manuale di tecniche di indagine

6 - il sistema di controllo della qualità dei dati

istat
istituto nazionale
di statistica

note e relazioni
anno 1989 n. 1

La preparazione del Fascicolo e il coordinamento redazionale dei testi sono stati curati da Mauro Masselli.
Autore:
— dei Capitoli: 1, 2, 3, 5, 6, 8, Mauro Masselli
4 Fernanda Panizon
7 Marina Signore
— dell'Appendice 2 del Capitolo 2
Giovanna D'Angiolini
— dell'Appendice 1 del Capitolo 5
Domenico Sabatini

Editing di
Mario Nanni e Claudio Antonio Pajer

Supplemento all'Annuario Statistico Italiano

ISSN: 0035-9856

abete grafica s.p.a. - Roma - Contratto n. 14762 del 6-8-1988 - copie 3.000

INDICE

CAPITOLO 3. LA RILEVAZIONE SUL CAMPO

CAPITOLO 4. LA REGISTRAZIONE

CAPITOLO 5. LA REVISIONE

PRESENTAZIONE

Il *Manuale di tecniche di indagine* la cui preparazione è stata curata dal Reparto Studi dell'Istituto, si configura come guida per la razionalizzazione delle operazioni di rilevazione ed è stato pure concepito quale strumento didattico da utilizzare ai fini della formazione dei funzionari dell'Istat. Poiché nell'effettuazione di indagini statistiche sono impegnati molti altri organismi pubblici e privati, si ritiene che esso possa costituire uno strumento utile anche per l'attività di questi organismi, in particolare di quelli che hanno un qualche ruolo nel sistema informativo socio-economico del Paese.

Il *Manuale* prende in esame i vari segmenti del *ciclo produttivo* nei quali si sviluppa normalmente ogni indagine statistica cogliendo aspetti che vanno dalla costruzione del disegno campionario al controllo della qualità dei dati, dall'analisi delle caratteristiche delle varie tecniche di indagine alla definizione di criteri standardizzati per la presentazione dei risultati. Pensato inizialmente per le indagini condotte con il metodo del campione, in particolare per quelle sulle famiglie, nella sua definitiva articolazione esso detta norme valide per fasi di lavoro riscontrabili nelle rilevazioni totali ed allarga pertanto il suo campo di applicazione che finisce per comprendere le generalità delle indagini.

La sua impostazione riflette il desiderio di colmare il divario fra il *libro di testo* ed il *manuale operativo*. Se da un lato infatti non si rinuncia al rigore della formalizzazione e si introducono spunti di innovazione sul piano metodologico, dall'altro si tengono ben presenti le esigenze del lavoro sul campo e risulta quindi ampio lo spazio riservato alle esemplificazioni.

Il *Manuale* consta dei seguenti fascicoli:

1. **Pianificazione della produzione di dati**
2. **Il questionario: progettazione, redazione, verifica**
3. **Tecniche di somministrazione del questionario**
4. **Techine di campionamento: teoria e pratica**
5. **Tecniche di stima della varianza campionaria**
6. **Il sistema di controllo della qualità dei dati**
7. **Le rappresentazioni grafiche di dati statistici**

In ogni caso va precisato che il *Manuale* non è da considerarsi completato in quanto è previsto che ai fascicoli programmati se ne aggiungano altri mano a mano che l'attività di ricerca avrà portato a termine l'esplorazione di aspetti per ora solo individuati.

CAPITOLO 1 - LA QUALITÀ DEI DATI ED IL SISTEMA DI CONTROLLO DELL'INDAGINE

1. Introduzione

Il termine *qualità dei dati* e quello ad esso speculare di *errore* non vengono utilizzati in letteratura in modo univoco; essi, infatti, assumono significati diversi in contesti ed autori differenti.

Le varie interpretazioni di tali concetti si riflettono sulla scelta delle metodologie e sulla delimitazione del campo d'intervento; è opportuno, pertanto, specificare in via preliminare i significati ad essi attribuiti nel presente volume ed illustrare le implicazioni che ne derivano.

Nel Manuale, l'indagine statistica viene considerata un processo produttivo che ha come obiettivo la produzione di *informazione* statistica; alle interrelazioni tra le operazioni del processo corrispondono quelle tra gli errori da esse generati.

La qualità dell'informazione è quindi garantita dal livello di controllo del processo, realizzato nelle diverse fasi e nei differenti momenti in cui questo è suddivisibile.

L'unitarietà del processo produttivo implica necessariamente una visione unitaria del *controllo*; per tale ragione si è tentato di definire la struttura di un «sistema» di controllo, capace di integrare metodologie e tecniche diverse.

Il processo di produzione e le tecniche di controllo in esso incorporate, generano un flusso di informazioni che opportunamente selezionate ed integrate costituiscono un *archivio* di indicatori di qualità; tale archivio rappresenta la base informativa del sistema di controllo dell'indagine.

2. La qualità dell'informazione statistica

Nel concetto di *qualità dei dati*, entrambi i termini, laddove non siano chiaramente definiti, possono generare ambiguità; che cosa è, infatti, la «qualità» e cosa sono i «dati» in una indagine statistica?

Il significato del termine *dati statistici* varia a seconda del contesto; a volte vengono considerate come «dati» le singole variabili di rilevazione, altre volte l'insieme delle informazioni elementari attribuibili all'unità statistica, altre volte ancora i risultati delle loro aggregazioni.

Nel Manuale verranno considerati, quale prodotto finale di una indagine statistica, tre livelli di informazione:

I) i *microdati*, ovvero i dati rilevati sulla singola unità;
II) i *macrodati*, ovvero il risultato di una qualsiasi funzione dei dati elementari;

III) i *metadati*, ovvero le informazioni di carattere qualitativo e/o quantitativo riguardanti le diverse operazioni effettuate.

Il loro complesso costituisce l'informazione statistica derivante da una rilevazione; ed è per tale ragione (e per evitare qualsiasi ambiguità) che è preferibile fare riferimento alla *qualità dell'informazione*, piuttosto che alla *qualità dei dati*.

In questa ottica, è necessario assumere una definizione di qualità che si adatti a ciascuno dei suddetti livelli dell'informazione prodotta.

Nell'accezione più ampia, si definisce *qualità di un prodotto* l'adeguatezza del medesimo all'uso per il quale è stato realizzato, ovvero la capacità di un prodotto di soddisfare le proprietà garantite dal produttore (O. Arkhipoff, 1986).

In un processo di produzione manifatturiera, le proprietà garantite implicitamente od esplicitamente dal produttore possono essere suddivise in due insiemi:

a) *garanzie di progettazione* ovvero i requisiti del bene prodotto (ad esempio, forma, dimensione, potenza e durata media di vita di una batteria elettrica);
b) *garanzie di tolleranza*, ovvero i limiti entro cui determinati requisiti possono variare (ad esempio, la durata minima di vita garantita per una batteria elettrica).

Analogamente, possiamo definire e specificare le proprietà di qualità di una indagine statistica, in riferimento sia alle proprietà complessive dell'indagine, sia all'accuratezza dei risultati forniti. In particolare, assumeremo che le proprietà di progettazione sono quelle che si riferiscono alla capacità dell'indagine di soddisfare la domanda proveniente dall'utenza, mentre quelle di tolleranza riguardano il processo di misurazione del fenomeno in studio.

Cosicché possiamo identificare come garanzie di progettazione:

a) la *tempestività*,
b) la *rilevanza teorica*,
c) la *rilevanza effettiva*,
d) la *trasparenza*,

e quali garanzia di tolleranza

e) la *precisione campionaria*,
f) la *precisione non-campionaria*.

La rilevanza teorica denota il raggiungimento degli obiettivi prefissati, ovvero l'adeguatezza dell'informazione prodotta alle necessità informative, mentre la rilevanza effettiva fa riferimento a quanto dell'informazione prodotta viene effettivamente utilizzato; essa dipende dalle modalità di elaborazione e di diffusione dei dati. La trasparenza indica la possibilità, per l'utente, di accedere a tutte le informazioni relative agli strumenti d'indagine utilizzati (definizioni, classificazioni, rete di rilevazione, questionario, piano ed errori di campionamento, indicatori di qualità etc.), necessarie ad un uso corretto dei dati. La tempestività si riferisce al periodo di tempo che intercorre tra la nascita delle esigenze informative e la disponibilità dei risultati; minore è tale lasso di tempo, maggiore è la validità e l'utilità delle informazioni prodotte.

La tolleranza riguarda la precisione dei risultati; la precisione può essere definita in termini di *distanza* tra il valore vero e la stima ottenuta. Tale differenza può essere dovuta

I) all'uso della tecnica campionaria, ovvero al fatto che le stime sono calcolate solo su una parte delle unità costituenti l'universo indagato, oppure
II) alle discrepanze, che si verificano nella pratica, tra l'indagine *ideale* e quella *reale*, cioè tra le operazioni programmate e quelle realizzate.

Si possono, quindi, caratterizzare le garanzie di tolleranza in termini di precisione campionaria e di precisione non campionaria; alcuni autori definiscono *precisione* la prima ed *accuratezza* la seconda.

I vari aspetti della qualità, pur logicamente distinti, sono di fatto interdipendenti; ad esempio, controlli minuziosi sulla rilevazione e sulla produzione di dati incidono sulla tempestività dell'informazione. Di fatto, gli aspetti tecnici della rilevazione (le garanzie di tolleranza) sono, in buona misura, subordinati alla politica dell'informazione (le garanzie di progettazione) ed all'organizzazione del lavoro stabilite dal produttore di dati statistici.

Il presente volume si limiterà a trattare i primi aspetti, ed in particolare quelli connessi alla prevenzione, misura e correzione degli errori non campionari, lasciando ad altre pubblicazioni l'approfondimento dei temi connessi ai secondi.

Misurare la qualità complessiva di una indagine non è un compito agevole. Teoricamente, la qualità può essere definitiva mediante un *vettore di garanzie ex ante* G = (a....f), cui è associato, ex post, un *vettore delle realizzazioni* G' = (a'....f'), e da un *vettore di qualità* M = m(G-G') che sintetizza le differenze riscontrate. In pratica, tuttavia, non è possibile quantificare le componenti

dei suddetti vettori, cosicché la valutazione si basa su un insieme di indicatori, quantitativi e qualitativi, ciascuno dei quali riferito ad un solo aspetto della qualità.

Inoltre, il richiamo a concetti quali *produttore, prodotto* ed *utilizzatore*, nella definizione di qualità, fà perdere alla stessa ogni possibile carattere di assolutezza, evidenziandone, al contrario, gli elementi di relatività: le *garanzie* non vengono determinate da standard teorici, bensì sono fissate in funzione dei costi/benefici derivanti dall'informazione prodotta da una determinata indagine.

3. L'errore totale

Nel caso delle garanzie di tolleranza, la qualità assume il significato di precisione che può essere espressa come funzione inversa dell'errore statistico; tanto minore è l'errore, tanto maggiore è la precisione dei risultati ottenuti.

Poiché l'errore è definito come differenza tra valore osservato e valore vero, il concetto di precisione si fonda sull'esistenza di quest'ultimo; il valore vero di una variabile può essere sempre postulato, ma il significato che gli si attribuisce determina l'estensione del campo degli errori non campionari.

Per talune variabili, infatti, (Hansen Hurwitz & Madow, 1953) è possibile definire precisamente il valore vero (ad esempio il sesso di una persona), mentre per altre lo si può individuare in relazione agli obiettivi dell'indagine (ad esempio la riduzione ad una scala discreta di misurazioni di una variabile continua); quando non si verificano tali situazioni (basti pensare a variabili attitudinali o di opinione), è ancora possibile definire il valore vero, ma solo come risultante del complesso delle operazioni necessarie all'effettuazione dell'indagine. Queste ultime, le *condizioni generali* di svolgimento dell'indagine, riguardano tanto l'eventuale disegno campionario (criteri di selezione delle unità campione e stimatori utilizzati) che il trattamento delle informazioni rilevate (definizioni, classificazioni, norme di rilevazione, di codifica, di revisione e di elaborazione).

Una definizione puramente operativa, oltre a rendere difficile l'attribuzione ad una variabile di un unico significato (questi, infatti, varierebbe a seconda delle condizioni generali) esclude dall'analisi gran parte degli errori, fino al caso limite in cui il valore osservato coincide con quello vero solo per effetto della definizione.

Al contrario, considerare il valore vero indipendente dalle condizioni generali, può portare ad estendere oltre misura, rispetto agli obiettivi, il concetto di errore; ad esempio, dovremmo considerare errato un valore discreto, relativo ad una variabile continua, anche se l'approssimazione è adeguata per gli scopi dell'indagine.

Distorsioni ed errori variabili

Gli errori sono usualmente classificati in due categorie: *distorsioni* ed *errori variabili*. Per poterli caratterizzare, e quindi formalizzare in un modello, si ipotizza che l'indagine sia ripetibile sotto le medesime condizioni generali; in questo caso, si assume che gli errori variabili sono distribuiti casualmente, con media nulla, e variano in ciascuna delle ipotetiche ripetizioni dell'indagine. Le distorsioni, invece, sono il risultato di fattori sistematici, dipendono dalle condizioni generali, sono costanti in tutte le ripetizioni ed hanno uno specifico «segno» rispetto al valore vero; distorsioni di tipo diverso possono presentare segni diversi e si sommano algebricamente.

L'errore campionario e non campionario

Gli errori possono verificarsi sia nei microdati, ovvero in una o più delle variabili afferenti alla singola unità statistica, sia nel calcolo di loro aggregazioni, ovvero nelle stime dei parametri della popolazione di studio (ad esempio i consumi medi nell'indagine sui bilanci di famiglia od il totale della popolazione nel censimento).

Nel primo caso la discrepanza tra il valore della generica variabile y_i osservata sulla i-esima unità ed il valore vero Y_i realmente posseduto dalla medesima, è imputabile al complesso delle operazioni di rilevazionie e trattamento dei dati (questionario, intervista, codifica, registrazione ed elaborazione dei dati); tali errori vengono definiti non-campionari o di misura (in senso lato).

Essi si ripercuotono nelle stime (i macrodati), mediante le operazioni di aggregazione dei microdati, necessarie al loro calcolo, ovvero mediante la funzione di sintesi $f(y_1, y_2, \ldots y_i \ldots y_n)$ delle informazioni elementari (media, frequenze relative ed assolute etc.). L'operazione di aggregazione viene effettuata sulle n unità rilevate, il cui numero può coincidere (rilevazioni totali) o meno (rilevazioni campionarie) con quello, N, della popolazione. In presenza di errori, la stima risulta diversa dal valore che si sarebbe ottenuto dai valori veri delle medesime unità, $f(Y_1, Y_2, \ldots Y_i \ldots Y_n)$.

Tuttavia, questa non è l'unica ragione per la quale la stima differisce dal parametro di interesse, $g(Y_1, Y_2, \ldots Y_i \ldots Y_N)$, calcolato sulle N unità della popolazione, dove quest'ultimo è stato indicato con il simbolo g(.) per evidenziare che la funzione di sintesi dei dati osservati, f(.), può anche non coincidere con quella del parametro (ad esempio lo stimatore rapporto utilizzato per la stima totale).

Tale differenza, infatti, può essere generata:

I) dall'*errore variabile di campionamento*, dovuto all'utilizzo della tecnica campionaria, per cui la stima differisce per effetto del «caso» dal valore vero nella popolazione;

II) dalla *distorsione dello stimatore*, ovvero dall'adozione di uno stimatore *non corretto*, la cui media, nell'universo dei campioni, non coincide con il parametro da stimare; formalmente $E[f(.)] \neq g(.)$.

Nelle indagini, quindi, l'errore, ovvero la discrepanza tra la stima ottenuta in una particolare rilevazione ed il valore vero nella popolazione, può essere concettualmente suddiviso in due parti: una imputabile alle diverse operazioni necessarie per l'effettuazione dell'indagine (errore non campionario) e l'altra derivante dal disegno di campionamento (errore campionario e distorsione dello stimatore):

[errore statistico] = [errore non campionario] +
[errore variabile campionario] +
[distorsione dello stimatore]

$$f(y_1\ y_2\ \dots\ y_n) - g(Y_1\ Y_2\ \dots\ Y_N) = f(y_1\ y_2\ \dots\ y_n) - f(Y_1\ Y_2\ \dots\ Y_n) + f(Y_1\ Y_2\ \dots\ Y_n) - E\left[f(Y_1\ Y_2\ \dots\ Y_n)\right] + E\left[f(Y_1\ Y_2\ \dots\ Y_n)\right] - g(Y_1\ Y_2\ \dots\ Y_N)$$

Per semplificare l'esposizione, possiamo riscrivere tale relazione in termini compatti come:

$$y - Y = \left[y - y^*\right] + \left[y^* - E(y^*)\right] + \left[E(y^*) - Y\right] \qquad (1.1)$$

dove l'operatore E è riferito all'universo dei campioni e

$$y = f(y_1\ y_2\ \dots\ y_n); \qquad y^* = f(Y_1\ Y_2\ \dots\ Y_n); \qquad Y = g(Y_i\ Y_2\ \dots\ Y_N)$$

La distinzione tra l'errore campionario e quello non campionario definisce, per contrapposizione, la loro sostanziale diversità: il primo dipende dalla variabilità del fenomeno in studio (ovvero dalla «realtà» esaminata), dal disegno di campionamento e dagli stimatori utilizzati, mentre il secondo è funzione degli aspetti organizzativi della rilevazione, del comportamento di una pluralità di soggetti e del contesto socio-culturale in cui si colloca l'indagine.

La misura dell'errore totale

La distinzione tra errore campionario e non, unitamente a quella tra distorsione ed errore variabile, porta a una doppia classificazione dell'errore statistico: errore variabile e distorsione campionaria, errore variabile e distorsione non campionaria.

I concetti di distorsione e di errore variabile non campionari possono essere intuitivamente illustrati ricorrendo all'esempio della rilevazione della variabile «reddito»; immaginando di intervistare più volte la medesima persona, od il medesimo campione di persone, è realistico pensare di ottenere di volta in volta risposte diverse (errore variabile), ma tutte sottovalutate rispetto al valore del reddito realmente percepito (distorsione).

Quanto sopra può essere tradotto in termini formali; ipotizzando la ripetizione dell'indagine sotto le medesime condizioni generali, è possibile esprime la differenza tra lo stimatore calcolato sui valori osservati e lo stesso calcolato sui valori veri come:

$$y - y^* = \left[y - E\,(y / c)\right] + \left[E\,(y / c) - y^*\right] \qquad (1.2)$$

$$= v + b$$

dove E (y/c) indica il valore medio, calcolato su tutte le ipotetiche ripetizioni dell'indagine, della stima ottenuta dai dati osservati, fissate le medesime unità campione.

La prima parte della (1.2) esprime l'errore variabile da indagine ad indagine, mentre la seconda esprime la distorsione, costante al variare delle ripetizioni dell'indagine.

Date le definizioni sopra riportate, possiamo assumere che v sia una variabile aleatoria con media zero e varianza data, pari a VNC, mentre b assumerà un valore costante B, nell'universo dei campioni:

$$E(v) = 0; \qquad Var(v) = E\,[v - E(v)]^2 = E\,(v^2) = VNC$$

$$E(b) = B$$

Sostituendo la (1.2) nella (1.1) si ottiene:

$$y - Y = [v + b] + [y^* - E(y^*)] + [E(y^*) - Y] \qquad (1.3)$$

Dalla (1.3), possiamo misurare l'errore totale dello stimatore y, mediante una qualsiasi funzione della differenza (y–Y); usualmente si ricorre alla «media quadratica»:

$$\begin{aligned} MSE(y) &= E\,(y - Y)^2 \\ &= E\,[v + b + y^* - E(y^*) + E(y^*) - Y]^2 \\ &= E\,[v]^2 + [y^* - E(y^*)]^2 + [B + D]^2 + 2\,cov(v,y^*) \\ &= VNC + VC + (B + D)^2 + 2\,cov(v,y^*) \end{aligned} \qquad (1.4)$$

dove con D si è indicata la distorsione dovuta allo stimatore utilizzato.

Nella (1.4) si è quindi espresso l'errore totale in funzione degli errori variabili non campionario (VNC) e campionario (VC), della distorsione dello stimatore (D) e non campionaria (B), della covarianza tra l'errore variabile non campionario e la stima.

Se si ipotizzano diversi tipi di errore, dovuti alle diverse operazioni dell'indagine (intervista, supervisione, registrazione, revisione, elaborazione ecc.) e si considera quello dovuto al disegno di campionamento uno di tali errori, la (1.4) può essere generalizzata nella (1.5):

$$MSE(y) = \Sigma_k\, V(y_k) + (\Sigma_k B_k)^2 + 2\Sigma\, cov(.,.) \qquad (1.5)$$

La (1.5), rappresenta l'errore totale come somma di distorsioni, varianze e covarianze derivanti dalle differenti fonti; in particolare, le covarianze che vi appaiono possono essere di tipo diverso: tra le misurazioni, tra queste ed il livello delle variabili, tra le fonti di errore etc.

La (1.5) costituisce, in forma del tutto generale, una rappresentazione dell'idea corrente che gli errori generati nelle diverse fasi si sommano, si elidono e si combinano per confluire infine nei risultati finali. Essa è valida sia per indagini campionarie che censuarie; per queste ultime, infatti, pur annullandosi le componenti relative al disegno di campionamento, permangono gli effetti degli errori non campionari.

Della (1.5) può essere data una rappresentazione geometrica, Figura 1.1, scomponendo l'errore totale mediante la successiva specificazione della parte variabile e della distorsione di livelli di errore via via più analitici (Kish, 1965; Singh e Chaddurray, 1986). Per semplificare la rappresentazione, sono stati utilizzati vettori ortogonali, ipotizzando quindi covarianze nulle; volendo introdurre ipotesi differenti, sarà necessario cambiare l'angolo di incidenza dei relativi vettori.

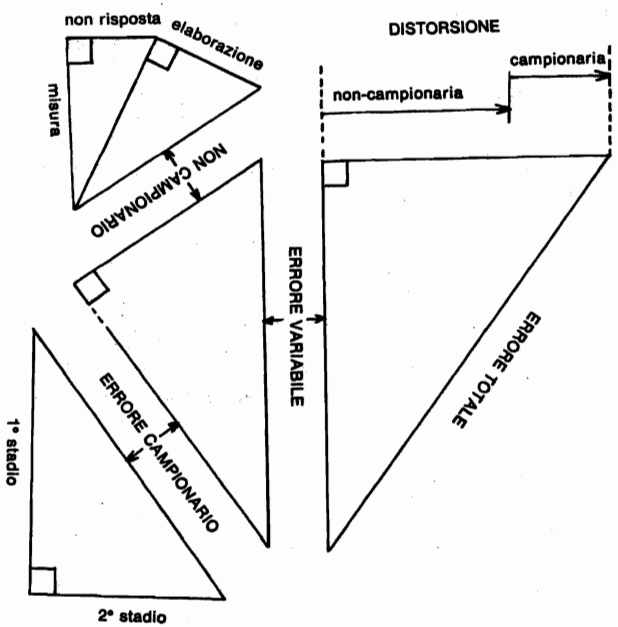


Figura 1.1 - Scomposizione dell'errore totale

In genere si può assumere che distorsioni ed errori variabili non campionari siano presenti in tutte le caratteristiche rilevate, anche se con peso diverso.

Le distorsioni non campionarie possono essere dovute (Kish, 1965) sia ai rispondenti, (ovvero valori rilevati e sistematicamente errati), sia ai non rispondenti, (ovvero le distorsioni indotte dalla mancata risposta ai quesiti o all'intera intervista); le distorsioni campionarie possono essere classificate in quelle relative all'uso di uno stimatore non corretto ma consistente ed in quelle dovute a stimatori non corretti e non consistenti. Nel primo caso la distorsione si annulla quando la dimensione del campione è sufficientemente elevata o nel caso dei censimenti (ad esempio lo stimatore rapporto), mentre, nel secondo, permane anche nel caso di indagini totali (ad esempio la mediana come stimatore della media in distribuzioni asimmetriche). In questo modo la distorsione dello stimatore dovuta alla dimensione dell'indagine, viene distinta da quella indipendente dal numero di unità rilevate.

La (1.5) ha un valore puramente descrittivo ed è stata ottenuta senza far ricorso, esplicitamente, a un modello di errore individuale.

Il riferimento ad un tale modello è, invece, necessario per esprimere l'MSE in termini analitici e quindi per ottenere gli stimatori delle varie componenti dell'errore totale; nel Capitolo 7, il modello viene sviluppato con riferimento alla stima della media (e quindi della percentuale e del totale).

