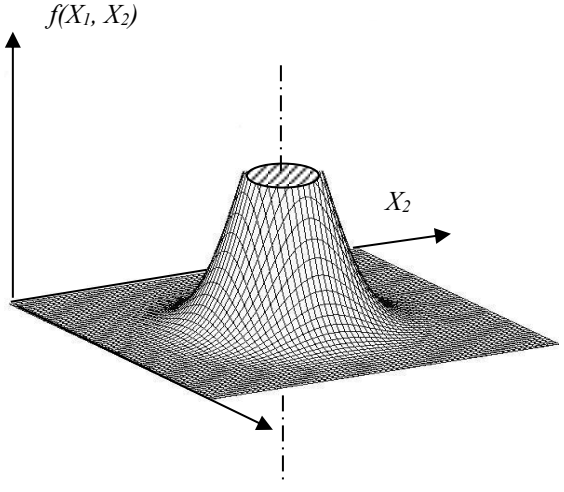



*Fig. 14.4 (b) – Gaussiana in tre dimensioni (attenzione, le curve di livello in alto assumono una forma più allungata nel nostro caso, come nel grafico successivo*

Abbiamo scelto di sezionare proprio sull'asse di rotazione ma il discorso vale per qualsiasi punto. Quello che dobbiamo definire ora non è più una media ed una varianza ma *un vettore di medie ed una matrice di varianze e covarianze*, con due medie quindi ed una matrice dove le varianze sono indicate sulla diagonale mentre le covarianze fuori dalla diagonale; ovviamente è una matrice simmetrica in quanto la covarianza tra una variabile ed un'altra variabile è la stessa se si inverte l'ordine, la covarianza tra $X_1$ ed $X_2$ è uguale alla covarianza tra $X_2$ ed $X_1$. Lo si può vedere chiaramente dalla figura che segue, (indicata anche nelle tre figure precedenti sul piano in alto), si ottiene sezionando con piani perpendicolari all'asse di rotazione della *Fig. 14.4*.

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$



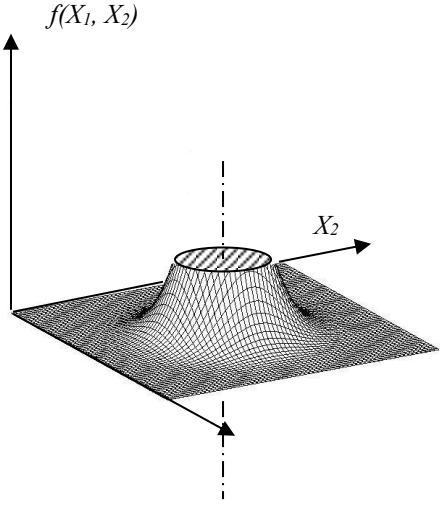




*Fig. 14.5 – Curve di livello o isolinee di una gaussiana tridimensionale*

Praticamente nella figura *14.5* abbiamo indicato quelle che vengono chiamate in cartografia *"isolinee"*, quelle che sulle carte geografie indicano le altimetrie dette anche *"curve di isolivello"* . Nel nostro caso *le curve di isolivello indicano la densità di probabilità*, quella che in due dimensioni abbiamo chiamato $p(x)$, ed ora chiamiamo $f(X_1, X_2)$. Quindi tagliando con piani paralleli all'asse di rotazione otteniamo delle ellissi concentriche man mano che aumenta la densità e ci avviciniamo al baricentro che sarebbero le due medie. L'inclinazione dell'ellisse come indicato nella figura *Fig. 14.5* è data dalla covarianza tra due variabili aleatorie, inclinazione che cambia a seconda che la covarianza tra le due variabili aleatorie sia positiva o negativa e quindi anche la posizione sul grafico cambierà a seconda di quest'ultima, mentre l'ampiezza dei due semiassi è data rispettivamente dalle due varianze, più aumenta la varianza di una variabile e più aumenta la larghezza dell'ellisse, viceversa nel caso contrario. Finora abbiamo trattato il caso di tre dimensioni ma il nostro ragionamento può estendersi a qualsiasi numero di dimensioni.

Alla luce delle nuove considerazioni quindi per quanto riguarda i nostri residui di regressione noi ipotizziamo una distribuzione *gaussiana multivariata* $\varepsilon \sim N\left(0, \sigma_\varepsilon^2 \cdot I\right)$ di media sempre zero, ogni residuo ha media zero e matrice di varianze e covarianze $\sigma_\varepsilon^2$ per *I,* dove *I* è la matrice identità, ossia che ha tutti uno sulla diagonale e zero fuori dalla diagonale; questo è un modo molto compatto per dire che ciascuna $\varepsilon$ è una gaussiana ad una dimensione di media *0* e varianza $\sigma_\varepsilon^2$ , sempre uguale su tutti gli $\varepsilon$ ed indipendenti l'uno con l'altro poiché abbiamo moltiplicato per *I* che ha zero fuori dalla diagonale, in pratica otteniamo una matrice di varianze e covarianze fatta in questo modo,

$$\begin{bmatrix} \sigma_\varepsilon^2 & 0 & 0 & 0 \\ 0 & \sigma_\varepsilon^2 & 0 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & 0 & \sigma_\varepsilon^2 \end{bmatrix}$$

sulla diagonale avremo sempre $\sigma_\varepsilon^2$ mentre fuori dalla diagonale sempre zero, ovvero i residui non sono correlati l'uno con l'altro, in altri termini $\text{cov}\left(\varepsilon_i, \varepsilon_j\right) = 0$ per ogni $i, j$.

L'ipotesi che abbiamo fatto in questo caso $\varepsilon \sim N\left(0, \sigma_\varepsilon^2 \cdot I\right)$ (con vettore di media zero) equivale a dire che *la varianza dei residui non cambia al variare dell'unità statistica* ed in più abbiamo ipotizzato che i residui non siano correlati l'uno con l'altro. Queste ipotesi hanno dei nomi un po' strani. Che la varianza sia sempre uguale al cambiare dell'unità statistica equivale a fare l'ipotesi di *omoschedasticità,* la cui versione contraria viene detta *eteroschedasticità*.

Ma quand'è che non c'è omoschedasticità? Riprendendo la nostra retta di regressione, se i residui inizialmente sono piccoli e poi aumentano di dimensione all'aumentare della variabile *X* otteniamo un tipico esempio di *eteroschedasticità,* e tra l'altro è abbastanza frequente, ovvero più aumenta la variabile e più si sbaglia, e più il residuo di regressione aumenta.

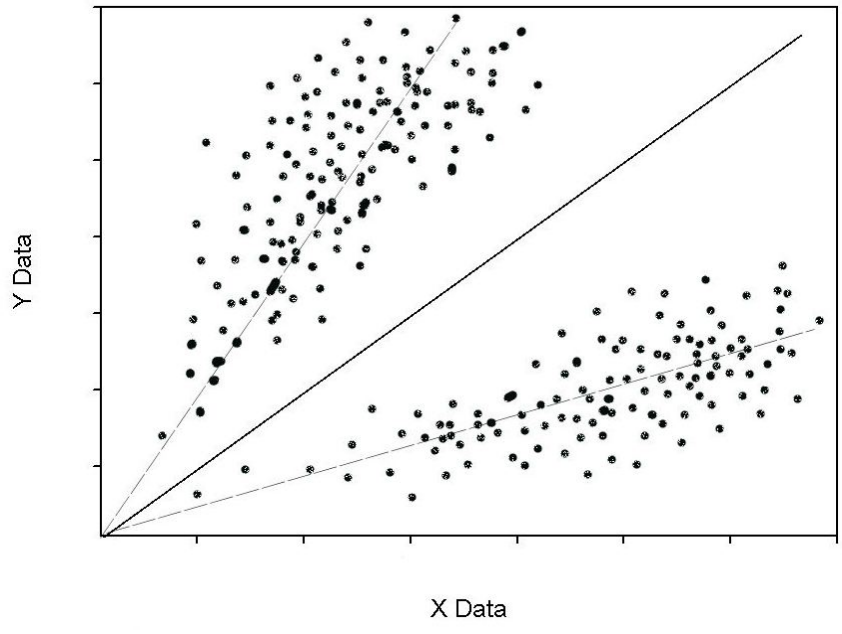


*Fig. 14.6(a) – Eteroschedasticità in due dimensioni*

L'altra ipotesi invece, ovvero della covarianza pari a zero, quando viene violata, ovvero quando non c'è indipendenza dei residui si dice che c'è correlazione seriale dei residui o *autocorrelazione dei residui.* Cosa si fa in questi casi? Come affrontiamo il problema quando le ipotesi che abbiamo fatto non vengono rispettate? Quando tutte le ipotesi vengono rispettate $\varepsilon \sim N\left(0, \sigma_\varepsilon^2 \cdot I\right)$, ovvero c'è *omoschedasticità ed indipendenza dei residui* allora lo stimatore

$$\hat{\beta}_{LS} = \left(x^T \cdot x\right)^{-1} \cdot x^T \cdot y$$

può essere utilizzato, ossia gode di tutte le proprietà quali correttezza, efficienza, etc. Quando invece *i residui, che continuano comunque ad essere di media zero* poiché è un'ipotesi che non ha senso violare, non hanno più la matrice di varianze e covarianze senza forma semplificata

$$\varepsilon \sim N\left(0, \sigma_\varepsilon^2 \cdot I\right)$$

ma una qualsiasi forma

$$\varepsilon \sim N\left(0, \Sigma\right)$$

quindi ci può essere sia eteroschedasticità ovvero varianze sulla diagonale che cambiano valori diversi sulla diagonale, in questo caso la stima $\hat{\beta}_{LS}$ diventa,

$$\hat{\beta}_{WLS} = \left(X^T \cdot \Sigma^{-1}\right)^{-1} \cdot X^T \cdot \Sigma^{-1} \cdot Y$$

poiché *viene pesata per la matrice di varianze e covarianze*, come prima solo che c'è l'inversa della matrice di varianze e covarianze generica dopo *X trasposta*, la matrice di varianze e covarianze serve per pesare i dati in modo diverso rispetto al precedente; vuol dire che se abbiamo dei residui che hanno una varianza diversa, quando andremo a minimizzare la somma dei quadrati dei residui *peseremo di più quelli che hanno una varianza inferiore*, tutto qui, ovvero li pesiamo per l'inverso della varianza, per $\Sigma^{-1}$,

quindi più hanno varianza e meno li pesiamo, è lo stesso discorso che riguarda gli intervalli di confidenza, un residuo peserà in modo diverso a seconda che la varianza sia molto alta o molto bassa, se una varianza è molto alta naturalmente un residuo avrà minor peso. Ancora, un errore di *1* in una popolazione che ha una varianza di *10*, essere distante dalla media di *1* in una varianza di *10* ha un significato, in una varianza di *100* ne ha un altro;

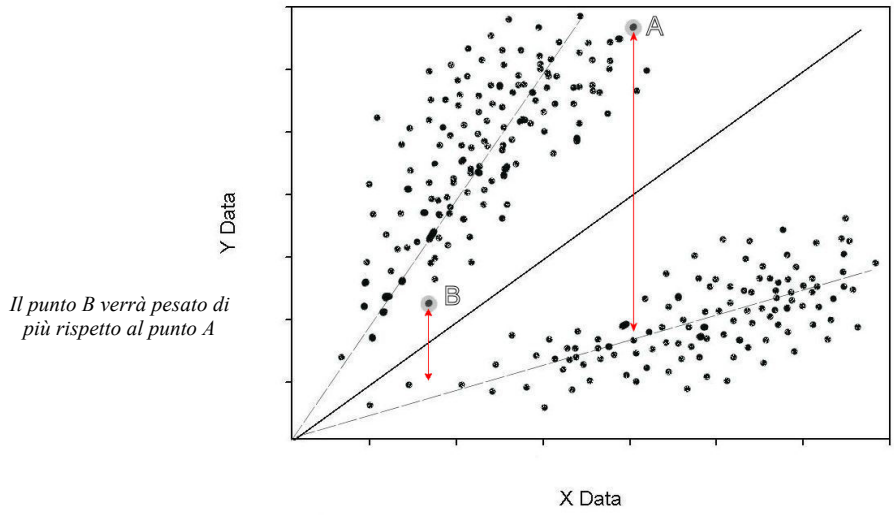



*Fig. 14.6(b) – Eteroschedasticità in due dimensioni*

In breve più è alta la varianza e meno pesa la distanza dalla media. *Quindi quando si rimuovono le ipotesi di indipendenza e di omoschedasticità bisogna pesare le matrici.*

Teniamo presente che per analizzare dati economici, queste due ipotesi sono entrambe rimosse quasi immediatamente, perché, supponendo che i nostri dati siano dati di serie storiche ad esempio abbiamo produzione, capitale e lavoro, ritornando sulla *Cobb-Douglas*, abbiamo *25* numeri che sarebbero i dati italiani di produzione, capitale e lavoro di *25* anni, dire che c'è indipendenza dei residui significa ipotizzare che da un anno all'altro, nessuna delle tre variabili, produzione, capitale e lavoro dipende dal valore assunto nell'anno precedente, ipotesi che in economia non sussiste.

Oppure se preferiamo facciamo una regressione su dei dati di finanza in borsa ed ipotizziamo che nel valore di un'azione da una realizzazione alla precedente ci sia indipendenza, un andamento completamente random, questo non è assolutamente ammissibile. L'indipendenza tra valori precedenti e valori successivi in economia è un'ipotesi quasi folle, quindi se non l'omoschedasticità, che comunque viene rimossa,

ipotizzare una varianza costante dei dati nel tempo significa ipotizzare una costanza di situazioni nel tempo, come dire che da prima dell'*11* settembre a dopo l'*11* settembre sia rimasto tutto uguale, oppure ipotizzare che la varianza dei redditi dagli anni *'70* ad adesso sia rimasta costante non ha senso, al massimo possiamo ipotizzare che cresca, che diminuisca, ma non che sia costante. In definitiva per fare regressione di dati economici queste ipotesi vengono eliminate e si usa la versione pesata dello stimatore.

Per concludere enunciamo la formula della varianza delle nostre stime nei due casi in cui valgono le ipotesi e nel caso in cui queste vengano a cadere,

$$V\left(\hat{\beta}_{LS}\right) = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot Y \cdot \sigma_\varepsilon^2$$

mentre *se cadono le ipotesi*,

$$V\left(\hat{\beta}_{WLS}\right) = \left(X^T \cdot \Sigma^{-1} \cdot X\right)^{-1}$$

come vediamo in questo secondo stimatore non c'è più $\sigma_\varepsilon^2$ perché è già inglobato nella formula. Quindi la varianza di queste stime è la prima parte dello stimatore.

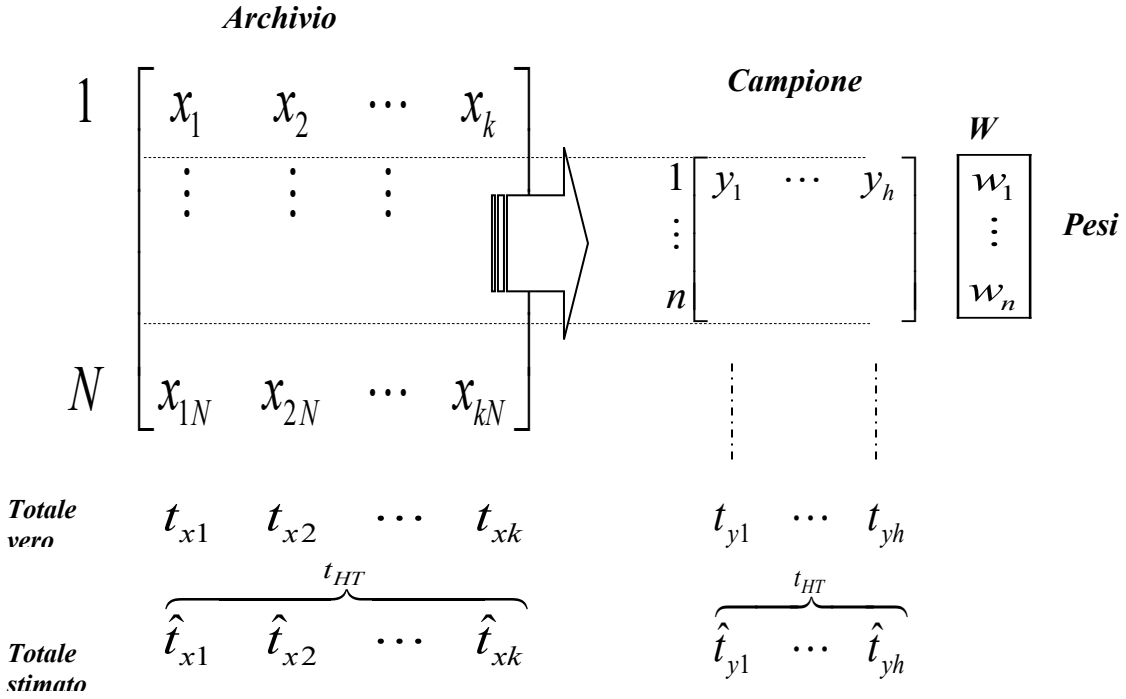
Ci fermiamo qui con la regressione multipla in quanto è argomento che viene trattato ampiamente in un'altra materia, in econometria, quanto utile a noi lo abbiamo trattato, il nostro scopo era riuscire a capire che il procedimento trattato, *con delle variabili ausiliarie, può aiutarci ad affinare le nostre stime sulla nostra variabile d'interesse e ci aiuta a ridurre la varianza delle stesse.*

## 15. Applicazioni della regressione multipla

### 15.1 Lo stimatore del quoziente

Vediamo ora a cosa serve la regressione multipla trattata precedentemente in termini di stime campionarie. Abbiamo il solito archivio di *N* elementi all'interno del quale conosciamo *K* ausiliarie e da questo archivio estraiamo secondo qualche criterio il nostro campione, da *1* ad *n* unità, in cui osserviamo le nostre *H* variabili d'interesse, quelle contenute nel questionario, calcoliamo i nostri pesi di riporto all'universo e

tramite somme pesate applichiamo lo stimatore di Horvitz-Thompson classico. Ora questi pesi li possiamo applicare alle *Y*, visto che conosciamo anche le *X* per le nostre unità, possiamo applicarli anche alle X, le nostre variabili ausiliarie,




e calcolare lo stimatore di Horvitz-Thompson per *X* allo stesso modo di come lo calcoliamo per le *Y* ossia ($\hat{t}_{x1}$ $\hat{t}_{x2}$ $\cdots$ $\hat{t}_{xk}$), ovviamente tenendo presente che anche per le *X* la stima è fatta solo sugli elementi del campione come per le *Y*, e che i valori veri delle *Y* sono ignoti.

D'altro canto, conosciamo anche i valori veri dei totali per le *X*, basta fare delle somme ($t_{x1}$ $t_{x2}$ $\cdots$ $t_{xk}$). Ora la domanda da porsi è: esiste una differenza tra i totali stimati ed i totali veri della *X*? Certamente, dovremmo avere molta fortuna affinché non ce ne sia troppa, c'è una probabilità molto bassa che questo avvenga, ovvero che i totali stimati della *X* siano uguali ai totali veri.

*Ma se c'è una differenza tra le stime ed i valori veri della X, è ragionevole ipotizzare che quest'ultima sia la stessa differenza che c'è tra i valori stimati ed i valori veri, ma ignoti, della Y?* Se esiste, come facciamo ad applicare una differenza

riscontrata sulla stima delle *X* alle *Y*? Come la applichiamo alle *Y*? Ebbene possiamo farlo attraverso il cosiddetto *stimatore del quoziente*:

$$\hat{t}_{y_1,Q} = \hat{t}_{y_1,HT} \cdot \frac{t_{x_1}}{\hat{t}_{x,HT}}$$

si prende lo stimatore di *Horvitz-Thompson*, e lo si moltiplica per il rapporto tra il valore vero di una variabile ausiliaria ed il valore stimato della stessa variabile; in pratica se la nostra stima, ad esempio, ha un errore del *10%* in meno rispetto al valore vero della popolazione, i totali stimati rispetto ai totali veri, allora alle *Y* dovremo togliere dalla stima calcolata il *10%* poiché ipotizziamo che il nostro errore percentuale in quota, così come lo abbiamo riscontrato sulle *X*, si riproponga tale e quale sulle *Y*, concettualmente significa estrarre un campione, assegnare agli elementi i valori delle *X*, riscontare un errore del *10%* rispetto ad i valori veri che conosciamo ed ipotizzare che se avessimo assegnato i valori delle *Y* invece che i valori delle *X* avremmo riscontrato lo stesso errore, quindi è ragionevole.

Ma quand'è che funziona questo ragionamento? Quali sono di fatto le ipotesi che facciamo affinché sia ragionevole che l'errore che si riscontra sulla *X* sia lo stesso che si riscontra in modo proporzionale *Y*? Le ipotesi sono:

1. La *X* è fortemente correlata con la *Y.*
2. La relazione sull'errore che si presenta sulla *X* e sulla *Y* è di tipo proporzionale.

Per quanto riguarda la prima ipotesi concettualmente vogliamo dire che se sbagliamo su una variabile sbagliamo allo stesso modo anche sull'altra, mentre la seconda ipotesi ci dice che l'errore, che si ripropone in modo moltiplicativo nello stimatore del quoziente, si ripropone in modo proporzionale sulla *Y*, che ci sia, come detto, una relazione proporzionale. Tradotto in termini grafici non solo ipotizziamo che i punti, come mostrato nel grafico seguente, ottenuti dalla relazione tra la *X* e la *Y* siano molto vicini alla retta di regressione e quindi ci sia una buona relazione, e che inoltre la relazione sia di tipo proporzionale poiché la retta passa per l'origine, ipotizziamo anche che ci sia

proporzionalità diretta, non c'è nessuna aggiunta additiva ma solo il fattore moltiplicativo.

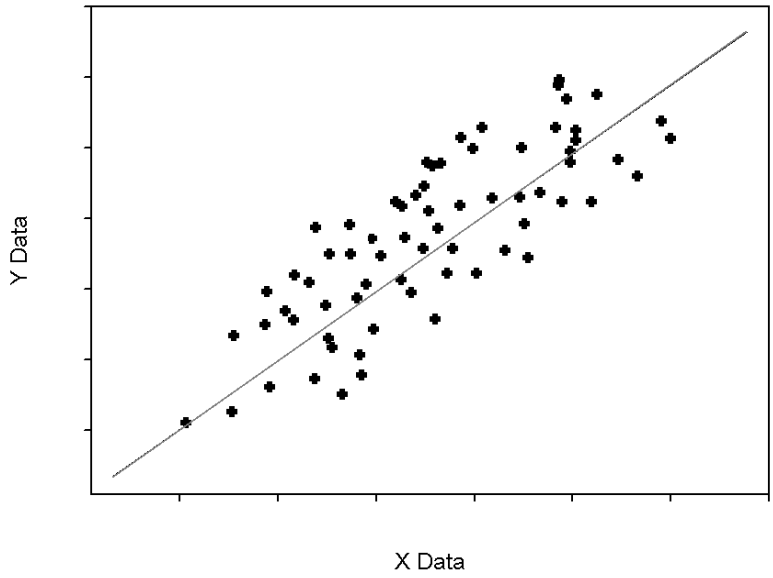



*Fig. 15.1 – Ipotesi di correlazione e proporzionalità delle X con le Y*

L'ipotesi che abbiamo fatto precedentemente è abbastanza ragionevole per le indagini sulle imprese ad esempio dove la variabile ausiliaria è quasi sempre il numero di addetti, molto proporzionale ad altre variabili, fatturato, etc. Tra l'altro è vero anche che l'ipotesi può non essere vera, la variabile *X* può essere si molto correlata con la *Y* ma molto probabilmente la retta di regressione non passerà per l'origine. Inoltre lo stimatore ha un grossissimo difetto, quello di *usare una sola variabile ausiliaria* quando in situazioni pratiche può capitare che di variabili ausiliarie se ne presentino a decine (per le imprese potrebbe anche essere sufficiente, la variabile addetti è abbastanza esaustiva).