Attraverso la specificazione del modello di errore è possibile adattare l'analisi e la stima degli errori, al livello di complessità desiderato. Ad esempio, se consideriamo un modello in cui l'errore di misura è dovuto solamente all'effetto rilevatore-rispondente (ipotesi semplificatrice, ma realistica in quanto le esperienze effettuate hanno dimostrato che tale fonte è causa di gran parte degli errori non campionari), è possibile esprimere l'MSE dello stimatore media in funzione di quattro componenti:

1) la varianza campionaria VCA/n,
2) la varianza semplice di risposta VSR/n, ovvero quella dovuta agli errori di misura dei singoli individui intervistati;
3) la varianza correlata di risposta (n−1)*VCR/n ovvero l'effetto del rilevatore;
4) la distorsione non campionaria B.

$$MSE(y) = VCA/n + VSR/n + (n-1)*VCR/n + B^2 \qquad (1.6)$$

Nella (1.6) non appare la distorsione D, dovuta al disegno campionario, poiché la media è uno stimatore corretto; la simbologia utilizzata mette in evidenza che mentre la varianza campionaria e quella semplice di risposta dipendono dall'ampiezza del campione, la componente correlata è indifferente a tale parametro, essendo praticamente uguale a 1 il rapporto (n−1)/n nelle indagini di medie/grandi dimensioni.

Ciò comporta che aumentando la dimensione del campione si riesce a ridurre l'errore di campionamento e quello semplice di misura ma non la componente correlata che rappresenta la gran parte dell'errore non campionario.

L'MSE(y), od alcune sue componenti, esplicitate in funzione delle principali fonti di errore, possono essere stimati utilizzando opportuni stimatori e tecniche di rilevazione (cfr. Capitolo 7).

4. L'indagine come processo di produzione

Una indagine statistica può essere assimilata ad un processo produttivo manifatturiero, in quanto, come questo, è costituita da un insieme di fasi ed operazioni interrelate; la produzione finale consiste nell'informazione statistica, come precedentemente definita, e la materia prima nell'informazione disponibile presso le unità di analisi. Quest'ultima può essere considerata come un *flusso produttivo* che viene trasformato nelle diverse fasi di lavorazione.

Tale flusso ed i legami logici intercorrenti tra le differenti operazioni (ad esempio tra la predisposizione del questionario e del piano di registrazione, tra questo e le procedure di compatibilità e correzione e l'elaborazione finale) definiscono la sequenza logica e temporale delle fasi.

La qualità dell'informazione prodotta dipende dal controllo che si riesce ad esercitare sulle operazioni e, quindi, considerare l'*indagine come processo produttivo*, facilita la classificazione e l'individuazione degli errori e fornisce una dimensione operativa ed organizzativa al loro controllo mediante il collegamento alle differenti fasi di lavoro.

Il processo di produzione dell'indagine può essere suddiviso a vari livelli di aggregazione e complessità; nello schema adottato si è cercato di considerare contemporaneamente gli aspetti organizzativi, di contenuto, di sequenza logica e temporale. Pertanto, si considerano come fasi dell'indagine:

- la progettazione
- la rilevazione
- la registrazione su supporto informatico
- la revisione e la codifica «centralizzate» del materiale grezzo
- l'elaborazione dei dati

- la validazione dei risultati
- la diffusione

Nella fase di progettazione, si mette a punto il «disegno dell'indagine», ovvero, sulla base delle risorse organizzative e finanziarie e delle conoscenze «a priori» del fenomeno indagato, si programmano le operazioni inerenti a tutte le successive fasi:

I) si definiscono gli scopi, i contenuti informativi, l'universo di studio, la tecnica e le unità di rilevazione, le unità di analisi e l'eventuale disegno campionario;
II) si articolano gli obiettivi nel questionario, nelle definizioni e nelle classificazioni;
III) vengono assunte tutte le decisioni riguardanti le successive fasi e si approntano i relativi piani di lavoro e di controllo;
IV) si verificano, con limitate indagini sul campo, i principali aspetti dell'indagine e si controlla la coerenza logica dei piani di lavoro relativi alle successive fasi.

Nella fase di rilevazione sono incluse tutte le operazioni che hanno per oggetto o sono effettuate dalla rete periferica: la selezione e l'istruzione dei rilevatori e dei supervisori, l'istruzione ed i contatti con gli organi periferici, la pubblicizzazione locale dell'indagine, la compilazione dei documenti aggiuntivi di rilevazione, la selezione delle unità campionarie, l'intervista, la revisione e la codifica effettuate in loco.

Il risultato di tale fase è costituito dai dati rilevati, o grezzi, generalmente presenti su supporto cartaceo (il questionario); essi, nella successiva fase di registrazione, sono trasferiti su supporto informatico e diventano quindi elaborabili.

La fase di revisione del materiale, consiste nella verifica, quantitativa e qualitativa, e nell'eventuale correzione dei dati grezzi; la sua posizione nel processo produttivo e l'estensione delle operazioni ad essa afferenti, sono strettamente connesse all'organizzazione del lavoro e alle risorse che supportano l'indagine. Si possono, pertanto, delineare due situazioni estreme. Nella prima, il materiale raccolto su supporto cartaceo viene controllato e corretto manualmente da «esperti»; in questo caso la fase di revisione precede logicamente e temporalmente quella di registrazione. Nel secondo caso tutte le operazioni di revisione e codifica sono svolte automaticamente da procedure informatiche e, quindi, la revisione segue la fase di registrazione. L'organizzazione concreta delle indagini si situa in modi diversi tra tali estremi; tuttavia poiché la tendenza è quella di una sempre maggiore compenetrazione tra lavoro di esperti e procedure informa-

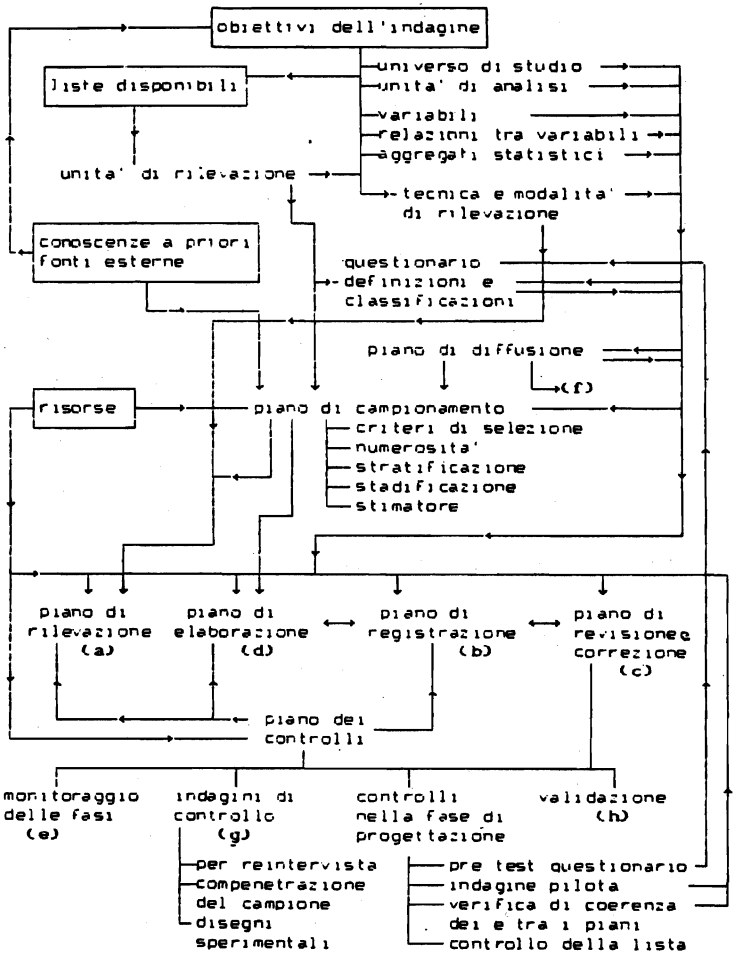
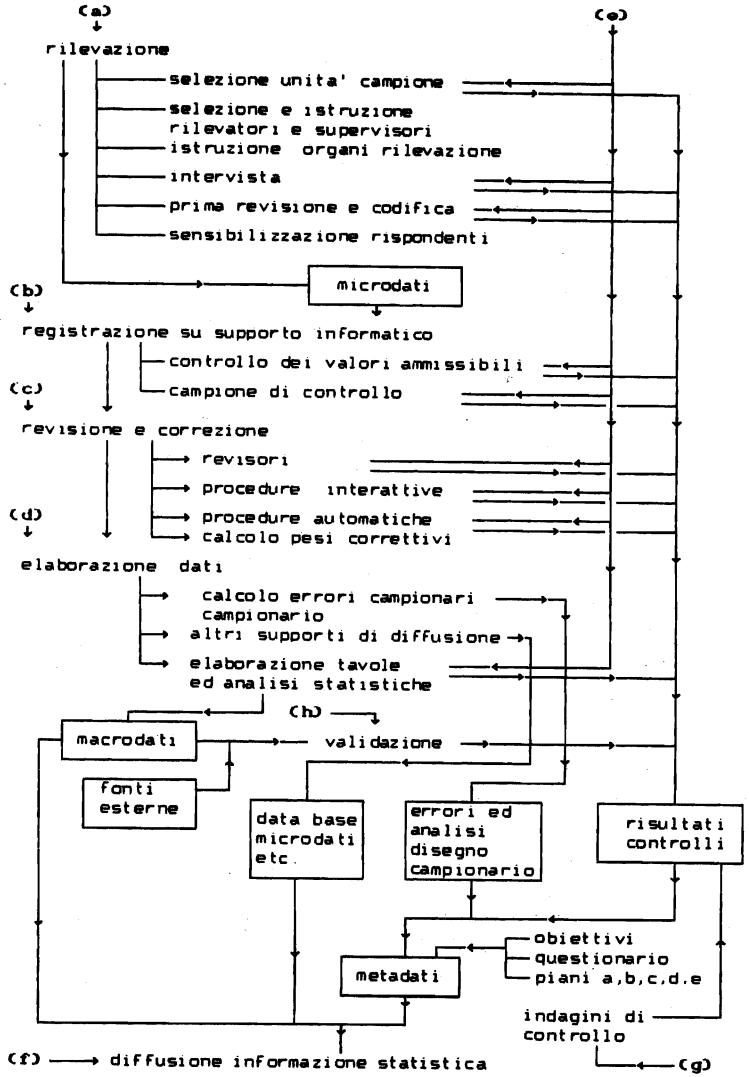


Figura 1.2 - Processo produttivo e sistema di controllo dell'indagine



segue Figura 1.2 - Processo produttivo e sistema di controllo dell'indagine

tiche, nel Manuale si farà riferimento ad una accezione lata della revisione, in cui sono incluse tutte le operazioni manuali od informatiche, di verifica e correzione dei dati. Da questo punto di vista i programmi di compatibilità e correzione rientrano nella fase di revisione, mentre a quella di elaborazione è delegato solo il compito di predisporre tavole, indicatori ed analisi statistiche dai microdati definitivi.

I macrodati, sotto forma di tavole od indicatori statistici, sono validati sulla base della loro coerenza interna, dell'analisi dell'eventuale serie storica o mediante confronto con fonti esterne.

I macrodati, i metadati (inclusi i risultati dei controlli effettuati) ed eventualmente gli stessi microdati, possono quindi essere diffusi agli utilizzatori finali.

Nello schema adottato non è stata prevista una fase di *controllo*; tale funzione, infatti, verrà considerata come un insieme organico di operazioni, inserite nelle altre fasi, che affianca il complesso dell'indagine e ne costituisce il *supervisore*.

Il processo produttivo sopra descritto, fà riferimento ad una indagine «tipo»; nelle situazioni concrete possono verificarsi dei cambiamenti nello schema utilizzato (ad esempio, nelle indagini telefoniche, la registrazione si fonde con la rilevazione dei dati), che però non ne inficiano la logica di fondo.

5. Gli errori non campionari

In ciascuna delle fasi e delle operazioni dell'indagine possono essere generati *errori non campionari*, che è possibile classificare con riferimento alla fonte dell'errore.

In realtà, date le interazioni tra operazioni e tra soggetti, l'errore è spesso dovuto alla combinazione di più fattori (ad esempio gli errori, nella fase di revisione, commessi dai revisori possono essere dovuti anche ad una insufficiente specificazione delle norme e del questionario). Tuttavia, pur essendo un modello semplificato, la classificazione che segue è sufficiente ad impostare, in termini operativi ed organizzativi, i controlli dei principali errori non campionari:

1) *fase di progettazione*
   - errori nella definizione degli obiettivi;
   - errori nella definizione del campo di rilevazione;
   - errori nella definizione delle unità di rilevazione e di analisi;
   - errori nella formulazione del questionario, delle definizioni e delle classificazioni;
   - errori nelle norme dei piani di lavoro;
   - errori nel coordinamento tra piani di lavoro;
   - errori di «rilevanza».

2) *fase di rilevazione*
   - errori dovuti ad insufficiente istruzione e assistenza alla rete di rilevazione;
   - errori nelle liste di selezione;
   - errori commessi nelle procedure di selezione delle unità campionarie;
   - errori dovuti ai rispondenti;
   - errori dovuti ai non rispondenti;
   - errori dovuti alla tecnica di indagine prescelta (nelle indagini dirette, al rilevatore ed al contesto dell'intervista);
   - errori dovuti ai supervisori;

3) *fase di registrazione*
   - errori dovuti agli operatori;

4) *fase di revisione*
   - errori dovuti ai revisori;
   - errori dovuti alle procedure informatiche;

5) *fase di elaborazione finale dei dati*
   - errori nei programmi di calcolo;
   - errori di rilevanza effettiva;

6) *fase di validazione*
   - errori di coerenza nelle tavole e negli indicatori;

7) *fase di diffusione*
   - informazioni agli utenti non rilevanti, non trasparenti e non tempestive.

Gli errori generati nella fase di progettazione, di elaborazione e di diffusione si riflettono sostanzialmente sulla «rilevanza» dell'indagine; tuttavia deficienze e discrepanze nella stesura del questionario, delle norme e dei differenti piani di lavoro influenzano le operazioni successive e quindi danno luogo ad errori di «precisione», che sono propri delle altre fasi dell'indagine.

Per le indagini condotte dall'Istat, in particolare quelle sulla popolazione, nel Prospetto 1.1 sono sintetizzate le principali operazioni con le relative fonti e tipo di errore.

Data la natura di flusso del processo produttivo, gli errori che hanno origine in una operazione si trasmettono a quelle successive sommandosi, combinandosi od elidendosi; nella Figura 1.3 è rappresentato tale procedimento per le operazioni e gli errori più rilevanti di una indagine.

Prospetto 1.1: operazioni dell'indagine, fonti e tipo di errore

operazioni	fonti	tipo di errore
scelta delle variabili, delle definizioni, delle classificazioni, delle unità	modello concettuale	rilevanza teorica
definizione del questionario	struttura lunghezza vocabolario quesiti retrospettivi proxy codifica	errori di misura
piano di diffusione		rilevanza effettiva e trasparenza
piani di lavoro		errori di misura
selezione PSU (*)	base statistica	calcolo probabilità di inclusione
selezione SSU (*)	base statistica lista supervisori	calcolo probabilità di inclusione ed errori di copertura
formazione elenchi ed assegnazioni	supervisori rilevatori	identificazione delle unità
rilevazione sul campo	supervisori rilevatori rispondenti	mancate risposte totali e parziali, incongruenze, errori di misura, effetto proxy, effetto ricordo
registrazione	operatori	errori di misura
revisione e correzione	revisori procedure automatiche e interattive	errori di misura, errori di identificazione
stime (*)	base statistica	calcolo fattori di espansione
elaborazione e validazione dei risultati	programmi	errori di calcolo rilevanza effettiva
diffusione		tempestività

(*) solo per indagini campionarie

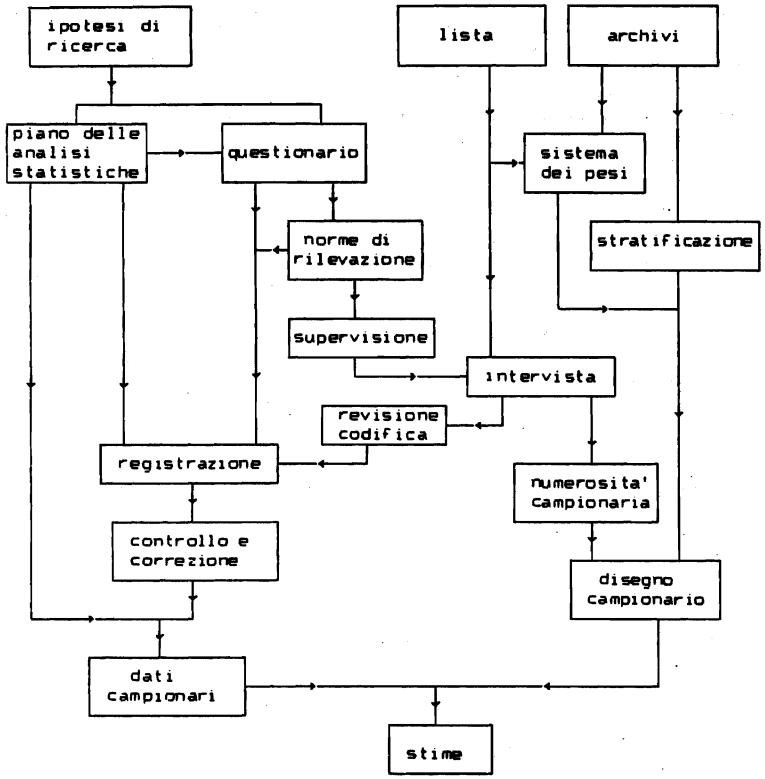


Figura 1.3 - Trasmissione degli errori tra le fasi e/o le operazioni di una indagine campionaria

Riguardo all'errore non campionario c'è, infine, da considerare una sua peculiarità; se da un lato esso costituisce un fattore di «disturbo» da rimuovere, dall'altro può essere considerato fonte di «informazione» sul complesso dell'indagine. Infatti, poiché qualsiasi rilevazione statistica è un modello a priori della realtà esaminata (imposto mediante le definizioni, il questionario, la codifica e le operazioni di correzione), l'errore non campionario contiene in sé una ambiguità: può essere sia «errore», nel senso proprio del termine, sia il rilevatore di una inadeguatezza nella formulazione del «modello implicito». Per tale ragione i risultati dei controlli possono divenire una fonte di informazione non solo sull'errore commesso, ma anche sulle relazioni tra modello concettuale utilizzato e realtà.

6. Gli effetti dell'errore non campionario sull'affidabilità delle stime

Nelle indagini campionarie su larga scala possiamo giovarci, per l'inferenza statistica sui principali parametri di studio (medi, totali, frequenze relative ed assolute), del teorema del limite centrale: esso ci assicura che la loro distribuzione, per n sufficientemente ampio, è approssimata dalla distribuzione *normale*. In base a tale teorema, possiamo calcolare gli intervalli di confidenza della stima ottenuta in funzione di prefissate probabilità; i limiti dell'intervallo vengono determinati nell'ipotesi di stimatori non distorti ed affetti dal solo errore campionario.

La presenza di errori non campionari, inducendo nei risultati delle distorsioni e/o un aumento della variabilità, conduce ad una erronea valutazione del livello di fiducia attribuito ad un determinato intervallo.

Presenza della sola distorsione

Supponiamo infatti che lo stimatore utilizzato sia distorto; allora lo stimatore $\hat{\mu}$ avrà media pari a $E(\hat{\mu})$, diversa dal *valore vero* μ.

La differenza $B = \mu - E(\hat{\mu})$ rappresenta la distorsione dello stimatore.

Se si ignora l'esistenza e/o l'entità della distorsione, si calcolerà l'intervallo di confidenza, al livello di fiducia, stabilito, come se fosse centrato su μ, mentre esso è, in realtà, centrato su $E(\hat{\mu})$. In termini formali, si farà, in maniera non corretta, la seguente asserzione

$$Pr\left[\hat{\mu} - t_{\alpha/2}\,\sigma_{\hat{\mu}} \leq \mu \leq \hat{\mu} + t_{\alpha/2}\,\sigma_{\hat{\mu}}\right] = 1 - \alpha$$

mentre l'asserzione corretta sarebbe:

$$Pr\left[\hat{\mu} - t_{\alpha/2}\,\sigma_{\hat{\mu}} \leq E(\hat{\mu}) \leq \hat{\mu} + t_{\alpha/2}\,\sigma_{\hat{\mu}}\right] = 1 - \alpha$$

Nella prima relazione, α risulta sovrastimata e deve essere sostituita da $\alpha' = \beta + \gamma$, dove β e γ sono le probabilità corrispondenti all'intervallo centrato su μ, (cfr. figura 1.4.).

Esprimendo α' in funzione di α, β e γ

$$\alpha' = \alpha + \left(\beta - \frac{\alpha}{2}\right) - \left(\frac{\alpha}{2} - \gamma\right) \quad ,$$

per la simmetria e l'unimodalità della funzione normale, si ha

$$\left(\beta - \frac{\alpha}{2}\right) > \left(\frac{\alpha}{2} - \gamma\right) \quad ;$$

ne risulta che $\alpha' > \alpha$ e quindi

$$1 - \alpha' < \alpha$$

ovvero che la probabilità effettiva dell'intervallo di confidenza è minore di quella presunta.

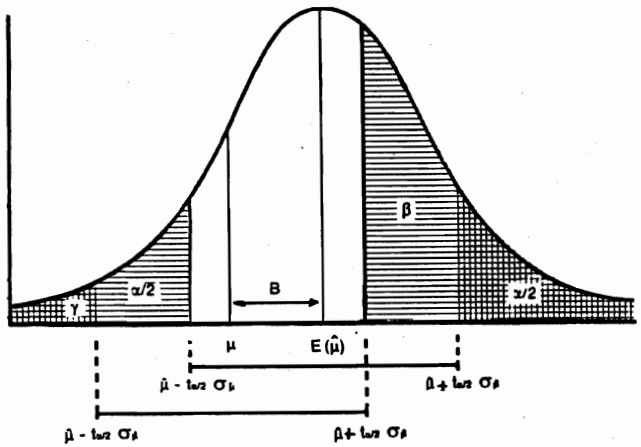


Figura 1.4

La determinazione analitica del nuovo livello di confidenza α', in funzione della distorsione B, si ottiene integrando la funzione di densità dello stimatore $\hat{\mu}$, separatamente dal due estremi dell'intervallo centrato su μ:

$$\beta = \frac{1}{\sigma_{\hat{\mu}}\,(2\pi)^{1/2}} \int_{\mu + t_{\alpha/2}\,\sigma_{\hat{\mu}}}^{\infty} \exp\left[-(\hat{\mu} - E(\hat{\mu}))^2 / 2\sigma^2_{\hat{\mu}}\right] d\,\hat{\mu}$$

Passando alla standardizzata, ponendo

$$z = (\hat{\mu} - E(\hat{\mu})) / \sigma_{\hat{\mu}}$$

si ottiene:

$$\beta = \frac{1}{(2\pi)^{1/2}} \int_{t_{\alpha/2} - B/\sigma_{\hat{\mu}}}^{\infty} \exp\left[-(z)^2/2\right] d\,z \qquad (1.7)$$

Con il medesimo procedimento si trova per γ:

$$\gamma = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{t_{\alpha/2} - B/\sigma_{\hat{\mu}}} \exp\left[-(z)^2/2\right] d\,z \qquad (1.8)$$

Mediante le (1.7) e (1.8), siamo in grado di calcolare α' in funzione del rapporto tra distorsione e s.q.m. dello stimatore (cfr. Tavola 1.1.).

Tavola 1.1 - Valori di α' in corrispondenza di $\alpha = 0.05$, in funzione del rapporto $B/\sigma_{\hat{\mu}}$

$B/\sigma_{\hat{\mu}}$	β	γ	α'	$1-\alpha'$
0.02	0.0238	0.0262	0.0500	0.9500
0.10	0.0197	0.0314	0.0511	0.9489
0.60	0.0052	0.0869	0.0921	0.9079
1.00	0.0015	0.1685	0.1700	0.8300
1.50	0.0003	0.3228	0.3231	0.6769

Effetti dell'aumento della varianza

Se si ipotizza che gli errori non campionari determinano solo un aumento della variabilità, ma non la distorsione dello stimatore, si può, con analogo procedimento, determinare α in funzione del rapporto tra le due varianze.

Sia ${}_m\sigma^2_{\hat{\mu}}$ la varianza di $\hat{\mu}$ affetta da errore di misura e:

$$\sigma^2_{\hat{\mu}} \,/\, {}_m\sigma^2_{\hat{\mu}} = k < 1 \quad \text{ovvero} \quad \sigma_{\hat{\mu}} = \sqrt{k}\,{}_m\sigma_{\hat{\mu}}$$

Si avrà quindi

$$\beta = \frac{1}{{}_m\sigma_{\hat{\mu}}\,(2\pi)^{1/2}} \int_{\mu + t_{\alpha/2}\,\sigma_{\hat{\mu}}}^{\infty} \exp\left[-(\hat{\mu}-\mu)^2 / 2\,{}_m\sigma^2_{\hat{\mu}}\right] d\,\hat{\mu}$$

$$\beta = \frac{1}{(2\pi)^{1/2}} \int_{\sqrt{k}\,t_{\alpha/2}}^{\infty} \exp\left[-(z)^2/2\right] d\,z \qquad (1.9)$$

ed analogamente per γ

$$\gamma = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{-\sqrt{k}\,t_{\alpha/2}} \exp\left[-(z)^2/2\right] d\,z \qquad (1.10)$$

La figura 1.5 illustra graficamente le considerazioni precedenti.

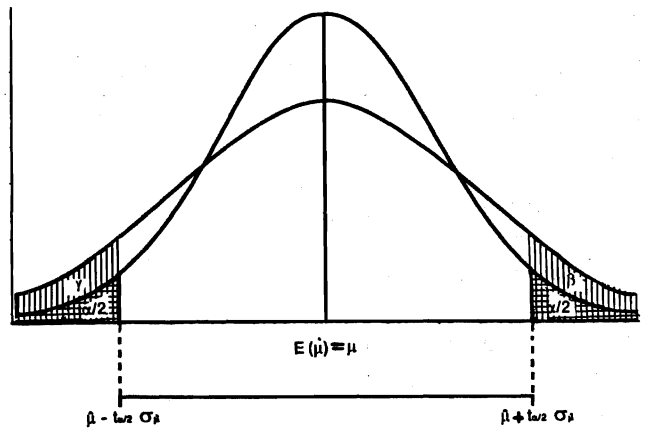


Figura 1.5

Nella Tavola 1.2, si è espresso il livello di confedenza α' in funzione del rapporto k, in corrispondenza ad un prefissato α pari a 0.05, calcolato mediante la (1.9) e la (1.10).

Tavola 1.2 - Valori di α', in corrispondenza di $\alpha = 0.05$, in funzione del rapporto k

k	β,γ	α'	$1-\alpha'$
0.10	0.2324	0.4648	0.5352
0.25	0.3365	0.6730	0.3270
0.50	0.4177	0.8354	0.1646
0.75	0.4554	0.9108	0.0892
1.00	0.4750	0.9500	0.0500

Effetti congiunti

Quando gli errori non campionari causano sia una distorsione, sia un aumento della varianza dello stimatore, le probabilità β e γ potranno essere ottenute mediante le:

$$\beta = \frac{1}{(2\pi)^{1/2}} \int_{B/\sqrt{k}\,\sigma_{\hat{\mu}} + \sqrt{k}\,t_{\alpha/2}}^{\infty} \exp\left[-(z)^2/2\right] d\,z$$

$$\gamma = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{-B/\sqrt{k}\,\sigma_{\hat{\mu}} - \sqrt{k}\,t_{\alpha/2}} \exp\left[-(z)^2/2\right] d\,z$$

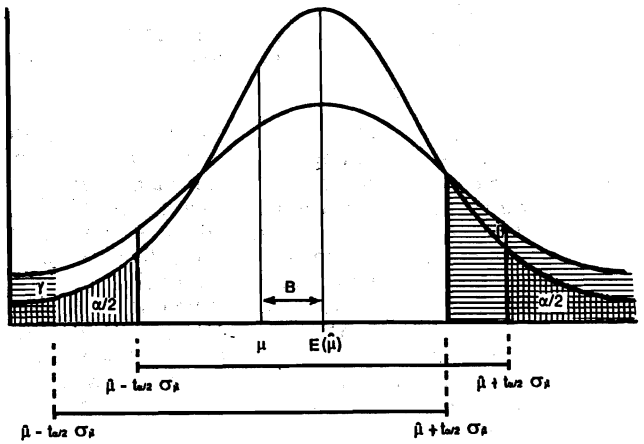


Figura 1.6

La figura (1.6) illustra, graficamente, le relazioni tra la probabilità presunta e quella corretta dell'asserzione sull'intervallo di confidenza.

7. Il sistema di controllo

L'analogia tra indagine e processo di produzione manifatturiero non comporta, meccanicamente, l'applicazione delle tecniche del controllo statistico di qualità, utilizzate in quest'ultimo ambito, al sistema di controllo dell'indagine. Tali metodi, infatti, si basano sulla riproducibilità e sulla serialità del singolo «pezzo», laddove, in una rilevazione statistica in campo economico o sociale, l'informazione raccolta è, in realtà, unica e non riproducibile sotto le medesime condizioni. L'individuazione dell'errore nella produzione manifatturiera è immediata, essendo predeterminato il prodotto tipo, e la variabilità riscontrata è attribuibile unicamente al processo; scopo del controllo è la riduzione, entro termini accettabili, di detta variabilità. Al contrario, in un'indagine statistica, il valore vero non è conosciuto a priori e la variabilità è insita nei fenomeni indagati; ciò complica l'individuazione e l'eliminazione dell'errore.

Il controllo statistico di qualità viene perciò definito diversamente nei due casi, (U.S. Department of Commerce, 1978):

I) metodo per controllare la qualità di un prodotto manifatturiero di larga scala; esso si prefigge di determinare ed eliminare le variazioni sistematiche o di ridurle ad un livello accettabile, dovuto al caso. Quando questo si verifica, il processo risulta sotto controllo.
II) osservazione e procedura utilizzata in ciascuna operazione di una indagine per prevenire o ridurre l'effetto dell'errore non campionario.

Secondo la definizione sopra riportata, l'oggetto del controllo statistico di qualità di una indagine è limitato all'errore non campionario; del resto la sostanziale diversità dei due tipi di errore si riflette nei diversi strumenti di controllo e, persino, in una differente possibilità di attuarlo. Il produttore dell'informazione statistica, infatti, ha la possibilità di controllare l'errore campionario e di minimizzarlo, date le risorse disponibili, mediante la scelta della numerosità del campione, dello stimatore, delle tecniche di stratificazione e stadificazione. Il controllo e la riduzione dell'errore non campionario, che dipende, invece, dall'operato di una pluralità di soggetti, si presenta complesso e pone delicati problemi di competenze, di responsabilità e di potere di intervento dell'Istituto.

Date le interrelazioni tra le operazioni dell'indagine e, di conseguenza, tra gli errori da esse generati, è opportuno inquadrare i singoli controlli in un insieme coerente ed organico; è conveniente, perciò, considerare un «sistema» di controlli, che affianchi il flusso produttivo e ne attui il monitoraggio. Le informazioni derivanti da quest'ultimo, potranno essere utilizzate per migliorare il sistema di produzione dell'indagine (se ripetuta), per la programmazione di indagini simili e per fornire valutazioni qualitative e quantitative sull'attendibilità dei risultati finali.

Rispetto al momento in cui avviene ed ai fini per cui si effettua, il controllo può essere considerato *preventivo*, se precede la rilevazione sul campo ed ha lo scopo di verificare e migliorare la programmazione dell'indagine; *contemporaneo*, se attuato durante le operazioni di rilevazione, revisione ed elaborazione con l'obiettivo di individuare e correggere l'errore; *successivo*, se finalizzato alla predisposizione del profilo dell'errore od alle analisi delle singole fasi.

Così definito, il «sistema dei controlli dell'indagine» ha per oggetto l'errore non campionario e per obiettivi:

I) la prevenzione dell'errore;
II) la correzione dell'errore;
III) il monitoraggio della fasi del processo di formazione del dato e la stima dell'errore totale.

Elementi costitutivi del sistema sono l'individuazione delle cause d'errore e la definizione dei livelli di controllo, la predisposizione delle fonti e l'organizzazione dell'informazione, i metodi di analisi e di correzione.

I «livelli» costituiscono un sottoinsieme delle fonti di errore, ovvero le operazioni e gli operatori (ad es. il rilevatore, il comune, la registrazione etc.) che sono effettivamente sottoposti a verifica.

In questo contesto, si precisa la distinzione tra livelli e fasi: la fase rappresenta, nel flusso logico-temporale della produzione, il punto in cui è possibile o conveniente effettuare il controllo (si potrebbe dire il «quando»), mentre il livello è l'operazione sulla quale il controllo viene esercitato (il «dove»). Cosicché il medesimo livello può essere controllato in fasi differenti e in ciascuna fase possono essere controllati più livelli.

Per ragioni organizzative e di costo, non è, generalmente, possibile sottoporre la singola rilevazione all'insieme dei controlli, su tutte le fasi e su tutte le possibili fonti di errore. Il numero ed il tipo dei controlli da effettuare, ovvero il sistema di controllo della singola indagine, devono, quindi, essere selezionati in funzione delle risorse disponibili, dei tempi di esecuzione dell'indagine e del livello accettabile di errore.

La prevenzione dell'errore

Le tecniche per prevenire l'errore sono fondamentali per ridurre l'errore totale; infatti i metodi di correzione risolvono solo parzialmente ed in modo non del tutto soddisfacente il problema.

L'eliminazione o la riduzione dell'errore di rilevanza è demandata alla fase di progettazione; in tale fase tuttavia si riduce anche l'errore di misura, definendo nel modo migliore lo strumento tecnico di rilevazione (il questionario) e programmando, in modo coerente ed esaustivo, l'intero complesso dell'indagine. In tale ambito vanno curate particolarmente le norme di istruzione e di assistenza alla rete periferica, predisponendo norme e manuali chiari e completi, con numerosi esempi esplicativi.

Per quanto riguarda il controllo della *rilevanza* un utile strumento è costituito dalla metodologia della *Progettazione Concettuale*, sia per definire il modello concettuale relativo alle esigenze conoscitive, sia per tradurlo nel questionario. Quest'ultimo può essere verificato con varie tecniche: il giudizio di esperti, il pret-test sul campo ed il test di alternative. Infine, l'intero complesso dell'indagine, od almeno gli aspetti salienti, può essere controllato preventivamente mediante una indagine pilota, versione ridotta dell'indagine «madre».

La correzione dell'errore

Per indagini di media-grande dimensione, gli errori che è possibile determinare e correggere, sono solo una parte di quelli presenti nel materiale rilevato.

Più precisamente, è possibile apportare correzioni per gli errori che si manifestano come mancate risposte, totali e parziali, e come incoerenze della singola variabile o tra variabili logicamente collegate; essi, infatti, sono identificabili in tempi utili, per non mutare le condizioni generali di svolgimento dell'indagine, e con costi economici ed organizzativi contenuti.

Assimilando le *coerenze fallite* alle *mancate risposte parziali*, possiamo definire due soli tipi di errore: le *mancate risposte totali* e quelle *parziali*.

A ciascuna tipologia corrispondono differenti tecniche di correzione: nel primo caso essa viene apportata a livello di stime, mentre nel secondo a livello dei micro-dati.

Per questi ultimi, data l'attuale organizzazione e la dimensione delle indagini, non è possibile ricorrere alla tecnica del ritorno presso le unità rispondenti in maniera diffusa e totale; cosicché, la sola strada praticabile rimane l'applicazione di norme e metodologie di revisione e correzione, influenzate dalle scelte soggettive del responsabile dell'indagine.

La correzione dei dati elementari, può essere attuata sia a livello periferico (prima revisione del materiale a cura dei rilevatori e dei supervisori), sia nella fase di revisione centralizzata, mediante l'utilizzo di personale specializzato e di procedure automatiche od interattive. I macrodati, invece, vengono rettificati applicando, in sede di elaborazione dei dati finali, pesi correttivi per le unità non rispondenti.

I controlli nella fase di registrazione, generalmente, non hanno l'obiettivo di correggere gli errori, ma solo quello di limitarne il numero ad una quota accettabile; una parte di tali errori, precisamente quelli che danno luogo a valori fuori campo e ad incongruenze logiche, saranno rimossi nella fase di revisione.

La stima dell'errore

L'errore non campionario rappresenta, secondo gli studi condotti in Italia ed in altri Paesi, la componente più rilevante dell'errore totale; la sua determinazione, quindi, è indispensabile per conoscere la reale precisione delle stime fornite da un'indagine. D'altro canto la conoscenza dell'entità degli errori di misura, presenti nelle varie fasi del processo di produzione, è il presupposto necessario per il suo miglioramento.

Per raggiungere i suddetti obiettivi, si utilizzando *modelli statistici*, i cui parametri possono essere stimati mediante indagini di controllo, ed *indicatori di qualità*, ottenuti dalle procedure standard dell'indagine.

L'utilizzo di modelli, permette di conoscere l'influenza dell'errore non campionario sulle stime, mentre gli indicatori si riferiscono ad aspetti particolari dell'indagine, (ad esempio, le procedure di correzione, le mancate risposte totali e parziali, l'errore di registrazione, ecc.) e non offrono una misura sintetica dell'errore di misura.

La prima tecnicha risulta più costosa in termini economici ed organizzativi, perché implica un ritorno sul campo oppure maggiore complessità nell'usuale organizzazione della rilevazione, mentre il calcolo degli indicatori richiede solo una attenta programmazione delle fonti informative.

Di norma, le stime ottenute dai modelli non sono disponibili per tutti i livelli di controllo coinvolti nella rilevazione, essendo basate su un campione dell'indagine principale, mentre gli indicatori sono calcolabili per tutti i livelli organizzativi e territoriali.

Le due strade non sono in alternativa, ma devono confluire in una unica strategia: le informazioni derivanti dalle differenti fonti, costituiranno un *archivio di qualità*, che potrà essere analizzato con le consuete metodologie univariate o multivariate, globalmente o per la fase/livello di controllo desiderato.

8. L'archivio di qualità

L'archivio di qualità è costituito dagli indicatori desumibili dalle indagini di controllo, dai documenti di rilevazione, dalle procedure standard dell'indagine principale e dai confronti con fonti esterne; in particolare essi possono derivare:

(a) dalla verifica di coerenza delle operazioni dell'indagine;
(b) dalle analisi dei dati relative alla fase di rilevazione, registrazione e revisione;
(c) dal confronto con i risultati aggregati ottenuti da altre indagini o dalla stessa in tempi diversi;
(d) dal confronto con micro-dati provenienti da altre indagini o dalla stessa in tempi diversi;
(e) dai risultati di indagini di controllo.

Mentre la disponibilità delle informazioni delle fonti (a)-(d) dipende dall'organizzazione dell'indagine, quelle relative alla fonte (e) implicano una ripetizione della rilevazione su sub-campioni e quindi un maggiore aggravio sia in termini economici che di organizzazione del lavoro.

Ciascuna di esse permette il calcolo degli indicatori per un livello gerarchicamente superiore; ad esempio le mancate risposte potranno essere riferite all'intervistatore, ovvero, aggregando per singolo comune, all'organizzazione di rilevazione comunale ed ai successivi domini territoriali.

La costruzione dell'archivio, tuttavia, non è un mero assemblaggio di indicatori provenienti dalle numerose fonti di informazione; essi devono costituire un insieme in grado di fornire informazioni non ridondanti per tutti i livelli sotto controllo. Inoltre, poiché il potenziale informativo dell'archivio risiede nei collegamenti che è possibile istituire tra indicatori relativi ad operazioni diverse per il medesimo livello, è necessario organizzarne la base informativa. A tale scopo, diviene rilevante l'affidabilità e la completezza del sistema di codici identificativi delle unità, al quale è demandato il compito di assicurare i collegamenti tra le varie informazioni.

La costruzione dell'archivio di qualità è facilitata laddove sia stato costruito il Sistema Informativo Statistico dell'indagine; in questo caso l'archivio rappresenta una delle funzioni del suddetto sistema.

La costituzione di archivi di qualità per gruppi omogenei di indagini, permette di sfruttare le sinergie, derivanti dal collegamento di tali archivi, per il controllo di *strutture* comuni alle diverse rilevazioni.

Un caso particolarmente rilevante, è costituito dalla *rete di*

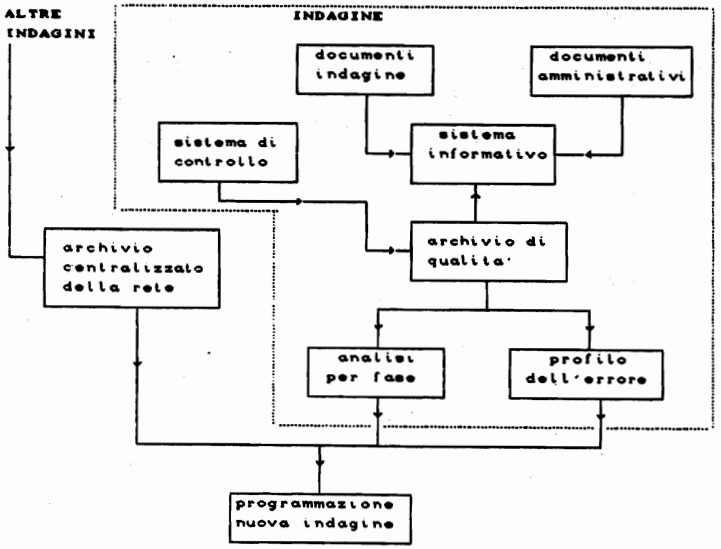


Figura 1.7 - Relazioni tra il sistema di controllo, il sistema informativo, l'archivio di qualità e l'archivio della rete di rilevazione

rilevazione per le indagini sulle famiglie; il relativo archivio centralizzato, adeguatamente alimentato, con indicatori provenienti dalle varie rilevazioni, permetterà di arricchire le informazioni desumibili dai singoli archivi e di determinare le situazioni anomale su cui concentrare gli sforzi organizzativi, invece di disperdere le risorse con interventi «a pioggia».

Nella Figura 1.7 vengono schematizzate le relazioni tra il *sistema di controllo*, il *sistema informativo statistico*, l'*archivio di qualità* e l'*archivio centralizzato della rete di rilevazione*, in relazione alle possibili utilizzazioni.

L'archivio di qualità non solo costituisce la base informativa per il controllo *a posteriori* dell'indagine, ma può essere anche utilizzato per la programmazione di indagini similari o della stessa indagine in tempi differenti.

In particolare, le informazioni contenute nell'archivio potranno risultare utili per una migliore allocazione delle risorse disponi-

bili e, poiché esse costituiscono la *memoria storica* della rilevazione, per la valutazione degli interventi correttivi, effettuati in precedenti occasioni.

Mediante l'archivio di qualità, per le indagini ripetute e per quelle in cui può essere ipotizzata la «portabilità» degli indicatori, i controlli *successivi* della singola rilevazione assumono il ruolo di controlli *preventivi* per la nuova rilevazione.

RIFERIMENTI BIBLIOGRAFICI

ARKHIPOFF O. (1986), *La qualité de l'information et sa precision*, Colloque de l'ISEOR

BAGLIONI P. (1983), *Livelli di produzione e di utilizzazione delle informazioni statistiche: aspetti generali e considerazioni specifiche sulla qualità dei dati ai fini della programmazione regionale*, in Atti della IV Conferenza Italiana di Scienze Regionali, Firenze.

BAILAR B.B. (1985), *Error profiles: uses and abuses* in «Statistical methods and the improvement of data quality», Edited by T. Wright, Academic Press, New York.

BAILAR B.B. (1985), *Quality issues in measurement*, International Statistical Review.

BIGGERI L., COLOMBO B. (1991), *Relazione sull'attività della Commissione Scientifica della S.I.S. sulla qualità dei dati*, Bollettino della S.I.S. n. 22, aprile 1991.

BROOKS C.A., BAILAR B.B. (1978), *An error profile: employment as measured by the current popolation survey*, Statistical Policy Working Paper n. 3, U.S. Department of Commerce - Office for Federal Statistical Policy and Standards, Washington D.C. - U.S. Government Printing Office.

COCHRAN W. (1977), *Sampling Techniques*, cap. 13, J. Wiley & Sons, New York.

COLOMBO B. (1979), *Sul concetto di qualità delle statistiche ufficiali*, in «Studi di statistica e di economia in onore di L. Lenti», Università degli Studi di Milano, Pavia e L. Bocconi, Milano.

COLOMBO B. (1983), *La qualità dei dati statistici* in Atti del Convegno 1983 della S.I.S., Trieste.

CORTESE A. (1991), *Linee direttive per l'illustrazione di contenuti e qualità dei dati statistici*, Bollettino della S.I.S. n. 22, aprile 1991.

CORTESE A., GIOMMI A. (a cura di) (1991), *Bibliografia di autori italiani*, Bollettino della S.I.S. n. 22, aprile 1991.

DALENIUS T. (1983), *Errors and other limitations of survey*, in «Statistical methods and the improvement of data quality», Edited by T. Wright, Academic Press, New York.

GIOMMI A. (a cura di) (1991), *Glossario dei principali termini su: "la qualità dei dati statistici"*, Bollettino della S.I.S. n. 22, aprile 1991.

GOTTARDO G. (1983), *Alcune considerazioni sulla valutazione della qualità dei dati provenienti da un'indagine campionaria in campo sociale*, in Atti del Convegno 1983 della S.I.S., Trieste.

I.N.S.E.E. (1985), *Rapport sur la qualité des travaux statistiques*, documento interno, Parigi.

KISH L. (1965), *Survey sampling*, cap. 13, J. Wiley & Sons, New York.

MANICARDI G., VENTURI M. (1988), *Analisi integrata di dati e funzioni nei Sistemi Informativi Statistici*, documento interno, Istat.

MASSELLI M. (1985), *La qualità dei dati nelle rilevazioni statistiche*, Rivista Italiana di Economia, Demografia e Statistica, Vol. 40.

MASSELLI M., SIGNORE M. (1989), *Il sistema di controllo delle indagini campionarie dell'Istat: linee di ricerca e principali contributi del Progetto Qualità dei Dati*, relazione alla Giornata sul campionamento statistico, Istat, Annali di Statistica, Serie 9ª, Vol. 10, Anno 120.
MASSELLI M. (1991), *Il profilo degli errori nell'indagine sulle forze di lavoro*, Bollettino della S.I.S. n. 22, aprile 1991.
MONTINARO M. (1988), *Un modello per la determinazione dell'error profile del commercio con l'estero*, in Atti della XXXIV Riunione scientifica della S.I.S., Siena.
MONTINARO M. (1991), *Il profilo dell'errore nell'indagine statistica del Commercio con l'Estero*, Bollettino della S.I.S. n. 22, aprile 1991.
OUTRATA E., CHINNAPPA N. (1989), *General survey function design at Statistics Canada*, Proceedings of the 47th Session of ISI, Parigi.
PARENTI G. (1983), *Sulla qualità dei dati statistici*, in Atti del Convegno 1983 della S.I.S., Trieste.
QUINTANO C., CALZARONI M., DINI P., MASSELLI M., POLITI M., TACCINI P. (1987), *Una ricognizione dell'error profile dell'indagine sul prodotto lordo*, in «Attendibilità e tempestività delle stime di contabilità nazionale», a cura di U. Trivellato, CLEUP Padova.
RYTEN J. (1988), *Errors in foreign trade statistics*, in Survey Methodology, Volume 14, n. 1, june 1988.
SINGH D., CHAUDHARY F.S. (1986), *Theory and analysis of sample survey design*, J. Wiley & Sons, New York.
Statistics Canada, (1976), *A compendium of methods of error evaluation in censuses and surveys*.
TERRA ABRAMI V. (1989), *Manuale di tecniche di indagine: Pianificazione della produzione dei dati*, Note e Relazioni, n. 1, ISTAT.
U.N. (1982), *National household survey capability programme. Non-sampling errors in household surveys: sources, assesment and control*, New York.
ZARKOVICH (1967), *Sampling methods and censuses*, «Volume II - Quality of statistical data», F.A.O.

CAPITOLO 2 - LA PROGETTAZIONE DELL'INDAGINE

1. La fase di progettazione

Nella fase di progettazione si formula il modello concettuale e si programmano i vari aspetti dell'indagine (sulla base delle esigenze conoscitive, delle risorse disponibili, delle informazioni a priori sul fenomeno in studio) ed, infine, si integrano le diverse operazioni in un unico meccanismo organizzativo.

Più specificatamente, si individuano:

— gli obiettivi dell'indagine;
— le variabili di rilevazione;
— l'universo di riferimento;
— le unità di analisi;
— le unità di rilevazione;

e si predispongono:

— il piano di campionamento;
— il piano di diffusione dei dati;
— il piano di rilevazione sul campo;
— il piano di registrazione su supporto informatico;
— il piano di revisione;
— il piano di elaborazione;
— il piano dei controlli.

Nella progettazione dell'indagine possono essere generati errori di rilevanza ed errori nella formulazione dei diversi piani di lavoro; questi ultimi, a loro volta, possono dar luogo ad errori di misura e di rilevanza, in sede di attuazione pratica. Un'ulteriore fonte di errore, dato che i diversi protocolli sono strettamente collegati, è costituita dalla mancata coerenza tra i piani di lavoro; infine, un'insufficiente programmazione del sistema dei controlli, può far mancare le funzioni di monitoraggio e correzione al processo produttivo.

La verifica del progetto d'indagine, quindi, deve essere condotta su due livelli: da un lato il controllo della validità del singolo piano, dall'altro la coerenza tra i differenti piani di lavoro.

Dal punto di vista della qualità dei dati, la progettazione costituisce un momento particolarmente delicato perché in tale fase (I) si prevengono di fatto gli errori non campionari e (II) viene programmato il controllo di tali errori.

La prevenzione costituisce un obiettivo tanto più rilevante, se si considera che la possibilità di modificare le norme in corso d'opera è scarsa ed estremamente costosa in termini organizzativi ed economici.

Data la sua complessità, il progetto d'indagine, generalmente, si configura come un processo iterativo che procede per verifiche successive. A tale scopo è possibile utilizzare alcune tecniche specifiche: la Progettazione Concettuale, i test del questionario, l'indagine pilota e la verifica di coerenza del progetto.