Quindi da questa prima soluzione semplicistica dobbiamo arrivare ad un qualcosa di un po' più generale, dobbiamo estendere il concetto. Riflettiamo sul fatto che se abbiamo parlato di regressione lineare semplice nella discussione precedente, probabilmente possiamo estendere il concetto sulla regressione multipla, utilizzare un'intercetta, utilizzare più variabili ausiliarie, ed una relazione che lega una *y* a più *x*.

Lo stimatore del quoziente nel modo in cui è stato espresso è una soluzione molto semplicistica anche se spesso utilizzato nella realtà, e di fatto cosa fa, non fa nient' altro che applicare quasi una forma di *PPS a posteriori*, invece di estrarre le unità con probabilità proporzionali alla nostra unica *X* che conosciamo, l'informazione che possiamo ottenere dalla *X* la riapplichiamo alla fine invece che all'inizio in fase di estrazione, per far ritornare il totale noto di una variabile ausiliaria, riapplichiamo con proporzionalità ad i valori di una *X*.

Ora *lo stimatore del quoziente di cui abbiamo parlato è una caso particolare dello stimatore di regressione* di cui stiamo per parlare, ovvero nel caso in cui forziamo l'intercetta ad essere zero ed usiamo una sola variabile ausiliaria, quindi lo abbandoniamo e cominciamo a trattare qualcosa di più generale.

## 15.2 Lo stimatore di regressione

Lo stimatore di regressione prende lo stimatore di Horvitz –Thompson

$$\hat{t}_{y,REG} = \hat{t}_{y,HT} + (t_x - \hat{t}_{x,HT})^T \cdot \hat{B}$$

con vettori $t_x = \begin{bmatrix} t_{x1} \\ t_{x2} \\ \vdots \\ t_{xk} \end{bmatrix}$ $\hat{t}_{x,HT} = \begin{bmatrix} \hat{t}_{x1,HT} \\ \hat{t}_{x2,HT} \\ \vdots \\ \hat{t}_{xk,HT} \end{bmatrix}$ $\hat{B} = \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \\ \vdots \\ \hat{B}_k \end{bmatrix}$

e ci aggiunge la differenza riscontrata tra i valori veri della popolazione sulle variabili ausiliarie e le rispettive stime di Hurvitz-Thompson sempre sulle variabili ausiliarie, mentre $\hat{B}$ è il vettore dei parametri di regressione che dovremo stimare.

Ma dov'è l'intercetta all'interno dell'equazione? Ancora non è stata inserita, l'avremmo inserita solo ed esclusivamente se tra le ausiliarie ce ne fosse stata una

sempre uguale ad uno, ricordiamo infatti dalla regressione multipla che mettere *l'intercetta significa inserire una variabile ausiliaria sempre costante su tutta la popolazione*, quindi non mettiamo di fatto l'intercetta perché sarebbe banale, infatti basterebbe aggiungere un'ausiliaria qualsiasi sempre costante su tutta la popolazione uguale ad uno, ed avremmo in pratica il numero di unità statistiche. Ma poiché sappiamo che la somma nota della popolazione è *N* e nel nostro campione è *n*, e che la differenza tra *N* ed *n*, cioè quanto manca da *N* ad *n*, è ciò che fa l'Horvitz-Thompson ovvero far tornare il numero di osservazioni, l'intercetta allora diventa uno qualsiasi dei parametri ( $\hat{B}_1 \quad \hat{B}_2 \quad \cdots \quad \hat{B}_k$ ) relativo alla variabile sempre costante ed uguale ad uno su tutta la popolazione, abbiamo messo quindi anche l'intercetta, bastava estendere un po' il concetto di variabile ausiliaria.

Ma quali sono le caratteristiche di questo stimatore? La prima caratteristica è che possiamo utilizzare un qualsiasi numero di variabili ausiliarie, non dobbiamo mettere l'intercetta, quindi è molto più generale del precedente e la *stima dei coefficienti di regressione viene fatta come abbiamo già visto tramite la stima dei minimi quadrati pesati,*

$$\hat{\beta}_{WLS} = \left( X^T \cdot \Sigma^{-1} \cdot X \right)^{-1} \cdot X^T \cdot \Sigma^{-1} \cdot Y$$

dove la matrice di varianze e covarianze viene messa non specifica ma generale perché se i campioni sono estratti in modo indipendente la matrice sarà si diagonale poiché non c'è covarianza proprio perché li abbiamo estratti in modo indipendente, però sulla diagonale potrebbe esserci eteroschedasticità;

$$\begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

ad esempio se il nostro campione è stratificato all'interno di ciascuno strato le varianze possono essere diverse e quindi ipotizzare che su ciascuna unità le varianze siano uguali non è molto lecito, per cui la varianza possiamo lasciarla libera di cambiare da unità ad unità. Possiamo tra l'altro trasformare in sommatoria la precedente espressione sviluppando i prodotti matriciali, dove $\pi_k$ rappresenta il peso di riporto all'universo delle *X*.

$$\hat{B} = \left( \frac{\sum\sum x_1 \cdot x_k^T}{\sigma_k^2 \cdot \pi_k} \right) \cdot \left( \frac{\sum x_k \cdot y_k}{\sigma_k^2 \cdot \pi_k} \right)$$

Quindi in definitiva facciamo una regressione tra ogni *y* e tutte le *x*, calcoliamo i $\hat{B}$ con una delle due formule ed a questo punto pesiamo con i $\hat{B}$ di regressione le differenze ottenute sulle ausiliarie, questa somma pesata di differenze ottenuta sulle ausiliarie la aggiungiamo allo stimatore di Horvitz-Thompson, quindi la cosa di per sé, tralasciando lo sproloquio di formule, significa semplicemente *prendere l'Hurvitz-Thompson ed aggiungerci la differenza osservata, differenza osservata che aggiungiamo in pratica mediando le differenze sulla base dei coefficienti di regressione,* cosa tra l'altro abbastanza ragionevole.

Un esempio forse può chiarire meglio; supponiamo che tra gli addetti osserviamo una differenza di *10* tra i valori stimati e reali, e che, supponendo una sola variabile ausiliaria, la mia *Y* abbia una regressione rispetto agli addetti del tipo $y = 2 + 3x$ , in questo caso sarebbe stupido aggiungere *10* alla nostra *Y* dal momento che conosciamo che tipo di relazione esiste tra la *X* e la *Y*, aggiungeremo allora *32* alla *Y*, perché sappiamo che una differenza di *10* sulla *X* equivale ad una differenza di *32* sulle *Y*. Lo abbiamo fatto con una sola *X* ed una sola *Y,* con più variabili il principio è lo stesso: usare lo stimatore di regressione significa *tradurre sulla base della regressione multipla* le differenze osservate sulle *X* nella stima della *Y*. Chiaramente più questa regressione si avvicina ai dati e quindi si avvicina alla realtà riscontrata e più il nostro applicare le differenze riscontrate sulle *X* alle *Y* corregge e ci fa avvicinare al valore vero della popolazione.

Tradotto in statistica di base significa che più il nostro $R^2$ è alto, più è vicino ad uno e più fare la nostra stima significa che il nostro modello ha colto l'errore sulla nostra *Y*. Al limite massimo se l' $R^2 = 1$, ovvero se la relazione di regressione è perfetta allora la nostra stima è per forza identica alla popolazione, vuol dire che le nostre ausiliare spiegano perfettamente l'andamento delle nostre variabili d'interesse, quindi dire che l'ausiliaria sfrutta la popolazione equivale a dire che la nostra variabile d'interesse sfrutta tutta la popolazione. Chiaramente questo è un limite applicativo che non succederà mai, ma già avere un $R^2 \geq 0.9$ distrugge, schiaccia la varianza in modo sensibile, schiaccia la varianza di stima. Quindi avere delle buone ausiliarie significa ridurre di molto la varianza, e ridurre la varianza non è un problema statistico, alla fine è semplicemente un problema di costi, di danaro, perché ridurre la varianza significa che la stessa precisione la possiamo ottenere con un numero di unità campionarie di gran lunga inferiore e quindi spendere meno; dimezzare la varianza con l'utilizzo delle ausiliarie significa ottenere la stessa precisione con la metà delle unità statistiche, e quindi spendere la metà. Quindi le formule che abbiamo trattato aiutano a spendere meno, ad ottenere la stessa qualità con una spesa inferiore.

### 15.2.1 Un difetto applicativo, troppa regressione

Sembra un procedimento ragionevole fino a questo punto ma *c'è un difetto*; il difetto è che lo *stimatore di Horvitz-Thompson* da un punto di vista organizzativo era comodissimo, perché noi statistici di solito calcolavamo i pesi di riporto all'universo, poi con il dato dell'indagine rilevato mettevamo vicini i pesi e qualsiasi informatico, pur non sapendo niente di statistica sapeva che qualsiasi operazione la doveva fare pesata, calcolo di frequenze, somme etc. bastava pesare per quel determinato valore. Adesso non più, se abbiamo cento variabili d'interesse dobbiamo fare cento regressioni, e le stime le dobbiamo fare noi. Ad esempio se dobbiamo fare cento stime per le *20* province d'Italia dovremo fare *2000* regressioni - ma siamo davvero cosi pazzi? Assolutamente no, poiché si può dimostrare che *calcolare lo stimatore precedente equivale a fare l'Horvitz-Thompson non sui classici pesi di riporto all'universo ma su dei pesi di riporto all'universo* $W^*$ che si ottengono trasformando i pesi di riporto all'universo iniziali tramite questa relazione,

$$W^* = y_k \cdot w_k$$

$$g_k = 1 + (t_x \cdot \hat{t}_{HT})^T \cdot \hat{T}^{-1} \cdot \frac{x_k}{\sigma_k^2}$$

dove abbiamo che,

$$\hat{T}^{-1} = \left( \frac{\sum x_1 \cdot x_k^T}{\sigma_k^2 \cdot \pi_k} \right)^{-1}$$

in pratica per ottenere $W^*$ prendiamo i vecchi pesi di riporto all'universo e li moltiplichiamo per un coefficiente di correzione, una volta moltiplichiamo per *2*,una volta per *0,5*, etc., e questo coefficiente di correzione è pari a $g_k$. Quindi lo stimatore di regressione si usa in questo modo: non faremo le *2000* regressioni, calcoliamo il nostro correttore $g_k$, perché è la stessa cosa che calcolare lo stimatore di regressione, i calcoli da fare sono gli stessi, la matrice è sempre la matrice beta, la differenza è che i calcoli li facciamo una volta sola, prendiamo i nostri pesi di Horvitz-Thompson li *risistemiamo con il nostro correttore* e ci sarà un informatico che ci darà il nuovo risultato, ovvero gli diamo i nuovi dati, i nuovi pesi, (tra l'altro a lui non interessa come abbiamo calcolato i nuovi pesi, a lui l'unica cosa che interessa è che qualsiasi operazione ci sia da fare, che sia una tabella a doppia entrata, che sia un totale, che sei a un valor medio, etc.) lui sa che deve pesare rispetto ai nostri nuovi dati, e noi quindi li calcoliamo semplicemente riaggiustando i vecchi pesi una sola volta. Ma oltre tutti gli obbiettivi detti finora, tra cui quello di *abbassare la varianza delle Y*, c'è un altro scopo per cui eseguiamo questa operazione.

### 15.2.2 Un pregio dei nuovi pesi

Questi pesi aggiustati hanno la seguente caratteristica, se non li applichiamo alle *X*, cioè se ci calcoliamo la quantità,

$$\sum w_k^* \cdot x_k = t_x$$

la somma pesata su una qualsiasi variabile ausiliaria, e i pesi sono aggiustati nel modo descritto dall'espressione precedente, l*e* $t_x$ *riproporranno esattamente il totale vero della popolazione*. In altri termini abbiamo aggiustato i pesi del loro Horvitz-Thompson in modo che i totali delle variabili ausiliarie siano rispettati e coincidano esattamente con i totali noti della popolazione. Questo procedimento quindi non solo ha il vantaggio statistico di ridurre la varianza ma ha anche l'indubbio vantaggio pratico di non farci fare figuracce, ovvero, ad esempio, quando andiamo a fare stime di variabili note ad esempio sul sesso, maschi e femmine, non tireremo fuori dalle nostre stime il *20%* dei maschi e l'*80%* di femmine quando tutti sanno che sono *50%* e *50%*, ma ci ritorna esattamente il totale noto della popolazione di *50%* e *50%*. Il nostro procedimento quindi serve per produrre delle tabelle in uscita dalla nostra indagine in cui se una variabile è nota per tutta la popolazione, le frequenze, i totali di quella variabile sono esattamente quelli noti, equivale quindi a imporre dei vincoli alle stime, ovvero alle stime note gliele imponiamo e ovviamente ci ritornano in modo esatto.

### 15.2.3 Altri difetti, pesi di riporto minori o uguali a zero

Il procedimento ha ancora altri difetti. Un peso di riporto all'universo ha la funzione di rappresentare quante unità sono descritte da ciascuna unità campionaria, quindi di per sé non può essere minore di uno, figuriamoci inferiore a zero. Inferiore ad uno non è possibile poiché se un'unità campionaria è finita nel campione certamente non può rappresentare la metà di se stessa, figuriamoci per valori negativi. Il problema è che con il nostro correttore esiste la possibilità che vengano, e di fatto nella realtà vengono, pesi di riporto all'universo $W^*$ che sono non solo inferiori ad uno ma spesso anche inferiori a zero, ovvero pur di rispettare tutti vincoli *il correttore trova pesi di qualsiasi genere su tutto l'asse reale*, può venire qualsiasi valore, cosa che ci disturba non poco dal punto di vista logico in quanto significa che l'unità che ha un peso, ad esempio, uguale a - 10, man mano che fornisce valori più elevati, nelle risposte sempre più si abbassa la stima. Riguardo ciò tratteremo successivamente degli stimatori che correggono questo aspetto,

questo difetto, ponendo dei vincoli tramite dei minimi e dei massimi pur rispettando il totale della popolazione.

### 15.2.4 Altri difetti; solo variabili quantitative?

*(Ricorda il caso dei salti di strato)*
Un altro difetto riguarda il tipo di variabile ausiliaria preso in considerazione è: non è possibile effettuare una regressione lineare fra variabili non quantitative. Quindi l'altro difetto è che lo stimatore prende solo delle variabili quantitative ausiliarie, l'esempio del sesso che abbiamo trattato precedentemente infatti interrompe l'applicazione del nostro procedimento, e tra l'altro molte delle variabili ausiliarie che abbiamo sono di tipo qualitativo, la regione di appartenenza, il codice di attività economica, il sesso, provincia o comune di residenza, professione, titolo di studio, etc. Come lo risolviamo questo problema?

Supponiamo di avere una variabile ausiliaria che assume tre modalità con il vettore non-lineare delle modalità della variabile qualitativa e la sua traduzione in una matrice delle variabili indicatrici. A questo punto nella matrice inseriamo tante colonne quante sono le modalità della variabile qualitativa ed in ogni colonna della matrice inseriamo delle variabili indicatrici espresse da zero ed uno che indicano rispettivamente se non è presente la modalità o se è presente. Queste sequenze di zero ed uno o queste colonne le inseriamo successivamente nella nostra matrice delle $X$, nella matrice delle ausiliarie e le trattiamo come tre ausiliarie diverse ovvero l'ausiliaria non è più qualitativa ma sempre trattata come *variabile binaria composta da zero ed uno*, come presenza o assenza di una modalità. In questo caso possiamo, una volta inserita la nostra *matrice di zero e uno* all'interno della matrice delle $X$, applicare le formule standard che conosciamo, poiché fare questo non ci dà alcun fastidio dato che il $\beta$ di regressione in questo caso significa "l'effetto della presenza di …." essere maschio, essere femmina, essere della regione Piemonte, etc., ovvero appartenere o non appartenere ad una determinata modalità. Il problema è che non possiamo inserire tutte le modalità quantificandole poiché in questo caso stabiliremmo un ordinamento mentre tenendole separate tramite degli zero e degli uno non creiamo nessun problema, non

introduciamo un ordinamento fittizio né una distanza euclidea, ma ci possiamo tranquillamente riadattare alle formule precedenti della regressione multipla.

Tra le altre cose i totali di queste tre variabili indicatrici sono le frequenze assolute nella popolazione delle tre modalità che poi di fatto è ciò a cui noi ci vogliamo vincolare; supponendo ad esempio che trattiamo come modalità maschi e femmine nella variabile "sesso", noi vogliamo che i pesi di riporto all'universo facciano ritornare il totale dei maschi ed il totale delle femmine, non c'interessa la relazione tra le due modalità, le trattiamo in modo indipendente come se fossero due variabili diverse; oppure trattando della provincia di residenza dovremmo inserire per *103* province, ovvero modalità, *103* variabili ausiliarie, dobbiamo trovare dei pesi che ci facciano tornare tutti i totali di unità per provincia.

$$X = \begin{bmatrix} A \\ A \\ B \\ C \\ A \\ B \\ C \\ A \\ A \end{bmatrix} \rightarrow \begin{bmatrix} \overset{A}{1} & \overset{B}{0} & \overset{C}{0} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

*Fig. 15.2 – Trasformazione di variabili qualitative*

Tutto il gioco sta nell'adattare la regressione multipla alle ausiliarie, è la regressione multipla che accetta solo ausiliarie, non è lo stimatore di regressione ma è proprio la teoria della regressione che vuole delle variabili ausiliarie quantitative.

Se ricordiamo l'argomento riguardante i salti di strato avevamo detto che c'era una stratificazione iniziale su cui avevamo estratto ed una stratificazione osservata, ed avevamo detto che esistevano degli stimatori che ci facevano rispettare sia i totali della stratificazione di entrata sia quelli della stratificazione di uscita, ebbene è proprio questo il caso, se noi inseriamo come ausiliaria la stratificazione osservata, i nuovi pesi di

riporto all'universo faranno sì che possiamo rispettare i totali della stratificazione osservata *ex-post*, tant'è che questo procedimento se lo utilizziamo con *una sola ausiliaria qualitativa data dai codici di strato* di destinazione viene detta *post-stratificazione* perché di fatto facciamo in modo che i nuovi pesi rispettino degli strati di destinazione, ovvero nel nostro caso i totali delle tre modalità *A*, *B*, *C*. *Non è una stratificazione fatta a monte ma è una stratificazione fatta a valle dell'osservazione perché la applichiamo alla fine, non in fase di estrazione campionaria ma in fase di stima.*

Ora l'ultima cosa che ci rimane da discutere è, come sempre, la varianza di questo benedetto stimatore di regressione, una formula abbastanza facile da ricordarsi poiché è molto simile a quella dell'Horvitz-Thompson nel caso in cui utilizzavamo la matrice delle probabilità di inclusione doppie, dove moltiplicavamo per $\breve{Y}_k \cdot \breve{Y}_L$, adesso invece moltiplichiamo per $(g_k \cdot \breve{e}_k)(g_L \cdot \breve{e}_L)$ dove $g$ è il correttore dei pesi e la $\breve{e}$ altro non è che il residuo di regressione, che per essere riportato all'universo deve essere diviso per $\pi_k$, quindi in definitiva abbiamo i vecchi pesi di Horvitz-Thompson $\pi_k$ moltiplicati per *g* che è il correttore appunto dei pesi.

$$\hat{V}(\hat{t}_{HT,REG}) = \sum\sum (g_k \cdot \breve{e}_k) \cdot (g_L \cdot \breve{e}_L)$$

$$\breve{e} = \frac{Y_k - \breve{Y}_k}{\pi_k} \qquad \textit{residuo di regressione}$$

$$g_k, g_L \qquad \textit{correttore dei pesi}$$

$$\pi_k \qquad \textit{vecchi pesi dell'Horvitz-Thompson}$$

Ciò che fa la grossa differenza tra il vecchio Horvitz-Thompson e questo stimatore in termini di varianza è che il vecchio Horvitz-Thompson usava il valore fisico della *Y,* qui invece usiamo un residuo di regressione. In un esempio grafico vediamo ad esempio che,

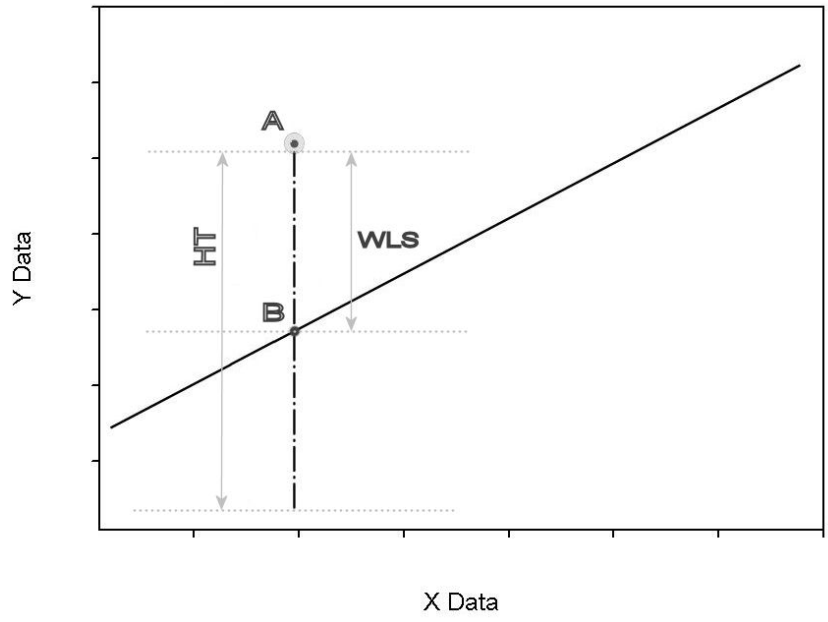



*Fig. 15.3 – I residui come pesi per il nuovo stimatore*

lo stimatore di regressione somma i quadrati del segmento più corto, mentre l'Horvitz-Thompson somma i quadrati del segmento più lungo, questo significa che se il punto sta esattamente sulla retta la varianza sarà uguale a zero, ovvero se $R^2 = 1$ i residui sono tutti uguali al zero e quindi la varianza della stima non esiste per principio. *Se invece* $R^2 = 0$ *applicare su tutto il valore della Y o sul residuo è esattamente la stessa cosa,* poiché i nostri residui non sono altro che le distanze o i residui dal valor medio, quindi la varianza non cambia, quindi ripetendo, se $R^2 = 0$ lo stimatore di regressione e lo stimatore di Horvitz-Thompson sono identici, se $R^2 = 1$ lo stimatore di regressione per principio avrà varianza zero ovvero spiega la popolazione con certezza. I problemi ed i metodi di calcolo di questa varianza a questo punto sono identici a quelli dell'Horvitz-Thompson poiché semplicemente invece che calcolarli sulle variabili originali li calcoliamo sui residui di regressione, anche qui avremo gli stessi problemi (matrice delle probabilità di inclusione doppie, metodo bootstrap, jack nife, etc.), quindi la varianza si calcola esattamente allo stesso modo, semplicemente, invece di mettere la *Y* nel calcolo della varianza mettiamo i residui di regressione rispetto alle ausiliarie.

### 15.2.5 Quando usiamo lo stimatore di regressione

Ma questo procedimento viene utilizzato sempre, spesso, oppure si usa semplicemente l'Horvitz-Thompson? In pratica lo stimatore di regressione è una pura riflessione teorico-mentale oppure di fatto nelle indagini reali si usa di più questo procedimento o l'Horvitz-Thompson che è sicuramente più semplice?

Fare una regressione non è molto complicato da un punto di vista pratico, il vero problema è avere delle variabili ausiliarie accettabili, di buona qualità dell'archivio. *Nelle applicazioni pratiche si usa di solito sempre e solo l'Horvitz-Thompson solo quando non ci si può fidare della qualità delle X,* perché potrebbero fare solo confusione. Quelle che vengono utilizzate quasi sempre in realtà sono le variabili qualitative, in pratica la nostra stratificazione, ovvero imporre vincoli per far tornare i totali noti nella popolazione come ad esempio sesso, età, professione etc., la qualità è prossima ad uno. Mentre utilizzare delle variabili ausiliarie quantitative è molto rischioso, per un'impresa ad esempio siamo sicuri che gli addetti hanno un'ottima qualità. Volendo fare un esempio opposto, quando facciamo un campione di aziende agricole, usare la superficie dell'azienda agricola che è facilmente reperibile al catasto è abbastanza rischioso, perché la qualità delle ausiliarie già alla fonte cioè al catasto è abbastanza compromessa, dati non aggiornati ed altri problemi.