La Progettazione Concettuale è contemporaneamente una tecnica di controllo ed una operazione del processo produttivo; il metodo produce un «modello concettuale» internamente coerente che genera un questionario la cui struttura è anch'essa coerente.

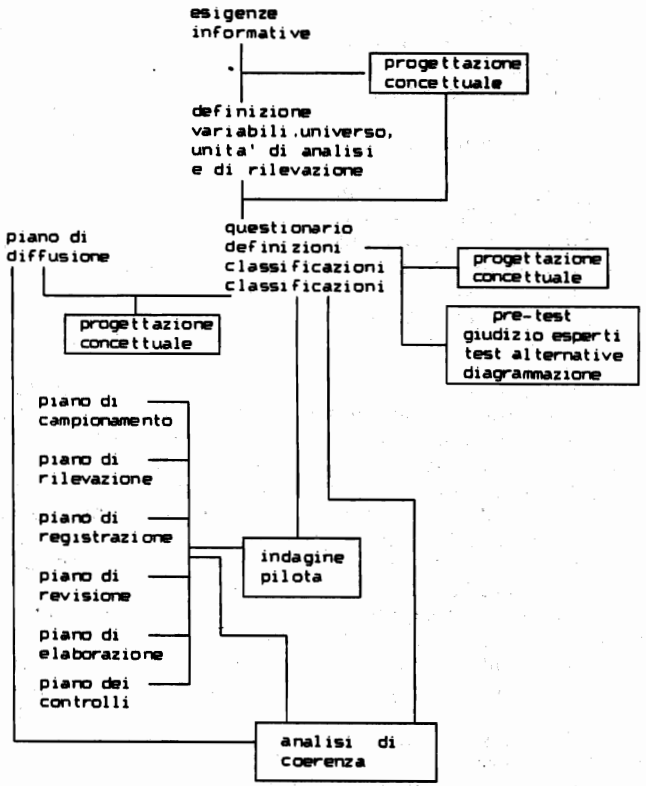


Figura 2.1 - La sequenza dei controlli nella fase di progettazione

La coerenza è un prerequisito, ma non è sufficiente a garantire la buona riuscita del questionario. Per altri aspetti (il vocabolario, le domande retrospettive, i quesiti delicati, il numero e la sequenza delle domande ecc.) è necessario ricorrere a differenti tecniche di controllo: il test sul campo, il giudizio degli esperti ed il test di alternative.

Il test del questionario può essere inserito nell'indagine pilota; quest'ultima, però, ha come obiettivo principale, la verifica sul campo della praticabilità delle norme e dei protocolli di tutte, o le più importanti, operazioni progettate per l'indagine «madre».

L'analisi di coerenza del progetto d'indagine, consiste, in una verifica logica, effettuata dal responsabile d'indagine, dei singoli piani di lavoro e dei reciproci legami.

La sequenza dei controlli da effettuare nella fase di progettazione, rispecchia la sequenza logica ed i legami tra le operazioni; essa è riportata nella Figura 2.1.

Nel diagramma, la Progettazione Concettuale appare più volte, per indicare le possibili applicazioni di tale tecnica; in realtà, una volta utilizzata per tradurre le esigenze informative in variabili di studio e per definire le entità coinvolte nell'indagine, diviene automatica la derivazione del questionario. Inoltre, gli schemi derivati dalla Progettazione Concettuale possono essere, utilizzati per la predisposizione del piano di diffusione.

Tale piano, poiché, in ultima analisi, costituisce la «specificazione operativa» degli obiettivi, precede logicamente ed influenza la formulazione degli altri protocolli. Il legame più ovvio è quello con il piano di elaborazione, ma anche il piano di rilevazione e di revisione ne sono influenzati; ad esempio nella procedura di compatibilità e correzione, o nelle norme di rilevazione per raccomandare particolare attenzione verso alcune variabili considerate strategiche per la tabulazione.

Qualora non venga utilizzata la Progettazione Concettuale, và comunque predisposta una completa documentazione, riguardante le suddette operazioni, in cui siano specificati i nessi logici e i motivi delle scelte effettuate.

Il questionario gioca un ruolo centrale nelle progettazione dell'indagine poiché è collegato con tutti i piani di lavoro, anzi ne costituisce in gran parte il «prius» logico; la sua verifica deve quindi precedere la predisposizione delle altre operazioni.

La validità delle norme e delle procedure, predisposte in via provvisoria, sarà sottoposta a test, direttamente sul campo, mediante l'indagine pilota; infine, la verifica di coerenza costituisce l'indispensabile controllo prima della rilevazione sul campo e delle fasi successive.

Se i risultati ottenuti dai suddetti controlli, portano a modificare le operazioni oggetto della verifica, sarà anche necessario cambiare i protocolli dei piani di lavoro ad esse collegati (cfr. Figura 2.2).

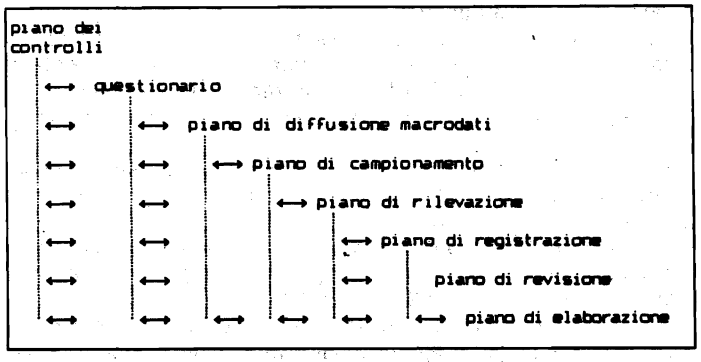


Figura 2.2 - Relazioni tra le operazioni della fase di progettazione

2. La progettazione concettuale

La specificazione degli obiettivi in variabili e nelle loro relazioni, il passaggio da queste ai quesiti ed alla struttura del questionario, l'identificazione della popolazione di riferimento e delle unità di rilevazione e di analisi, non è altro che la costruzione di un *modello* della realtà esaminata.

Errori di rilevanza possono sorgere se tale modello non viene esplicitato ed enunciato per tale, ma risulta definito solo implicitamente ed imperfettamente dalle operazioni di cui sopra.

L'uso di un *modello concettuale* che costringa a definire in maniera chiara e precisa i concetti coinvolti, aiuta a prevenire tali errori.

Mediante la metodologia denominata *Progettazione Concettuale*, i contenuti informativi di una rilevazione statistica possono essere individuati e rappresentati in maniera formale, indipendentemente dalle specifiche problematiche; la tecnica permette di definire le relazioni tra entità, gli attributi delle stesse e le strutture gerarchiche tra entità.

La documentazione sulle definizioni e la rappresentazione formale degli schemi concettuali costituiscono il patrimonio informativo dell'indagine e permettono il controllo della rilevanza, teorica ed effettiva, dell'informazione prodotta.

Inoltre, poiché attraverso il modello concettuale è possibile rappresentare le relazioni tra le diverse entità ed i loro attributi, gli schemi prodotti possono essere utilizzati per la stesura del questionario e per definire una parte delle regole di compatibilità, mediante le quali vengono determinate le incongruenze logiche nei dati raccolti.

L'uso delle tecniche di progettazione concettuale è diffusamente trattato nel *Manuale per la progettazione concettuale di dati statistici* Istat 1987.

3. La redazione del questionario

Sotto il termine di *questionario* si intende qualsiasi strumento utilizzato per la raccolta dei dati presso l'unità di rilevazione. In questo contesto sono, quindi, considerati questionari anche i modelli per la raccolta di informazioni amministrative, lo schema di domande per le indagini telefoniche o di quelle effettuate mediante i computer portatili.

Un'adeguata formulazione del questionario è cruciale nel processo di produzione, poiché esso è lo strumento mediante il quale i dati rilevati presso le unità vengono trasformati in «informazione» disponibile per le successive elaborazioni.

La predisposizione del questionario può generare errori di rilevanza e di misura, in modo particolare in funzione:

- del vocabolario utilizzato
- della sequenza dei quesiti
- delle norme di compilazione
- dei quesiti retrospettivi
- di risposte proxy
- di quesiti particolarmente delicati

I problemi connessi alla predisposizione, alla somministrazione ed al test sul questionario sono analiticamente trattati nel Fascicolo 2 del *Manuale di tecniche di indagine*; nel presente capitolo ci si limiterà a richiamare bevemente le considerazioni più rilevanti ai fini del controllo di qualità.

In termini generali il contenuto di un questionario può essere distinto in tre parti:

1. quesiti per la rilevazione delle variabili di studio
2. codici identificativi
3. quesiti per il controllo dell'intervista.

È opportuno che tale distinzione venga formalizzato nella fase di progettazione, poiché le tre tipologie giocano ruoli diversi

e subiscono trattamenti differenziati nel corso delle fasi successive. Rimandando agli appositi capitoli qui basti citare, ad esempio, che per i codici identificativi se ne consiglia una attenta verifica nella fase di registrazione e l'eventuale modificazione solo dopo la fase di revisione quantitativa; per i quesiti relativi alla qualità (notizie sull'intervista e sul rilevatore), invece, non devono essere previste modificazioni, in quanto comporterebbero perdita di informazione.

Le variabili di studio

Estraendo dal modello concettuale un *albero*, ed eventualmente dei sub-alberi, di aree omogenee di informazione, è possibile tradurre le variabili, precedentemente identificate, nella struttura e nei quesiti del questionario.

La sequenza dei quesiti deve essere resa il più lineare possibile, evitando i rimandi a domande, o blocchi di domande, precedenti e segnalando chiaramente con accorgimenti grafici (colori, frecce ecc.) gli eventuali salti dipendenti da domande *filtro*.

È conveniente che la sequenza rispecchi l'ordine implicito nei legami logici tra le variabili; nel caso di un questionario particolarmente gravoso, è opportuno che i quesiti *strategici* vengano posti all'inizio del modello, per evitare che il decrescente interesse del rispondente incrementi l'errore di misura delle informazioni più importanti da rilevare.

I quesiti possono dar luogo a risposte *aperte* o *chiuse*; nel primo caso si accetta qualsiasi risposta fornita dal rispondente, mentre nel secondo lo si costringe a scegliere tra un predeterminato numero di risposte.

Le risposte precodificate hanno il pregio di essere più facilmente e più rapidamente trattabili nelle fasi successive di registrazione e di revisione dei dati e contribuiscono a ridurre l'effetto ricordo. Tuttavia esse scontano una perdita di informazione che è tanto più grave quanto più mancano informazioni a priori sul fenomeno da rilevare. In questo caso l'indagine pilota effettuata mediante questionario con quesiti *aperti* può costituire una base di informazioni per *chiudere* i medesimi.

La precodificazione implica la scelta:

— del tipo di classificazione da adottare,
— del livello di disaggregazione della codifica.

Nella scelta della classificazione è opportuno rifarsi agli standard Istat; laddove sia necessario ricorrere a classificazioni più *fini* occorre prevedere la possibilità di ricostruzione dei suddetti standard. Nella definizione del livello di disaggregazione, bisogna tener conto che la maggiore informazione derivante da classificazioni più analitiche sconta anche una maggiore imprecisione nelle risposte fornite.

Nell'articolare le classificazioni, soprattutto nella riduzione di variabili continue a variabili intervallo, occorre tenere presente gli obiettivi per cui le variabili sono rilevate. Se tra di essi vi è la costruzione di indici statistici, occorre esplicitarli in una lista analitica per poter controllare l'adeguatezza della codifica adottata.

I codici identificativi

È stata richiamata più volte, nel corso del Capitolo 1, l'importanza dei codici identificativi per il sistema di controllo; essi infatti sono il prerequisito per l'individuazione delle unità e delle loro relazioni (ad esempio, gli individui appartenenti alla medesima famiglia) e costituiscono il legame delle informazioni relative alla singola unità nelle diverse fonti informative dell'indagine (il questionario, i documenti aggiuntivi di rilevazione, i file). Poiché rappresentano relazioni tra unità e tra archivi, gli identificatori costituiscono un «sistema». Nel contesto del controllo di qualità, tale sistema deve essere predisposto sulla base di una preventiva scelta dei livelli di controllo desiderati e del quadro completo degli archivi e dei documenti di rilevazione da utilizzare.

In genere, un codice identificativo è il risultato della *concatenazione* di singoli identificatori, ciascuno dei quali riferito ad un tipo di unità statistica, coinvolta nella rilevazione oppure rappresentante un dominio territoriale od un dominio di studio, (ad esempio il comune, la regione, il ramo di attività economica).

Mediante il concatenamento di tali codici è possibile:

— distinguere ciascuna unità statistica dalle altre dello stesso tipo, (ad esempio un determinato comune dai rimanenti, un determinato individuo dagli altri individui);
— assegnare ciascuna unità ad una unità di ordine superiore (ad esempio l'individuo alla famiglia, questa al comune, ecc.);
— porre in relazione due unità dello stesso tipo o di tipo diverso, ma non incluse l'una nell'altra, (ad esempio la famiglia principale con quella coabitante, l'area con il rilevatore);
— riferire la singola unità ad una lista esterna alla rilevazione, permettendone, quindi, l'identificazione sul territorio od in più occasioni d'indagine (ad esempio il codice comunale riferito alla lista dei comuni italiani, il codice familiare della struttura longitudinale riferito alla lista di selezione).

L'articolazione dei codici sul questionario, dato il ruolo svolto dal medesimo nell'indagine, costituisce il fulcro del sistema di identificazione delle unità; a questo livello possiamo, quindi, distinguere i codici in:

1. identificatori che collegano il questionario ad unità di ordine superiore (ad esempio al rilevatore, al comune ecc.);

2. identificatori che collegano due o più questionari inerenti ad unità diverse ma tra le quali è presente una relazione logica (ad esempio il questonario della famiglia principale e di quella coabitante);
3. identificatori interni al modello di rilevazione per collegare informazioni relative alla stessa unità di analisi in parti diverse del modello (ad esempio le variabili demografiche individuali, raccolte in una parte comune, alle rimanenti informazioni presenti sui fogli individuali).

L'identificatore del questionario viene esteso a tutte le unità contenute nel medesimo; tuttavia, tale codice può non essere sufficiente al riconoscimento delle unità di ordine inferiore nei documenti aggiuntivi di rilevazione o quando trasposte su supporto informatico.

Ad esempio la vacanza attribuita ad un determinato componente, perfettamente riconoscibile su supporto cartaceo, può non esserlo più nel file se non si introduce un ulteriore codice di identificazione e di collegamento tra tali unità.

In aggiunta a quelli sopra citati, si devono, quindi, prevedere altri due gruppi di identificatori:

4. identificatori per la trasposizione su supporto informatico;
5. identificatori per i modelli aggiuntivi di rilevazione.

Il sistema dei codici deve assicurare a ciascuna unità elementare un unico identificatore che la renda riconoscibile in tutte le fonti di informazione; in questo modo è possibile utilizzare, congiuntamente, informazioni diverse per il medesimo livello di controllo e calcolare i relativi indicatori di qualità.

Le variabili di controllo dell'intervista

Nel questionario di rilevazione devono essere raccolte alcune informazioni riguardanti le modalità di svolgimento dell'intervista; la scelta del numero e del tipo di tali quesiti deve essere parsimoniosa ed efficiente, nel senso che non si deve sovraccaricare il questionario con quesiti di cui non sia stata preventivamente stabilita l'utilizzazione.

Le informazioni di controllo dell'intervista possono essere suddivise in due gruppi:

— informazioni da cui derivare indicatori di qualità
— informazioni necessarie ad indagini di controllo (successive o contemporanee).

In termini generali, i dati del primo tipo sono quelli riguardanti la situazione dell'intervista (ovvero quanti componenti erano presenti, l'eventuale rispondente o risposta proxy), il giorno, l'ora e la durata, i conteggi riassuntivi delle unità di analisi, o degli eventi, contenuti nel questionario, le valutazioni sull'accoglimento e la partecipazione all'intervista o a parti di essa.

I dati necessari per effettuare le indagini di controllo, e, spesso, per analizzarne i risultati, sono di tipo diverso:

I) codici identificativi di unità gerarchicamente superiori a quella oggetto di studio, già standardizzati e presenti nel questionario (ad esempio il codice di area e di rilevatore da utilizzare per la compenetrazione del campione);
II) quesiti aggiuntivi e codici ad hoc, generalmente non presenti sui questionari (ad esempio il quesito relativo al possesso di telefono da utilizzare per indagini di controllo mediante reintervista telefonica);
III) caratteristiche di studio identificative dell'unità (ad esempio sesso e data di nascita per poter effettuare le reinterviste).

I quesiti retrospettivi

Tra le variabili di studio, un caso particolare è rappresentato da quelle rilevate mediante i quesiti retrospettivi per mezzo dei quali si chiede agli intervistati di riportare tutti gli eventi di un certo tipo (nascite, morti, malattie, periodi di vacanza, spese) avvenuti in un determinato lasso di tempo precedente.

Tali quesiti possono comportare errori dovuti (I) all'omissione di eventi oppure (II) ad una erronea collocazione temporale degli stessi.

L'intervallo temporale rispetto al quale si richiedono le informazioni si definisce *periodo di riferimento*; esso può essere *fissato* oppure *mobile* anche se di durata costante. Ad esempio, un periodo di riferimento di una settimana è fissato se è compreso tra due date prefissate (1 gennaio - 7 gennaio), mentre è mobile se riguarda i sette giorni precedenti la data dell'intervista. In questo caso, infatti, se le interviste non vengono condotte tutte nel medesimo giorno, i dati raccolti si riferiranno ad un periodo di tempo variabile.

Al periodo di riferimento sono associati due concetti: il periodo dell'indagine ed il periodo di ricordo.

Il periodo dell'indagine è l'arco di tempo durante il quale si svolge effettivamente la rilevazione sul campo, mentre il periodo di ricordo è l'intervallo temporale trascorso tra la data in cui si è verificato l'evento e quella in cui si chiede di ricordarlo.

Generalmente, il periodo di ricordo è collegato a quello di riferimento dell'indagine; in questo caso il primo deve essere mi-

nore od uguale al secondo, se quest'ultimo è mobile, mentre può essere superiore se il periodo di riferimento è fissato. Ad esempio, se si chiedono le spese effettuate in una settimana di riferimento fissata e l'intervista avviene nella settimana successiva, allora il periodo di ricordo può essere anche di due settimane.

Le relazioni tra il periodo di riferimento, fissato e mobile, e gli altri due sono illustrati graficamente nella Figura 2.3.

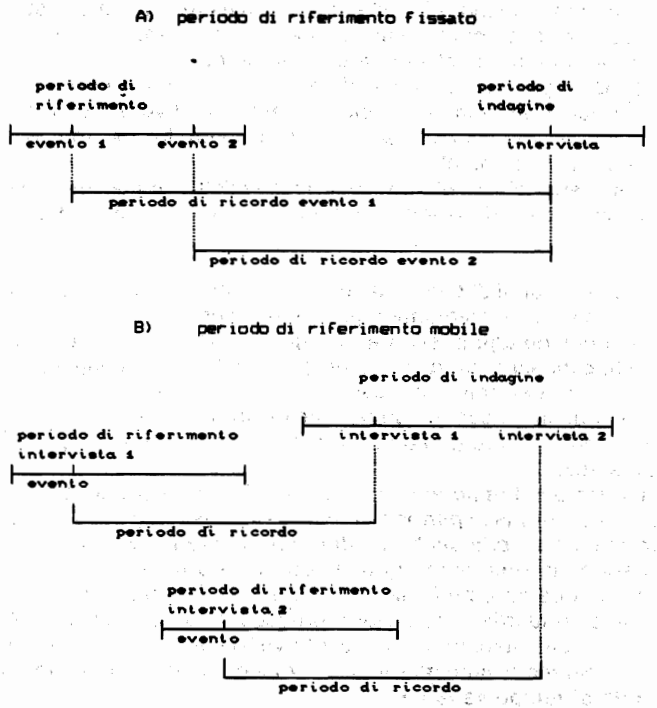


Figura 2.3 - Relazioni tra i periodi di riferimento,d'indagine e di ricordo

In alcune rilevazioni il periodo di ricordo può non essere collegato ai periodi di riferimento dell'indagine; ad esempio quando si chiede la data dell'ultimo evento verificatosi invece di chiedere se si sono verificati eventi nel periodo stabilito. In questo caso il periodo di ricordo può variare notevolmente da rispondente a rispondente.

Per quanto concerne l'omissione degli eventi, le esperienze condotte in altri paesi hanno mostrato che esso è collegato, da un lato, al periodo di ricordo e, dall'altro, all'importanza dell'evento per il rispondente.

Per prevenire tale errore si dovrebbe ridurre il periodo di ricordo e quindi anche il periodo di riferimento; in tal modo però si riduce il numero degli eventi rilevabili ed aumenta l'errore campionario delle stime basate su un numero fisso di interviste. È quindi conveniente diversificare i periodi di riferimento in funzione della frequenza degli eventi.

Gli errori nella collocazione temporale di un evento possono riguardare sia uno spostamento all'indietro che uno in avanti (effetto telescopio), rispetto al momento reale in cui si è verificato; essi possono causarne o l'erronea collocazione temporale all'interno del periodo, ovvero l'esclusione di un evento, oppure la presenza di un evento precedente o successivo nel periodo stesso.

Lo spostamento all'interno del periodo di riferimento può non produrre effetti sul numero complessivo degli eventi ma può produrne relativamente ad eventuali domini di studio temporali; ad esempio, può verificarsi che il numero e l'ammontare delle spese in un anno sia stimato correttamente, ma i dati relativi ad intervalli mensili siano distorti.

Negli altri casi, invece, può verificarsi una distorsione delle stime: una sovrastima se l'effetto *telescopio* è predominante, una sottostima nel caso contrario.

In termini generali, si può comunque affermare che periodi di riferimento e di ricordo più brevi diminuiscono gli errori di omissione; essi, però, possono causare maggiori errori di collocamento temporale, se il periodo non è chiuso, ed un aumento dell'errore di campionamento delle stime.

Per prevenire e ridurre gli errori non campionari relativi ai quesiti retrospettivi si ricorre a tecniche differenti.

Riguardo agli errori dovuti alla collocazione temporale dell'evento, si dovrebbe limitare la possibilità degli spostamenti, ovvero *chiudere* gli estremi del periodo di riferimento rispetto al passato; ad esempio, l'intervallo corrispondente alla vita dell'intervistato fino al momento dell'intervista è un intervallo *chiuso*. Un periodo di riferimento fissato, invece, ha entrambi gli estremi aperti, mentre uno mobile (ad esempio la settimana precedente l'intervista) ha solo l'estremo inferiore aperto.

Una tecnica per *chiudere* gli estremi dell'intervallo consiste nell'usare un periodo di riferimento mobile e di reintervistare, dopo un certo periodo, i rispondenti, in merito agli eventi verificatisi nell'intervallo tra due interviste.

Per prevenire l'omissione di eventi, si ricorre al metodo, noto nella letteratura come *aiuto alla memoria*; esso consiste nello stimolare il ricordo evitando domande aperte e fornendo una lista delle possibili risposte oppure delle indicazioni *chiave*. Ad esempio, in una indagine sulle letture, è preferibile non porre il quesito nella forma *quali riviste o quotidiani ha letto nell'ultima settimana*, ma presentare al rispondente un elenco, chiedendo di segnalare quali letture ha fatto nel periodo di riferimento. Tale metodo riduce gli errori di omissione, ma può causare un effetto *telescopio*.

Per aumentare l'affidabilità delle informazioni rilevate, si può adottare la tecnica di chiedere al rispondente di riportare solo gli eventi documentati (ad esempio mediante scontrini fiscali, conti correnti ecc.). Tale metodo consente, tra l'altro, di eliminare gli errori di collocazione mediante la data riportata sui documenti, ma non garantisce rispetto all'omissione degli eventi.

Quanto detto evidenzia come la scelta dei periodi di riferimento sia piuttosto delicata e richieda sperimentazioni ad hoc che forniscano informazioni sulle distorsioni associate a periodi di diverso tipo e di diversa lunghezza.

Le risposte proxy

Al momento dell'intervista, può accadere che le informazioni, riguardanti l'unità designata, presente od assente, vengano fornite da altra unità. Ad esempio, nelle indagini Istat sulla popolazione, generalmente basate su interviste individuali a tutti i membri della famiglia campione, in assenza di uno di essi, è previsto che le notizie che lo riguardano vengano fornite da altro componente; oppure può accadere che un componente si sostituisca, nella risposta a taluni quesiti delicati, al familiare intervistato.

Deve, quindi, essere stabilita la regola di comportamento dell'intervistatore di fronte alla possibilità di *risposte proxy*, dove, con tale espressione si intende l'accettazione di informazioni non dal rispondente designato ma da altra unità.

Quattro sono le possibili alternative:

— non accettare la risposta proxy
— accettarla per tutti i quesiti
— accettarla limitatamente ad alcuni quesiti
— accettarla solo dopo un certo numero di ritorni

In questo ultimo caso devono essere fissati il numero (in genere non superiore a tre) e le modalità dei ritorni (ad es. intervista diretta o telefonica).

L'accettare le risposte proxy si fonda sul presupposto che le unità *vicine* siano in possesso di notizie attendibili riguardanti il rispondente.

Laddove, invece, per mancanza di informazione o per reticenza, il valore rilevato diverge da quello che avrebbe fornito l'unità designata, viene generato un errore non campionario, che è funzione del tipo di variabile indagata (effetto proxy).

La scelta di una delle strategie sopra riportate, deve essere attentamente valutata, ricorrendo nel caso ad una verifica sul campo, esaminando i relativi vantaggi e svantaggi.

Ad esempio, non accettare, del tutto od in parte, le risposte proxy e prevedere dei ritorni, presenta il vantaggio di aumentare l'affidabilità dei dati raccolti, ma dilata i tempi della rilevazione, aumenta il carico di lavoro degli intervistatori (e quindi il costo), produce un incremento dell'errore campionario e di quello non campionario, dovuto alla mancata intervista di una quota delle unità designate.

Tuttavia, il rifiuto delle risposte fornite da altre unità, dipende, in ultima analisi, dagli scopi dell'indagine, ovvero di quanto preminente sia l'interesse ad una rilevazione accurata nel caso di caratteristiche notevolmente soggette all'effetto proxy.

4. Il controllo del questionario

Numerose sono le scelte che devono essere effettuate, come si è detto nel paragrafo 1, prima di arrivare alla redazione definitiva del questionario. Tali scelte possono essere suffragate da esperienze effettuate precedentemente per la medesima indagine o per indagini similari; tuttavia, è sempre opportuno, anche nei casi più semplici, condurre una verifica sperimentale su una o più versioni provvisorie del modello di rilevazione.

A tale scopo si ricorre a differenti tecniche: la Progettazione Concettuale, il giudizio degli esperti, la validazione qualitativa in *laboratorio*, la diagrammazione del questionario, il pre-test sul campo, con o senza reintervista, e il test di alternative.

Le varie tecniche si differenziano riguardo agli obiettivi, ai costi economici, alle implicazioni organizzative, alla tempestività ed alla complessità delle relative metodologie di analisi dei risultati. La scelta di uno o più metodi di verifica del questionario dipende, oltre che dall'obiettivo del controllo, anche dai costi, in termini sia economici che organizzativi, in relazione agli obiettivi conoscitivi dell'indagine ed alla loro rilevanza.

Nel caso di ristrutturazione del questionario, prima di passare alla verifica sul campo, è conveniente analizzare gli indicatori di mancata risposta parziale, dei valori fuori campo e delle incongruenze logiche (cfr. Capitolo 3 e 4); tali informazioni, infatti, possono fornire utili indicazioni sulla validità dei quesiti utilizzati.

La progettazione concettuale

È stato già ricordato che l'utilizzazione della Progettazione Concettuale nel ricavare, da generici obiettivi conoscitivi, la definizione del modello concettuale e da quest'ultimo il questionario, ne garantisce la coerenza della struttura (definizioni, classificazioni, unità di analisi e di rilevazione e loro relazioni).

Derivare il modello di rilevazione dagli schemi prodotti, costituisce l'uso più efficiente di tale tecnica; tuttavia, questa può anche essere utilizzata per verificare la strutturazione di un questionario già predisposto.

Esempi di applicazione della suddetta metodologia a questionari già predisposti, sono contenuti nel citato *Manuale di progettazione concettuale di dati statistici*.

La diagrammazione del questionario

La diagrammazione del questionario può essere applicata tanto al controllo di versioni provvisorie che quale ultima verifica della redazione definitiva. La tecnica consiste nella rappresentazione della sequenza dei quesiti mediante un diagramma di flusso, sostanzialmente dello stesso tipo di quelli utilizzati nell'ambito dell'elaborazione dati; la visualizzazione delle relazioni, così prodotta, permette di verificare la *linearità* della struttura del questionario e di determinare l'eventuale presenza di norme contraddittorie o lacunose.

Il controllo che viene esercitato, quindi, è del tutto *formale* e non entra nel merito di altri aspetti del questionario (vocabolario, lunghezza, quesiti delicati, retrospettivi, proxy, ecc.).

Per mezzo del diagramma, può essere rappresentato il *flusso* dei quesiti, oppure le selezioni operate dalle domande per identificare l'appartenenza del rispondente a particolari subpopolazioni; nel primo caso le relazioni sono derivate dalle regole di compilazione, contenute nello stesso modello o nel manuale di istruzione per i rilevatori, mentre, nel secondo, dall'ordine dei quesiti che viene stabilito per esplicitare le definizioni.

Nel primo esempio dell'Appendice 1, è stato rappresentato il flusso dei quesiti del questionario individuale dell'indagine forze di lavoro, derivato dalle norme di compilazione; nel secondo esempio, invece, è stata diagrammata, esplicitando le definizioni, la sequenza delle domande necessaria al calcolo di uno degli aggregati (gli occupati).

Il giudizio degli esperti e le tecniche di laboratorio

Prima dei controlli sul campo, è opportuno verificare la redazione provvisoria del modello di rilevazione mediante il giudizio di esperti di settore, per individuare eventuali lacune od imprecisioni nel contenuto del questionario, e di esperti della comunicazione, per scegliere i migliori requisiti formali e di somministrazione del questionario.

In questa fase, inoltre, disponendo di adeguate professionalità e di un Centro specializzato, è possibile utilizzare tecniche di *laboratorio* che consistono nell'intervistare in profondità un gruppo o singoli individui appartenenti alla popolazione oggetto di indagine. Gli intervistatori, esperti nella comunicazione, utilizzeranno un questionario strutturato, o semplicemente una traccia di quesiti, e registreranno le reazioni dei rispondenti.

Il pre-test del questionario

Una volta stabiliti i contenuti informativi, e redatta una versione provvisoria del questionario, il medesimo è sottoposto a verifica sul campo mediante somministrazione ad un campione di unità appartenenti all'universo indagato.

Il pre-test viene utilizzato per scopi diversi: controllare il vocabolario, la sequenza, i quesiti che pongono problemi particolari, i concetti, le definizioni, la lunghezza, le istruzioni per la compilazione del questionari e chiudere eventuali domande aperte. È importante che gli obiettivi del pre-test siano anticipatamente chiariti, per adeguarvi la numerosità campionaria e gli strumenti di rilevazione e di analisi.

Nella predisposizione di un questionario, sono di norma necessarie più redazioni provvisorie del medesimo, ciascuna delle quali dovrebbe essere sottoposta ad una verifica sul campo: il pre-test, quindi, si configura come un procedimento iterativo, in cui la prova precedente viene utilizzata per modificare alcuni aspetti che sono successivamente controllati sul campo, fino a raggiungere risultati soddisfacenti.

Spesso, però, ragioni di bilancio o di tempestività non rendono possibile una strategia articolata su più verifiche di campo del modello di rilevazione, cosicché il pre-test si riduce ad una sola prova, riguardante l'insieme degli aspetti sopra considerati.

Per particolari obiettivi, ad esempio, per valutare l'effetto proxy ed il numero più conveniente di ritorni, o per stabilire il periodo di riferimento dei quesiti retrospettivi, è opportuno effettuare reinterviste di controllo.

Il campione utilizzato per il pre-test è generalmente di ridotta numerosità, sia per permettere un'analisi accurata del materiale raccolto, sia per contenere i costi ed i tempi di esecuzione dell'indagine. Nello stesso tempo, però, le unità campionate devono rispecchiare la massima variabilità delle condizioni della rilevazione e delle caratteristiche strutturali dell'universo in esame; per tali ragioni si ricorre ad un campione ragionato, di località e di unità di rilevazione.

Esperienze internazionali indicano dimensioni del campione variabili da 50 a 500 unità in funzione del tipo di verifica e dell'eterogeneità della popolazione sotto osservazione; riguardo agli obiettivi, ad esempio, il controllo della sequenza e del vocabola-

rio richiede meno interviste della chiusura delle domande, il test sul complesso del questionario o su *scale* di preferenza, invece, necessitano del massimo.

Anche la scelta del tipo di rilevatori da utilizzare per il pretest dipende dagli obiettivi. È opportuno che un unico controllo del questionario venga condotto con il medesimo personale impiegato nella rilevazione madre, per simulare al meglio le condizioni effettive di rilevazione; è preferibile invece che controlli più approfonditi di singoli aspetti siano condotti mediante rilevatori particolarmente selezionati o dagli stessi responsabili della redazione del questionario.

L'istruzione dei rilevatori dovrà essere particolarmente accurata, poiché viene loro richiesto un lavoro aggiuntivo ed una maggiore attenzione rispetto ad un «normale» rilevatore. Essi, infatti, devono prendere in considerazione e riferire su tutti quegli aspetti ed impressioni soggettive, che emergono nel corso dell'intervista, relative:

— alla completezza e alla correttezza del questionario rispetto agli obiettivi;
— alle difficoltà riscontrate dagli intervistati ed al loro atteggiamento di fronte all'indagine;
— alla semplicità di gestione da parte dell'intervistatore dello strumento «questionario».

Per la raccolta di tali informazioni, che avviene durante o dopo l'intervista, possono essere utilizzate tecniche diverse. Nel corso dell'intervista mediante (I) la registrazione della stessa su *nastro magnetico*, (II) la presenza di un supervisore che compila un questionario aggiuntivo od un brogliaccio informale sull'andamento dell'intervista e sulle reazioni del rispondente (III) quesiti riservati all'intervistatore ed inseriti nello stesso modello di rilevazione; dopo l'intervista, ma a stretto ridosso della stessa, nel corso di riunioni in cui si chiede agli intervistatori ed ai supervisori di compilare un questionario o produrre e discutere una relazione sulle interviste effettuate.

L'analisi dei risultati del pre-test viene condotta mediante l'esame a vista dei questionari, da parte di esperti e l'elaborazione di indicatori di mancata risposta parziale e di incongruenza logica.

Generalmente, tale analisi, indica l'insorgere dei problemi, ma non fornisce soluzioni atte a rimuoverli; per quest'ultimo obiettivo è necessario, sulla base dei risultati del pre-test, formulare alternative diverse e sottoporle a verifica sul campo, mediante un test di alternative.

Il test di alternative

Il test consiste nel sottoporre a verifica, su campioni bilanciati, più redazioni, generalmente due, del questionario che differiscono per un aspetto (ad esempio, la sequenza delle domande, la formulazione di quesiti, i periodi di riferimento temporali). Le differenti versioni vengono somministrate, con le medesime modalità, a campioni indipendenti, i quali, però, sono simili, tra loro, per la dimensione e, alla popolazione, per la struttura di alcune caratteristiche rilevanti per l'indagine (ad esempio la struttura per sesso e classi di età). Ciò equivale a condurre un esperimento, mantenendo fissi i fattori che influenzano la variabile di risposta.

L'analisi dei risultati verrà condotta mediante la comparazione dei risultati nei subcampioni, sintetizzati con gli indicatori più opportuni in relazione all'aspetto sotto controllo (ad esempio, la percentuale di mancate risposte parziali e gli indici di distribuzione delle risposte per valutare l'efficienza del questionario o di differenti classificazioni). La dimensione del campione viene stabilita in relazione ai livelli di affidabilità desiderati per l'esperimento.

5. L'indagine pilota

L'indagine pilota si differenzia dal pre-test del questionario in quanto persegue un obiettivo più ampio di quest'ultimo, ovvero la verifica di tutti gli aspetti della rilevazione.

Essa è condotta mediante un campione probabilistico e costituisce una *versione ridotta* dell'indagine principale; tutte le procedure devono essere sottoposte a controlli particolarmente accurati, allo scopo di identificare gli eventuali errori.

Si può quindi affermare che l'indagine pilota è meno *estesa*, ma più *approfondita* rispetto all'indagine madre; per suo mezzo si raccolgono non solo le caratteristiche oggetto di studio (allo scopo di stimare la variabilità dei fenomeni e quindi determinare, in mancanza di altre fonti, la numerosità campionaria) ma anche, e soprattutto, le informazioni concernenti l'organizzazione dell'indagine. A tale scopo è conveniente associare alla pilota un corpo selezionato di *supervisori* e prevedere modelli ad hoc e relazioni per il controllo delle procedure ai vari livelli e fasi.

Gli obiettivi dell'indagine pilota possono essere riassunti in:

- verifica definitiva del questionario, delle classificazioni e delle definizioni;
- verifica delle definizioni delle unità di rilevazione e dello stato delle liste;
- verifica delle modalità dell'intervista;

- verifica della rete di rilevazione e dei collegamenti tra *centro* ed organi periferici;
- verifica dei documenti accessori di rilevazione;
- verifica delle modalità di selezione dei rilevatori e dei supervisori, e dei relativi manuali e norme di istruzione;
- verifica del calendario di rilevazione;
- verifica dei piani di codifica, registrazione, revisione ed elaborazione dati;
- verifica del sistema di identificazione delle unità e della completezza, rispetto agli obiettivi fissati, delle fonti per il *sistema informativo statistico*;
- stima della variabilità dei fenomeni oggetto di studio per determinare la numerosità campionaria.

6. I modelli ausiliari

In tale categoria sono inclusi tutti i modelli compilati dagli organi periferici come ausilio alle operazioni di rilevazione sul campo: le assegnazioni dei rilevatori, gli elenchi dei rilevatori, i questionari sul rilevatore, gli elenchi delle unità campione, gli elenchi delle unità non rispondenti, ecc.

Tali documenti hanno il compito di agevolare il compito degli organi periferici, ma costituiscono anche una fonte di informazione per il calcolo di indicatori di qualità.

Essi, quindi, vanno inseriti tra le fonti informative del Sistema di Controllo; nei predisporli si dovrà tenere conto della loro utilizzazione a fini statistici.

Tra i suddetti modelli, assumono particolare importanza, per le informazioni contenute, gli elenchi delle unità non rispondenti. In essi deve essere prevista la causa della mancata collaborazione e le principali caratteristiche dell'unità, se disponibili da fonte diversa dall'intervista (ad esempio nel caso delle famiglie, il numero, il sesso e l'età dei componenti risultanti nella scheda anagrafica).

Le motivazioni della mancata intervista, riportate nel succitato documento di rilevazione, devono essere esplicitate in funzione degli indicatori che si intende calcolare per analizzare il fenomeno (vedi Capitolo 3) e della possibilità di raccogliere sul campo le informazioni necessarie. Inoltre, per assicurare l'omogeneità dei dati rilevati, l'individuazione delle possibili fonti e le modalità della richiesta delle informazioni devono essere espressamente e chiaramente riportate nelle istruzioni del rilevatore.

Tranne che per i rifiuti, le informazioni necessarie per individuare la motivazione della mancata intervista devono essere richieste ad altri, ad esempio i vicini di casa, la persona trovata all'indirizzo, il portiere dello stabile ecc.. È tuttavia improbabile, soprattutto nelle grandi città, che l'intervistatore riesca a classificare in maniera analitica le unità non intervistate; è quindi opportuno limitare la classificazione standard a poche, ma accertabili, modalità:

A) unità presenti all'indirizzo riportato sull'assegnazione del rilevatore
   - A1) che rifiutano l'intervista (rifiuti);
   - A2) con le quali non è stato possibile stabilire alcun contatto (non a casa);

B) unità irreperibili all'indirizzo.

Per uno studio più analitico del fenomeno, eventualmente da programmare per cicli di rilevazioni con un particolare addestramento degli intervistatori e dei supervisori, si suggerisce di suddividere la modalità «irreperibile» all'indirizzo nelle seguenti:

B) unità non presenti all'indirizzo
   - B1) per decesso
   - B2) per trasferimento nello stesso comune
   - B3) per trasferimento fuori del comune (eventualmente distinguendo tra estero ed altro comune italiano)
   - B4) per altre cause, voce residuale in cui confluiscono le unità che non è stato possibile classificare altrimenti (ad eempio indirizzo errato, persona sconosciuta all'indirizzo, ecc.).

Nella classificazione sopra riportata, la distinzione tra *trasferimenti* dentro e fuori dello stesso comune, ha lo scopo di accertare l'appartenenza dell'unità alla popolazione oggetto di rilevazione, poiché, nel caso delle indagini Istat sulla popolazione, il comune rappresenta il livello territoriale di selezione delle famiglie residenti. Tuttavia, nel caso che il disegno di campionamento preveda, come penultimo stadio, una diversa unità, i trasferimenti dovrebbero essere riferiti a quest'ultima.

Nell'eventualità che vengano previste sostituzioni delle unità non rispondenti, deve essere riportato nel modello un codice identificativo, che permetta di istituire un legame tra unità sostitutiva e sostituita; inoltre, nel questionario, le famiglie sostitutive devono essere identificate mediante un codice, per renderne possibile l'individuazione in sede di elaborazione dei dati.

È opportuno riportare su supporto informatico, i documenti aggiuntivi, per facilitare i controlli nella fase di revisione quantitativa e per il calcolo degli indicatori di qualità.

APPENDICE

1. Esempi di diagrammazione del questionario

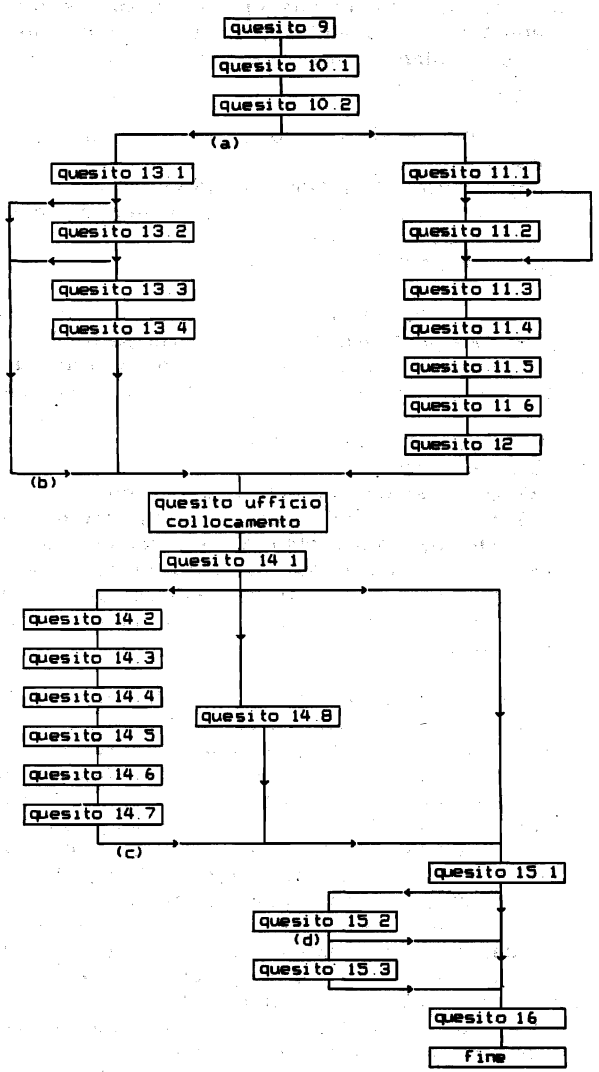


Figura 2.4 - Diagramma della sequenza dei quesiti per il questionario individuale dell'indagine forze di lavoro

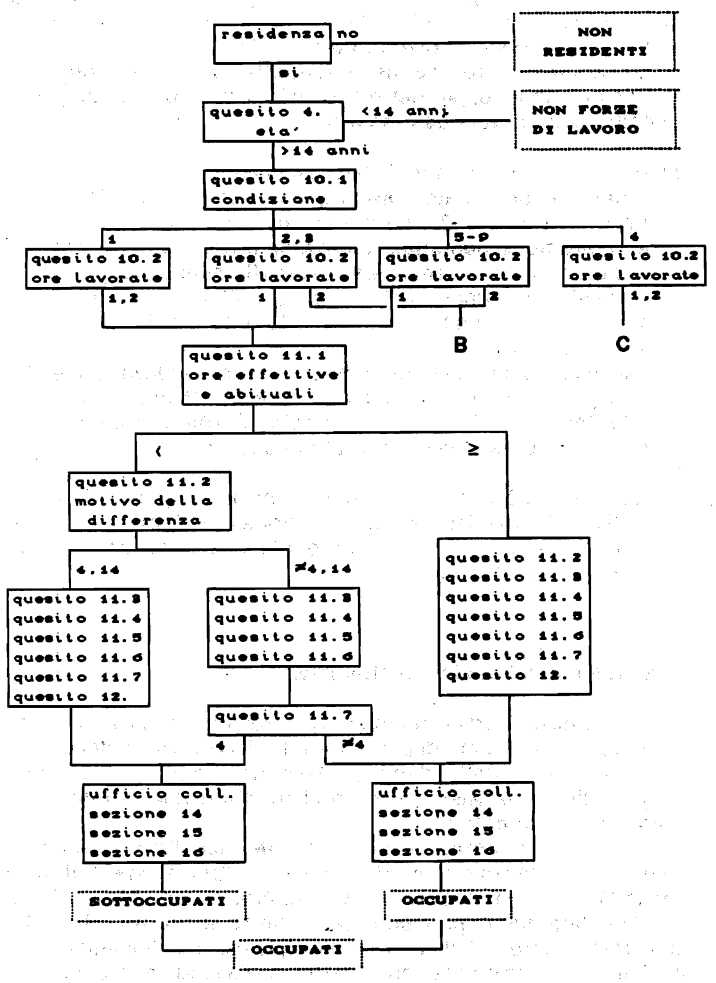


Figura 2.5 - Diagramma della sequenza dei quesiti necessaria per individuare la condizione di «occupato» nell'indagine forze di lavoro

A titolo di esempio si riporta, nella figura 2.4, la diagrammazione del flusso dei quesiti del questionario individuale utilizzato nell'indagine sulle forze di lavoro 1988.

La rappresentanzione, nella suddetta forma, delle relazioni interne del modello, permette di individuare alcune incongruenze nella stesura del questionario:

(a) non esiste un rimando diretto dal quesito 10.2 al quesito 13; il salto è specificato solo in quest'ultima domanda;
(b) nei quesiti 13.1 e 13.2 si rimanda al quesito 14, cosicché non è chiaro se la domanda relativa all'ufficio di collocamento debba o meno rimanere escluso dalla sequenza;
(c) non esiste un rimando diretto dal quesito 14.7 alla domanda 15; il salto è possibile solo dopo la lettura delle avvertenze al quesito 14.8;
(d) esiste un'apparente contraddizione tra l'avvertenza al questito 15.1 (se «SI» rispondere sia al 15.2 che al 15.3) ed il rimando alla domanda 16 dopo la 15.2; tale contraddizione viene rimossa solo dall'avvertenza al quesito 15.3

Nella figura 2.5, invece, la diagrammazione riguarda il flusso dei quesiti che caratterizza la condizione di OCCUPATO; schemi simili possono essere ottenuti per le altre condizioni indicate dalle «uscite» B e C.

2. Il sistema dei codici identificativi

Il sistema dei codici identificativi, associato ad una rilevazione statistica, ha lo scopo di permettere la ricostruzione del contenuto informativo dell'indagine stessa; nell'ambito del Sistema di Controllo, è alla base dei controlli quantitativi nella fase di revisione (cfr. Capitolo 5).

Il controllo quantitativo del materiale raccolto riguarda non solo la consistenza quantitativa in senso stretto (ad es., il fatto che in ciascun comune sia stato intervistato un numero di famiglie uguale a quello previsto) ma anche la ricostruibilità dei legami tra informazioni di natura diversa (ad es., in un'indagine sulle famiglie e sui loro componenti, che le notizie relative ad ogni componente siano associabili alle notizie relative alla famiglia di appartenenza). La mancata esecuzione di questo secondo tipo di controllo, infatti, renderebbe difficoltosi i successivi controlli logici (ad es., che ogni famiglia abbia un capofamiglia).

La controllabilità di entrambi questi aspetti dipende dal modo in cui è definito il sistema dei codici associato al file contenente i dati provenienti dall'indagine, una volta posti su supporto informatico.

Nel seguito ci soffermeremo su come vengono definiti tali codici, a partire dalla considerazione dei codici che compaiono sul materiale cartaceo (in pratica, sul modello di rilevazione), descrivendo prima l'insieme dei codici che possono teoricamente essere definiti, e quindi illustrando come la rappresentazione fisica adottata per i dati su supporto informatico determini l'insieme dei codici praticamente presenti nel materiale registrato.

Come è noto, in un'indagine statistica a ciascuna unità di rilevazione di un dato tipo (ad es., a ciascuna famiglia, o a ciascuna impresa) corrisponde un modello di rilevazione (ad es., il foglio di famiglia). In ogni modello compaiono uno o più codici identificativi, composti generalmente di sottocodici: ad esempio, in un modello potrà comparire un codice così composto: codice Istat comune + numero d'ordine del questionario, ed eventualmente un secondo codice: codice Istat comune + codice area di circolazione + codice rilevatore + numero d'ordine del questionario.

Scopo di questi codici è:

— identificare il singolo modello, permettendo di conteggiare le unità rilevate (ad esempio le famiglie);
— permettere l'associazione tra unità rilevate e loro raggruppamenti significativi, definiti nella fase di progettazione dell'indagine: strati o grappoli di unità definiti nel piano teorico di campionamento (è il caso di associazioni come famiglia-comune, comune-strato di comuni), raggruppamenti territoriali (ad es., famiglia-sezione di censimento), raggruppamenti d'interesse per le successive analisi statistiche sui dati;
— associare al modello il rilevatore e altri organi della rilevazione.