Ma c'è un caso in cui questo tipo di stimatori viene utilizzato quasi obbligatoriamente, ovvero nella trattamento delle mancate risposte.

### 15.2.6 Introduzione al trattamento delle mancate risposte totali e parziali

Iniziamo quindi con l'introdurre le mancate risposte che riprenderemo successivamente. Cominciamo con il distinguere due categorie di mancate risposte: *mancata risposta totale è mancata risposta parziale.*

Per *mancata risposta totale* si intende quando l'unità statistica inserita nella campione non ha risposto per niente al questionario mentre si tratta di *mancata risposta parziale* quando l'unità statistica inserita nel campione ha risposto, ma non a tutte le variabili inserite nel questionario, ovvero la riga relativa alla nostra unità statistica contiene dei dati mancanti. Le due situazioni sono diverse anche da un punto di vista logico ossia un caso è la mancata risposta totale per la palese non-intenzione di partecipare all'indagine,

un altro rarissimo è la perdita del questionario a causa del servizio postale. La mancata risposta parziale invece può essere dovuta si ad una mancanza di volontà di rispondere proprio a quella domanda ma può essere dovuta anche ad un errore di compilazione del questionario, nel senso che l'unità statistica non si è accorta di aver saltato una domanda, cosa che può accadere soprattutto per questionari molto complessi oppure con molti salti da una sezione ad un'altra, domande filtro, etc. oppure peggio ancora molti dei dati mancanti presenti nelle mancate risposte parziali li abbiamo introdotti noi stessi, infatti come ricordiamo all'inizio del corso abbiamo detto che *la predisposizione di un questionario deve rispettare delle regole*, il cosiddetto *piano di compatibilità*. Se due variabili non rispettano una di queste regole un modo per fargliele rispettare è porre come dato mancante una delle due variabili, ad esempio se abbiamo un'unità statistica che per professione fa il pensionato ed ha un'età di 14 anni o poniamo dato mancante la professione o l'età o anche ambedue così non sbagliamo, una delle due variabili è sbagliata e sicuramente, cosa che vedremo nel controllo e correzione dei dati successivamente, dobbiamo considerare una variabile come dato mancante per errore di incoerenza. Quindi per ora le mancate risposte parziali le accantoniamo un attimo e parliamo di mancate risposte totali.

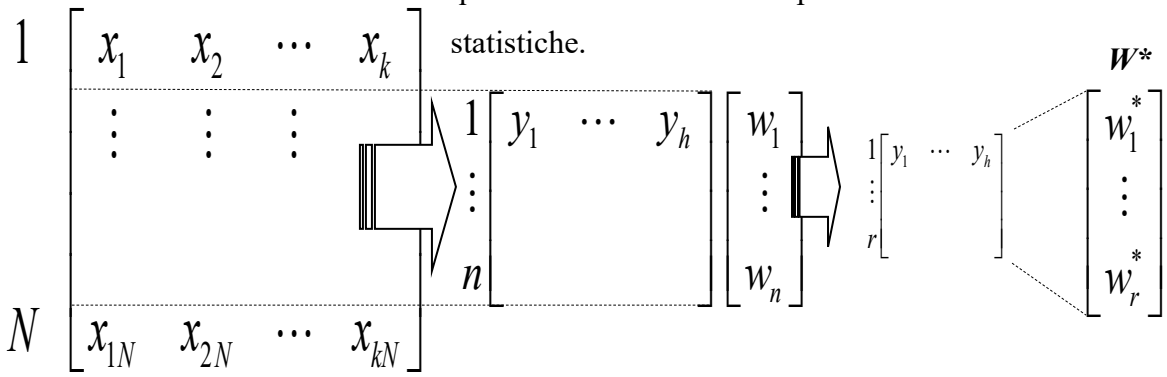
## 16. Le mancate risposte totali

Per il trattamento delle mancate risposte totali, quindi per fare la stima ci sono *2* vie (per trattamento s'intende ovviamente un trattamento alla fine del processo, poiché ovviamente si deve fare di tutto per aumentare il tasso di risposta, dare gadget, sollecitare, etc., soluzioni pratiche di cui non parliamo): la prima è *riponderare*, ovvero modificare i pesi di riporto all'universo per tenere conto che il nostro campione è più piccolo di quello che avremmo voluto, di quello progettato, mentre la seconda soluzione è *l'imputazione dei dati*, ovvero inventiamo dei dati ma naturalmente in modo ragionevole, esistono opportuni metodi per farlo; ad esempio per imputazione dei dati si intende che per le aziende che non ci hanno risposto in un'indagine, prendiamo i dati dell'azienda più vicina in termini di variabili ausiliarie, e li imputiamo alle aziende che non hanno risposto, in questo caso *l'imputazione è da donatore*, facciamo conto che

abbia risposto al posto di quella che non ha risposto. Per ora comunque parliamo solo della *ponderazione*.

## 16.1 La riponderazione

Come si fa la riponderazione? Abbiamo il nostro solito archivio, estraiamo il campione di $n$ elementi e di questo campione teorico che abbiamo utilizzato per spedire $n$ questionari otteniamo risposta solo da $r$ unità statistiche.




***dove*** $W^* = \begin{bmatrix} w_1 \\ \vdots \\ w_r \end{bmatrix} * g = \begin{bmatrix} w_1^* \\ \vdots \\ w_r^* \end{bmatrix}$ ***e*** $n > r$

Il problema ora è che ci mancano dei pesi di riporto dal campione ovvero i pesi delle unità che non hanno risposto ($n - r$), quindi dobbiamo correggere i pesi delle unità che hanno risposto, e questo lo facciamo riponderando, ovvero tramite dei pesi $W^*$ che tengono conto delle unità che mancano, ottenuti moltiplicando i vecchi pesi delle unità che hanno risposto per un correttore $g$ che banalmente è uguale a

$$g = \frac{n}{r}$$

come prima soluzione, quindi abbiamo osservato la metà del campione teorico ed abbiamo moltiplicato i rispettivi pesi per un correttore. Sembra troppo semplice ma è anche la soluzione più accettabile in quanto è basata su un'ipotesi fondamentale, che *le unità che non ci rispondono sono equidistribuite su tutto il campione*, ossia la probabilità di rispondere o no è uguale per tutte le unità, se è vera questa ipotesi, ovvero *equidistribuzione* delle mancate risposte, facciamo il nostro calcolo tramite il correttore

ed aggiustiamo i nostri pesi. Nel nostro caso poiché non ci ha risposto la metà del campione dovremo moltiplicare per due, poiché ogni unità che ha risposto ripropone quanto è stato estratto casualmente.

Il problema è che questa ipotesi raramente sussiste in quanto è molto più probabile che non rispondano, ad esempio in un campione d'imprese le piccole imprese piuttosto che le grandi, oppure in un campione sulle famiglie, sui redditi, è assai più probabile che non rispondano i redditi estremi, in quanto è più difficile, ad esempio, fare rispondere un senza-tetto oppure Berlusconi, quindi presupponiamo che la probabilità di non partecipare all'indagine, *quindi la probabilità di mancata risposta sia una qualche funzione delle nostre variabili ausiliarie*, ad esempio essere ricchi o poveri, essere una grande, media o piccola impresa. Allora come calcoliamo il nostro $g$ per correggere i pesi iniziali e trovare dei nuovi pesi di riporto all'universo?

Banalmente ad esempio utilizzando il nostro stimatore di regressione troviamo quel $g$ che fa tornare le ausiliarie che abbiamo scelto dall'archivio in modo tale che se i vincoli che abbiamo imposto, ad esempio, sulle piccole imprese sono pochi, il correttore sarà più ampio, supponendo che abbia risposto una sola piccola impresa su *100* allora i pesi delle piccole imprese li moltiplicherò per *100*, se delle grandi imprese mi hanno risposto tutte allora il correttore sarà uguale ad uno così mi ritornerà il totale della popolazione. Quindi in questo modo la correzione da mancata risposta la facciamo tenendo conto di tutte le ausiliarie che abbiamo, che dovrebbero quantomeno discriminare gruppi di popolazione con probabilità di mancata risposta o di risposta diversa. Questo è il sistema più utilizzato per la correzione da mancata risposta, stimatore di regressione o stimatori simili che accettano vincoli su totali noti della popolazione.

Ad esempio volendo fare un'indagine sugli individui, *l'età, il sesso e la professione sono dei grossi discriminanti per mancata risposta totale*, come anche ad esempio sulle classi estreme in un'indagine sull'uso di droga probabilmente saranno soprattutto i giovani a rispondere di meno oppure gli anziani perché non si fidano, magari si sentono infastiditi nel sentire la parola "droga", quindi in quest'ultimo caso *riproporzioniamo*. Siccome tra giovani ed anziani supponiamo che ci abbia risposto *1* su *10* mentre nelle classi medie *1* su *2*, i pesi li potremmo correggere ad esempio moltiplicando i pesi dei giovani e degli anziani per *10* e nelle classi intermedie per *2*, in

un modo molto banale per non fare calcoli, ma potrebbe accadere anche di avere una variabile quantitativa come ad esempio gli addetti ed in questo caso la correzione andrà fatta variare da unità ad unità poiché la variabile ausiliaria addetti cambia da unità ad unità. Quindi se *lo stimatore di regressione può essere utilizzato* o no per correggere, per schiacciare la varianza, una cosa sicura è che *nella realtà pratica viene utilizzato molto quantomeno per correggere le mancate risposte totali*, per questo motivo bisogna costruire bene un archivio, ed avere delle buone ausiliarie è fondamentale, non tanto per estrarre il campione tramite uno dei metodi che conosciamo (*PPS*, stratificato, etc.) ma perché ci serviranno alla fine per correggere le mancate risposte totali, perché se un'unità statistica non ci ha risposto possiamo correggere ed avere un qualcosa di più ragionevole piuttosto che farla rappresentare da una qualsiasi altra unità, almeno qualcosa possiamo sapere.

*Una Studentessa*: l'indagine non viene compromessa se ci sono troppe mancate risposte e quindi un'eccessiva correzione?
*Risposta*: la mancata risposta rovina sempre "le uova nel paniere", però il problema è che c'è per forza e più c'è mancata risposta più il nostro campione è andato "a pallino", non tanto perché si è ridotto il campione quanto perché la selezione del campione potrebbe non essere più random; ad esempio da *4* milioni di imprese ne prendiamo *4000* e sappiamo che le abbiamo prese in modo casuale poiché siamo noi ad avere il controllo sulla selezione, ne siamo sicuri, la mancata risposta invece non è più sotto il nostro controllo ma è sotto il controllo di chi risponde, e quindi che ne sappiamo se le mancate risposte si presentano in modo casuale oppure seguono un modello deterministico proprio per modificarci la stima. Quello che può fare uno statistico è utilizzare le ausiliarie, ovvero quello che sa delle unità che non hanno risposto per cercare in qualche modo di fare meno casino possibile.
Se ci chiedessero di fare un'indagine e di selezionare ad esempio *500* maschi e *500* femmine, effettuando un campione per quote dovremmo intervistare unità statistiche fino a quando non otteniamo risposta esattamente dal numero di unità richieste selezionate sulla base del sesso, *500* e *500*, in questo modo banalmente non abbiamo mancate risposte, ma abbiamo una truffa in quanto per avere il numero richiesto di maschi e femmine magari abbiamo dovuto intervistare *50.000* unità; e gli altri *49.000*

cosa ne pensavano? Quindi non è un campione casuale e non è un modo per limitare la mancata risposta, la mancata risposta va limitata con tanti accorgimenti tecnici durante altre fasi, non è possibile risolvere con delle formule problemi pratici, ma solo limitare i danni.

Volendo menzionare alcuni numeri l'ISTAT che ha l'obbligo di risposta per legge ha un tasso di mancata risposta intorno al *60%*, di cui la stragrande maggioranza sono piccole imprese e questo è drammatico poiché non sono equidistribuite.

Nelle indagini sulle famiglie la mancata risposta dipende molto dal tipo d'indagine, ad esempio nell'indagine sui consumi, sui bilanci di famiglia, dove si pretende ad esempio di registrare tutti gli scontrini, il tasso di risposta è intorno al *70%*.

In altre indagini come quelle ad esempio riguardanti le forze lavoro, se si è occupati o no, siamo intorno ad un tasso di risposta dell'*80-85%.*

Per quanto riguarda le indagini sulle aziende agricole abbiamo un tasso di risposta intorno alla *95%* in quanto l'ISTAT ha ideato una gran furbata, in quanto i rilevatori sono di fatto gli stessi assessori del comune in cui si trova l'azienda agricola, (quindi non solo non gli sparano ma... "si accomodi dottore,un po' di grappa?"), c'è un po' di effetto rilevatore ma il tasso di mancata risposta è bassissimo.

Sulle indagini telefoniche i tassi di mancata risposta sono naturalmente altissimi, non fosse altro perché non c'è contatto diretto, nel senso che viene imputata per mancata risposta anche l'unità statistica momentaneamente assente.

## 17. Gli stimatori di calibrazione

La volta scorsa ci siamo lasciati con un dubbio che era la possibile, anzi la quasi possibile esistenza di pesi di riporto all'universo che poco ci piacevano perché negativi, cosa che non è solo un disturbo formale ma crea da un punto di vista pratico applicativo un problema di robustezza delle stime. Per robustezza intendiamo, ricordando che gli stimatori possiedono diverse proprietà, una delle proprietà delle stime che non enunciamo in modo formale ma in modo intuitivo. *Uno stimatore si dice robusto se dipende poco dalla presenza di dati anomali*, facciamo un esempio, la media aritmetica è un tipico esempio di stima non-robusta poiché se c'è anche un solo numero che abbiamo sbagliato ad inserire nel calcolo la media aritmetica se ne va "a pallino",

supponiamo abbiamo *100* dati compresi tra zero ed uno, se inseriamo un valore maggiore di uno ad esempio *15*, solo questo numero falsa completamente la nostra media.

La mediana visto che divide in due la distribuzione, se c'è un dato anomalo, ad esempio un estremo eccessivo, non ne modifica affatto il valore, il valore della mediana, quindi la mediana gode della proprietà di robustezza. Questo è il motivo per cui in tantissime applicazioni, specialmente nelle indagini, la robustezza è fondamentale perché errori di vario tipo, di compilazione dati, compilazione del questionario, etc., sono spesso presenti, quindi usare delle stime che sono poco sensibili a questi errori, ovvero robuste è sempre positivo. Lo stimatore di regressione con i pesi negativi fa si che dove sono stati messi pesi strani, pesi negativi, se c'è un dato anomalo proprio dove sono stati messi i pesi negativi, più che fare stime stiamo dando i "numeri al lotto". Questo è il motivo per cui la cosa non ci piace, non solo perché formalmente vedere un peso negativo ci disturba ma anche perché è pericolosissimo avere pesi negativi.

Un altro problema applicativo riguardo la proprietà della robustezza lo abbiamo già trattato precedentemente a proposito dei *salti di strato*, un'impresa piccola ad esempio che è stata campionata come grande o viceversa costituisce un problema di robustezza, perché abbiamo un peso piccolo e ne applichiamo uno grande e l'unità anomala da sola falsa tutta la nostra stima, questo è un tipico esempio di robustezza.

Come risolviamo questa situazione? Cosa cerchiamo? *Cerchiamo dei metodi di stima che rispettino tutte le caratteristiche buone e positive dello stimatore di regressione*, per buone e positive intendiamo il fatto che possiamo forzare lo stimatore a rispettare il totale e quindi ci possiamo aggiustare un po' come vogliamo, tutto si traduce nel *trasformare il vettore dei pesi* e da un punto di vista organizzativo tutto si riconduce a livello informatico a somme e a medie pesate. Quindi queste caratteristiche positive, di riduzione della varianza e possibilità di far rispettare i pesi dei totali noti rimanendo su stimatori lineari, ovvero somme e medie pesate, quindi far sì che tutto venga ricompreso nel modificare i pesi, queste sono le caratteristiche positive. Il modo di modificare i pesi invece non ci piace poiché *i pesi vengono modificati in modo da non avere dei limiti*, ne minimi ne massimi, non abbiamo possibilità di controllare un minimo od un massimo di correzione dei pesi di riporto all'universo, cosa che invece esigiamo per non avere pesi o eccessivamente bassi o addirittura negativi o anche

eccessivamente alti che generano in entrambi i casi problemi di mancanza di robustezza delle stime, perché il problema di avere pesi negativi sì ci crea problemi da un punto di vista formale ma anche pesi eccessivamente alti ci creano problemi dallo stesso motivo, se proviamo un peso eccessivamente alto su un'unità statistica, se quell'unità ha un errore potremmo incorrere in alcuni problemi. Vogliamo allora poter intervenire sui minimi e sui massimi che possono essere attribuiti ai pesi di riporto all'universo, quindi abbandoniamo la regressione multipla e ne manteniamo gli aspetti positivi. Come? Andando a cercare dei *nuovi pesi di riporto all'universo tali che minimizzino una distanza*, che per ora non la specifichiamo e la supponiamo di qualsiasi genere $G$, tra il vettore dei pesi iniziale ed il vettore dei pesi corretto.

$$\begin{cases} Min & G\left[W; g \cdot W\right] \\ \sum w_i^* \cdot X_{i,j} = t_{x \cdot j} & \left(\forall j = 1 \ldots K\right) \end{cases}$$

***dove*** $$W^* = g \cdot W$$

Ora visto che i pesi $W^*$ per il modo in cui sono stati calcolati rappresentano una caratteristica positiva dello stimatore di regressione li manteniamo, ma li correggiamo cercando di allontanarci il meno possibile dai pesi iniziali, e tali che se applicati ad una variabile ausiliaria ci rispettino il totale noto; vogliamo modificare i pesi di riporto all'universo però meno ci allontaniamo dai pesi iniziali dell'Horvitz-Thompson e meglio è, poiché rappresentano i pesi con cui abbiamo estratto le unità e quindi non ci vorremmo allontanare tanto, d'altra parte esigiamo il rispetto dei totali della popolazione, altra caratteristica positiva dello stimatore di regressione. A questo punto quindi vogliamo trovare dei moltiplicatori dei pesi iniziali tali che si rispettino i totali e tali che si allontanino il meno possibile dai pesi dell'Horvitz-Thompson.

Prima ancora di vedere come si trovano questi moltiplicatori, diciamo che questi stimatori sono un caso ancora più generale dello stimatore di regressione, poiché se per distanza $G$ utilizziamo una distanza euclidea

$$G = \sum \left( w_i - w_i^* \right)^2$$

il risultato è esattamente lo stimatore di regressione, quindi *lo stimatore di regressione non è che un caso particolare di soluzione di questo problema*, *se usiamo una funzione di distanza specifica che è la distanza euclidea*. Ed a quanto ci risulta è anche l'unico caso in cui questi stimatori hanno una soluzione analitica, ovvero in cui la $G$ ha una formula, in un qualsiasi altro caso bisogna minimizzare per via numerica ovvero non abbiamo una formula ma il procedimento ci fornisce tutti i numeri, tutti i correttivi, ma non abbiamo una formula, con qualsiasi altra distanza bisogna risolvere questo sistema di minimizzazione vincolata dal calcolo numerico.

Gli stimatori di cui stiamo parlando si chiamano *stimatori di calibrazione o anche di ponderazione vincolata* (nel senso che ponderiamo i dati con quei pesi vincolati a rispettare i totali della popolazione). Come sempre le variabili ausiliarie, come nella regressione, possono essere sia quantitative che qualitative ritrasformate in variabili indicatrici, nel senso che le variabili possono essere ad esempio anche codici di regione o altro, si possono fare rispettare totali anche di frequenze assolute.

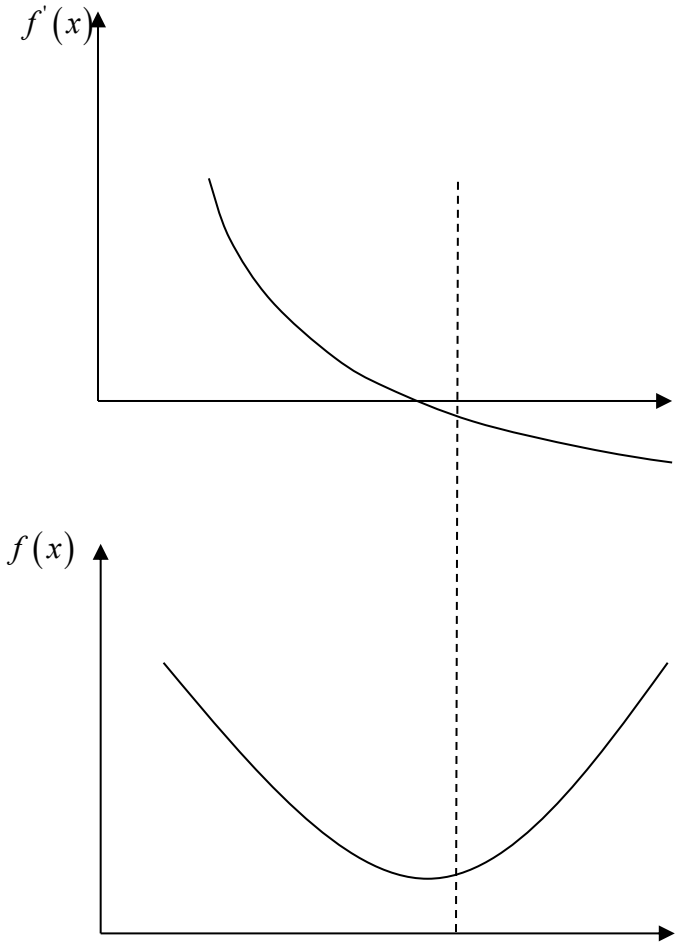
Ora la nostra richiesta di poter porre dei limiti di minimo e di massimo a questi ipotetici pesi la esaudiamo in termini di funzione di distanza della $G$, ovvero *la distanza euclidea ci permette di ottenere qualsiasi valore dei pesi che sia negativo o anche estremamente positivo perché è una distanza che è definita su tutto l'asse reale*, la distanza euclidea può andare da meno a più infinito, e possiamo quindi imporre dei vincoli utilizzando delle funzioni di distanza che fondamentalmente troncano le distanze dai requisiti minimi, ossia andando oltre certi valori la $G$ non verrà utilizzata, la distanza la prenderemo infinita, quindi giocando sulla $G$ che siamo liberi di specificare a nostro piacere purché sia una distanza, possiamo ottenere dei correttori che rispettino anche tante altre caratteristiche anche se quella che ci interessa fondamentalmente è la richiesta che abbiamo fatto precedentemente.

## 17.1 Introduzione al calcolo numerico

Ora prima di andare a specificare le varie funzioni di distanza utilizzate, anche se fondamentalmente è una la funzione che tronca rispetto ai requisiti minimi, ci dobbiamo

chiedere come si può risolvere un problema di minimizzazione per via numerica; interrompiamo momentaneamente il discorso riguardante gli stimatori ed apriamo momentaneamente una parentesi riguardante *il calcolo numerico*, poiché fino ad ora non abbiamo ancora parlato di come si trovano i minimi ed i massimi di una funzione per via numerica. Tra l'altro l'algoritmo precedente può essere facilmente portato in Excel.

Studiando matematica generale abbiamo imparato che per trovare una minimo di una funzione prima si fa la derivata e si eguaglia a zero e fin qui nessun problema, il problema è semplicemente che provare i valori per cui si annulla la derivata può risultare particolarmente complesso, non è sempre semplice, allora il problema di calcolare o trovare il minimo per via numerica non è tanto quello di calcolare la derivata, ma vedere quando questa derivata  raggiunge il suo punto di minimo, perché calcolare la derivata di per sé implica solamente l'applicazione di una regola. Sapendo che la derivata si esprime genericamente



$$\frac{f(x+\varepsilon)-f(x)}{\varepsilon} \qquad \text{con } \varepsilon \to 0,$$

possiamo esprimere la stessa regola come

$$\frac{f(x+\varepsilon)-f(x-\varepsilon)}{2\cdot\varepsilon}$$

con ad esempio $\varepsilon = 1\cdot E^{-10}$ per fare eseguire il calcolo ad un computer. E' come se ordinassimo al computer di dirci che differenza c'è tra $f(x+\varepsilon)$ e $f(x-\varepsilon)$ e dividerla per l'altezza $2\cdot\varepsilon$ e poi al posto di $\varepsilon$ inseriamo $1\cdot E^{-10} = 0,00000000001$, che significa dividere per $10^{10}$. A questo punto è il computer che calcola numericamente la derivata. Naturalmente il computer esegue il calcolo in un singolo punto, non abbiamo a disposizione tutta la funzione. Quindi la funzione sia calcolandola per via analitica sia calcolandola punto per punto per via numerica, dato una qualsiasi *X* la derivata la possiamo calcolare. Il punto è portarla a zero. Per via analitica basta eseguire dei calcoli e non ci dovrebbero essere particolari problemi, per via numerica invece potrebbero sorgere altri problemi.

Per farlo di solito si usano dei procedimenti o algoritmi di calcolo tra cui il più veloce e famoso è *l'algoritmo di Newton* che fondamentalmente è molto banale ma sicuramente efficacissimo. L'algoritmo va a cercare la tangente alla funzione della derivata prima, individua il punto di intersezione con l'asse delle ascisse e prende questo nuovo punto per aggiornare la *X*, ovvero una nuovo punto di partenza generando una processo

$f'(x)$

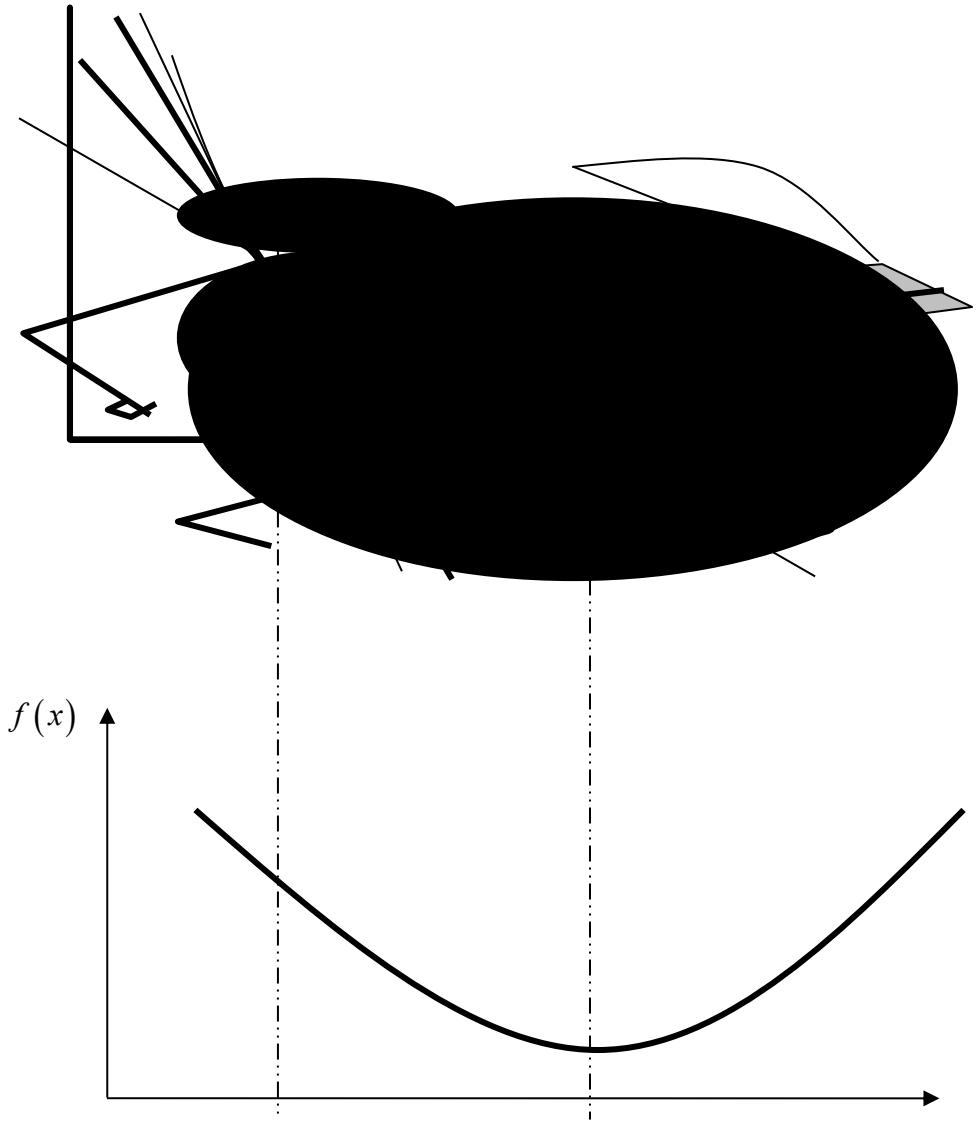



*Fig. 17.1 – Il metodo di Newton*

iterativo che porta al continuo affinamento del valore della *X* di partenza, e si vede bene dal grafico che dopo poche iterazioni raggiunge il punto in cui la funzione della derivata prima interseca l'asse delle ascisse. Se il punto viene oltrepassato l'algoritmo è progettato in modo da riportarci indietro fino al punto *E* che da un punto di vista grafico nel nostro caso viene raggiunto alla quinta iterazione del processo, possiamo pensare al punto *E* come ad un punto di equilibrio. Solitamente anche per funzioni molto complesse dove la complessità risiede di più nella forma grafica e non nella funzione stessa è possibile usare l'algoritmo di Newton ovvero quelle funzioni che si avvicinano all'asse molto lentamente e che toccano il punto d'incrocio solo dopo molte iterazioni,

cosa che comunque potrebbe darci dei problemi. Di solito comunque per le funzioni normali bastano sei o sette iterazioni per arrivare al punto d'incrocio con l'asse delle ascisse in cui si annulla la derivata. Ma che difetto ha questo algoritmo? *Non ci garantisce di trovare un ottimo globale*, ovvero questo algoritmo *serve per trovare gobbe* ma se i punti di minimo sono globali o locali non possiamo saperlo. Come soluzione o dimostriamo che la funzione è concava o convessa e che quindi ha una sola gobba, oppure partiamo da diversi punti e cerchiamo di vedere se otteniamo sempre lo stesso risultato. Nel grafico seguente infatti abbiamo diversi punti di minimo ed il risultato cambierà a seconda del punto di partenza poiché la funzione è multi-modale.

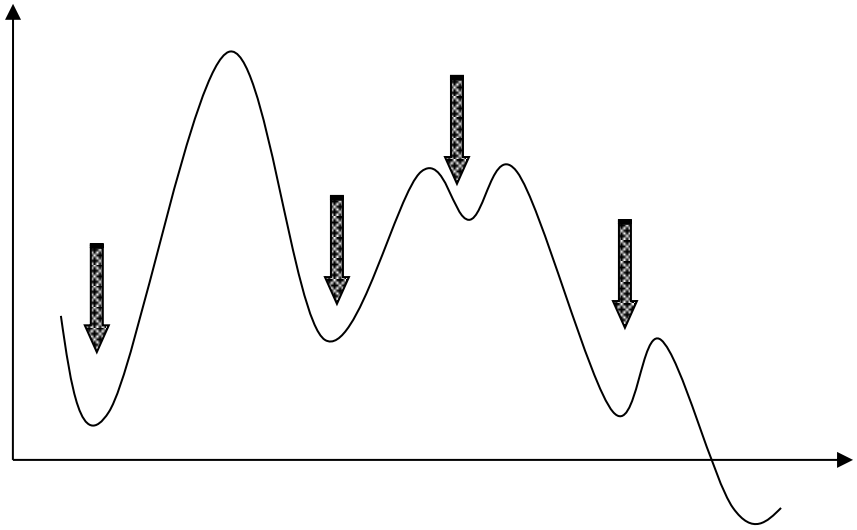


*Fig. 17.2(a) – La ricerca dei minimi locali*

Per risolvere il problema quindi prendiamo diversi punti di partenza, ovvero tanti $x_0$ di partenza e vediamo se ci torna sempre lo stesso risultato, in caso contrario prenderemo il risultato più basso. Ciò succede perché questo algoritmo è efficientissimo ma ha il grande difetto che può solo scendere lungo la funzione, possiamo immaginarlo come una pallina che viene lanciata lungo la funzione da un punto qualsiasi e che si ferma in una concavità, è un esempio poco ortodosso ma che rende bene l'idea.

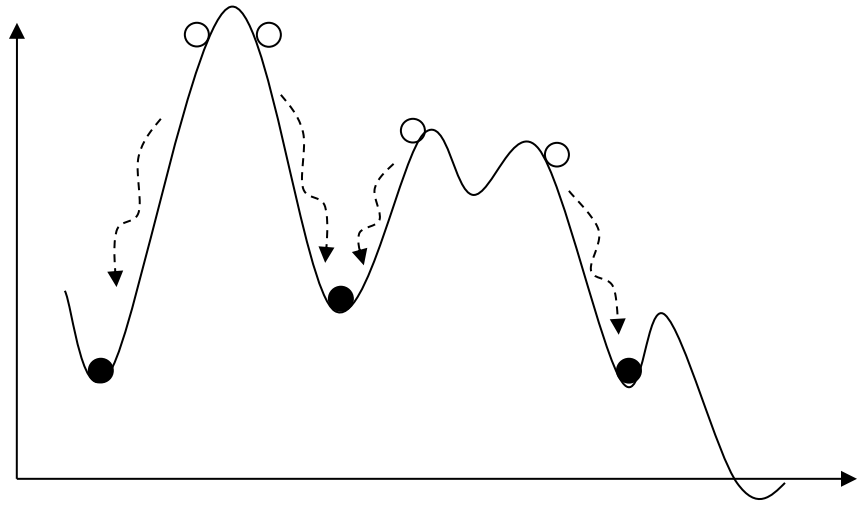


*Fig. 17.2(b)*

Da un punto di vista analitico l'algoritmo è la derivata seconda che quando andremo calcolare i pesi di ponderazione vincolata, o a minimizzare la funzione di distanza, è relativamente semplice da calcolare in modo analitico, quindi inseriamo direttamente i valori, oppure le facciamo calcolare direttamente al programma. Fino ad ora abbiamo parlato di un procedimento ad una dimensione, per più dimensioni la logica è lo stessa, semplicemente la $X$ non è più ad una dimensione ma sarà un vettore di $K$ dimensioni.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

Supponendo di essere in tre dimensioni con la terza che rappresenta la funzione che vogliamo minimizzare, il nostro $x_0$ sarà un punto sul piano, le derivate di cui parliamo saranno le derivate parziali, e possiamo tranquillamente spostarci sul nuovo $x_1$**.** Dato che probabilmente non tutti hanno trattato le derivate in più dimensioni, diciamo solo che, dato un vettore, la derivata prima sarà di nuovo un vettore che si chiama *Jacobiamo* della funzione mentre la derivata seconda sarà una matrice che si chiama *Hessiana*. In

pratica la formula precedente diventa un vettore, uguagliato ad un altro vettore meno *Jacobiano* fratto *Hessiano*, semplicemente quelli che erano valori diventano vettori e matrici. Questo è quanto ci serve per minimizzare.

Un altro algoritmo che vediamo solo da un punto di vista grafico e che potrebbe essere più veloce del metodo di Newton è il *Metodo della Secante*.

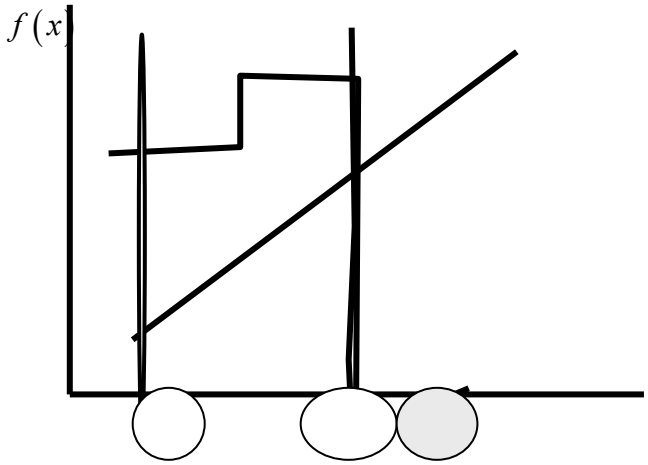



*Fig. 17.3 – Il metodo della secante*

Abbiamo sempre la nostra funzione, dobbiamo raggiungere il minimo in cui invece di andare a cercare le tangenti partiamo da due punti iniziali $x_0$ ed $x_1$, e di questi non andiamo a cercare il punto in cui la tangente incrocia l'asse come facevamo prima, ma calcoliamo la seccante per questi due punti, e poi andiamo a vedere dove la secante incrocia l'asse, il nostro $x_2$, ovvero ogni volta invece di tener conto solo dell'ultimo punto trovato si tiene conto degli ultimi due e questo velocizza ancora di più il processo.

A questo punto, dato che siamo in argomento, introduciamo anche un metodo per calcolare l'integrale di una funzione che genera numeri casuali.

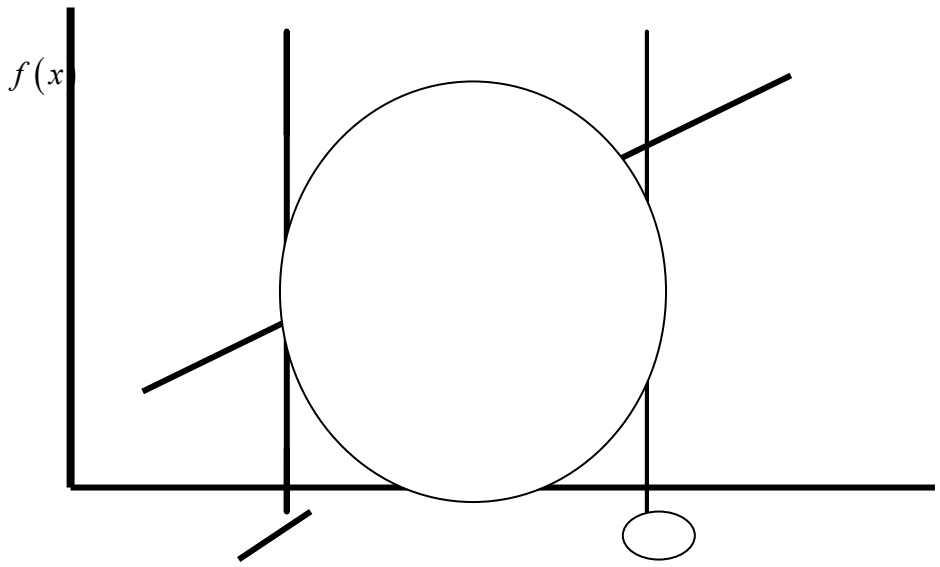



*Fig. 17.4 – Il calcolo di un'integrale*

Supponiamo di avere una qualsiasi funzione e di volere definire l'area al di sotto della curva compresa nell'intervallo compreso tra *A* e *B*, ovvero l'integrale definito della funzione, quindi generiamo un numero casuale compreso tra A e B e ne calcoliamo $f(x)$, generiamo un altro numero casuale e ne calcoliamo il rispettivo valore $f(x)$ e cosi via quante volte vogliamo, supponiamo n-volte. Naturalmente

$$\frac{\sum f(x_i)}{n}$$

è l'altezza media della funzione e l'area quindi la calcoliamo moltiplicando per l'intervallo compreso tra *A* e *B* ossia

$$\frac{\sum f(x_i)}{n} \cdot (B - A)$$

Detto in altre parole è come se individuassimo un rettangolo che contiene la funzione da integrare nell'intervallo scelto. A questo punto scegliamo un certo numero di punti a caso all'interno del rettangolo, diciamo *n*, e contiamo quanti punti si trovano sopra, e quanti sotto la funzione, diciamo $n_1$. L'integrale cercato è dato dall'area del rettangolo (che conosciamo) moltiplicata per la frazione di punti sotto la funzione. Se i numeri

casuali che utilizziamo sono veramente casuali possiamo andare avanti ad estrarre numeri fino a misurare l'integrale.

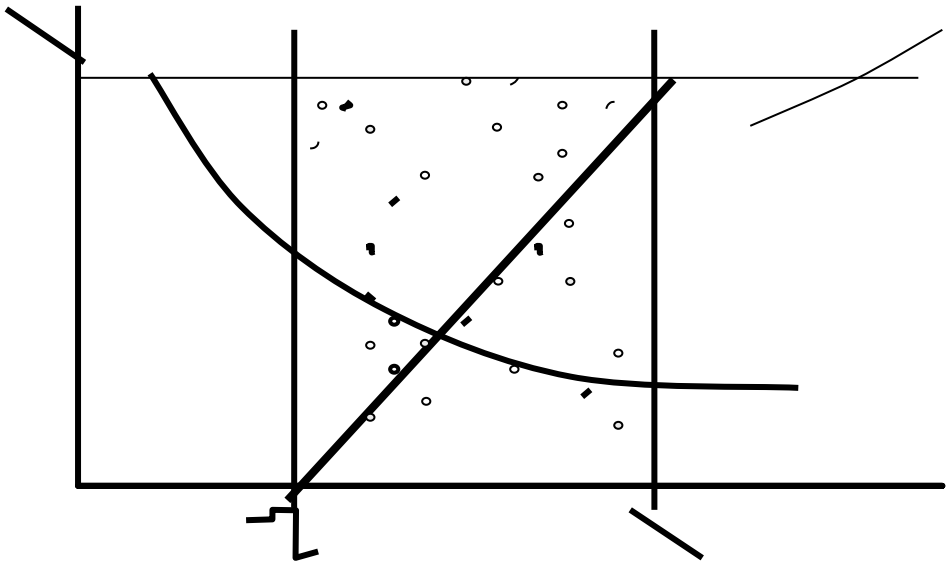


*Fig. 17.5 – Il metodo Montecarlo*

$$P \cdot (B - A) \cdot \max f(x) \qquad \textbf{\textit{dove}} \qquad P = \frac{n_1}{n}$$

*Il metodo Montecarlo* ha un gran pregio ovvero essendo basato su un campione è l'unico che *ci da anche una stima di varianza*, perché effettivamente la media che abbiamo calcolato è una stima e quindi anche l'area.

### 17.2 I non-metodi di attribuzione delle mancate risposte

Ritorniamo ora a parlare di mancate risposte che dividiamo in mancate risposte parziali e mancate risposte totali; per quanto riguarda le *mancate risposte totali* avevamo detto che ci sono due possibili sistemi, usare la ponderazione vincolata che è uno dei possibili metodi *d'imputazione*, quindi il metodo generale è intervenire sui pesi dei rispondenti per tener conto dei non rispondenti oppure l'altra possibile soluzione è *imputare*, inventarsi le risposte dei non rispondenti naturalmente con un determinato criterio. Ma i

due metodi sono molto differenti? Volendo fare un esempio didattico, abbiamo *10* elementi di archivio, ne estraiamo cinque e ne rispondono quattro, noi possiamo sia inventarci i numeri del quinto elemento sia modificare i pesi dei primi quattro in modo da tener conto dell'elemento che non ha risposto - *le due scelte sono così diverse*? Senza le mancate risposte le cinque unità che abbiamo selezionato avrebbero avuto pesi di riporto all'universo uguali a *2,2,2,2,2* ovviamente.

| | *w* |
|---|---|
| 1 | 2 |
| 2 | 2 |
| 3 | 2 |
| 4 | 2 |
| 5 | 2 |

Se interviene una mancata risposta, o inventiamo i numeri da attribuire all'unità secondo qualche criterio oppure modifichiamo i pesi in modo da riponderare.

| | *w* |
|---|---|
| 1 | 2 |
| 2 | 2 |
| • | |
| 4 | 2 |
| 5 | 2 |

Si può dimostrare banalmente che a livello logico, tra la ponderazione e l'imputazione non c'è molta differenza se non da un punto di vista formale, di procedimento. Avremmo infatti potuto assegnare alla terza unità il valore della seconda unità e quindi i risultati sarebbero stati uguali per ambedue i metodi.

| | *w* |
|---|---|
| 1 | 2 |
| 2 | 4 |
| • | |
| 4 | 2 |
| 5 | 2 |

Oppure avremmo potuto assegnare a tutte le unità uno stesso valore, la media dei rispondenti.

| | *w* |
|---|---|
| 1 | 2,5 |
| 2 | 2,5 |
| • | 2,5 |
| 4 | 2,5 |
| 5 | 2,5 |

Quindi se da un punto di vista pratico sono due cose diverse, *da un punto di vista teorico è facilmente e banalmente dimostrabile che qualsiasi criterio d'imputazione ha il suo corrispettivo criterio di ponderazione* e viceversa, quindi le due cose non sono diverse nel risultato o nella logica ma sono diverse nella pratica operativa, poiché imputare comunque significa inventarsi dei numeri.

Abbiamo *due metodi* per inventarci dei numeri fra i più semplici ed i più utilizzati: il primo ci dice di copiare completamente i numeri da un'altra unità statistica, dal punto di vista informatico dobbiamo copiare un intero Record. Naturalmente è chiaro che le informazioni andranno prese dall'unità statistica più vicina. Nel nostro archivio di solito abbiamo delle variabili ausiliarie per cui possiamo calcolare una distanza di qualche genere rispetto all'unità statistica che non ci ha risposto, a questo punto non ci resta che assegnare all'unità che non ci ha risposto le informazioni dell'unità che ha risposto e che ha minima distanza rispetto all'unità non-rispondente, viene scelta un'unità come candidata secondo il nostro criterio di distanza per donare i suoi dati, infatti questo metodo viene detto*"imputazione da donatore"*, donatore che può essere scelto o con criterio di minima distanza delle ausiliarie oppure, invece che selezionare il donatore più vicino selezioniamo a caso uno dei rispondenti, quest'ultimo è sempre un metodo d'imputazione da donatore, l'unica differenza è che scegliamo a caso l'unità statistica, *imputazione da donatore casuale*. Il metodo d'imputazione secondo la media aritmetica di cui abbiamo parlato precedentemente viene detto "*imputazione da valor medio"*.

La terza possibilità prende il nome di "*imputazione modellistica*", nel senso che tutte le varie $y_i$ che non hanno risposto all'indagine e che quindi non conosciamo, sulla base dei rispondenti si può presupporre che ciascuna di queste sia una certa funzione delle nostre $k$ variabili ausiliarie,

$$y_1 \ldots\ldots\ldots\ldots y_H$$

$$y_1 = f\left(x_1, x_2, \ldots\ldots\ldots x_k\right)$$

intendiamo ad esempio una regressione lineare multipla, diciamo ad esempio che $y_i = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 \ldots\ldots\ldots$, ma avremmo potuto definire una qualsiasi altra funzione, in ogni caso abbiamo stabilito che esiste una relazione ed una forma funzionale della relazione tra le *Y* osservate e le *X* ausiliarie. Questa relazione la stimiamo utilizzando solamente le unità rispondenti e dopo aver stimato i parametri $\beta_0, \beta_1, \ldots\ldots\ldots, \beta_k$, poiché le *X* ausiliarie già le abbiamo, possiamo utilizzare questa funzione per calcolare le stime, perché sono stime di fatto, per le unità non-rispondenti di cui non conosciamo la *Y* dato che non hanno risposto, ma conosciamo certamente la *X*.