Come già osservato, nei dati registrati sono presenti altri codici, oltre ai codici associati a ciascun modello di rilevazione, ai quali è affidata la ricostruibilità dell'informazione complessivamente portata dal modello.

Il modello di rilevazione, infatti, contiene informazioni su una o più unità di analisi (ad es., componenti della famiglia), ed eventualmente su uno o più fenomeni osservati (ad esempio le vacanze, o i ricoveri ospedalieri) che, dopo la fase di registrazione, si trovano organizzate in uno o più file. Generalmente nei piani di elaborazione dei dati vengono definiti diversi tipi di file, ciascuno riportante informazioni di tipo diverso. Ad esempio, un'indagine sui consumi delle famiglie potrà dare luogo dal punto di vista informatico a tre file: dati sulle famiglie, sulle persone, sulle spese mensili delle famiglie. In tal caso, le informazioni contenute in ciascun modello si troveranno suddivise su più record:

nell'esempio, ciascun modello relativo ai consumi di una famiglia corrisponderà a tre insiemi di record dei tre tipi suddetti: un record con i dati sulla famiglia, alcuni record con i dati su ciascuna persona, altri con i dati su ciascun capitolo di spesa.

La ricostruzione dell'informazione contenuta nell'intero modello è possibile solo associando tra loro i diversi record mediante i codici presenti su di essi: ciò rende determinante l'esatta definizione e il controllo di tali codici.

I distinti tipi di file che possono essere definiti per ciascuna indagine corrispondono a ben identificabili sottoinsiemi delle informazioni raccolte nel modello, e precisamente agli insiemi di informazioni relativi a distinte unità d'analisi e fenomeni. I codici che dovranno essere definti dipenderanno perciò dai legami logici sussistenti tra tali sottoinsiemi di informazioni.

Quanto detto evidenzia come, per poter definire il sistema dei codici, devono essere individuati e descritti tutti gli insiemi di informazioni concettualmente distinti che compongono l'indagine, e i legami logici tra di essi.

In pratica si tratta di individuare, per ciascuna indagine:

a) le aggregazioni d'interesse delle unità rilevate, che determinano la composizione del/dei codici di modello;
b) i sottoinsiemi distinti di informazioni logicamente individuabili nel modello, che potranno trovarsi distinti fisicamente, dal punto di vista informatico, in diversi file, a ciascuno dei quali corrisponderà un codice dipendente dal codice di modello;
c) i legami tra tutti questi tipi di unità, che potranno trovarsi rappresentati fisicamente da codici di corrispondenza presenti sui diversi record.

Questa attività deve essere svolta dal responsabile dell'indagine nella fase di progettazione della stessa, e sarà facilitata dall'uso di metodologie di progettazione concettuale già in tale fase.

Gli insiemi di informazioni componenti un'indagine statistica

Nella progettazione di un'indagine statistica, vengono definiti:

— il collettivo o i collettivi che si intende osservare (ad es., persone, famiglie, imprese). Questi definiscono la/le unità d'analisi principali dell'indagine. Può avere interesse, per chi progetta l'indagine, condurre analisi statistiche su raggruppamenti di tali unità d'analisi, che possiamo considerare unità d'analisi di tipo aggregato (ad es., famiglie come aggregazioni dei loro componenti). Inoltre, possono essere definiti uno o più fenomeni d'interesse specifico dell'indagine (ad es., sport e vacanze degli italiani, lavoro);

— il piano teorico di rilevazione, nel quale vengono definiti, ai fini della rilevazione, diversi livelli di raggruppamento di unità d'analisi: ai livello più basso, la/le unità di rilevazione (ad es., famiglie e convivenze possono costituire unità di rilevazione rispetto all'unità d'analisi persone), ai livelli superiori tutti i raggruppamenti richiesti dal disegno campionario oppure di tipo territoriale: strati o grappoli di unità di rilevazione, aree di rilevazione, sezioni di censimento;
— l'organizzazione sul campo, nella quale vengono definiti gli organi della rilevazione, ai quali è demandata l'effettuazione pratica delle operazioni di raccolta dei dati: rilevatori, Comuni, UPS, ecc..

Quindi, in una generica indagine si possono individuare, *quattro tipi di insiemi di oggetti enumerabili*:

1) *unità d'analisi*, quelle principali e quelle ottenibili come risultato di aggregazioni;
2) *fenomeni, o eventi, osservati*
3) *unità di rilevazione, e loro raggruppamenti, territoriali e no*, che abbiano significato statistico con riferimento ai piani teorici di rilevazione: *strati, aree di rilevazione, sezioni di censimento*
4) *organi della rilevazione*, che compongono l'organizzazione sul campo.

Nell'indagine potranno essere definite una o più unità d'analisi, uno o più fenomeni osservati, una o più unità di rilevazione, uno o più organi della rilevazione.

Mentre si rimanda per una più completa trattazione delle unità di rilevazione e unità di analisi al cap. 2 del Manuale di tecniche d'indagine, Fascicolo 1, è opportuno precisare, in questo contesto, alcuni concetti:

— con il termine *unità di analisi* si definisce l'insieme di elementi che compone il collettivo che interessa osservare ai fini dell'indagine: individui, imprese, ecc., in pratica il collettivo al quale sono riferite le notizie raccolte con l'indagine;
— chiamiamo anzitutto *unità di rilevazione* l'insieme di elementi sul quale vengono raccolti i dati: a ciascun elemento di questo insieme corrisponderà un modello di rilevazione (nelle indagini su popolazione e famiglie si tratterà in genere dell'insieme delle famiglie, ma potrà in alcuni casi trattarsi dell'insieme degli individui). Nelle rilevazioni a più stadi vengono poi definiti insiemi di unità che costituiscono grappoli delle

corrispondenti al modello: chiameremo anche questi insiemi unità di rilevazione. Ad esempio, in un'indagine campionaria a due stadi, con unità di primo stadio i comuni e di secondo stadio le famiglie, le famiglie costituiscono l'unità di rilevazione di più basso livello, corrispondente al modello, e sono considerate raggruppate in comuni: in questo caso chiameremo unità di rilevazione tanto le famiglie che i comuni. Infine, considereremo in questo terzo gruppo di insiemi di oggetti enumerabili anche tutti i raggruppamenti delle diverse unità di rilevazione che in una data indagine è possibile definire, e cioè: strati e, a livello territoriale, aree di circolazione o sezioni di censimento.

Un insieme di elementi può essere contemporaneamente unità di rilevazione e di analisi: se, in un'indagine su popolazione e famiglie, la famiglia è unità di rilevazione e al tempo stesso su di essa vengono richieste notizie, essa costituirà anche un'unità di analisi.

In generale *uno stesso insieme di oggetti può comparire in un'indagine con diverse funzioni*: il comune, o l'area di rilevazione, può comparire tanto come un raggruppamento di unità d'analisi (ad es., di famiglie) rilevante per l'analisi statistica, quanto come un organo della rilevazione. In questo caso, il relativo codice può comparire come componente di diversi codici di modello. Per quanto detto, a ciascuna unità di rilevazione e di analisi, e a ciascun organo della rilevazione, corrisponde un insieme enumerabile di elementi (ad es., l'insieme dei comuni, l'insieme delle imprese, o dei componenti le famiglie, l'insieme dei rilevatori).

Per ciò che riguarda i *fenomeni* osservati, si fa notare che in alcuni casi il fenomeno osservato in una data indagine è definito unicamente come una caratteristica delle unità di analisi, mentre in altri casi dà luogo ad un evento che, sia pure osservato sulle unità d'analisi, può essere descritto e soprattutto enumerato indipendentemente: un esempio del primo caso è il lavoro (non può definirsi un evento «lavoro» conteggiabile indipendentemente dalle persone), esempi del secondo caso sono le vacanze, i ricoveri ospedalieri, ecc. Perciò non in tutte le indagini esisteranno classi enumerabili del secondo tipo sopra elencato, corrispondenti cioè a fenomeni (eventi).

Per ciò che riguarda i *legami logici* che possono esistere tra le classi di oggetti dei diversi tipi, conviene rappresentarli con lo schematismo grafico adottato nella seguente Figura 2.6. In essa ciascuna classe di oggetti enumerabili è rappresentata con un nodo, e i legami di corrispondenza tra le diverse classi di oggetti con archi che collegano i corrispondenti nodi.

L'unità di rilevazione di più basso livello, alla quale corrisponde il modello, è indicata con un nodo pieno. Per semplicità non sono stati indicati i raggruppamenti di unità di rilevazione e gli organi della rilevazione. Un singolo arco indica una relazione 1-n tra il nodo di sinistra e quello di destra, un doppio arco una relazione m-n. Come si osserva nella Figura 2.6, nelle indagini statistiche sono generalmente definite catene di relazioni 1-n tra le unità di rilevazione, da queste alle diverse unità d'analisi e da queste agli eventi. Relazioni m-n possono sussistere tra le diverse unità d'analisi e tra unità d'analisi ed eventi osservati su una unità d'analisi diversa.

Questo tipo di rappresentazione è particolarmente utile per visualizzare il meccanismo di costruzione dei codici identificativi.

Legami logici tra classi di oggetti e codici identificativi

I *legami logici* tra le classi di oggetti enumerabili che compongono l'indagine, determinano il modo in cui è costruito il codice identificativo attribuito agli elementi di ogni classe di oggetti: la costruzione dei codici identificativi è basata su un meccanismo di *propagazione dei codici* che tiene conto di questi legami logici.

In generale, il meccanismo di propagazione dei codici in un'indagine statistica è il seguente: i codici si propagano lungo le catene di relazioni 1-n, rappresentate nella Figura 2.6, e perciò dal raggruppamento di più alto livello, a quello di livello più basso, fino all'unità di rilevazione di più basso livello, corrispondente al modello, quindi da questa alle unità d'analisi e da queste agli eventi.

Ciò vuol dire, ad esempio, che il codice identificativo dell'unità d'analisi sarà composto dai codici identificativi delle diverse unità di rilevazione (ad esempio, in un'indagine a due stadi sulla popolazione, con unità di rilevazione comuni e famiglie, il codice identificativo del singolo individuo dovrà comprendere il codice di comune e il codice della famiglia). Ad ogni passaggio lungo i legami rappresentati viene aggiunta una componente del codice di livello più basso, che serve a distinguere gli oggetti della data classe di oggetti (si tratterà in pratica di un numero d'ordine).

Ad esempio si avrà, nella situazione descritta:

Codice comune: preso da liste esterne;

Codice famiglia: codice comune + numero d'ordine famiglia (numero di questionario);
Codice individuo: codice famiglia + numero d'ordine individuo;
Codice fenomeno: codice famiglia + numero d'ordine fenomeno;
oppure
codice individuo + numero d'ordine fenomeno

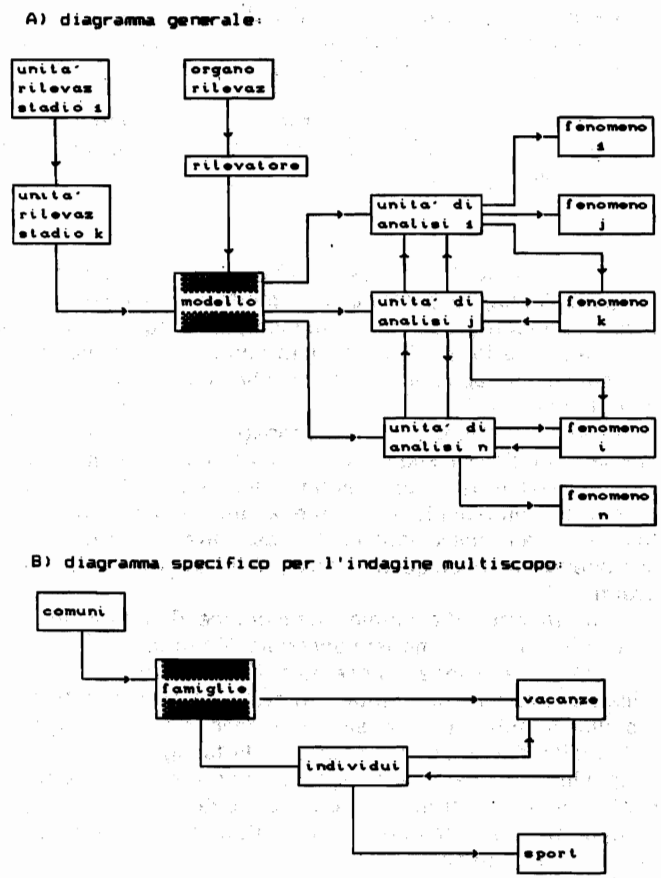


Figura 2.6 - Diagramma del meccanismo di propagazione dei codici identificativi in una indagine statistica

secondo l'unità d'analisi sulla quale è osservato l'evento, famiglia oppure individuo (ad esempio, il fenomeno «spese» sarà osservato sulle famiglie, il fenomeno «ricoveri ospedalieri» sugli individui).

Va osservato che tale propagazione è possibile proprio perché le relazioni comune-famiglia, famiglia-individuo, etc. sono 1-n tra il membro di sinistra e quello di destra.

A ciascuna classe di oggetti potrà essere attribuito più di un codice, sempre costruito nello stesso modo: ad esempio, la famiglia potrà avere un secondo codice, utile per controllare l'organizzazione sul campo, composto da codice comune + codice rilevatore + numero d'ordine della famiglia.

Rappresentazione logica e rappresentazione fisica dei dati

A ciascuna classe di oggetti enumerabile di uno dei tipi descritti corrisponderà un *insieme di dati logicamente distinto*, con un proprio codice identificativo costruito come descritto.

Dal punto di vista informatico *ciascuno di questi insiemi di dati logicamente distinti potrà dare luogo o meno ad un file separato*. Nel caso positivo avremo un file, e perciò un tipo-record differente, per ogni insieme di un dato tipo: ad es., in un'indagine a due stadi sulle vacanze effettuate dai componenti le famiglie, saranno definiti a livello fisico un file per i comuni, uno per le famiglie, uno per i componenti, uno per le vacanze. In caso contrario, in uno stesso tipo-record potranno trovarsi rappresentati insiemi di oggetti logicamente distinti: nell'esempio presentato sopra, uno stesso tipo-record potrà contenere le notizie sulle famiglie e i loro componenti, o sui componenti le famiglie e le loro vacanze.

Di conseguenza, *gli insiemi di codici effettivamente presenti nel materiale registrato, sui quali verranno svolti tutti i controlli quantitativi, dipendono tanto dai legami logici tra le diverse classi di dati* di ciascuno dei tipi suddetti presenti nella data indagine, *quanto dal tipo di rappresentazione fisica* adottata per esse a livello informatico.

I *legami logici* che possono esistere tra le classi di oggetti di vario tipo sono quelli rappresentati schematicamente nelle figure in allegato e che, come già osservato, determinano il meccanismo di formazione dei codici.

Circa la *rappresentazione fisica* adottata per le diverse classi di oggetti, consideriamo per fissare le idee i due casi estremi:

— ad ogni classe di oggetti corrisponde un tipo-record fisico;
— tutte le classi di oggetti sono contenute in campi multipli di uno stesso record fisico.

Nel *primo caso*, per ogni tipo-record corrispondente ad una data classe di oggetti, sarà definito un identificativo composto come descritto, a livello logico, nel paragrafo precedente.

Occorre osservare, però, che nelle applicazioni statistiche non ha interesse, in genere, poter selezionare il singolo record di ciascun tipo-record. L'unica funzione degli identificativi è quindi quella di consentire il *puntamento tra i singoli record di tipi record diversi*, lungo le catene descritte a livello logico nel paragrafo precedente: in pratica, poter associare a ciascun record corrispondente ad una unità di rilevazione i relativi record corrispondenti a unità d'analisi e a questi i record corrispondenti a eventi osservati. Il controllo di questi puntamenti sarà realizzato attraverso il controllo della corrispondenza tra le diverse componenti dei codici nei diversi tipi-record (ad es., codice comune e codice modello dovranno essere gli stessi sul tipo record «famiglia» e sul tipo record «individuo»).

Va notato che una conseguenza dell'osservazione precedente, sull'irrilevanza nelle applicazioni statistiche della possibilità di selezione di un singolo record, è che il tipo-record corrispondente alla classe di oggetti situata alla fine della catena (tipicamente un evento) può anche non avere un identificativo di più basso livello (un numero d'ordine). Lo avrà nei casi in cui è necessario rappresentare un puntamento tra tipi record indipendente dalla catena di propagazione dei codici e cioè una relazione, del tipo rappresentato nelle figure dal doppio arco, tra un nodo unità d'analisi e un nodo fenomeno: questo caso rientra nel controllo dei puntamenti tra classi di oggetti (vedi oltre).

Si osserva ancora che anche il *numero* di record, di un certo tipo record, effettivamente associati ad un altro tipo record dovrebbe essere controllato (nel corso del controllo sui codici o, nel caso in cui non esistono i numeri d'ordine, mediante conteggio dei record fisici), in tutti i casi nei quali esistono campi che riportano questa informazione, per esempio a scopo di controllo (ad es. esiste nel tipo-record «famiglia» un campo «numero dei componenti»).

Infine, ovviamente, tutti i tipi record dovranno avere un identificativo di più basso livello, da controllare, in quei casi, eccezionali rispetto alla norma, nei quali interessa comunque trattarli individualmente.

Per quel che riguarda il secondo caso (classi di oggetti contenute in campi multipli di uno stesso record fisico), una sua esemplificazione concreta si ha quando esiste un unico tipo-record associato all'unità di rilevazione finale, o all'unità d'analisi, all'interno del quale la classe di oggetti corrispondente all'unità d'analisi, o, rispettivamente, all'evento osservato è rappresentata fisicamente come un campo multiplo. In questo caso è evidente che *l'associazione tra classi di oggetti è realizzata a livello fisico*, e non è necessario definire identificatori di più basso livello (numeri d'ordine) e controllare puntamenti. Valgono comunque anche in questo caso le tre osservazioni, fatte in precedenza, circa i casi in cui può darsi la necessità di definire un codice identificativo di più basso livello o comunque di controllare il numero di campi (controllo di puntamenti tra tipi record indipendenti dalla catena di propagazione dei codici, esistenza di campi che riportano informazioni sulla numerosità di altri campi, interesse per il trattamento del dato individuale).

Le situazioni reali circa la rappresentazione fisica delle classi di oggetti costituiranno combinazioni delle due situazioni estreme sopra descritte, e risulteranno determinati di conseguenza i tipi di controlli quantitativi da effettuarsi.

Come anticipato, può essere necessario in un'indagine rendere possibili, attraverso la definizione di appropriati codici di corrispondenza, *puntamenti tra classi di oggetti*.

Può avvenire cioè che, indipendentemente dalla catena di propagazione dei codici sopra descritta, possano esistere altri legami logici tra classi di oggetti, rappresentati nelle figure con doppi archi perché di tipo m-n: è il caso del legame tra individui e vacanze, «componenti la famiglia che hanno partecipato ad una vacanza». Questi saranno sempre *rappresentati a livello fisico da puntamenti*, cioè da corrispondenze tra codici associati a tipi-record o a campi di uno stesso record, qualsiasi sia la rappresentazione fisica adottata per le classi di oggetti collegate. Tali puntamenti tra classi di oggetti dovranno naturalmente essere sempre controllati.

RIFERIMENTI BIBLIOGRAFICI

AA.VV. (1985), *Special issue on questionnaire design*, Journal of Official Statistics, n. 2, Statistics Sweden.

BARCAROLI G., FORTUNATO E., MAGALOTTI A., MANICARDI G., VACCARI C. (1987), *Manuale per la progettazione concettuale di dati statistici*, Istat.

BARCAROLI G., D'ANGIOLINI G. (1988), *La progettazione concettuale dei sistemi informativi statistici*, Istat, documento interno.

MASSELLI M., DE MARCHIS M.A., SIGNORE M., DI PIETRO E. (1988), *Obiettivi e metodi di controllo dell'indagine pilota - Indagine sulla storia lavorativa*, documento interno Istat.

NARGUNDKAR M.S., PLATEK R. (1989), *Qualitative methods in questionnaire design*, I.S.I. Proceedings of 47th session, Paris.

WORLD FERTILITY SURVEY (1980), *Basic Documentation: n. 4*.

CAPITOLO 3 · LA RILEVAZIONE SUL CAMPO

1. La fase di rilevazione sul campo

La fase di rilevazione sul campo include la raccolta delle informazioni presso il rispondente e tutte le operazioni a questa accessorie, realizzate dall'organizzazione periferica: la manutenzione e l'approntamento delle liste o di altri moduli organizzativi della rilevazione (ad esempio la suddivisione del territorio in aree), la selezione delle unità, la predisposizione delle assegnazioni dei rilevatori e dei documenti ausiliari di rilevazione, la scelta e l'istruzione dei rilevatori, la pubblicizzazione locale dell'indagine, l'attività di supervisione, la codifica dei quesiti aperti.

La caratteristica principale di questa fase, è che essa, al contrario delle altre, è solo parzialmente sotto il diretto controllo dell'Istituto e che non è possibile modificare significativamente, durante lo svolgimento delle operazioni, le procedure prestabilite. Ciò comporta che la *qualità* dei risultati delle operazioni sul campo, è strettamente dipendente dalle scelte operate nella fase di progettazione e riguardanti il questionario, le norme, i manuali di istruzione, il calendario e l'organizzazione; in questa fase, quindi, si riflettono le insufficienze derivanti da quella di programmazione.

In aggiunta agli errori *importati* dalla fase precedente, quella di rilevazione ne genera di propri. Essi rappresentano la gran parte dell'errore totale e sono imputabili al comportamento delle unità coinvolte: gli organi di rilevazione e di supervisione, i rilevatori ed i rispondenti.

Una parte dei suddetti errori, che possono essere considerati *di misura* in senso lato, può essere identificata e corretta nella fase di revisione; una seconda parte, invece, non è individuabile mediante le usuali analisi del materiale raccolto, ma solo per mezzo di opportune tecniche di stima e non è, generalmente, suscettibile di correzione.

2. Gli errori di rilevazione ed i loro effetti

Gli errori che possono verificarsi nella fase di rilevazione sul campo, sono quelli derivanti dalla tenuta degli archivi e delle liste, dalle operazioni di selezione delle unità campionarie, dalla compilazione dei documenti di rilevazione e dall'assegnazione di codici di identificazione dalle unità, dalle unità non rispondenti e dalla misurazione non corretta dei fenomeni oggetto di studio,

dall'eventuale codifica dei quesiti non precodificati. Gli errori di lista, di selezione e di identificazione delle unità, sono imputabili alle strutture organizzative periferiche, responsabili degli archivi e delle operazioni sul campo, mentre gli altri sono ascrivibili al complesso delle relazioni che si instaurano, al momento del primo contatto e dell'intervista, tra intervistatore e rispondente, nonché al tipo di assistenza e di controllo esercitati dal supervisore sul gruppo di rilevatori affidatogli.

In particolare, durante l'intervista, gli errori possono derivare dai seguenti fattori: l'influenza dei rilevatori sui rispondenti, la reticenza, la difficoltà a ricordare, la scarsa motivazione a rispondere, la mancanza di informazioni dei rispondenti, il condizionamento su questi esercitato dalla presenza di terze persone, la raccolta da altra unità delle informazioni riguardanti quella designata ed, infine, la disattenzione e la trascuratezza nella compilazione del questionario.

Il complesso degli errori di misura influenza i risultati finali, inducendo, nelle stime, distorsioni o variabilità aggiuntiva rispetto a quella propria del fenomeno.

Nei successivi paragrafi, verranno trattati gli effetti, il controllo e la correzione dei principali tra i suddetti tipi di errore.

Gli errori derivanti dai quesiti retrospettivi e dalla risposta proxy sono propri della fase di rilevazione sul campo; tuttavia si è preferito, in un'ottica *operativa*, trattarli nella fase di progettazione (Capitolo 2), dato che la loro prevenzione richiede tecniche e regole specifiche nella predisposizione del questionario. Nel paragrafo 4 del presente capitolo, invece, sono succintamente riportate le tecniche che ne permettono la stima, che è stata inclusa tra i parametri di controllo della fase di rilevazione sul campo.

Gli errori di selezione e di lista

La *lista* costituisce il supporto fisico contenente l'elenco delle unità di rilevazione e le informazioni necessarie alla loro individuazione; per suo tramite è quindi possibile procedere alla selezione delle unità, nel caso d'indagine campionaria, ed alla loro rilevazione sul campo.

Spesso, la lista funge anche da *archivio* di informazioni, utilizzato per il conteggio delle unità (ad esempio le anagrafi per valutare la consistenza della popolazione) o per il loro raggruppamento in strati (ad esempio la codifica della circoscrizione amministrativa).

Rispetto alla popolazione obiettivo, definita nella fase di progettazione, la lista dovrebbe risultare:

— completa, ovvero dovrebbe contenere tutte e solo le unità di rilevazione designate;

— aggiornata, ovvero non dovrebbe riportare duplicazioni e ciascuna unità dovrebbe essere distinguibile dalle altre ed individuabile sul territorio;

— informativa, ovvero dovrebbe contenere per ciascuna unità le caratteristiche stabilite dal piano di rilevazione o di campionamento (ad esempio la dimensione della famiglia per procedere ad una sostituzione o le variabili necessarie alla stratificazione dei comuni).