Esiste una quarta ed ultima categoria di modalità d'imputazione che si chiama "*imputazione multipla*", che però esprimiamo solo nei criteri generali perché è una cosa un po' complicata anche se nei principi generali è abbastanza semplice. Sappiamo che per i non rispondenti c'è una distribuzione di probabilità data da $P(y_{NR} / y_R, y_{NR})$, in altre parole esiste una relazione tra le *y* e le *x* oppure tra le *y* e le *y*, poiché la probabilità condizionata si può esprimere con "sapendo che…". Avendo già a disposizione delle informazioni possiamo in pratica riuscire a fare un qualche calcolo sui non rispondenti, dire probabilità condizionata o funzione di….è la stessa cosa, sapere in pratica come cambia un fenomeno conoscendone un altro, questo è il principio della probabilità condizionata di cui la regressione lineare è un esempio poiché *può essere applicata appunto solo se esiste un andamento lineare*. La regressione multipla in pratica stima in qualche modo le probabilità condizionate $\hat{P}\left(y_{NR} / y_R, y_{NR}\right)$ e genera a caso le *Y* secondo

queste probabilità condizionate, ovvero $y_{NR} \sim P$. In altri termini poiché siamo in grado di generare numeri da una distribuzione uniforme o da qualsiasi altra distribuzione se *conosciamo una distribuzione di probabilità*, *di probabilità condizionate,* possiamo generare dati da questa distribuzione di probabilità, in pratica il metodo genera le mancate risposte a caso secondo la nostra distribuzione di probabilità. Siamo sicuramente in grado di calcolare le probabilità condizionate in qualche modo per poi generare ed estrarre dalla distribuzione delle probabilità condizionate. Questo vuol dire che *se ci siamo generati i dati per le non-rispondenti* a questo punto abbiamo un campione completo di *n* per tutte le nostre variabili. A questo punto però poiché le mancate risposte sono state generate in modo casuale da una distribuzione di probabilità, come abbiamo generato un campione di *n* elementi ne possiamo generare infiniti di campioni di *n* elementi ed ogni volta avremo risultati diversi.

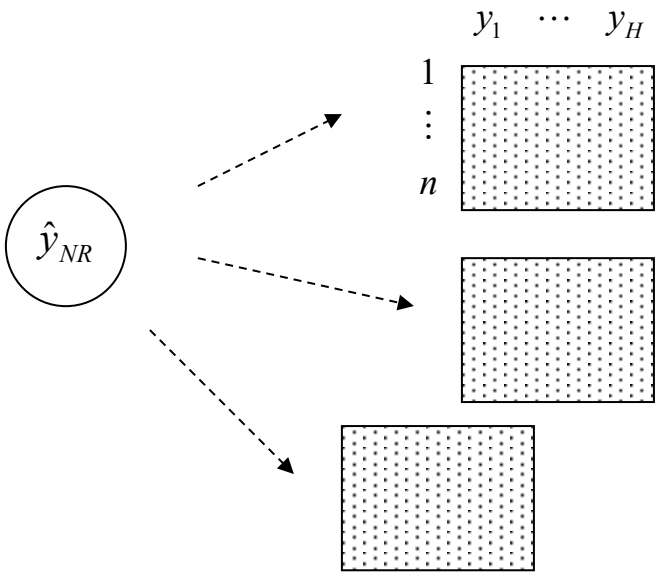



*Fig. 17.6 – L'imputazione multipla*

Tra l'altro su ognuno di questi campioni completi possiamo farci delle stime, ad esempio sul primo possiamo fare la stima del totale di Horvitz-Thompson, nel secondo, etc…, se facciamo questa operazione di imputazione random *1000* volte avremo *1000* stime ovviamente diverse, poiché i numeri casuali generati sono diversi avremo una

distribuzione dei nostri stimatori di Horvitz-Thompson e la stima finale sarà naturalmente la media delle stime.

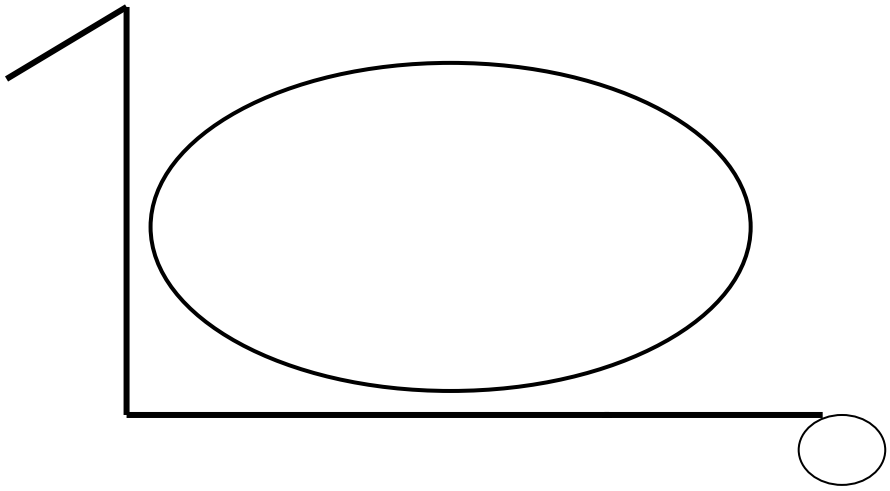


*Fig. 17.7 – La distribuzione di probabilità che otteniamo*

Questo metodo d'imputazione ha una *grandissimo pregio*, *ci permette anche di calcolare la varianza* dovuta all'imputazione, quindi di ciascuna stima oltre alla varianza campionaria che calcoliamo con i metodi standard siamo in grado anche di calcolarci la varianza dovuta al fatto che abbiamo imputato i dati.
Chiaramente la variabilità delle nostre stime dipenderà certamente da quanti dati dobbiamo imputare, ovvero se dobbiamo imputare ad una sola unità. Anche se ripetiamo il procedimento di stima per diversi campioni, la variabilità non sarà elevata, la stima varierà poco, la distribuzione avrà varianza prossima a zero, se invece bisogna imputare per molte mancate risposte alla fine sarà maggiore la varianza dovuta all'imputazione di quella dovuta al campione. E qui concludiamo *i quattro non-metodi di imputazione dei dati*, non di ponderazione, di cui abbiamo visto il metodo principale che è quello di ponderazione vincolata, che sono i più semplici.

## 17.3 I domini di stima

Ora se con la riponderazione o con l'imputazione si riesce sempre a trovare un modo per assegnare un valore ragionevole alle mancate risposte, perché si usa l'uno o l'altro metodo?

Ci sono diverse esigenze pratiche tra cui il fatto che l'imputazione è dispendiosa sia da un punto di vista spaziale e sia per quanto riguarda il tempo per effettuare l'operazione d'imputazione quando abbiamo un elevato numero di dati; d'altro canto ha un pregio dovuto solo ed esclusivamente ai domini di stima. *Ma cos'è un dominio di stima?* Ad esempio vogliamo effettuare stime regionali e la nostra mancata risposta si trova in Abruzzo. Se le altre nostre mancate risposte non si trovano in Abruzzo l'unico modo che abbiamo di non perdere il codice del dominio di stima che abbiamo nell'archivio insieme a molti altri codici di domini di stima (quindi di fatto bisogna rispettare più domini di stima), è di imputare un valore di un'unità statistica vicina che si trova in Abruzzo, a differenza della riponderazione in cui spesso non possiamo, perché non abbiamo i dati, *non c'è nessuna altra unità che rispetti tutti i domini di stima*. Nell'imputazione invece rispettare i domini di stima è automatico, poiché i codici di attività economica, della regione, della provincia, etc. già li sappiamo dall'archivio, imputiamo le *Y* e sul nostro dominio rifacciamo la stima. Nella riponderazione invece non solo il vettore dei pesi che è unico deve rispettare i pesi di riporto all'universo ma deve tener conto anche delle mancate risposte e deve fare in modo che tutti i domini di stima siano ugualmente rispettati dalla popolazione per numerosità e dimensione, ci sono troppi vincoli. *Imputare i dati invece è molto più semplice* perché ad esempio per la ricerca del più vicino possiamo procedere *in ordine gerarchico* lungo i domini di stima; se in un'indagine sulle imprese ci manca una risposta nel codice di attività economica dei calzolai andiamo a prendere il valore più vicino nello stesso codice di attività economica, tra i calzolai, se non lo troviamo andiamo ad un livello gerarchico superiore, etc., nell'imputazione è molto semplice stabilire le regole di ricerca di un campione, ad esempio ci manca un dato su un'impresa dell'Abruzzo andremo a prendere un valore di un'impresa nel Molise invece che di un'Impresa della Sardegna. *Quindi il motivo principale per cui si preferisce l'imputazione alla riponderazione è il problema dei domini che spesso sono talmente piccoli che la riponderazione può diventare molto complessa.*

La realtà pratica operativa vuole, ma non è sempre vero, che l'imputazione di mancate risposte totali nelle indagini sulle famiglie o sugli individui non vada mai rispettata poiché i domini sono molto aggregati, molto ampi, come il sesso ad esempio. Per le imprese invece i domini di stima di solito sono molto piccoli. *Nelle indagini sulle famiglie di solito viene fatta la riponderazione*; nelle indagini censuarie si usa esclusivamente l'imputazione, sulle grandi indagini di solito viene usata sempre l'imputazione perché sono le indagini strutturali sulle imprese e i domini di stima sono molto dettagliati, pensiamo a tutti i codici di attività economica; mentre la riponderazione solitamente viene usata per le indagini congiunturali, poiché di solito il dominio di stima è uno solo, il totale a livello nazionale o per settori di attività economica.

*Domande*: dopo aver definito i pregi e i difetti del metodo d'imputazione e di riponderazione dobbiamo trovare i pregi e i difetti delle quattro categorie del metodo d'imputazione, tenendo conto che i pregi e i difetti si bilanciano, nel senso che ognuna ha un suo perché, un suo motivo, infatti in indagini reali vengono usate di solito a seconda del bisogno tutte e quattro le categorie.

### 17.4 Come si limita l'incongruenza dei pesi

Vediamo ora la formula della funzione di distanza di cui abbiamo parlato precedentemente; naturalmente non è necessario impararla a memoria ma sapere cosa significa.

$$G = \left( \frac{W_{ks}^{*}}{W_{ks}} - L \right) \cdot \ln \left( \frac{\frac{W^{*}}{W} - L}{1 - L} \right) + \left( U - \frac{W_{ks}^{*}}{W_{ks}} \right) \cdot \ln \left( \frac{U - \frac{W_{ks}^{*}}{W_{ks}}}{U - 1} \right)$$

Questa funzione se si vuole commentare non è altro che il rapporto tra i due pesi *meno il correttore minimo* per il logaritmo del correttore minimo *più il correttore massimo* meno il rapporto tra i due pesi per il logaritmo del correttore massimo. È un

modo per bilanciare la distanza tra minimo e massimo e non farlo andare oltre, ma quello che è importante è che questa formula fa si che $U$ ed $L$ diano i limiti "upper and lower" del rapporto tra i pesi prima e dopo la "cura dimagrante". L'unica cosa che possiamo aggiungere è che se i moltiplicatori già di per sé con lo stimatore di regressione si trovano nei limiti di "upper and lower" più o meno questa funzione di distanza da dei risultati che sono praticamente identici a quelli che si ottengono usando la funzione di distanza euclidea. Quindi questa funzione viene utilizzata, interviene, modifica i risultati, quando i correttori che si ottengono con la distanza euclidea sforerebbero i limiti di minimo e massimo, se già da soli sono all'interno delle nostre richieste si ottengono più o meno gli stessi risultati.

### 17.2.1 Complementarità dei pregi e difetti dei metodi d'imputazione

Torniamo ora alle mancate risposte. Avevamo visto la questione dell'imputazione e ci eravamo lasciati con delle domande riguardanti le quattro grosse tipologie di metodi d'imputazione, i pregi e difetti di ogni tipologia e indicare un esempio, un caso particolare sull'utilizzo di una o l'altra tipologia. Quindi riepilogando le quattro tipologie abbiamo:

1. Imputazione da donatore;
2. Imputazione da valor medio;
3. Imputazione modellistica;
4. Imputazione Multipla;

***Imputazione da donatore.***

*Difetti:* l'imputazione da donatore è di una *lentezza estrema*, supponiamo di avere l'indagine con le nostre solite *30.000* imprese, per stabilire il valore da attribuire anche ad una sola mancata risposta dobbiamo individuare l'unità che plausibilmente ha la minima distanza dalla nostra unità e non ha risposto. Ma questo vorrebbe dire che per conoscere l'unità statistica che ha la minima distanza rispetto alla nostra mancata risposta dobbiamo calcolarci tutte le distanze tra le varie unità, che significherebbe eseguire su *30.000* imprese esattamente $30.000^2 = 900.000.000$ calcoli, ovvero distanze

euclidee, per conoscere l'unità statistica più vicina dobbiamo conoscere tutte le distanze logicamente.

*Pregi*: al contrario dell'imputazione modellistica, l'imputazione da donatore *è assolutamente oggettiva*, una volta stabilita la funzione di distanza dei pesi tutti saranno propensi a scegliere l'unità più vicina, quindi questo metodo è difficilmente attaccabile in questo senso poiché non dipende dalla scelta di chi effettua il calcolo, ed in quanto oggettivo il metodo è anche ripetibile.

***Imputazione da valor medio.***

*Pregi***:** è il metodo *più veloce* in assoluto, calcolare una media di valori infatti richiede pochissimo tempo, una volta inseriti i dati in un computer basta cliccare un tasto.

*Difetto*: il difetto del valor medio è insito nella sua semplicità, già partendo dal fatto che otteniamo tutti valori identici nelle mancate risposte, quindi abbiamo una *flessibilità nulla* e per quanto possa essere accettabile il valore medio non si può ritenere plausibile riportare alla realtà gli stessi pesi per tutte le unità statistiche, *le unità non sono state discriminate* in nessun modo.

***Imputazione modellistica.***

*Pregi*: poiché abbiamo detto che imputiamo il non-rispondente sulla base di determinate relazioni, se abbiamo molte ausiliarie e le relazioni sono effettivamente forti, e se sono relazioni economicamente sane, *tutta la nostra soggettività di fatto è coerente con la realtà*, questa diventa la stima più precisa, ovvero se la nostra conoscenza, la nostra soggettività di fatto può essere applicata in modo più efficace a questo metodo e ci permette una certa flessibilità, questa diventa la stima più ragionevole.

*Difetti*: la stima modellistica, che significa fare regressione etc., ha un grosso difetto, *ognuno può stabilire una relazione diversa* per legare le variabili ausiliarie alle risposte. L'imputazione modellistica essendo un modello economico che va ad imputare i dati, ognuno si stabilisce il suo, quindi è molto soggettiva.

***Imputazione multipla.***

*Difetti***:** i difetti sono moltissimi, tra cui ce n'è uno già visto per l'imputazione da donatore ovvero *la lentezza,* perché bisogna fare più campioni, in più occupa moltissimo spazio.

*Pregi*: permette una *minore incertezza nella stima d'imputazione*, e con il fatto che andavamo a campionare dalle distribuzioni condizionate, questo metodo di fatto è come se fosse un modello, quindi *c'è previsione e flessibilità*, poiché a seconda di come stabiliamo la distribuzione di probabilità condizionata possiamo inserire tutte le ipotesi e teorie economiche che vogliamo, quindi è la più precisa ed indubbiamente la più flessibile; inoltre volendo fare una differenza rispetto all'imputazione modellistica vediamo che in quest'ultima si stima sì il modello in un primo momento, ma poi lo applica in modo deterministico, non inserisce la casualità della stima, quindi *l'imputazione multipla è anche quella che da un punto di vista empirico è anche la più corretta* proprio perché ripropone l'incertezza del fatto che i numeri ce li siamo inventati, e quindi anche *le stime che facciamo mantengono l'incertezza* di questo fatto. L'unico metodo che tiene conto in modo formale dell'incertezza è proprio quest'ultimo metodo.

### 17.2.2 Quando vengono usati questi metodi?

***Imputazione da donatore.***

È un metodo che tipicamente viene utilizzato nelle statistiche ufficiali e permette di *standardizzare a livello internazionale la stima*. Di solito è preferibile in indagini campionarie anche se grazie all'avvento di software sempre più avanzati e ad alcuni espedienti operativi, nell'ultimo censimento dell'industria è stato usato il metodo da donatore.

***Imputazione da valor medio.***

È un metodo che tipicamente viene utilizzato nelle statistiche ufficiali e permette di *standardizzare a livello internazionale la stima*. Di solito è preferibile in indagini censuarie.

***Imputazione modellistica.***

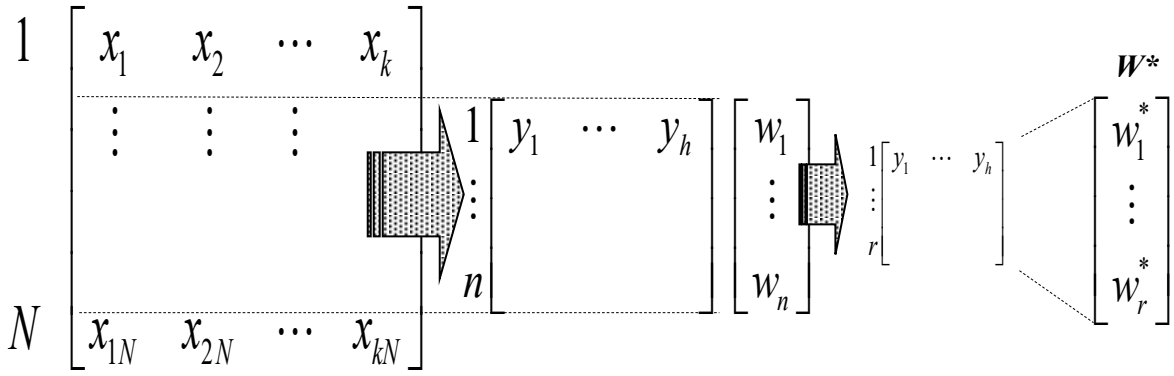
L’imputazione modellistica *non viene utilizzata spesso all'interno di qualsiasi indagine che riguardi dati ufficiali*, perché a seconda del modello che scegliamo possiamo ottenere risultati diversi; se ad esempio l'Istat volesse abbassare il prodotto interno lordo al di sotto del 3% potrebbe scegliere la regressione giusta che da risultati coerenti con questo obiettivo. Quindi è una categoria d'imputazione utilizzata più che altro da istituti che non fanno stime ufficiali e che hanno dei ricercatori preparati che si occupano di determinati settori, nella fattispecie economici.

***Imputazione multipla***

*Viene utilizzata per campioni molto piccoli*. Inoltre il nostro obiettivo non è la velocità o l'operatività ma *essere sicuri di essere scientificamente corretti* sotto tutti i punti di vista. In indagini reali questo metodo non viene utilizzato per quanto noto, viene utilizzato da chi fa indagini che abbiano un fondamento di precisione e serietà, probabilmente da ricercatori, ma è assolutamente inapplicabile nelle indagini sulla forza lavoro, bilanci famigliari, prezzi, etc., poiché se usiamo questo metodo ad esempio su un'indagine riguardante *30.000* imprese con *1000-3000* mancate risposte potrebbe anche scoppiare il computer, perché dovremmo fare calcoli nell’ordine di grandezza di *30.000.000*. È un metodo ancora molto teorico che non è ancora possibile utilizzare da un punto di vista pratico-applicativo.

## 17.5 Un caso particolare della riponderazione

Sulla riponderazione invece che è l'altro caso possibile ci limitiamo ad esporre e specificare meglio un caso particolare molto più semplice per il fatto che il 90% di imputazione delle mancate risposte viene fatta con quest'ultimo. Ricordando lo schema

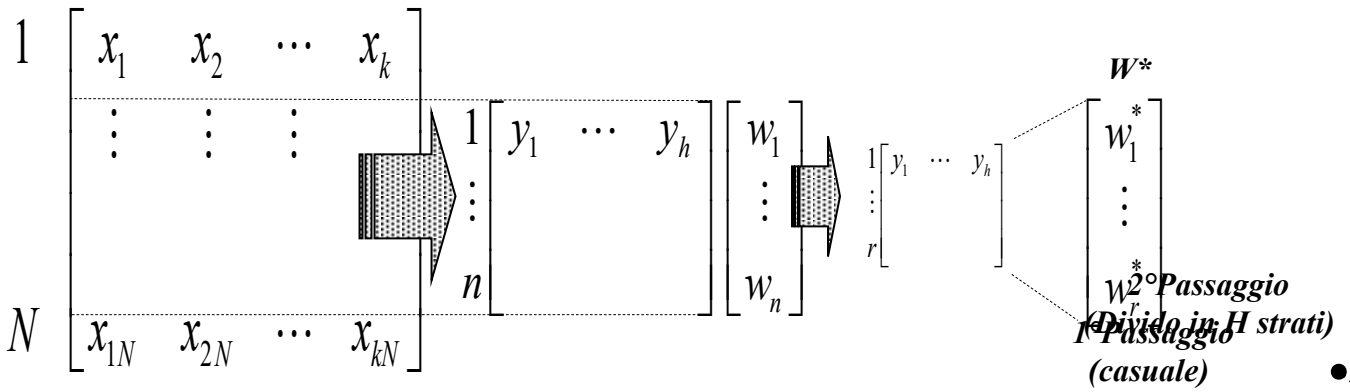


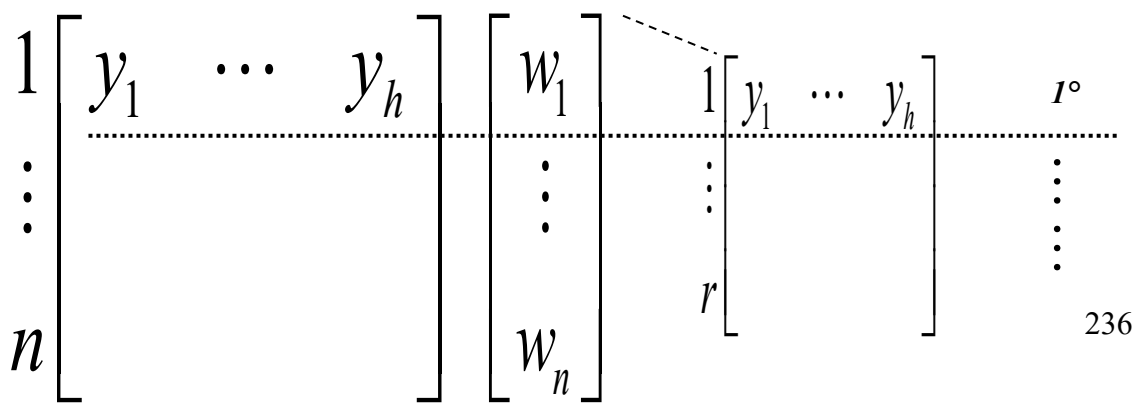
***2°Passaggio***

***1°Passaggio***

avevamo ipotizzato che sia il primo passaggio sia il secondo passaggio fossero determinanti in maniera casuale, stavamo ipotizzando che in pratica anche *il secondo passaggio fosse determinato in maniera casuale* e quindi con probabilità di mancata risposta uguale per tutte le unità, ovvero sia chi ha risposto sia chi non ha risposto ha la stessa probabilità di non rispondere. D'altra parte avevamo già parlato di un modo per discriminare le unità statistiche per la loro probabilità di non rispondere con la ponderazione vincolata tenendo conto delle ausiliarie, insomma vedere chi risponde di più e chi risponde di meno. Il fatto è che comunque lo stimatore ottenuto tramite la ponderazione vincolata è abbastanza complesso.

Allora per risolvere il problema in modo pratico e sbrigativo e quindi riconoscere che alcune unità rispondono più di altre, *nel secondo passaggio dividiamo in parti il nostro campione*, facciamo delle ipotesi sulle percentuali di unità che rispondono di più o di meno, e quindi, successivamente *la riponderazione la effettuiamo in base agli strati ottenuti in base alle nostre ipotesi sulle probabilità* che le unità hanno di rispondere o non rispondere.



***H***

Nel secondo passaggio in pratica dividiamo il nostro campione iniziale in *H* strati e la riponderazione la facciamo invece che uguale su tutte le unità ovvero

$$W^* = \frac{n}{r} \cdot W$$

Uguale solo all'interno di ogni singolo strato ovvero facciamo l'operazione

$$W^* = \frac{n_h}{r_h} \cdot W \text{ con } h = 1 \ldots\ldots H$$

al variare di *H* per ogni singolo strato. Questo metodo viene detto *"riponderazione per gruppi omogenei di mancata risposta"* ovvero si ipotizza che la probabilità di mancata risposta sia non-uguale su tutta la popolazione ma solo all'interno di uno stesso strato. In ogni caso se il campione è stratificato niente ci vieta che la stratificazione di questa estrazione campionaria sia diversa da quella che vogliamo, poiché nel primo caso stratifichiamo per omogeneità della variabile *Y* che andiamo a rilevare, nel secondo passaggio invece stratifichiamo per omogeneità di mancata risposta che in teoria è diversa.

## 18. Le mancate risposte parziali

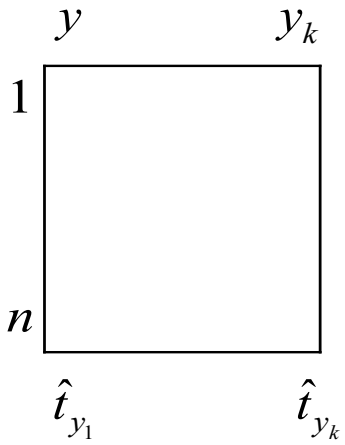
Come abbiamo detto precedentemente per le mancate risposte parziali dobbiamo sempre tener presente che non sono dovute sempre all'intenzione di non rispondere da parte di chi compila il questionario, ma spesso sono generate addirittura da noi. Ovvero ci ritornano i questionari, registriamo i dati e quando andiamo ad analizzare troviamo degli errori, che possono essere dovuti ad una cattiva progettazione del piano di compatibilità o a dati anomali che molto spesso è meglio sostituire con un dato mancante per via del fatto che sono poco verosimili. Possiamo avere in pratica mancate risposte reali, errori solo di compilazione o mancate risposte prodotte artificialmente da noi, che possono essere dovute a vincoli di compatibilità oppure ad individuazione di dati anomali, quindi

c'è una categoria di mancate risposte parziali artificialmente introdotta da noi e tra l'altro la maggior parte delle mancate risposte appartengono a quest'ultima categoria, errori che nell'elaborazione dei dati noi abbiamo introdotto. Quindi prima ancora di dire cosa ci mettiamo al posto del dato mancante dobbiamo discutere sul come mettiamo i dati mancanti, quindi prima ancora del trattamento della mancata risposta parziale bisogna discutere sul come noi fisicamente operiamo o interveniamo manualmente o automaticamente sui dati. Questo processo si chiama *Editing* oppure *controllo e correzione dati* che di fatto occupa il *95% del tempo* di chi fa le indagini, qui non stiamo facendo delle stime, e se ci dovessimo fermare alla ricerca di una stima supponendo che i dati siano tutti perfetti e puliti, per fare un'indagine potremmo anche mettere davanti ad un computer uno scimpanzè che pigia i tasti e termineremmo l'obiettivo della nostra ricerca.

## 18.1 Il macro-editing

Il problema è che dati perfetti non esistono, ripulire ad esempio una campione di *30.000* imprese non può essere fatto dando solo un'occhiata ai dati, quindi più o meno, poiché la soggettività di come si ripuliscono i dati è molto ampia, vediamo di definire alcune categorie di operazioni utili a questo scopo.

La prima che viene chiamata *macro-editing* *significa lavorare sui macro valori, gli aggregati,* magari suddivisi per regione, per settore di attività economica, etc., ovvero *sulle stime che sintetizzano gli aggregati* e non sui dati individuali delle singole unità statistiche, quindi dagli aggregati dobbiamo individuare le stime poco soddisfacenti,che eventualmente nascondono problemi, confrontando gli aggregati con fonti esterne, dati degli anni precedenti riguardanti la stessa indagine, etc.



Ad esempio il valore aggiunto del settore manifatturiero dal *1995* alla *1996* aumenta del *130%* dopo aver fatto le stime, naturalmente un risultato poco plausibile, quindi *partiamo dai risultati ed individuiamo quali risultati, quali aggregati ci convincono poco*, tra l'altro non avremo un'infinità di risultati come quando parlavamo di variabili ausiliarie e di regressione multipla ma *al massimo un numero di risultati pari al numero delle variabili*, oppure tanti quanti sono i domini principali. Quindi questo meccanismo funziona predisponendo delle tabelle che ci danno le nostre stime $\hat{t}_{y_1}$ confrontate con delle fonti esterne $F_E$ o dati precedenti $D_P$, magari facendo qualche statistica come ad esempio la percentuale di variazione $\%_{0_{var}}$, si selezionano le caselle, gli aggregati che ci convincono poco, ed a questo punto, *dall'individuazione degli aggregati problematici chiediamo al meccanismo di estrarre le unità che più pesano sulle celle*, ovvero sulla cella come evidenziato nel grafico.

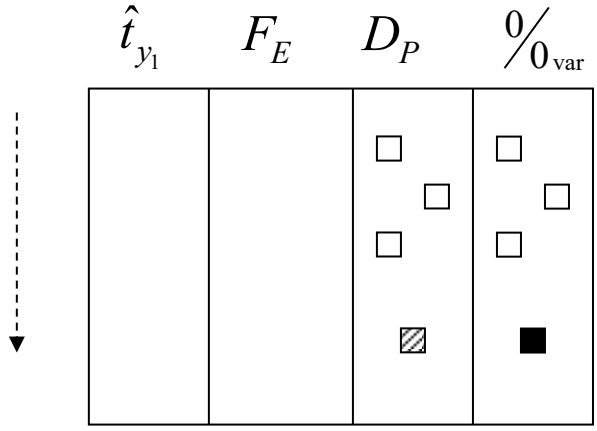



*Fig. 18.1 – Il Macro-editing*

Ad esempio parlando del valore aggiunto del settore manifatturiero, quali sono le imprese del settore manifatturiero lo sappiamo, prendiamo le prime *10* che hanno,

supponiamo, il prodotto tra il valore aggiunto ed il peso di riporto all'universo più elevato, le più importanti nella zona che ci ha creato problemi ovvero se c'è stato un aumento delle valore aggiunto del *130%* vorrà dire che l'aumento deve nascere dalle unità statistiche che pesano di più, se ci fosse stato un aumento del 2000% su unità che non pesano nulla non l'avremmo visto sulla stima. Quindi prendiamo le prime *10-15* unità ed andiamo a vedere quali di queste hanno avuto un aumento, in quale è finito uno zero di troppo, etc., quindi *il macro-editing parte dalle stime per identificare degli aggregati sulla base dei quali selezionare le unità statistiche da correggere.* Quindi si parte dalle stime per tornare indietro fino alle unità, invece di guardarle tutte e *30.000* ne guardiamo solo alcune, quelle che secondo la nostra analisi hanno creato problemi.

## 18.2 Il micro-editing

Dopo questa prima ripulitura che tra l'altro specialmente nelle indagini sulle imprese prevede che le grandi imprese siano ricontrollate sempre, nel secondo passaggio si procede al contrario, ovvero si usa *il micro-editing*, ovvero correggere direttamente i dati di un'unità, *di solito si basa su rappresentazioni grafiche*, quasi esclusivamente di due tipologie, o *uni variate,* ovvero istogrammi per singola variabile, o *bivariate* che sono finalizzate fondamentalmente o meglio quasi esclusivamente all'individuazione di dati anomali. Ad esempio immaginiamo una distribuzione del reddito contro consumi come nel grafico seguente, queste rappresentazioni grafiche permettono, come vengono fatte attualmente sui computer, di tenere grafici legati l'uno all'altro.

***DISTRIBUZIONE UNIVARIATA*** ***DISTRIBUZIONE BIVARIATA***

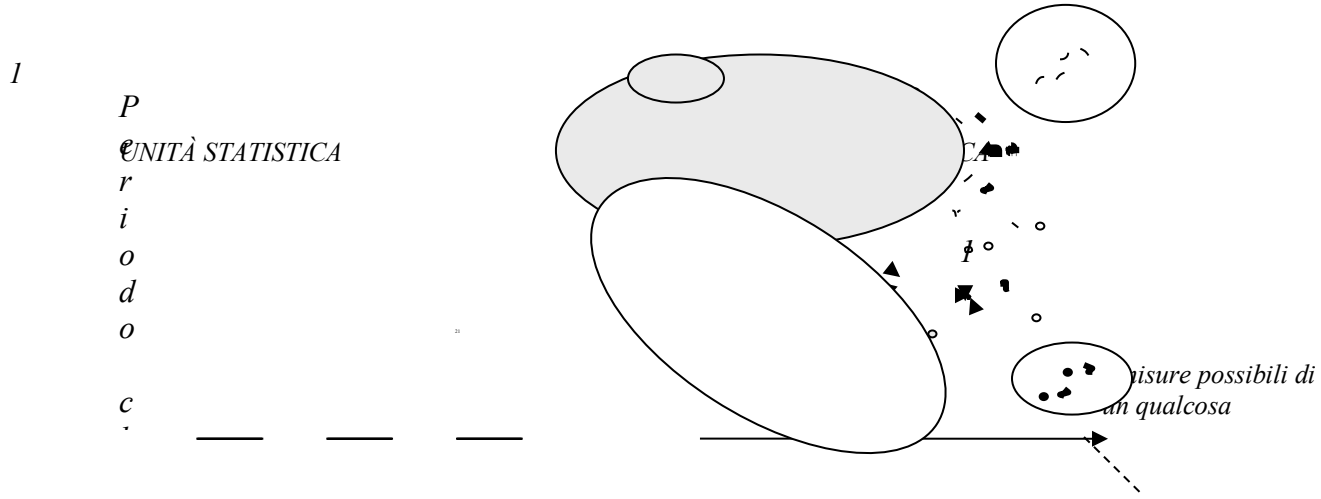



*Fig. 18.2 – Il Micro-editing, partendo dalla distribuzione univariata*

Come si vede dal grafico precedente, dopo aver selezionato delle classi di reddito molto elevate nella distribuzione *univariata* che indica le distribuzioni di frequenze del reddito, potrebbe interessarci andare a vedere se questi ricchi consumano molto o poco, se hanno questo valore particolare solo sul reddito o anche sui consumi. In pratica dobbiamo verificare se è possibile che persone con valori altissimi di reddito consumano molto o poco, naturalmente se consumano poco i dati saranno anomali. Se invece si trovano nella parte alta della distribuzione *bivariata* vuol dire che a tanto reddito corrisponde tanto consumo e abbiamo spostato le unità in posizioni particolari sia sui redditi sia sui consumi, allora sarà difficile che un dato anomalo si riproponga su più variabili, che su più unità ci sia un *10* in più o in meno rispetto ai valori dei redditi.

Di operazioni grafiche di questo tipo naturalmente ne facciamo più di una, facciamo più incroci tra più variabili che sappiamo essere correlate in qualche modo, e facciamo istogrammi di frequenze per quasi tutte le variabili, diamo un'occhiata, selezioniamo i gruppi fuori dalle code e analizziamo i dati incongruenti, oppure possiamo operare anche in senso contrario, partendo dal bivariato, procedimento utile nelle indagini in cui abbiamo la stessa variabile in due anni diversi. Ad esempio in un'indagine sulle imprese, vogliamo analizzare il prodotto interno lordo in due anni diversi.

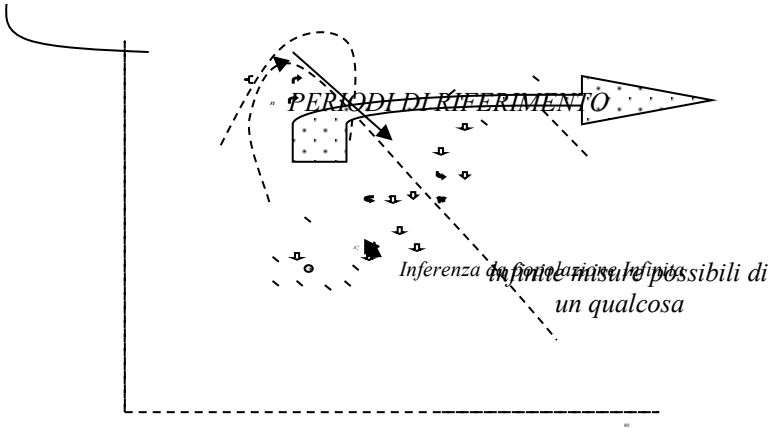



*Fig. 18.3 – Il Micro-editing, partendo dalla distribuzione bivariata*

La correlazione sarà quasi uguale ad uno poiché i dati si trovano quasi tutti lungo la retta, mentre c'è un gruppo fuori dalla retta. In tutti questi casi, sia di *macro-editing* e che di *micro-editing*, a meno di riprendere il questionario fisico dell'unità che abbiamo selezionato, non inventiamo dei numeri. *In tantissimi casi l'unica cosa che facciamo è individuare le unità, non sostituire il numero.*

## 18.3 Il piano di compatibilità

Ora supponendo che nelle due fasi precedenti abbiamo sistemato i problemi più grandi, nella terza ed ultima fase lanciamo a tappeto una procedura che ripulisce tutto, stiamo parlando in pratica *del piano di compatibilità*. Solo dopo aver sistemato le unità che creano problemi sulle stime, eseguiamo un'operazione automatica, ovvero *cerchiamo di eliminare le incoerenze formali*. Il piano di compatibilità individua a tappeto tutte le imprese, tutti i record, tutte *le unità statistiche che violano uno o più vincoli*, che hanno problemi; fino ad ora abbiamo visto soltanto dove sono le incompatibilità, ora le unità statistiche che creano problemi le dividiamo in due, una parte di unità statistiche la risottoponiamo a controllo, ricominciamo da capo, ovvero ricontrolliamo quelle unità che hanno violato delle regole del piano di compatibilità, mentre il resto di unità le trattiamo in modo automatico, ovvero facciamo un'individuazione per errore automatico, ordiniamo in pratica al software di correzione di mettere i dati mancanti in modo da farmi rispettare tutte le regole del piano di compatibilità. In pratica per

ciascuna unità noi abbiamo *K* variabili, $y_1, y_2, ....., y_k$, che hanno dei vincoli, ad esempio come più volte detto l'età può non essere compatibile con la professione, un pensionato di *14* anni è poco verosimile. Anche in e questo caso è come se dicessimo al software di correzione di *trovare il numero minimo di variabili da porre come dato mancante tali da rispettare tutti i vincoli*.

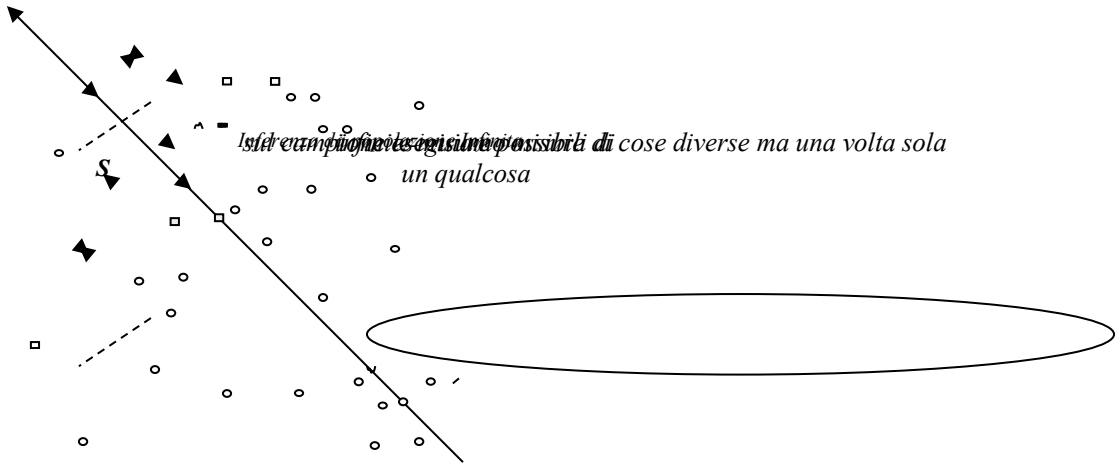



*Fig. 18.4 – Dati mancanti*

È come se, guardando il grafico precedente, dicessimo al software di considerare come dato mancante non sia *A* che *B*, ma solo *B***,** ovvero *il numero minimo di dati mancanti che rispettano tutti i vincoli* del piano di compatibilità; come si trova poi il numero minimo è un problema di informatica, *l'algoritmo di Cernikova*. Questo algoritmo in pratica mette come dati mancanti tutti quelli che deve mettere in modo tale da rispettare perfettamente il piano di compatibilità, in modo da ripulire pienamente l'indagine. Quindi dopo aver effettuato tutte le nostre operazioni, dopo aver ripulito tutto, otteniamo una sorta di groviera, un certo numero di dati mancanti - cosa facciamo a questo punto?

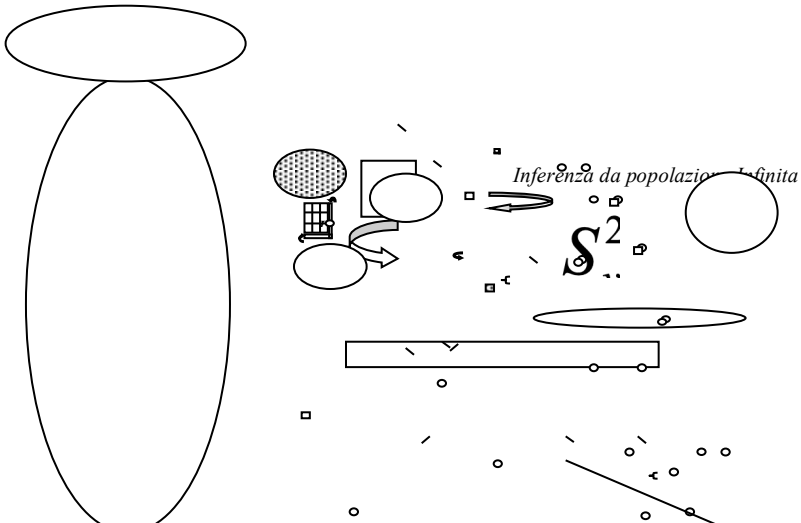


*Fig. 18.5 – Dati mancanti ed imputazione*

Riponderiamo? No, riponderare aveva senso nelle mancate risposte totali perché i pesi rimanevano uguali comunque per tutte le variabili, qui invece tra i dati mancanti non esiste una relazione lineare, non è detto che quando ad un'unità manca una variabile all'altra manchi la stessa, può darsi che a qualche unità manchi una variabile e ad un'altra ne manchi una diversa, e quindi nel migliore dei casi dobbiamo calcolare tanti vettori dei pesi quante sono le variabili - e quando dobbiamo incrociare due variabili quale peso usiamo? *Quindi la riponderazione sulle mancate risposte parziali non è possibile*, *dobbiamo imputare per forza. Ma va bene ogni metodo d'imputazione?* C'è un problema pratico, logico, che ci spinge ad usare una sola tipologia d'imputazione tra quelle che abbiamo trattato, ovvero dopo aver ripulito i nostri dati - cosa ci garantisce che dopo aver imputato un valore in un modo piuttosto che in un altro il piano di compatibilità venga rispettato? Dobbiamo quindi trovare un sistema d'imputazione che ci deve permettere di rispettare il piano di compatibilità. Scegliamo l'imputazione da donatore poiché quest'ultima tipologia d'imputazione prendendo ad esempio il grafico precedente, il valore lo va a prendere nell'unità con il valore più vicino, ad esempio il punto *A* nel caso il dato mancante si trovi in B, che logicamente se rispettava prima il piano di compatibilità lo rispetta anche adesso, dopo aver imputato. *Per variabili del piano di compatibilità*, è bene ribadirlo, intendiamo ad esempio quattordicenne, pensionato, divorziato, etc. *Quindi il metodo del donatore,* se non possiamo dire che è il più utilizzato per imputare le mancate risposte totali possiamo dire che *sicuramente è il più utilizzato per imputare le mancate parziali,* in special modo per indagini su individui e famiglie. In indagini su individui e famiglie non viene quasi mai imputata la mancata risposta totale, di solito si ripondera, ma viene sempre imputata la mancata risposta parziale da donatore, tra l'altro aggiungiamo che il donatore in questo caso, se sulle mancate risposte totali il vicinato si dovrà basare su delle variabili ausiliarie, qui il vicinato lo andiamo a cercare tra le rispondenti stesse, le *Y*, poiché ne abbiamo poche risposte mancanti. Nelle indagini sulle famiglie e gli individui il donatore viene ricercato di solito a distanza zero, ovvero andiamo a cercare la prima unità statistica che

ha gli stessi valori delle variabili della mancata risposta per le altre variabili, la stessa combinazione di riposte.

*Di queste tre fasi qual è quella che occupa più tempo?* La prima fase indubbiamente, e se così non è vuol dire che non si capisce nulla del fenomeno che s'indaga ed è qui che si trovano gli errori davvero madornali, poiché in questa fase la conoscenza va nel dettaglio, mentre la seconda fase già è più standardizzata, la conoscenza del fenomeno è relativa, mentre quella che richiede meno tempo e difficoltà più che altro da un punto di vista informatico, in automatico, è il calcolo dei minimi con il piano di compatibilità; stessa cosa per l'imputazione, anche l'imputazione è automatica. Di solito comunque questo processo in tre fasi è iterativo, viene ripetuto tre o quattro volte.

Dalle *10.000* unità statistiche in su per eseguire questo processo in tre fasi occorrono di solito *9-10* mesi, con circa una trentina di persone che ci lavorano.

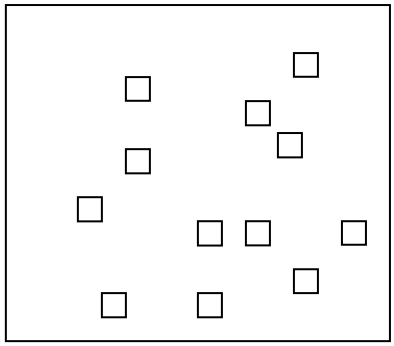
Per concludere diciamo che senza un piano di compatibilità l'indagine si può buttare nella spazzatura, è meglio fare un cattivo campione che un piano di compatibilità sbagliato; se poi facciamo un buon questionario ed un buon campione ma non ci siamo interessati di correggere i dati, i risultati non avranno un gran valore. In altre parole più che concentrarsi sui metodi di stima che in definitiva sono dei calcoli bisogna concentrarsi sui momenti reali dell'indagine, quelli cruciali, come appunto il questionario e quindi il piano di compatibilità, il micro-editing o il macro-editing, le stime in pratica saranno accettabili, fermo restando che altre fasi cruciali per l'indagine siano state eseguite in modo corretto.

# 19. Migliorare la stima

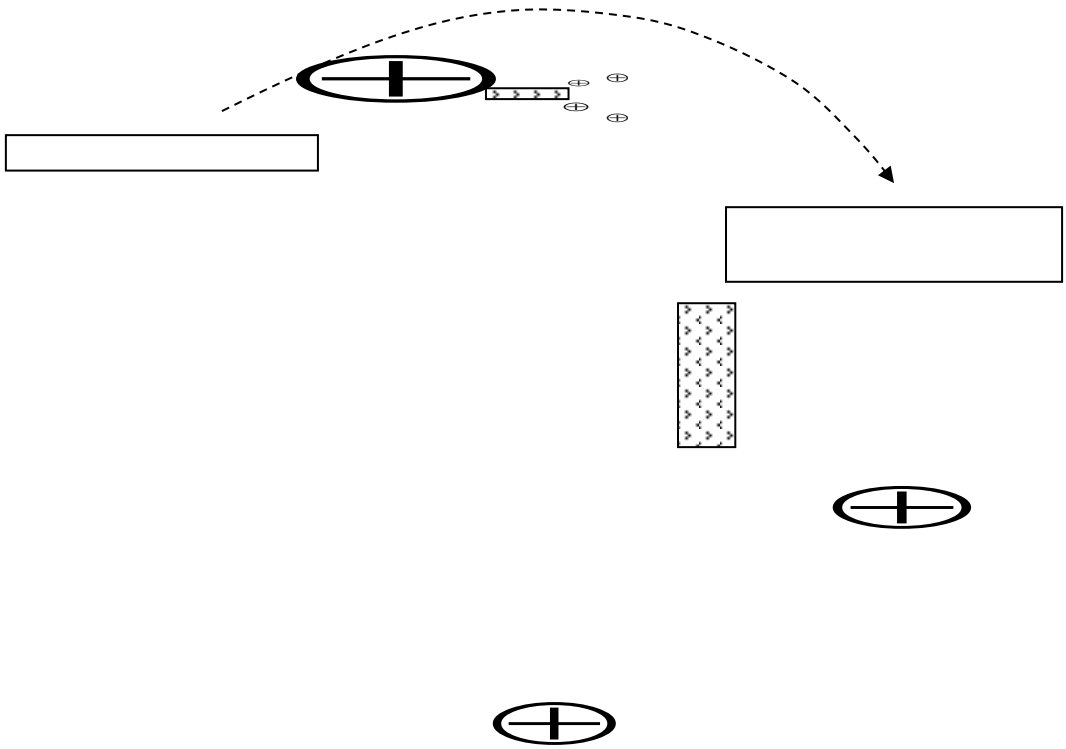
## 19.1 Le stime per domini, pianificati e non pianificati

Vediamo ora le stime per domini. Di solito quelle che produciamo sono tabelle di stime in cui tutta la nostra teoria dei campioni è ricompressa in una singola cella della tabella seguente, ad esempio il reddito per le classi di altezza della famiglia e magari per posizione professionale del capofamiglia, per regione di appartenenza. Cerchiamo di

entrare nel dettaglio e quindi di produrre le nostre stime, i nostri $\hat{t}_y$ di fatto possono essere ciascuna di queste caselle.