Per la singola indagine, si può costruire una lista ad hoc, oppure può essere utilizzato un elenco di unità preesistente (ad esempio i registri anagrafici o le liste elettorali per le famiglie e gli individui).

La predisposizione di una lista, mirata all'indagine ed esente da errori, o quantomeno con errori casuali e di lieve entità, risulta estremamente dispendiosa in termini economici e di tempo; in pratica si tratterebbe di effettuare un censimento negli ambiti territoriali designati (ad esempio, nel caso tipico di una indagine campionaria a due stadi sulla popolazione, occorrerebbe effettuare il censimento delle famiglie nei comuni campione).

Risulta, quindi, più economico e pratico utilizzare liste già esistenti, anche se esse raramente si adeguano ai succitati requisiti. In pratica, le liste disponibili vengono aggiornate con cadenze proprie, diverse dalla data di riferimento delle indagini, e riportano non solo le unità oggetto di rilevazione, ma anche duplicazioni ed unità non incluse nella popolazione di riferimento, mentre non contengono una quota di quelle designate.

Le unità oggetto di rilevazione, sono dette *includibili* e vengono designate mediante un *criterio di includibilità*, che deriva dalla definizione della popolazione di riferimento.

Gli errori di lista possono causare distorsioni nei risultati finali, in funzione (I) della distribuzione e della quota delle unità mancanti, (II) dal tipo di indagine condotta (esaustiva o campionaria) e (III) dall'uso della lista come archivio per il calcolo dei coefficienti di espansione.

Il mancato aggiornamento contribuisce alla mancata risposta totale mediante l'irreperibilità sul campo delle unità, dovuta ad erroneo indirizzo, ed incrementa le duplicazioni e la quota delle unità non includibili incluse e di quelle includibili non incluse.

Le unità includibili mancanti non hanno alcuna possibilità di essere selezionate ed intervistate; se esse sono distribuite casualmente nella popolazione di riferimento, la struttura della lista rispecchierà quella dell'universo, in caso contrario, invece,

subpopolazioni particolari sfuggiranno alla rilevazione. Tali unità possono essere, in pratica, difficilmente individuate, in quanto richiedono od un confronto con altri archivi più aggiornati od un censimento effettuato sulla medesima base territoriale di riferimento della lista; a tutti gli effetti, esse possono essere considerate come mancate risposte totali.

Le unità non includibili incluse, possono essere individuate al momento della selezione, depurando la lista dalle unità che non rispondono al criterio di includibilità, oppure nel corso della rilevazione sul campo, esplicitando il suddetto criterio in un quesito filtro del questionario; le duplicazioni possono essere colte verificando la lista o, a posteriori, confrontando le assegnazioni dei rilevatori.

Nelle indagini totali, tali unità possono essere completamente individuate, cosicché questi errori non producono alcun effetto sui risultati finali.

Nel caso delle indagini campionarie, invece, l'individuazione sarà limitata alla quota di unità rilevate e ciò non è sufficiente ad eliminare gli effetti degli errori di lista. Questi si riflettono sulla probabilità di inclusione nel campione e sulla stima dei coefficienti di espansione all'universo, se la lista viene anche utilizzata come archivio. Inoltre, se le unità includibili vengono individuate nel corso della rilevazione, la presenza di unità non includibili comporta la riduzione della numerosità campionaria programmata. Il ripristino della dimensione desiderata, può avvenire mediante la sostituzione delle unità non includibili, oppure sovradimensionando il campione di una quota di tali unità, stimata da fonti esterne.

Infine, nelle indagini campionarie, anche l'inosservanza delle norme di estrazione delle unità, od una loro insufficiente articolazione, può causare distorsioni. Mediante l'operazione di selezione dalle liste, infatti, si assegnano alle unità prefissate probabilità di inclusione nel campione; qualsiasi intervento estraneo alle norme di estrazione (ad esempio la tendenza a *saltare* le famiglie numerose in favore di quelle di ridotta dimensione), causa una distorsione nel meccanismo probabilistico del disegno di campionamento.

Le mancate risposte totali

Le mancate risposte totali sono costituite dalle unità di rilevazione per le quali non è stato possibile raccogliere informazioni nel corso dell'intervista per cause diverse: errori di lista, incapacità di convincimento da parte del rilevatore, rifiuto o impossibilità di reperimento.

Le mancate risposte totali producono, sui risultati finali, due

effetti: da un lato riducono la numerosità campionaria e quindi incrementano il relativo errore di campionamento, dall'altro inducono distorsioni nelle stime, se il meccanismo che le genera è, come avviene generalmente nella realtà, non casuale.

In questo caso, possiamo concettualmente dividere la popolazione in due strati, i rispondenti ed i non rispondenti, di numerosità diversa e con differenti caratteristiche (ad es. medie, totali, proporzioni etc.); la distorsione è allora funzione della quota di non rispondenti e della differenza tra i parametri relativi ai due insiemi.

A titolo di esempio si consideri il caso della stima della media della generica caratteristica Y.

La popolazione, costituita da N unità con media pari a $\bar{Y}$, può essere divisa in due strati: quello relativo alle unità non rispondenti, di numerosità N_{NR}, e quello dei rispondenti, di numerosità N_R. Siano quindi $W_R = N_R/N$ e $W_{NR} = N_{NR}/N$ i pesi delle due subpopolazioni, $\bar{Y}_R$ e $\bar{Y}_{NR}$ le rispettive medie.

L'estrazione di un campione dalla popolazione si riduce, in realtà, alla selezione di unità appartenenti alla subpopolazione dei rispondenti; la stima $\bar{y}_R$ ottenuta dai dati campionari sarà, allora, corretta ripetto a tale sub universo, ovvero $E(\bar{y}_R) = \bar{Y}_R$, ma presenterà una distorsione, rispetto alla media complessiva, data dalla:

$$B(\bar{y}_R) = E(\bar{y}_R - \bar{Y}) = E(\bar{y}_R - W_R\bar{Y}_R - W_{NR}\bar{Y}_{NR})$$

$$= W_R(\bar{Y}_R - \bar{Y}_{NR}) \qquad (3.1)$$

La (3.1) mostra che la distorsione B è funzione della quota dei rispondenti nella popolazione e della differenza tra le medie delle due subpopolazioni; essa è valida anche nel caso di indagini totali.

L'errore totale dello stimatore media, calcolato mediante la (3.2), è la risultante della somma della varianza dello stimatore, relativo al sub universo dei rispondenti, $V_R(\bar{y}_R)$ e del quadrato della distorsione:

$$MSE(\bar{y}) = E(\bar{y}_R - \bar{Y})^2 = E(\bar{y}_R - W_R\bar{Y}_R - W_{NR}\bar{Y}_{NR})^2$$

$$= V_R(\bar{y}_R) + B^2(\bar{y}_R) \qquad (3.2)$$

La stima dell'errore di campionamento, basata sui risultati campionari, è anch'essa distorta.

Sia infatti v_R la stima della varianza campionaria di $\bar{y}_R$; essa sarà corretta, ma solo rispetto al parametro della popolazione di riferimento, ovvero:

$$E(v_R) = V_R(\bar{y}_R)$$

La distorsione di v_R, data dalla,

$$\begin{aligned} B(v_R) &= E(v_R - V) = E(v_R - W_R^2 V_R - W_{NR}^2 V_{NR}) \\ &= (1 - W_R^2) V_R - (1 - W_R)^2 V_{NR} \\ &= W_{NR}\left[(1+W_R) V_R - (1-W_R) V_{NR} \right] \end{aligned} \qquad (3.3)$$

risulterà negativa, quindi con una sottostima della varianza $V(\bar{y})$, per:

$$V_R / V_{NR} < W_{NR} / (1+W_R)$$

Le mancate risposte parziali

Per mancate risposte parziali si intende l'assenza di risposta ad uno o più quesiti, nell'ambito del questionario compilato, che può derivare da rifiuto od incapacità a rispondere.
Ad esse possiamo assimilare i valori non ammissibili e le incongruenze logiche tra risposte a quesiti differenti. Queste ultime si manifestano come contraddizioni, rispetto a relazioni *sostanziali* o *formali*, tra i valori assunti dalle relative variabili. Le relazioni formali sono date dalle norme di compilazione del questionario, mentre quelle sostanziali sono relazioni tra variabili, implicite nella realtà esaminata. Al verificarsi di una incoerenza tra quesiti, ed in assenza di un criterio oggettivo per stabilire quale risposta sia vera e quale falsa, dovremmo considerare tutte le variabili coinvolte come informazioni mancanti.

Ai fini della correzione, le diverse tipologie di mancate risposte parziali possono essere trattate alla stessa stregua; dal punto di vista del controllo, invece, i tre gruppi presentano significati diversi.

I valori fuori campo, nel caso di risposte non precodificate, possono essere prevalentemente addebitati al rilevatore o all'eventuale codificatore, mentre le incoerenze logiche tra variabili ed i rifiuti a rispondere, all'interazione tra questionario, rilevatore e rispondente.

Gli effetti prodotti da tali errori, sono simili a quelli dovuti alle mancate risposte totali, riferiti però alle singole variabili di studio: distorsione e riduzione della numerosità campionaria.

In realtà, la distinzione tra mancata risposta *totale* e *parziale* deriva da una decisione soggettiva del ricercatore, non da parametri oggettivi, validi per qualsiasi rilevazione.

La mancata risposta, infatti, può essere considerata un *continuum*, i cui limiti sono costituiti, da un lato, dalla mancata intervista e, dall'altro, dal modello correttamente compilato in ogni sua parte; la soglia di accettabilità di un questionario viene stabilita in funzione dell'utilità delle risposte fornite per gli obiettivi conoscitivi dell'indagine.

Ad esempio, un questionario in cui tutti i quesiti, tranne quelli *strategici*, siano stati compilati, può essere considerato come mancata risposta totale, perché l'informazione raccolta non è utile agli scopi stabiliti.

Gli errori di codifica

Nel caso di un modello di rilevazione, parzialmente o totalmente, non precodificato, è necessario tradurre l'informazione rilevata in codici alfanumerici, adatti all'elaborazione dei dati. La codifica dei quesiti aperti richiede che il personale addetto interpreti l'informazione riportata in chiaro sul questionario, ricerchi nella classificazione fornitagli l'appropriato codice e lo trascriva negli appositi spazi.

In tali operazioni, possono essere generati errori sia di tipo casuale (ad esempio banali errori di trascrizione), ovvero errori sistematici derivanti dall'insufficiente istruzione impartita o dall'attitudine dei codificatori ad interpretare in modo del tutto personale la classificazione; questi ultimi errori inducono distorsioni nei risultati finali.

Gli errori di identificazione

Durante le operazioni di trascrizione o di apposizione di codici identificativi nei documenti accessori di rilevazione (ad esempio le assegnazioni) o nel questionario, i supervisori, o i rilevatori, possono commettere errori per semplice trascuratezza o per mancata comprensione e rispetto delle norme.

Nelle indagini sulle famiglie, i codici affetti da errori dovuti alla fase di rilevazione, sono, generalmente, gli identificatori a livello subcomunale, per i quali è necessario l'intervento dei supervisori o dei rilevatori. Tali errori si trasmettono alla successiva fase di registrazione e sommandosi a quelli commessi durante tale operazione possono, nei dati su supporto informatico:

— rendere non distinguibili due o più unità e quindi creare una duplicazione (ad esempio individui con lo stesso codice);

— compattare più unità in una sola (ad esempio le famiglie di due comuni vengono riferite ad uno solo);
— suddividere una unità in due o più unità e quindi creare unità fittizie (ad esempio i componenti di una famiglia vengono divisi in due unità familiari);
— invalidare il legame tra unità (ad esempio tra il record della famiglia principale ed il record della famiglia coabitante).

Tali effetti possono manifestarsi singolarmente, o, più frequentemente, in combinazione tra loro, compromettendo, ad esempio, la ricostruzione dell'eventuale struttura longitudinale, la stima delle probabilità di inclusione, la conduzione di reinterviste o causando false assegnazioni nelle compenetrazione del campione.

3. La prevenzione degli errori

La prevenzione degli errori di rilevazione deriva, in parte, da un costante lavoro di controllo e di indirizzo dell'Istituto, in parte è esercitabile al momento della raccolta dei dati della singola indagine. In tal caso, i controlli preventivi devono essere *mirati* alle unità coinvolte: il controllo e l'assistenza agli organi periferici e la pubblicizzazione dell'indagine verso i rispondenti.

Controllo e assistenza agli organi periferici

Il controllo degli organi periferici si attua mediante visite ispettive, il cui scopo è verificare che le differenti operazioni siano state condotte secondo le norme stabilite; esse devono, quindi, venire espletate al momento in cui l'operazione viene svolta (ad esempio al momento della selezione delle unità campionarie).

Un controllo totale, dato il numero di unità coinvolte, può risultare però troppo oneroso; per tale ragione, è utile compilare una *mappa di rischio* ed indirizzare gli sforzi verso quelle realtà che risultano più *sospette*.

La definizione della *mappa* può essere basata su informazioni di tipo sostanzialmente qualitativo (come quelle desunte dai rapporti degli ispettori e degli Uffici Regionali), e sull'analisi delle informazioni quantitative, disponibili da rilevazioni precedenti della stessa indagine o di indagini diverse, cioè sugli indicatori contenuti nell'archivio di qualità dell'indagine e nell'archivio centralizzato della rete.

Dallo stesso archivio è possibile dedurre indicazioni in merito alle prestazioni dei rilevatori, cosicché, laddove sia possibile stante l'attuale normativa, si può intervenire anche nella fase di selezione degli intervistatori.

L'assistenza sul campo viene attuata, in primo luogo, con l'istruzione dei supervisori, dei rilevatori e degli eventuali codificatori, mediante i manuali precedentemente redatti. Questi, in via generale, possono seguire il seguente schema:

- definizione degli obiettivi dell'indagine;
- definizione dell'unità di rilevazione e dei criteri per la sua identificazione;
- la tecnica di primo contatto con l'unità di rilevazione;
- la tecnica dell'intervista;
- la struttura generale del questionario;
- la struttura di eventuali fogli individuali e dei criteri di selezione dei rispondenti;
- il ruolo delle domande filtro nei collegamenti tra blocchi;
- la presentazione dei quesiti, delle loro relazioni in ciascun blocco o sezione di domande;
- il controllo delle principali coerenze;
- il sistema di identificazione;
- il ruolo dei quesiti di controllo e di quelli sull'intervista;
- uso dei documenti accessori di rilevazione.

Il manuale e le istruzioni devono essere corredati di numerosi esempi esplicativi sulle situazioni *dubbie* prevedibili.

Nelle riunioni di istruzione, inoltre, devono essere previste esercitazioni pratiche sul questionario e sui documenti accessori di rilevazione; in particolare possono essere utilizzati modelli precedentemente compilati con errori, da rintracciare e da discutere con i partecipanti alla riunione.

Deve essere infine previsto, e comunicato alla rete periferica, l'ufficio responsabile dell'assistenza a livello centrale, al quale è possibile rivolgersi durante il periodo di rilevazione.

La pubblicizzazione dell'indagine

La politica dell'immagine dell'Istituto, realizzata con mezzi di comunicazione di massa, quali radio, televisione e giornali a diffusione sia nazionale che locale, può contribuire a creare nel paese un *clima* genericamente favorevole all'attività dell'Istat.

Tali effetti si riflettono positivamente sulla singola indagine, ma è comunque necessario utilizzare mezzi più *mirati* alle specifiche unità di rilevazione.

I mezzi di pubblicizzazione possono essere i più vari; tra essi si ricordano:

— la pubblicità effettuata dagli organi periferici;
— il coinvolgimento di organizzazioni od associazioni delle unità appartenenti alla popolazione oggetto di indagine;
— la lettera di presentazione del sindaco o di altra autorità a livello locale;
— la lettera di presentazione dell'indagine da parte del Presidente dell'Istituto.

4. Il controllo degli errori

Il controllo della fase di rilevazione sul campo, può essere attuato mediante:

I) la stima degli errori di misura, effettuata esplicitando *modelli* dell'errore ed utilizzando adeguate tecniche di rilevazione;
II) l'analisi di indicatori di qualità, ottenuti dalle procedure standard dell'indagine.

I modelli misurano direttamente la qualità dei risultati, stimando le distorsioni e le variabilità imputabili alle fonti sotto controllo (gli intervistatori, l'effetto ricordo, l'effetto delle risposte proxy, ecc.). Al contrario, gli indicatori e le analisi delle informazioni provenienti dalle procedure d'indagine (le mancate risposte totali e parziali, l'errore di identificazione delle unità, la situazione dell'intervista), possono essere considerati come parametri *approssimati* per la valutazione dei dati prodotti.

La stima dell'errore di misura necessita di indagini aggiuntive di controllo, generalmente condotte su di un campione della rilevazione principale, o di particolari schemi di campionamento che richiedono risorse, finanziarie ed organizzative, aggiuntive rispetto alla rilevazione madre. Al contrario gli indicatori di cui al punto (II) risultano più *economici* dei precedenti, in quanto il loro calcolo implica solo la razionalizzazione delle procedure esistenti. Essi, inoltre, sono ottenibili per tutti i *livelli di controllo* coinvolti nella rilevazione sul campo (ad esempio gli uffici comunali ed i rilevatori), mentre le stime possono essere riferiti solo al campione da cui provengono.

Infine, il calcolo e l'analisi sono, indubbiamente, più semplici e tempestivi per gli indicatori di qualità che non per le stime dirette dell'errore.

Le informazioni necessarie per la costruzione degli indicatori e per le analisi, sono, generalmente, disponibili dopo l'elaborazione dei dati dell'indagine principale e di eventuali indagini di controllo; ad esempio, le mancate risposte e le incongruenze logiche dai risultati della revisione, le stime dell'errore totale di misura o dell'effetto ricordo e proxy, dalla compenetrazione del campione o dall'analisi dei risultati di rilevazioni aggiuntive. Nell'attuale organizzazione delle indagini, quindi, tali controlli non sono contestuali alla fase di rilevazione, ma si configurano come controlli «successivi».

La stima dell'errore totale di misura

Come è stato già osservato, l'errore di misura rappresenta la componente più rilevante dell'errore totale e, quindi, la sua stima è indispensabile per conoscere la reale precisione dei risultati.

Gli errori che non danno luogo a incongruenza logiche o a valori fuori campo, non sono determinabili sulla base dei soli risultati dell'indagine; la loro identificazione e quantificazione richiede tecniche particolari.

I modelli matematici, gli stimatori, le tecniche ed i problemi pratici della loro utilizzazione verranno diffusamente trattati nel Capitolo 7; nel prospetto 3.1, sono, succintamente, riportate le componenti dell'errore di misura e le tecniche necessarie alla loro stima, nel caso, sufficientemente realistico, in cui è ipotizzabile che esso sia dovuto ai rispondenti, ai rilevatori ed alle loro interazioni.

A tale riguardo, si ricorda (cfr. Capitolo 1) che l'errore totale è composto dalla differenza tra la media dello stimatore (calcolata sull'universo dei campioni) ed il *valore vero*, cioè la distorsione, e da una parte variabile. Quest'ultima (varianza totale) è pari alla somma della varianza campionaria, della varianza semplice di risposta (che misura l'errore dovuto al solo rispondente) e della varianza correlata di risposta (che misura, invece, l'influenza del rilevatore sulle risposte fornite). La stima di ciascuna componente (o combinazione di componenti) richiede una adeguata tecnica di indagine.

La stima della distorsione e quella delle diverse componenti della parte variabile, non possono essere ottenute con la medesima tecnica: infatti, mentre la prima richiede un processo di misurazione più preciso di quello dell'indagine originaria, allo scopo di appurare il valore *vero*, la seconda si basa su una replicazione indipendente dell'indagine, sotto le stesse condizioni generali, variando solo il fattore da controllare (il rilevatore).

La reintervista con riconciliazione consiste in un ritorno presso un subcampione di unità che vengono reintervistate da un ri-

levatore più esperto o dal precedente assistito dal supervisore; l'intervistatore ripropone i medesimi quesiti (eventualmente utilizzando una versione più dettagliata del questionario e con un maggior numero di domande di controllo) avendo a disposizione le risposte precedentemente fornite e, in caso di discordanza, accerta la risposta *vera*. Se, inoltre, si accertano anche le ragioni della differenza, tale metodo permette di attribuire al rilevatore o al rispondente le differenze riscontrate.

La reintervista senza riconciliazione viene condotta con intervistatori diversi da quelli dell'indagine (il fattore di controllo), ma dello stesso grado di abilità e preparazione; in questo modo ci si assicura dell'indipendenza delle due replicazioni e dell'equivalenza delle condizioni essenziali.

A differenza dei due metodi sopra citati, la compenetrazione del campione non implica una reintervista delle unità; il campione dell'indagine principale viene casualmente diviso in subcampioni di uguale numerosità (ciascuno dei quali costituisce un campione rappresentativo della popolazione di origine) che vengono assegnati a rilevatori diversi.

Prospetto 3.1 - Componente dell'errore di misura e relativi metodi di stima

	METODI DI STIMA		
COMPONENTE STIMATA	reintervista con riconciliazione	reintervista senza riconciliazione	compenetrazione del campione
Distorsione	si		
Varianza totale			si
Varianza di risposta		si	
Varianza di risposta semplice		si	
Componente correlata			si
Varianza campionaria	si	si	si

Nel prospetto 3.1 sono riassunti i metodi di stima sopra richiamati, in funzione della componente di errore determinabile.

La stima dell'effetto proxy e dell'effetto ricordo

Tali tecniche, generalmente, non permettono di stimare l'errore dovuto a cause particolari, come l'effetto proxy e l'effetto ricordo, per la cui stima sono necessari disegni sperimentali.

Per valutare l'entità dell'errore dovuto alla risposta fornita da altri, è necessario programmare un disegno sperimentale, che preveda il ritorno presso l'unità non rispondente (si veda ad es. K.W. Haase, 1972). Tale metodo permette, inoltre, l'analisi dei fattori che influiscono sull'entità degli errori (ad esempio la natura delle domande, le caratteristiche individuali ecc.).

Oltre tale metodo possono essere utilizzate altre tecniche per determinare e quantificare l'effetto proxy:

— il confronto con i medesimi dati provenienti da altra fonte e relativi all'unità non rispondente;
— indagini condotte su due campioni provenienti dalla stessa popolazione in uno solo dei quali sono ammesse risposte proxy;
— la reintervista con riconciliazione effettuata su di un campione di unità non rispondenti.

La stima dell'errore dovuto ad omissione, od errata datazione di eventi, si basa su modelli matematici; per maggiori approfondimenti si rimanda ad esempio a S. Sudman & N.M. Brandbun, 1973. Il modello dei due autori tiene conto dell'effetto congiunto degli errori di omissione e di spostamento in avanti dell'evento, che è l'errore di datazione commesso più frequentemente; il modello è stato testato con risultati soddisfacenti dai due autori, confrontando la stima ottenuta con l'errore osservato, calcolato mediante dati esterni, in alcune indagini.

Gli indicatori di qualità

Gli indicatori di qualità sintetizzati in sei Prospetti e le analisi segnalate nel seguito, costituiscono indicazioni generali; è compito del responsabile dell'indagine scegliere i metodi, i livelli di controllo e le informazioni più adeguati nelle specifiche condizioni organizzative della rilevazione.

Nei Prospetti, per ciascun indicatore di qualità, sono state individuate le fonti cui è imputabile l'errore esaminato e sulle quali è, in genere, possibile esercitare direttamente azioni correttive,

ovvero la rete di rilevazione (comuni e rilevatori) e le metodologie dell'Istituto (norme, questionario, procedure informatiche ecc.).

A tale riguardo, tuttavia, si precisa che la fonte è stata identificata sulla base della preponderanza di responsabilità; in realtà la complessità delle interazioni tra cause diverse non permette, in molti casi, di distinguere nettamente le insufficienze dell'organizzazione centrale e quella periferica, la responsabilità del rilevatore da quella del comune e/o del rispondente.

Gli indicatori possono essere utilizzati sia come controllo delle fonti di errore cui sono stati riferiti, sia come parametri di qualità se calcolati per il complesso dell'indagine o per domini rilevanti di studio (strati, regioni, gruppi omogenei di comuni o rilevatori ecc.).

Come parametri di controllo, gli indicatori relativi ai rilevatori possono essere logicamente estesi alle unità di ordine superiore che esercitano attività di supervisione.

Gli indicatori dell'errore di lista e di mancata risposta totale

Il campione od il complesso delle unità (nel caso di indagini totali) selezionate ed intervistate, rappresentano, a causa degli errori di lista e dei problemi che sorgono nella fase di rilevazione, una popolazione diversa da quella definita nella fase di progettazione.

Il processo di riduzione dell'universo *teorico* all'universo *effettivo*, genera alcune subpopolazioni che assumono, ai fini del controllo di qualità, significati diversi e che possono essere classificate in due gruppi: gli *errori di lista* e le *mancate risposte totali*.