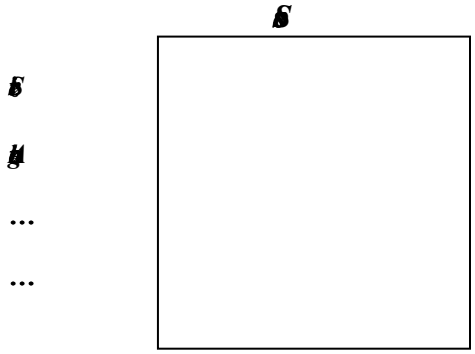


Ciò che mettiamo per riga e per colonna in queste tabelle solitamente a due dimensioni, sono dei codici che nascono da variabili di cui disponiamo nell'archivio, o che abbiamo rilevato nel campione. Ad esempio se facciamo le stime di reddito o di consumo famigliare suddivise per tipologie di spesa, consumo famigliare in pesce, alimentari, intrattenimento, automobili, etc., questo è un esempio di stima di una variabile suddivisa per codici di un'altra variabile che abbiamo rilevato per via campionaria, non ce l'abbiamo nell'archivio, oppure al contrario per regione di residenza della famiglia, questo è un tipico esempio di variabile che possediamo all'interno dell'archivio, la regione di residenza la conosciamo. I codici di queste variabili rispetto cui dividiamo le nostre stime si chiamano domini di stima. La distinzione fondamentale dei domini è tra *domini pianificati* e *domini non pianificati. Un dominio pianificato* è un pezzo di popolazione perchè ciascuna variabile di fatto partiziona la popolazione, è un pezzo di popolazione di cui noi abbiamo tenuto sotto controllo la numerosità, mentre un *dominio non pianificato* è un pezzo di popolazione in cui il numero di unità campionarie che ci cascano è totalmente fuori dal nostro controllo, ovvero *la numerosità all'interno del dominio è una variabile aleatoria*.



Facciamo un esempio: abbiamo un campione d'imprese e stratifichiamo per regione, per attività economica, classe di addetti, etc., e successivamente possiamo produrre una tabella di cui vogliamo sapere e fare la stima della produttività per settore di attività economica, sulle colonne ci mettiamo il sesso del conduttore dell'azienda per vedere se le donne imprenditrici sono più brave dei maschi. Questa tabella è a due dimensioni, una per righe ed una per colonne; in queste due dimensioni - quali sono i domini pianificati e quali no? E' pianificato il settore di attività economica non perché ce l'abbiamo nell'archivio ma perché avendo stratificato abbiamo controllato il numero di unità che finiscono in ciascuno strato, quindi le numerosità per ciascun settore di attività economica le abbiamo prefissate noi, quindi per definizione ed esclusione *il dominio non pianificato è il sesso* del conduttore dell'azienda. Tuttavia questa tabella potremmo anche non avere la possibilità di produrla poiché nulla ci garantisce di estrarre imprese condotte da donne, semplicemente perché le imprese condotte da donne sono circa il *5%*, potremmo non avere il numero sufficiente per produrre delle stime. In ogni caso anche se il numero di conduttori di azienda divisi per sesso avesse una percentuale di *50%* e *50%*, poiché il nostro campione non è forzato a rispettare nulla,

potrebbe capitare di avere tutti uomini o tutte donne, *la numerosità campionaria è completamente aleatoria*, totalmente libera, fuori dal nostro controllo, ed ha una probabilità che capitino tutti maschi o tutte donne.



Quindi possiamo fare una tabella di questo genere? Con i domini non pianificati? Sì. Quando abbiamo dei domini pianificati, fare queste tabelle rientra nelle regole e giacché sicuramente il campione è già stratificato fare le stime per strato è più agevole e garantisce una maggiore qualità, tra l'altro le avremmo comunque dovute fare Quindi il dominio stratificato non lo trattiamo perché rientra in ciò che abbiamo visto finora. Ma quando nelle tabelle inseriamo dei domini non pianificati qual è il reale problema che incontriamo - si possono applicare ugualmente tutte le formule viste fino ad ora? Vediamo il problema graficamente, abbiamo un archivio e supponiamo di avere tre classi di attività economica, i domini di stima , poi facciamo la nostra rilevazione e dividiamo il nostro archivio secondo il sesso del conduttore  in due parti. Chiaramente le due parti nella realtà non saranno divise in modo così esatto ma più sparpagliate all'interno dell'archivio. Il dominio in base al sesso non l'abbiamo pianificato ma dobbiamo fare le stime per sesso - come facciamo questa stima? I pesi di riporto all'universo non li tocchiamo, avremmo pesi sempre uguali in ogni strato.

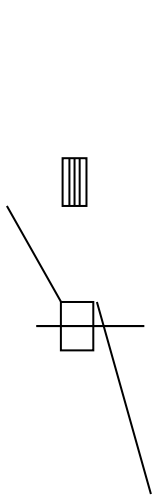


*Fig. 19.1 – Operare con i domini non pianificati*

Quando andiamo a fare le stime per maschi e femmine calcoliamo dei nuovi pesi di questo tipo?

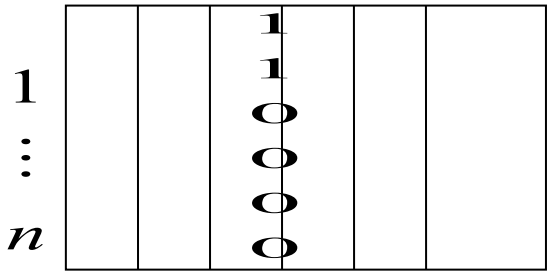
$$\frac{N_M}{n_M} \text{ ed } \frac{N_F}{n_F}$$

Teniamo i vecchi pesi o ne calcoliamo altri per il nuovo dominio? Il problema è che quanti maschi e femmine ci hanno risposto lo sappiamo, ma non sappiamo quanti maschi e femmine ci sono all'interno della popolazione, allora i pesi di riporto all'universo devono restare per forza quelli vecchi, e la stima la facciamo facendo la somma pesata nello stesso modo in cui facevano prima, è solo che invece che all'interno dello strato sommiamo all'interno dei maschi o delle femmine, e chiaramente i pesi cambieranno al cambiare dell'unità poiché sommiamo al cambiare dell'unità, perché stiamo cambiando strato, perché i dati sono mischiati. Quindi i pesi di riporto all'universo dell'Horvitz-Thompson non è detto che debbano essere cambiati.

.

$$\sum_M w_k \cdot y_k$$

Abbiamo in pratica sempre una somma pesata, abbiamo comunque un vettore di pesi di riporto all'universo, è solo che la nostra somma pesata non la facciamo solo all'interno dello stesso strato ma la estendiamo a tutti i maschi o a tutte le femmine. Lo stimatore, rimanendo sull'Horvitz-Thompson, non cambia, al punto che, visto che tra le variabili abbiamo anche il sesso, che ci genera una sequenza di zero e di uno, possiamo anche farci una stima del numero di aziende condotte da maschi ad esempio facendo la somma dei pesi per i maschi.




Poiché la somma dei pesi indica quante imprese sono rappresentate da ciascuna unità campionaria, se noi sommiamo le imprese che sono rappresentate da quelle campionate e condotte dal sesso maschile, abbiamo una stima di quante imprese sono condotte da maschi. *Badiamo bene che è una stima del numero della popolazione* perché per i domini non pianificati è ovvio che la numerosità totale non la conosciamo altrimenti la pianificheremmo, useremmo la numerosità totale per forzare la numerosità del campione.

$$\hat{N}_M = \sum_M w_k$$

Ma allora se possiamo fare la stima qual è il problema che possiamo incontrare nel fare i domini non pianificati? Se è fuori controllo la numerosità campionaria sicuramente *sarà fuori controllo anche la varianza*, sicuramente, ad esempio se per caso in un campione di *30.000* imprese solo tre imprese sono condotte da donne, la varianza fatta su questi tre numeri sarà infinita. A questo punto neanche la stima ha più senso. Chiaramente la varianza di questo stimatore non è quella del campione stratificato, perché la varianza sarà più complessa, poiché *usare i domini non-pianificati equivale a fare un post-*

*stratificazione*, che è un'operazione successiva, è una cosa un po' più complicata della varianza dello stratificato, perché deve tener conto del fatto che oltre alla sua varianza c'è anche una variabilità nell'*n*, non solo nell'*N*. Comunque quello che bisogna sapere è che la varianza del dominio non-pianificato non solo è fuori dal nostro controllo ma solitamente è molto più alta di quelle sotto il nostro controllo perché, anche se ci capitano molte numerosità campionarie, il fatto che la numerosità campionaria sia fuori dal nostro controllo ci genera un'aleatorietà che ci fa alzare la varianza, quindi il problema dei domini non-pianificati non è un problema di calcolo ma è che *la varianza è sempre troppo alta,* cioè i domini non pianificati hanno sempre varianze troppo alte per poter produrre delle stime.

Nella realtà quando si fa un'indagine vengono prodotte delle tabelle con delle stime all'interno, e non si guarda se un dominio è pianificato o no. Chi svolge le indagini in modo un po' più serio per i domini pianificati da solo un'occhiata per vedere se i vincoli sulle numerosità sono rispettatati, e per i non pianificati va a vedere se queste numerosità, comunque aleatorie, sono così squilibrate da aver creato varianze troppo alte, ed allora interviene, altrimenti lascia tutto com'è ma controlla i non pianificati. Restando ancora nelle soluzioni pratico-operative, nella terza versione, quella che trattiamo noi, se il dominio non pianificato è molto importante e quindi possiamo produrre buone stime nel dominio non pianificato abbassandone la varianza dobbiamo intervenire per schiacciare la varianza in un qualche modo, in un qualche modo che non riguardi più l'indagine poiché tutto quello che potevamo lo abbiamo fatto, abbiamo già utilizzato le ausiliarie, etc., abbiamo già fatto tutto, quindi alla fine ci vengono delle varianze troppo alte - come facciamo ad abbassare esogenamente, artificiosamente, le varianze di un'indagine che ormai è arrivata alla fine? Come facciamo ad abbassarle in modo deontologicamente corretto? Fino a quando stavamo all'interno del nostro progetto avevamo molti modi per abbassare la varianza, potevamo, aumentare la numerosità campionaria in fase iniziale quando pianificavamo il campione, potevamo, dopo la riconsegna dei questionari utilizzare lo stimatore di regressione tramite la relazione con le ausiliarie della mia variabile per abbassare la varianza, quindi per abbassare la varianza che in definitiva significa utilizzare meno unità campionarie, potevamo intervenire sia in fase di disegno dell'indagine utilizzando il piano di campionamento più opportuno, *PPS*, stratificato, etc., oppure in fase di stima

utilizzando la regressione, ponderazione vincolata, etc., poiché hanno comunque sempre una varianza più bassa dell'Horvitz-Thompson. Ma alla fine nonostante tutti i nostri sforzi in fase di pianificazione e di stima vediamo che in fase di produzione di dati abbiamo domini non pianificati che ci danno delle varianze troppo alte - cosa facciamo?

## 19.2 La stima per piccole aree

La soluzione sta sempre nell'uso di variabili ausiliarie ma non contenute nell'archivio, ovvero non note per ciascuna $X$ ma note per il dominio di nostro interesse. Si capisce bene che ottenere delle variabili ausiliarie per imprese o individui è un discorso, ne chiediamo poche ed è difficile ottenerle, mentre per grandi aggregati, che sarebbero i nostri domini di stima, la situazione cambia. Ad esempio sappiamo che la nostra produttività è legata da una qualche correlazione all'età del conduttore, magari i giovani sono più produttivi e gli anziani di meno, successivamente andiamo a cercare l'età dei maschi e delle femmine italiani, andiamo in pratica a cercare dati aggregati per dominio di stima fuori dalla nostra indagine che abbiano una qualche correlazione con le nostre stime. Queste problematiche di solito si creano quando il dominio è di stampo territoriale, infatti quello che stiamo introducendo è la "Stima per piccole aree", anche se la sua denominazione non indica necessariamente un ambito solo territoriale, magari sarebbe stato più opportuno chiamarle "stime per domini", poiché il metodo non sarebbe cambiato ad esempio utilizzando come dominio non la regione o la provincia ma il settore di attività economica; vengono chiamate in questo modo perché sono nate per risolvere un problema di stime di dettaglio territoriale, stime di dettaglio territoriale che nascono da una determinata realtà empirica ovvero che qualsiasi indagine deve essere significativa quindi bisogna aver prefissato il limite del coefficiente di variazione ad un certo livello territoriale, più viene fissato ad un basso livello e più la numerosità campionaria aumenta, sarà diverso ad esempio accettare l'*1%* d'errore a livello nazionale, regionale, provinciale o comunale, a livello comunale diventa quasi un censimento. Ma il problema è che potrebbe succederci che in un'indagine il campione lo facciamo a livello nazionale perché ci costa poco ma l'utente che ci ha commissionato l'indagine potrebbe richiedere anche le stime regionali, sarebbe difficile spiegare perché abbiamo fatto stime solo a livello nazionale, allora dobbiamo cercare un modo per

risolvere la situazione. Ci arrampichiamo sugli specchi quando ci viene chiesta la stima regionale ma non ce l'abbiamo perché non l'abbiamo pianificata, l'abbiamo fissata a livello nazionale, però abbiamo un'infinità di dati e variabili ausiliarie a livello regionale dalle varie fonti. Ad esempio la stima per il livello di ricerca e sviluppo delle imprese sicuramente sarà correlata con qualcosa che conosciamo per ciascuna regione, quantomeno con il numero di laureati, ricercatori presenti in ciascuna regione, etc., oppure la spesa ad esempio sarà correlata con il reddito medio di una determinata regione, in ricerca e sviluppo c'investono i ricchi naturalmente e non i poveri, quindi facciamo le nostre correlazioni con i dati aggregati semplici da trovare a livello regionale. Successivamente diciamo che le nostre stime di Horvitz-Thompson, ottenute per ciascun dominio, sono una qualche funzione delle nostre variabili ausiliarie note, quelle che abbiamo trovato a livello aggregato. Per funzione intendiamo alla fine una regressione, lineare o non lineare, ma la sostanza non cambia.

$$\hat{y}_{HT,i} = f\left(x_i\right) + \varepsilon_i$$

Come si calcola la varianza lo vedremo successivamente poiché è un processo molto lungo, per ora ci soffermiamo sul fatto che abbiamo inserito due casualità, la $\hat{y}_{HT,i}$ è già una stima, ha un suo errore campionario $\sigma^2_{HT,i}$ dovuto al fatto che abbiamo campionato da *N* ad *n*. Poi *il secondo errore casuale è dovuto al fatto che la nostra relazione lineare non è perfetta* ma ci sono degli scostamenti casuali. Questi due effetti del caso sono di tipo completamente diverso, la prima casualità è quella canonica che abbiamo trattato in questo corso, la novità di questo corso ovvero la casualità dovuta al campionamento da popolazioni finite mentre l'altro, l'$\varepsilon_i$ *è un errore di misura da popolazioni infinite* - ma allora come conciliamo le due cose quando i due errori casuali variano in modo completamente indipendente? Con quale modello? L'insieme di queste due cose *si chiama modello da superpopolazione*, *abbiamo due popolazioni infinite* e quindi, una superpopolazione. Vedremo che una volta che vengono rimosse le ipotesi di base per popolazioni finite ed infinite le stime che vedevamo a statica di base ed in teoria dei campioni diventano distorte, avremo formule molto lunghe e complicate anche perle stime più banali. In ogni caso da un punto di vista logico non dovremmo

avere particolari difficoltà ma solo da un punto di vista formale, basta capire cosa è diventato più complicato in realtà.

$$\hat{y}_{HT,i} = f\left(x_i\right) + \varepsilon_i$$

$$dove \qquad f\left(x_i\right) = a + b \cdot x \cdot \varepsilon_i$$

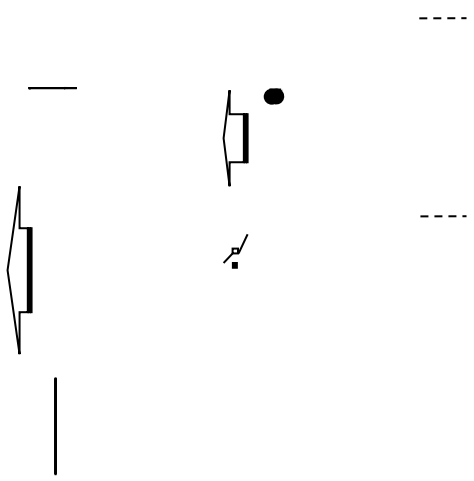


*Fig. 19.2 – L'approccio da superpopolazione*

Il campionamento da superpopolazione anche se complica un po' le formule da un punto di vista pratico è veramente un svolta perché, se trovare le ausiliare come facevamo precedentemente a livello d'impresa era molto complicato, di ausiliarie a livello aggregato ne troviamo moltissime, questo vuole dire che quando andiamo a fare le stime abbassare la varianza alla fine è una cosa banale se sappiamo utilizzare le superpopolazioni, certamente troveremo almeno una variabile che è correlata con le stime. Quindi arrivati alla fine, detto in parole povere, di *"colpi al cerchio ed alla botte"* ne possiamo dare quanti ne vogliamo in questo modo; tra l'altro dal punto di vista del calcolo, le formule saranno lunghe ma non è da questo che dipende il tempo del lavoro, dipende dalla quantità di dati che vogliamo elaborare.

Se prima eravamo costretti a lavorare con le imprese ora siamo sulle stime, avremo quindi meno dati da elaborare. Guardando il grafico precedente supponiamo di avere *4* punti di stima, abbiamo la nostra ausiliaria aggregata molto correlata con le nostre stime, ed abbiamo le nostre stime per ciascun dominio, a questo punto una regressione la sappiamo fare. Ma conosciamo solo le stime? No, per ciascuna stima conosciamo anche la variabilità, rappresentata dalle barre nel grafico in cui l'ampiezza sarebbe l'intervallo di confidenza, più o meno la varianza. Ora siccome la numerosità campionaria era fuori dal nostro controllo, in alcuni domini gli intervalli di confidenza ovvero le barre sono più larghe, mentre in altri ci sono finiti numerosi punti e quindi gli intervalli di confidenza si sono ristretti. Quindi in definitiva per ciascuna stima abbiamo le varianze campionarie ed in più per ciascuna stima abbiamo un errore $\varepsilon_i$ osservato rispetto a quanto si allontana dalla retta di regressione, quanto quel modello di regressione quella funzione, per quella determinata stima di quel dominio lo rappresenta. Quindi noi di fatto ogni volta che facciamo una regressione o un modello qualsiasi una misura riguardante il campione ed una misura riguardante il modello teorico di fatto ce l'abbiamo, allora poiché abbiamo sia una stima fatta sul campione che una stima proveniente dal modello, abbiamo due stime, una per ciascun dominio ma ne vogliamo una sola che intuitivamente potrà essere la media ,ovvero la stima finale $\hat{y}_F$, sarà una media tra la stima proveniente dalla indagini campionarie $\hat{y}_I$ ed una stima proveniente dal modello $\hat{y}_M$. Ma la media sarà quella semplice? Dividiamo per due? No, poiché qualche volta l'incertezza sarà grossa e qualche volta ci sarà un'incertezza piccola, quindi la media la facciamo pesata,

$$\hat{y}_F = \alpha_i \cdot \hat{y}_I + (1-\alpha_i) \cdot \hat{y}_M$$

$$dove \qquad \alpha_i = g\left(\varepsilon_i, \hat{\sigma}^2_{\hat{y}}\right) \qquad e \qquad 0 < \alpha_i < 1$$

con un peso che è una qualche funzione dell'errore di regressione e della varianza campionaria, dove $\hat{y}_F$ rappresenta una combinazione lineare convessa. *In pratica vogliamo fare una media delle due stime con pesi che dipendono dall'incertezza delle due*, ovvero se la varianza campionaria è molto bassa, se la sue stima è molto precisa, il

peso $\alpha_i$ dovrà essere molto alto per equilibrare le due stime, ovvero vogliamo che la media che cerchiamo mantenga la qualità dell'indagine. Viceversa se una regressione ha un $R^2$ molto alto, e quindi spiega bene i valori, vogliamo che il peso $\alpha_i$ sia molto basso. Quale delle due stime funziona meglio lo decidiamo sulla base della variabilità, *la variabilità nella regressione si chiama somma dei quadrati dei residui*, se tutti i punti si trovano sopra la regressione vuol dire che spiega bene il fenomeno e quindi sarà attendibile, vince nell'essere utilizzata per calcolare la media, per la variabilità campionaria viceversa. Il peso quindi dovrà avere determinate caratteristiche,

$$\hat{\alpha}_i = \frac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_\varepsilon + \hat{\sigma}^2_{y_{HT,i}}}$$

ovvero se la varianza dell'indagine è alta vogliamo che $\hat{\alpha}_i$ diminuisca, in questo modo

se

$$\hat{\sigma}^2_{y_{HT,i}}$$

sale il peso si annulla, quindi la varianza di regressione la mettiamo al numeratore per ottenere l'effetto opposto ed al denominatore per fare in modo che il rapporto, quando la varianza dell'indagine è nulla, sia uguale ad uno. Per ora a questo risultato ci siamo arrivati in modo informale, in ogni caso la sostanza è questa, *le formule lunghissime non si trovano come abbiamo visto nei calcoli precedenti, ma si troveranno nel tipo di varianza che otterremo*, poiché abbiamo varianze stimate, medie pesate con pesi stimati a loro volta, etc. Questo $\hat{\alpha}_i$ si chiama *"shrinkage factor"*, significa che "stiamo stirando", tirando la nostra stima ad andarsene sopra la regressione e più la regressione funziona più ci stiriamo sopra la variabilità, altrimenti non la tocchiamo. Naturalmente $\hat{\alpha}_i$ cambia al variare del dominio, avremo tanto più stiramento verso la retta quanto più la varianza campionaria è alta, questo processo in pratica ci riporta verso la retta di regressione.

## 19.3 Alcuni dubbi – quale varianza c'interessa?

*Digressione:* ci sono almeno tre piani di campionamento che hanno i pesi di riporto all'universo sempre uguali su tutte le unità? Il campione casuale semplice, lo stratificato proporzionale prendendo sempre la stessa proporzione in ogni strato, *PPS* con la variabile *X* uguale su tutte le unità ed il campionamento a grappoli. Ora questi piani di campionamento che hanno i pesi di riporto all'universo sempre uguali hanno tutti la stessa varianza? Avere lo stesso peso di riporto all'universo significa avere le stesse probabilità d'inclusione, del primo ordine però, il punto chiave è che nella varianza entrano le probabilità d'inclusione del secondo ordine che possono essere diverse. Un esempio classico si ha nel campionamento sistematico in cui abbiamo lo stesso vettore dei pesi di riporto all'universo e quindi delle probabilità d'inclusione del primo ordine ma il *coefficiente di correlazione intraclasse* potrebbe non essere uguale a zero. Lo stratificato a maggior ragione, poiché un campione stratificato ha una varianza sicuramente minore del campione casuale semplice, anche se ha lo stesso vettore dei pesi di riporto all'universo, avrà varianza inferiore perché sono diverse le probabilità d'inclusione doppie. Il campionamento a grappoli ha una varianza superiore al campione casuale semplice, ritorna il discorso, anche se non è stato trattato, del *coefficiente di correlazione intraclasse*. Ovvero se c'è un minimo di dipendenza all'interno dei grappoli e quindi i gruppi non sono completamente casuali, già questo fa aumentare la varianza.

Ricordando un esempio di campionamento fatto nelle prime lezioni di questo corso, se prendiamo un campione di palazzi ed intervistiamo tutte le famiglie, se c'è un'omogeneità tra le famiglie, di fatto intervistiamo famiglie molto simili, ad unità molto simili, quindi il contributo alla varianza del campione è basso, intervistiamo molte unità ma che sono tutte uguali tra loro. In pratica avremo una varianza minore all'interno del campione ed una maggiore tra i campioni, ma la varianza dello stimatore è la varianza tra un campione ed un altro, il campionamento a grappoli in pratica è uno stratificato al contrario. Ricominciando dal funzionamento dell'inferenza, guardando il grafico precedente, supponiamo di avere il nostro archivio, la nostra popolazione,

prendiamo un pezzo, un campione e facciamo la stima, prendiamo un altro pezzo e facciamo la stima, etc., possiamo prendere tutti i campioni che vogliamo e farci una stima sopra, ciascuno con le proprie probabilità di essere estratto, in modo da ottenere una distribuzione di probabilità delle nostre stime. Quindi quando parliamo di varianza intendiamo la varianza dello stimatore, della distribuzione di probabilità; quindi è la varianza calcolata sulla gaussiana in alto che c'interessa non la varianza all'interno del campione, ovvero se la varianza all'interno del campione è molto bassa, poiché abbiamo preso unità molto omogenee, abbassare la varianza all'interno del campione significa logicamente alzare quella tra le stime, tra i gruppi, tra i pezzi o in qualsiasi modo vogliamo chiamarli.

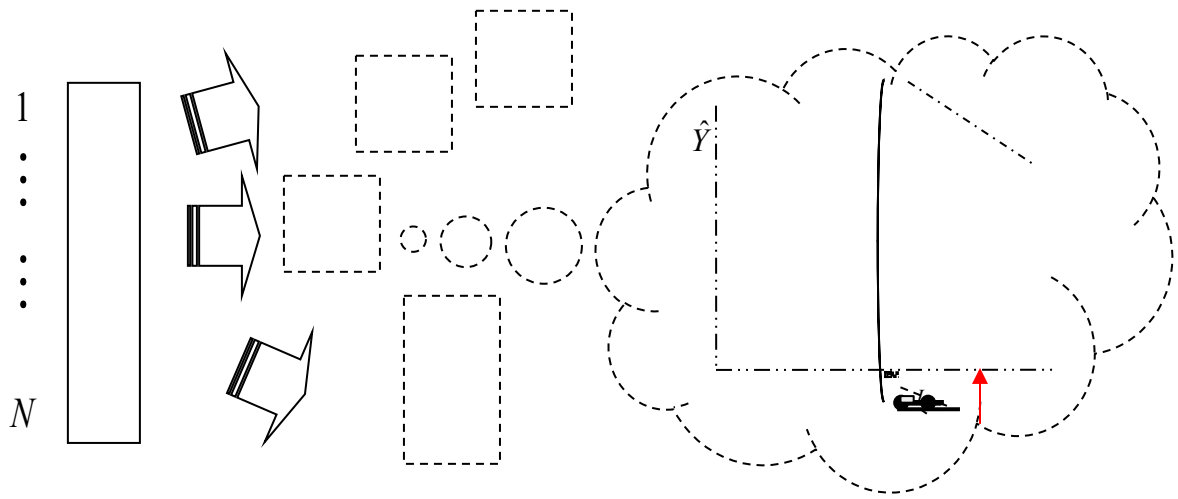



*Fig. 19.3 – La distribuzione di probabilità di un'estrazione*

### 19.4 La varianza nella stima per piccole aree

Riprendendo il discorso precedente ci siamo lasciati con il dubbio riguardante il modo in cui calcolare la regressione con due tipi di casualità, dovuta all'indagine ed al modello, ovvero come calcolare l' $\hat{\alpha}$ , e questo si fa stimando dei *modelli ad effetti misti*, i cosiddetti *Mixed Model*, effetti che sono in pratica gli effetti in un modello di regressione, ovvero gli errori; ora siccome nel nostro caso di errori ne abbiamo due, uno

che viene dal campione e l'altro dalla superpopolazione, la regressione precedente va stimata in un modo un po' particolare, non vanno più bene i minimi quadrati ma ci vuole qualcosa di un po' più complesso poiché abbiamo due fonti di errore, quella campionaria e quella da superpopolazione, sarebbero questi i *Mixed Model*. *Questi modelli ci danno i coefficienti* $\alpha$ e $\beta$ di regressione e la varianza di regressione, una volta che abbiamo questi dati poi stimiamo l'alfa che abbiamo visto precedentemente. Chiudendo il discorso sulle stime per piccole aree, queste stime si chiamano $\hat{y}_{EBLUP}$, fatte con l'alfa stimato per la stima campionaria ma potremmo metterci una qualsiasi stima proveniente da popolazioni finite, più quella proveniente dalla regressione tra i due.

$$\hat{y}_{EBLUP} = \hat{\alpha} \cdot \hat{y}_{CAMP} + \left(1 - \hat{\alpha}\right) \cdot \hat{y}_{REG}$$

*L'MSE* di questa stima sarà uguale a tre componenti, di cui diciamo solo la prima poiché le altre sono formule molto complicate; ovvero come abbiamo visto nelle pagine precedenti,

$$MSE\left(\hat{y}_{EBLUP}\right) = \hat{\alpha} \cdot V\left(\hat{y}_{CAMP}\right) + C_2 + C_3$$

nel fare la combinazione tra stime dirette campionarie e modellistiche della regressione noi di fatto spostavamo il punto verso la regressione in funzione delle due varianze, questo vuol dire che se una delle due varianze è già molto bassa, non la tocchiamo, l'alfa diventa prossimo ad uno nella combinazione lineare, rimane la stessa varianza. Nell'*MSE* finale $C_2$ rappresenta *la varianza di distorsione*, $C_3$ altro, poiché queste stime sono distorte, delle due componenti che non trattiamo una è la distorsione, ovvero la varianza dell' $\hat{y}_{EBLUP}$ è funzione della varianza della stima e del singolo alfa su ciascun dominio, ricordando che se facciamo stime su più domini l'alfa cambia su ogni dominio, ad esempio se facciamo stime provinciali avremo 103 alfa, ovvero nel fare queste stime $\hat{y}_{EBLUP}$ più l'alfa è basso e più spostiamo le stime sulla retta di regressione, e più spostiamo sulla retta le stime e più schiacciamo la varianza.

Ora, per concludere tutti gli argomenti di lezione, non ci rimane altro, dopo aver trattato tutte le stime possibili, trattare la diffusione dei dati ed i controlli di qualità,

poiché di solito ad ogni indagine, in quest'ultimo caso, sarebbe opportuno effettuare un controllo su come è stata svolta l'indagine.

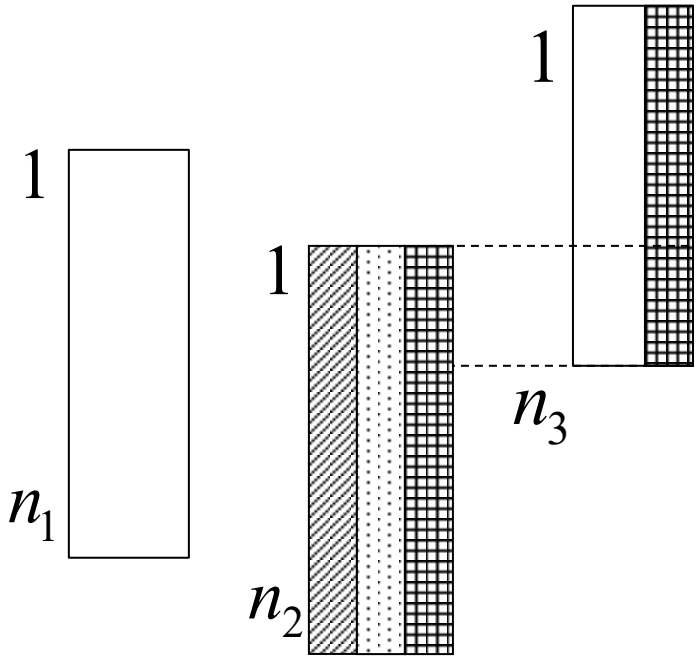
## 20. Diffusione dei dati e controlli di qualità

### 20.1 La diffusione dei dati

Cominciamo con la diffusione dei dati; fondamentalmente ci possiamo limitare alla diffusione informatica o digitale ed a quella cartacea. Su quella cartacea non ci sono particolari cose da dire. L'Istat è stata una delle prime a diffondere i dati in modo digitale, anche a livello europeo, si risparmia molto denaro per la stampa volumi e le indagini escono quasi un anno prima per aver evitato le numerose iterazioni dovute alle correzioni, controllo bozze, etc., il fatto è che nella diffusione dati via internet si apre una categoria di problemi, poiché non necessariamente produciamo tabelle fisse, standard, al contrario dei volumi nelle indagini cartacee dove inseriamo grafici già completi. Immaginiamo ad esempio in un'indagine in cui rileviamo $K$ variabili quante combinazioni di tabelle possibili possiamo fare; ora quando abbiamo fatto il questionario abbiamo detto che bisognava fare affiancato al questionario un piano di compatibilità. Allo stesso modo da subito viene fatto *un piano di tabulazione*, ovvero in parallelo, molto prima che facciamo l'indagine, una volta che abbiamo il questionario, *elenchiamo le tabelle da produrre*. Immaginiamo ad esempio quante tabelle possono venire fuori da un'indagine in cui abbiamo un migliaio di variabili, facendo anche semplicemente tabelle a doppia entrata, quindi dobbiamo elencare all'inizio tutte le tabelle da produrre. La prima modalità di diffusione si basa su un elenco di tabelle fisso, *un elenco di tabelle però fisse*. Il punto è che questa cosa però non è proprio il massimo che si possa fare; teoricamente potremmo mettere in linea tutta l'indagine con tutti i pesi di riporto all'universo e chi si collega si fa la sua tabella, nelle combinazioni di $K$ variabili a due a due ogni utente che si connette si sceglie la combinazione che gli piace.

Detta così sembra semplice ma è molto complicata, perché innanzitutto bisogna mettere l'indagine con i dati elementari per singola unità statistica disponibili sul web, magari non ad accesso illimitato ma comunque sul sito ci devono stare, questo è il primo problema. L'altro problema è che ogni volta che viene richiesta una tabella

quest'ultima deve essere realizzata fisicamente, ovvero bisogna fare una tabella di stime, questo vuol dire che se *1000* persone si collegano contemporaneamente sul sito potrebbero crollare il server. Quindi siamo passati da tabelle elencabili in qualche modo, a *tabelle mobili* poiché ognuno si sceglie la propria. *La terza soluzione*, di flessibilità totale, sta nel mettere tutte le indagini insieme nello stesso server e queste tabelle mobili si posso fare con una variabile presa da un'indagine ed una variabile presa da un'altra indagine. Questo significa che ci sarà sullo stesso server un certo numero di campioni, con indagini che naturalmente incidono sempre sulle stesse unità statistiche,




dove possiamo chiedere non solo una tabella di due variabili ma possiamo chiedere una tabella con una variabile che sta in un campione ed una variabile che si trova in un altro campione, poi il software si occupa di incrociare, prendere i pezzi in comune, e ricalcolare i pesi di riporto in modo da far tornare i valori. Questa cosa, di accesso *a più data-base* diversi, agganciabili o non agganciabili, informaticamente viene detta *data-where-house*.

Naturalmente le unità che abbiamo inteso nel grafico sono tutte unità agganciabili tramite codici. La prima versione che è la più semplice la troviamo su tutte le indagini dell'Istat, la seconda non viene fatta per famiglie ed individui ma viene fatta solo sulle imprese così come anche la terza, per un motivo molto semplice, perché *l'archivio delle*

*famiglie e degli individui non lo possiede l'Istat*, si trova presso i comuni, quindi sarebbe difficile agganciare le indagini ed i dati relativi. Quello che vogliamo dire e che è fondamentale nella diffusione via digitale, non sono solo i dati di cui stiamo parlando ma anche quelli che vengono chiamati *Meta-Dati*, ovvero l'altro gran pregio che ha la diffusione digitale delle indagini e che è possibile diffondere il metodo ed il criterio con cui le singole stime sono state calcolate, un contorno che comunque deve essere appoggiato a ciascuna tabella prodotta; ad esempio il questionario stesso è un meta-dato, oppure il tipo di indagine o stimatore che si è utilizzato. Poi abbiamo un aspetto non indifferente di stampo empirico; le tabelle di cui parliamo possiamo produrle tranquillamente senza alcuna difficoltà? Facciamo un esempio, supponiamo che chiediamo i dati sull'import-export per paese di destinazione, dove vengono mandate le merci e per prodotto, e su questa tabella esiste un'impresa che dall'Italia esporta in Afganistan dei cuscinetti a sfera. Magari viene fuori che i cuscinetti a sfera serviranno per il lancio di missili, quindi si rende opportuna una denuncia al ministero della difesa. Successivamente l'impresa decide di denunciarci per violazione della privacy, ovvero non possiamo diffondere dati a meno che questi dati non siano contenuti in pubblici registri, come sesso, età, stato civile, etc., qualsiasi altra cosa non la possiamo diffondere se è agganciabile all'unità statistica. Magari facendo un'indagine sulla ricerca e sviluppo nella provincia di Pescara risulta che la spesa in *R&S* è *100.000* euro, se però d'imprese che fanno ricerca e sviluppo ce n'è una sola è come se avessimo detto il nome, quindi la carcerazione è automatica. Non deve essere possibile agganciare nessun dato ad un nome ed un cognome. Come soluzione tecnica a questo problema gli istituti di statistica stabiliscono come *soglia per non diffondere le singole celle della tabella, le tre unità statistiche*, perché ad esempio diffonderne due significa che comunque un'impresa sa dell'altra, quindi si stabilisce questa *regola del tre*. Cosa che tra l'altro non è così banale, perché a questo punto abbiamo la nostra tabella a doppia entrata ed una casella ha una frequenza minore di tre; cosa facciamo? Mettiamo dato mancante? Assolutamente no poiché nella tabella comunque abbiamo il totale, la frequenza marginale, quindi per differenza possiamo ricostruirci il dato mancante, per cui bisogna, ogni volta che abbiamo un dato mancate, come si vede nel grafico successivo, *cancellare il dato mancante* sia per la frequenza marginale sulle righe che quella sulle colonne, dobbiamo togliere il totale di riga e di colonna, poi quale sia

l'algoritmo che minimizza il numero di punti in modo da ottenere sempre un quadrato non lo trattiamo, in ogni caso non si può cancellare ed andare avanti senza problemi.

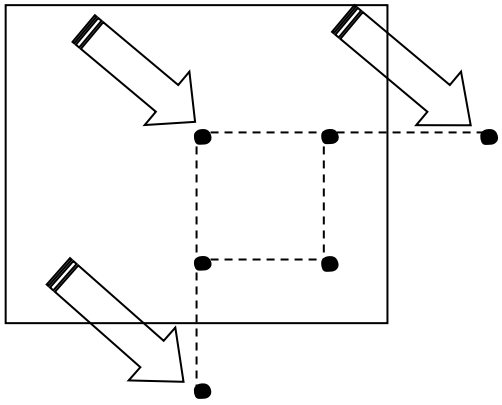
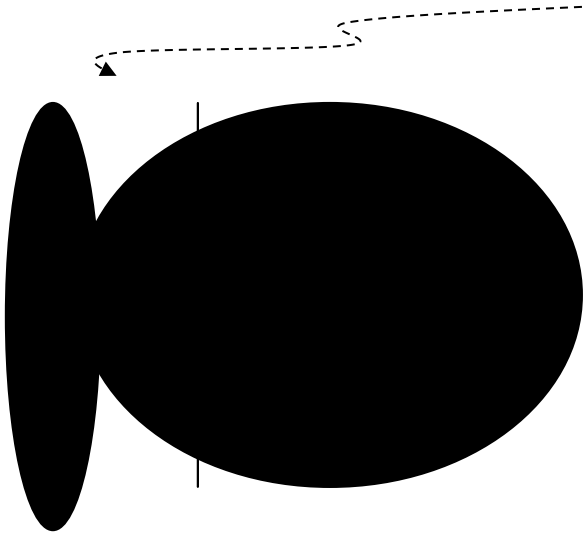


*Fig. 20.1 – La regola delle tre unità statistiche*

Fino ad ora abbiamo parlato di *macro-dati*, per quanto riguarda invece i *micro-dati*, che comprendono in pratica tutto il file dell'indagine, la legge non ci dice che non possiamo diffonderlo, *l'importante è che chi possiede quel file non possa agganciare i dati ad un nome ed un cognome*, e questo lo si fa sempre con la regola del tre di cui abbiamo parlato precedentemente. Supponiamo di avere la nostra indagine con le nostre variabili, e si presuppone che di queste variabili una parte sia disponibile al pubblico ed il resto no. Queste ultime sono quelle che creano problemi, sono quelle rispetto alle quali non ci deve essere mai una frequenza inferiore a tre. Ad esempio una ipotetica signora *G* non la potremmo identificare se diciamo che è una donna, ha oltre *65* anni ed è divorziata; se poi però diciamo che ha subito un particolare intervento chirurgico, la signora *G* può denunciarci perché potrebbe rappresentare un caso unico all'interno dei nostri dati. La regola del tre quindi vale per le cosiddette variabili chiave, quindi interveniamo su queste ultime, ripuliamo i dati da possibili elementi che possano rendere riconoscibili le nostre unità.



Questi dati di dominio pubblico vengo detti "*Public using files*" , *PUF*. In alternativa al metodo statistico precedente c'è anche un modo per accedere ad i dati grezzi. In questo caso la violazione della privacy è impedita per via fisica, ovvero siamo noi stessi ad elaborare i dati all'Istat, dove i computer sono senza supporti che permettano di portare i dati fuori dall'Istat. Una volta terminata la nostra elaborazione un impiegato controlla che la nostra analisi non contenga nessun dato elementare e ce lo stampa. Il sito su cui possiamo fare la richiesta per questi dati grezzi si chiama *ADELE*.

## 20.2 I controlli di qualità e gli errori non-campionari

Passiamo ai controlli di qualità. Supponendo di essere arrivati alla fine della nostra indagine, di aver fatto tutte le operazioni, ma non sappiamo dire dove sono gli errori non-campionari. Come si misurano gli errori campionari lo sappiamo, sarebbe la varianza dei nostri stimatori; l'obiettivo del controllo di qualità è misurare quelli che sono chiamati errori non-campionari. Di errori *non-campionari* ne abbiamo visti alcuni, ad esempio gli errori di *sovracopertura* e *sottocopertura*. Tra l'altro gli errori non-campionari sono molto più grandi di quelli campionari. Ad esempio il censimento, per quanto risulti laborioso e costi cento volte un'indagine classica, per molti è molto peggio di un'indagine campionaria poiché nei censimenti *anche se l'errore campionario è zero*, gli errori non-campionari crescono esponenzialmente. Quindi

abbiamo errori di archivio come l'errore di sottocopertura, *l'errore di misura* ne è un altro, nel senso che se ad esempio consegniamo i questionari e chiediamo il reddito, le unità potrebbero risponderci *3000* euro invece che ad esempio *3.648* euro, questo è un errore di misura, che può essere volontario o non volontario. *L'effetto rilevatore* è un altro errore non-campionario oppure ancora *l'effetto memoria*, che è dovuto al fatto che facciamo delle domande troppo retrospettive; *la mancata risposta*, totale o parziale, genera comunque errori non-campionari anche se abbiamo visto tutto il procedimento per risistemarli, e a questo proposito anche *il tasso d'imputazione* diventa un errore non-campionario, concettualmente diverso per la mancata risposta totale e la mancata risposta parziale e quindi va trattato in modo differente. Come vediamo ce ne sono parecchi di errori non-campionari. Alcuni di questi siamo in grado di calcolarli e misurarli a costo zero, come per le imputazioni, per l'errore di sovracopertura, però ci sono elementi che non siamo assolutamente in grado di misurare a meno che non facciamo delle indagini apposite.

Il nostro obiettivo in queste indagini non è solo misurare, ma fare delle indagini di controllo di qualità così come la TOYOTA fa i controlli di qualità sulle proprie automobili così lo statistico fa il controllo di qualità sulle proprie stime garantendo che i dati abbiano un certo grado di ragionevolezza e rispetti una serie di parametri di qualità. Vogliamo quindi migliorare la qualità della nostra indagine. Questo è il caso tipico in cui se il rilevatore della nostra indagine sa che non è controllato in nessun modo non si impegnerà molto nel proprio lavoro. Quindi le indagini di controllo di qualità vanno a coprire la sottocopertura delle unità statistiche, ed è questa la fase più complessa. *Per gli errori di misura* invece si fa un sottocampione dal nostro campione e si somministra ex-novo il questionario anche se spesso non completo cioè non su tutte le variabili. Queste ultime di solito sono telefoniche, si cerca di spendere il meno possibile, non si può raddoppiare la spesa e in contemporanea, oltre a cercare di coprire l'errore di misura, si fa anche il controllo sull'effetto intervistatore; "è venuto da lei l'intervistatore?".

*Come si procede per stimare la sottocopertura?* Potremmo agganciare più archivi come abbiamo detto prima - *ma come la calcoliamo?* Semplicemente utilizziamo l'unico archivio che non ha sottocopertura, ovvero il territorio per mezzo del campionamento diretto, ad esempio dividiamo l'Italia e intervistando confrontiamo se le unità sono contenute nel nostro archivio. Se non sono contenute, vuol dire che

appartengono alla nostra sottocopertura. Quindi facciamo un'indagine su base territoriale in cui l'unico obiettivo è vedere se l'unità c'è o non c'è nell'archivio, bisogna verificare in pratica l'esistenza dell'unità statistica. Una seconda possibilità riguarda metodi che sono detti di *"cattura e ricattura"* e sono utilizzati di solito per la popolazione. Questo metodo nasce dalla stima delle popolazioni animali, perché non esiste un archivio, ad esempio si potrebbe usare per sapere quanti pesci ci sono in un lago, non abbiamo l'archivio dei pesci per nome e cognome cui poter chiedere cose. C'è un sistema che si basa in definitiva su un'idea molto stupida, ad esempio selezioniamo una zona di un lago, prendiamo i pesci con una rete e li marchiamo in qualche modo, in modo che se ricapitano nella rete, possiamo saperlo. Buttiamo la rete una prima volta, marchiamo i pesci una prima volta e li ributtiamo in acqua; lanciamo la rete una seconda volta e vediamo quanti pesci marcati abbiamo ripescato. A questo punto è intuitivo capire che la stima della numerosità totale della popolazione si otterrà giocando *sulle probabilità d'inclusione doppie dei due campioni*. Numericamente se la popolazione di pesci è *10.000* e i nostri campioni sono formati da *10* unità, certamente la probabilità di riprendere gli stessi sarà molto bassa e con molta probabilità i pesci saranno diversi. Se poi riprendiamo i *10* pesci marcati nel primo stadio del campionamento certamente la numerosità totale non si allontanerà molto da *10* pesci. Quindi questa cosa che nasce da popolazioni totalmente ignote si è dimostrata funzionare benino anche per popolazioni note. Ad esempio nel caso degli individui dobbiamo estrarre a caso persone, naturalmente non sull'archivio ma sul campo per poi confrontarla con la numerosità dell'archivio e capiamo quanta sottocopertura è presente nel nostro archivio.

A scopo goliardico - come si farà mai la stima della numerosità di una popolazione di elefanti? Usiamo l'unica ausiliaria correlata praticamente a uno con il numero di animali, ovvero contiamo i cumuli di escrementi che si individuano sul territorio perfettamente riconoscibili dai biologi.