I primi riducono la numerosità campionaria (se non è prevista la sostituzione delle unità non intervistate), o l'universo indagato (nel caso di indagini a carattere censuario), e incrementano l'insieme delle mancate risposte totali; queste ultime, come è stato precedentemente mostrato, possono dar luogo a distorsioni nei risultati finali.

Le unità appartenenti alla popolazione teorica, possono o meno far parte della lista; a sua volta quest'ultima può contenere unità non includibili o duplicazioni.

L'insieme delle unità includibili non incluse è difficilmente quantificabile poiché sarebbe necessario fare ricorso ad altra lista, possibilmente più aggiornata e precisa di quella utilizzata, o ad un microcensimento. In questo caso, è possibile stimare l'errore confrontando i dati delle due fonti; il confronto può essere istituito a livello di singola unità o di risultati aggregati, in funzione dell'esistenza o meno di un codice identificativo comune alle due liste che ne renda possibile l'accoppiamento.

La differenza tra i risultati censuari e la numerosità degli iscritti in anagrafe, calcolata a ridosso del censimento, e l'indagine

di confronto censimento/anagrafe, condotta nel 1981 sulla base di un campione di fogli di famiglia e delle corrispondenti informazioni anagrafiche (A. Cortese 1983), costituiscono esempi di applicazione di tale tecnica.

L'aggregato delle unità non includibili, mediante l'utilizzazione del criterio di includibilità al momento dell'intervista o nell'analisi della lista, può essere stimato o completamente enumerato, a seconda del tipo di indagine (totale o parziale) e del controllo esercitato (se sull'intero archivio o solo sulle assegnazioni dei rilevatori); la verifica della lista è necessaria anche per determinare le eventuali duplicazioni.

Con una parte delle unità sarà possibile stabilire un contatto diretto, a seguito del quale alcune accetteranno mentre altre rifiuteranno l'intervista. Non sarà invece possibile intervistare le unità che, pur abitando all'indirizzo segnalato, risulteranno irreperibili dopo ripetuti tentativi e quelle a cui corrisponde un indirizzo errato; queste ultime, a seconda del tipo di errore, possono risultare o meno, includibili.

Il processo appena descritto ed i metodi per la determinazione dei differenti gruppi di unità sono diagrammati nella Figura 3.1.

In particolare, per il documento di rilevazione aggiuntivo, sono state considerate due situazioni informative.

La prima è relativa al caso in cui si suppone di poter ottenere, da altre unità, le informazioni necessarie a classificare quelle con indirizzo errato nei seguenti quattro gruppi, rilevanti per l'analisi delle mancate interviste e dell'errore di lista: *trasferite dentro il comune, trasferite fuori comune, decedute*, e *irriperibili*.

La seconda, invece, rispecchia il caso (frequente soprattutto nelle città di medie-grandi dimensioni) in cui non si dispone di tali informazioni e quindi le unità *con indirizzo errato* vengono assimilate alle *irreperibili*.

Il suddetto documento fa ormai parte delle procedure standard dell'indagine; utilizzandone le informazioni, è possibile, in modo semplice ed economico, calcolare alcuni indicatori di qualità.

Rapportando le unità a qualsiasi titolo non intervistate al numero programmato d'interviste, si ottiene l'indicatore «grezzo» dell'errore complessivo della fase di rilevazione; depurandone il numeratore dagli errori di lista, ovvero sostituendo la somma dei *rifiuti* e delle unità *non a casa*, verrà calcolato l'indicatore «grezzo» di mancata risposta.

Quest'ultimo rappresenta il peso delle interviste non effettuate rispetto al numero «atteso» di interviste e, quindi, include al denominatore anche gli eventuali errori di lista. Depurando il denominatore da tale quantità, si ottiene un indicatore «netto» di mancata intervista.

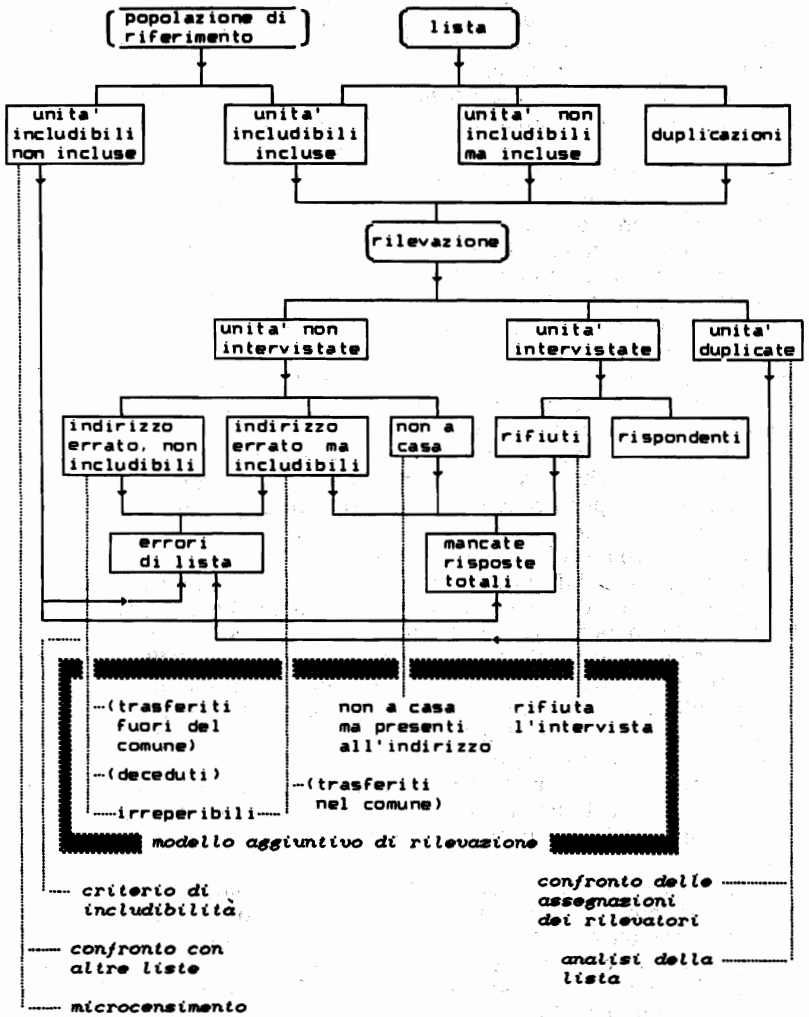


Figura 3.1 - Processo di formazione e metodi per la quantificazione delle mancate risposte totali e degli errori di lista

Prospetto 3.2 - Indicatori relativi alle Mancate Risposte Totali

INDICATORI	SIGNIFICATO	FONTE DI ERRORE		
		comuni	rilevatori	ISTAT
N_L / N	Lista	si	—	—
$(N_{L1} + N_{L2} + N_{L3}) / N_L$	aggiornamento della lista	si	—	—
N_{L4} / N_L	tenuta della lista	si	—	—
$(N_{L1} + N_{L2}) / N$	non includibilità	si	—	—
N_{NR} / N	errore complessivo di rilevazione	si	si	—
$(N_{RF} + N_{NAC}) / N$	lordo di mancata risposta	—	si	—
$(N_{RF} + N_{NAC}) / (N - N_L)$	netto di mancata risposta	si	si	—
$N_{RF} / (N_R + N_{RF})$	netto di rifiuto	—	si	—
$N_{NAC} / (N_R + N_{NAC} + N_{RF})$	netto di mancata intervista	—	si	—

N = numerosità programmata

N_R = numero di unità rispondenti

N_{NR} = numero di mancate risposte

N_{RF} = numero di rifiuti

N_{NAC} = numero di unità non a casa

N_L = numero di errori di lista

N_{L1} = numero di decessi

N_{L2} = numero di trasferiti nel comune

N_{L3} = numero di trasferiti fuori comune

N_{L4} = irreperibili

$N = N_R + N_{NR}$

$N_{NR} = N_L + N_{NAC} + N_{RF}$

$N_L = N_{L1} + N_{L2} + N_{L3} + N_{L4}$

Tali indicatori sono relativi alle prestazioni della rete periferica nel suo complesso e quindi il livello più adeguato di controllo è quello comunale. Per valutare l'operato dei rilevatori, è conveniente scomporre la mancata intervista nelle due componenti di *rifiuto* e di *non a casa* e calcolare i relativi tassi netti, utilizzando a denominatore, rispettivamente, il totale dei *rispondenti* e dei *rifiuti* e la somma dei *rispondenti*, dei *rifiuti* e dei *non a casa*.

L'errore di lista è rappresentato dalle unità che presentano l'indirizzo errato; rapportando il loro numero all'ampiezza del campione od alla numerosità della lista (per le indagini totali) si ottiene il relativo tasso. L'errore può essere ulteriormente analizzato, se si dispone delle necessarie informazioni, in una parte dovuta all'*aggiornamento*, (rappresentato dalle modalità *trasferiti nel* e *fuori del comune* e *deceduti*) ed in una parte residua (modalità *irreperibili*) che approssima la *tenuta* della lista; i relativi tassi (di aggiornamento e di tenuta) avranno come denominatore il numero di unità con indirizzo errato. Sempre utilizzando la medesima disaggregazione, è possibile calcolare un tasso di non includibilità che presenta al numeratore la somma dei *deceduti* e dei *trasferiti fuori del comune* ed al denominatore il numero programmato d'interviste.

Le stesse notizie, invece, non possono costituire (se non introducendo ipotesi «forti» sulla distribuzione delle subpopolazioni non direttamente determinate) la base informativa per stimare le quote e le caratteristiche dei rispondenti e dei non rispondenti nella popolazione teorica (secondo lo schema utilizzato nel paragrafo 2). Infatti, la rilevazione sul campo non coglie le unità includibili non incluse nella lista e non è possibile discriminare gli aggregati *non a casa* e *trasferiti nel comune* nelle suddette subpopolazioni; infine, gli *irreperibili*, non sono classificabili né come rispondenti / non rispondenti, né come includibili / non includibili.

Nel Prospetto (3.2) sono sintetizzati i livelli di controllo ed i relativi indicatori sintetici discussi nel paragrafo.

Per quanto riguarda le unità, si fa riferimento alle unità di selezione, in genere, per le indagini Istat sulla popolazione, le famiglie.

Oltre al calcolo dei tassi per la verifica dell'operato della rete periferica, le informazioni riportate sul modello aggiuntivo o sul foglio di anagrafe, possono essere utilizzate per condurre analisi più approfondite riguardo all'ubicazione e le caratteristiche delle unità non intervistate e non rispondenti.

Si potrà controllare, quindi, che l'omissione di unità non sia correlata a qualche fattore di distorsione dei risultati; ad esempio alla dimensione della famiglia, alla lontananza dal centro cittadino o alla residenza in zone «difficili».

Studi più analitici sull'errore di mancata risposta possono essere condotti analizzando le relazioni tra le caratteristiche delle unità non rispondenti e le modalità di effettuazione dell'intervista (ad esempio il giorno della settimana e l'ora). In questo modo è possibile determinare le tipologie delle subpopolazioni maggiormente a rischio e le tecniche di raccolta più efficienti per le unità appartenenti a tali gruppi.

Nel caso siano previste sostituzioni di non rispondenti, il confronto tra le caratteristiche delle unità sostituite e sostituitive, mediante tabulazione incrociata sottoposta a test di indipendenza e simmetria, permette il controllo dell'operazione di sostituzione, determinando se essa sia stata eseguita secondo le norme indicate (ad esempio famiglie della stessa ampiezza o territorialmente *vicine*).

Disporre di informazioni sulla consistenza dei differenti aggregati di unità non intervistate o non rispondenti, è utile non solo per una valutazione della qualità dei dati raccolti e dell'attività della rete periferica, ma anche per la programmazione di indagini future e per le possibili azioni correttive nella tenuta della lista.

La stima della copertura del censimento

La *copertura* del censimento costituisce un caso particolare di mancata risposta totale, caratterizzato dalla non esistenza a priori di numerosità di confronto; uno degli obiettivi dell'operazione censuaria è, infatti, l'aggiornamento degli archivi anagrafici.

Per la medesima ragione, questi non possono essere utilizzati, quale liste di selezione, per l'indagine campionaria di controllo effettuata a stretto ridosso del censimento; si ricorre, quindi, ad un campione di aree (sezioni di censimento) in cui si conteggiano, una seconda volta, tutte le unità in esse contenute.

Nell'indagine di controllo devono essere utilizzati i rilevatori più esperti, per assicurare una migliore qualità dei dati raccolti, e ciascuno di essi deve essere impegnato in un'area diversa dalla precedente, per garantire l'indipendenza delle due operazioni.

Nell'analisi dei risultati, si possono concettualmente suddividere le unità in quattro sub-universi differenti, a seconda della presenza od assenza riscontrata nelle due rilevazioni; più precisamente:

— unità presenti all'indagine ed al censimento (N_{11})
— unità presenti all'indagine ma non al censimento (N_{12})
— unità presenti al censimento ma non all'indagine (N_{21})
— unità non presenti sia al censimento che all'indagine (N_{22})

Mediante il confronto tra l'indagine di controllo ed il censimento è possibile stimare direttamente la consistenza dei primi tre gruppi ($\hat{N}_{11}$, $\hat{N}_{12}$, $\hat{N}_{21}$), ed indirettamente quella di N_{22} ipotizzando l'indipendenza tra le due operazioni:

$$\hat{N}_{22} = (\hat{N}_{12} * \hat{N}_{21}) / \hat{N}_{11}$$

In tal modo è possibile stimare N, ovvero il numero totale delle unità, come somma delle stime $\hat{N}_{ij}$

$$\hat{N} = \hat{N}_{11} + \hat{N}_{12} + \hat{N}_{21} + \hat{N}_{22},$$

il tasso di copertura del censimento

$$\hat{T}_c = (\hat{N}_{11} + \hat{N}_{21}) / \hat{N}$$

e la varianza della stima $\hat{N}$

$$\hat{\sigma}^2_{\hat{N}} = (\hat{N} * \hat{N}_{22}) / \hat{N}_{11}$$

mediante la quale determinare l'intervallo di confidenza per N.

Gli indicatori di mancata risposta parziale

Durante l'intervista, le interazioni tra questionario, rilevatore e rispondente possono determinare, per ciascun quesito, rifiuti od impossibilità a rispondere, risposte non dovute od incongruenze logiche sulla base delle norme di compilazione del questionario, valori non ammissibili rispetto al campo di variazione prestabilito.

La classificazione corrispondente è diagrammata nella Figura 3.2.

La distinzione tra risposte dovute e non dovute è funzionale alla costruzione di indicatori che rappresentano i diversi aspetti dell'errore e può essere ottenuta in due modi: (I) mediante l'analisi della risposta al singolo quesito, (II) mettendo in relazione le risposte fornite a più quesiti.

Nel primo caso occorre distinguere l'assenza di risposta derivante dai *rifiuti* e dai *non so* da quella per *risposta non dovuta*; ciò può essere realizzato prevedendo nel questionario una codica specifica per le suddette modalità. Tale metodo assicura il riconoscimento delle cause della mancata risposta ma può favorire la tendenza a non rispondere.

I codici che distinguono la risposta non dovuta dalla mancata, unitamente alle informazioni sui valori fuori campo e alle incongruenze riscontrate, forniscono la base di calcolo per gli indicatori riportati nel Prospetto 3.3.

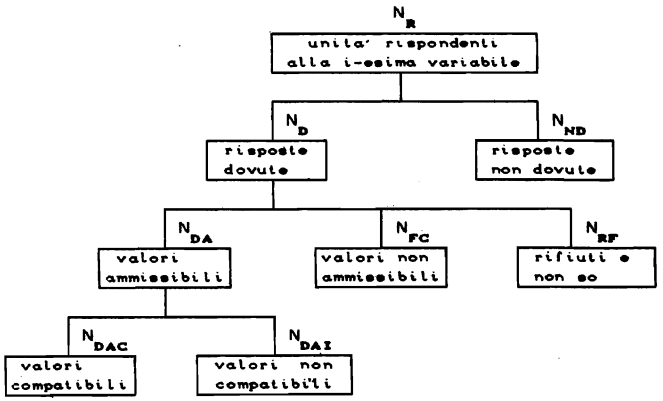


Figura 3.2 - Classificazione delle risposte all'i-esimo quesito

Il rapporto tra la somma delle risposte non dovute e di quelle utilizzabili (valori ammissibili e compatibili) ed il numero dei rispondenti, costituisce l'indicatore della *compilazione* del quesito.

Rapportando i valori ammissibili e compatibili alle risposte dovute, si ottiene l'indicatore dell'*efficacia dell'intervista*; al contrario, la somma, al numeratore, dei valori fuori campo, dei rifiuti e dei codici ammissibili ma incompatibili, fornisce la misura della *mancata risposta parziale*. Il fenomeno *rifiuto* è misurato dal rapporto tra i rifiuti e le risposte dovute, mentre, il numero dei codici ammissibili ma risultati incompatibili su quello dei codici ammissibili costituisce un indicatore di *incompatibilità*.

Tali indicatori vengono, di solito, elaborati dai dati registrati su supporto informatico e, quindi, includono l'errore di registrazione.Essi, tuttavia, non perdono di significato e di efficacia nell'analisi se possiamo ipotizzare

(a) che la parte preponderante dell'errore sia attribuibile alla fase di rilevazione;
(b) che l'errore di registrazione sia costante nei diversi indicatori calcolati.

Prospetto 3.3 - Indicatori relativi alle Mancate Risposte Parziali per la generica variabile - calcolo basato sui codici

INDICATORI	SIGNIFICATO	FONTE DI ERRORE		
		comuni	rilevatori	questionario
$(N_{ND} + N_{DA}) / N_R$	compilazione del quesito	—	si	si
N_{DAC} / N_D	efficacia dell'intervista	—	si	si
$(N_{DAI} + N_{FC} + N_{RF}) / N_D$	mancata risposta	—	si	si
N_{RF} / N_D	rifiuto	—	si	si
N_{DAI} / N_{DA}	incompatibilità	—	si	si

N_R = numero di rispondenti
N_{ND} = numero di risposte non dovute
N_D = numero risposte dovute
N_{RF} = numero di rifiuti
N_{FC} = numero di valori fuori campo
N_{DA} = numero di valori ammissibili
N_{DAC} = numero di valori ammissibili compatibili
N_{DAI} = numero di valori ammissibili incompatibili

Se non si utilizzano i suddetti codici, si dovrà ricorrere alle correzioni dovute alle regole *formali* dei piani di *compatibilità* che, mettendo in relazione la risposta fornita al singolo quesito con quella presente in altri quesiti, riconoscono la mancata risposta dovuta.

In questo modo, però, la distinzione tra risposta dovuta e non dovuta viene a dipendere dalla procedura ed è perturbata dagli errori nelle variabili, generati da altre fonti. Il riconoscimento dei diversi gruppi si baserà sulle variazioni intervenute tra il file *sporco* e quello *pulito*, durante la fase di revisione. Il confronto avverrà mediante collegamento tra i record relativi alla medesima unità; a tale scopo deve essere previsto un codice identificativo *esatto*, ad esempio un progressivo di record apposto mediante programma informatico sul materiale proveniente dalla registrazione e conservato immutato in tutte le successive fasi di elaborazione dei dati.

L'accoppiamento tra record «sporchi» e «puliti» può dar luogo ad una delle sei situazioni riportate nel Prospetto 3.4, e le unità, per ciascuna delle variabili del questionario, rimarranno quindi raggruppate nelle relative sette classi.

Prospetto 3.4 - Confronto tra i valori della singola variabile della generica unità nel file sporco e pulito

FILE SPORCO	FILE PULITO	
	Blank	Valori significativi
Valori non ammissibili	NB	NV
Blank	BB	BV
Valori significativi	VB	VV MM

Anche i valori diversi da *blank* e dai fuori campo, possono essere modificati dai piani di compatibilità e quindi danno luogo a due insiemi: i valori modificati (MM) e quelli immutati (VV).

Dal Prospetto 3.4 possiamo riclassificare le unità nei seguenti gruppi, significativi per la costruzione degli indicatori sintetici di qualità, per ciascuna variabile del questionario:

I) i valori fuori campo	(NB + NV)
II) i rifiuti	(BV)
III) le incongruenze	(VB + MM)
IV) le risposte dovute nette	(BV + VV + MM)
V) le risposte dovute lorde	(BV + VV + MM + NV)
VI) le risposte nette	(BB + VV)

L'indicatore generico della qualità del materiale *disponibile*, è dato dal rapporto tra le *risposte nette* ed il numero delle unità rispondenti; depurando il denominatore dai *valori fuori campo*, si ottiene il medesimo indicatore per il materiale *raccolto*.

Addebitando i *valori fuori campo*, nei quesiti precodificati, alla fase di registrazione, il rapporto tra questi ed il numero di unità

rispondenti, costituisce il *tasso minimo di errore di registrazione*; minimo, in quanto non sono comprese in esso le modificazioni in valori ammissibili, non identificabili con tale tecnica. Tale tasso non può quindi sostituire il controllo della fase di registrazione, ma ne costituisce una verifica basata sul complesso dei dati.

Il tasso di rifiuto, per il singolo quesito, può essere calcolato rapportando *i rifiuti* sia alle *risposte dovute nette*, sia a quelle *lorde*; in questo secondo caso la stima è conservativa, in quanto addebita l'insieme NV unicamente all'errore di registrazione. I due valori costituiscono quindi gli estremi del campo di variazione del tasso di rifiuto *reale*.

Prospetto 3.5 - Indicatori relativi alle Mancate Risposte Parziali per la generica variabile - calcolo basato sui risultati delle procedure di revisione

INDICATORI	SIGNIFICATO	FONTE DI ERRORE		
		comuni	rilevatori	questionario
RN / N_R^*	qualità materiale disponibile	—	si	si
RN / (N_R^* – VFC)	qualità materiale raccolto	—	si	si
VFC / N_R^*	errore minimo registrazione	—	—	si
INC / RDN	incongruenza	—	si	si
RF / RDN	rifiuto (netto)	—	si	si
RF / RDL	rifiuto (lordo)	—	si	si
(RF + INC) / RN	efficacia intervista	—	si	si
(RF + INC) / RDN	efficacia raccolta	—	si	si
(RF + INC + VFC) / N_R^*	efficacia indagine	—	si	si

		N_R^* unità rispondenti
NB + NV	=	VFC valori fuori campo
BB + VV	=	RN risposte nette
BV + VV + MM	=	RDN risposte dovute nette
BV + VV + MM + NV	=	RDL risposte dovute lorde
BV	=	RF rifiuti
VB + MM	=	INC incongruenze

Quale indicatore dell'efficacia della tecnica di raccolta, possiamo considerare il rapporto tra la somma dei *rifiuti* e delle *incongruenze* e le *risposte dovute nette*; l'efficacia del complesso dell'indagine è, invece, data dalla somma dei *rifiuti*, delle *incongruenze* e dei *valori fuori campo*, rapportata al numero di unità rispondenti.

L'analisi dei tassi dei Prospetti 3.3 e 3.5 può risultare difficoltosa, dato il numero delle variabili presenti su di un questionario ad obiettivi plurimi. Per ridurre la mole dell'informazione da valutare, possono essere calcolate delle medie (semplici o ponderate con il numero delle risposte dovute) sul complesso delle variabili o su sottoinsiemi rilevanti di esse.

Le unità considerate sono quelle a cui sono riferite le informazioni da cui sono calcolati i tassi: generalmente gli individui.

I tassi relativi alla qualità del materiale raccolto, ai rifiuti e all'efficacia della raccolta e dell'indagine, possono essere utilizzati per il controllo della rete periferica; gli indicatori della qualità del materiale disponibile e dell'errore minimo di registrazione, analizzati per tutte le variabili, per domini territoriali o sul complesso dei dati raccolti forniscono indicazioni sull'affidabilità delle stime ottenute.

A livello totale, inoltre, l'analisi dell'indicatore dell'efficacia della tecnica di raccolta, per singolo quesito o gruppi di quesiti, possono rivelare l'esistenza di ambiguità o di errori nella struttura o nelle norme di compilazione del questionario.

Sulla mancata risposta parziale possono essere condotte ulteriori analisi ponendo in relazione i tassi di rifiuto e le caratteristiche strutturali delle unità; sull'efficienza della tecnica di raccolta mediante l'esame delle relazioni tra gli indicatori di efficacia ed i dati concernenti la situazione in cui si è svolta l'intervista.

Gli indicatori dell'intervista

Le modalità di intervista, (il numero dei ritorni, la durata, il giorno e l'ora di effettuazione ed i rispondenti proxy) possono essere analizzate a fini:

— di controllo dell'operato dei rilevatori;
— di controllo della reale numerosità campionaria dei rispondenti;
— di analisi delle caratteristiche dei non rispondenti;
— di ricostruzione delle tipologie di situazioni dell'intervista.

Il carico di lavoro, la durata media dell'intervista e la percentuale di risposte proxy, calcolate per rilevatore e per ufficio periferico, possono essere utilizzati per il controllo della rete; la percentuale di rispondenti è un indicatore della reale dimensione campionaria (Prospetto 3.6).

L'analisi delle caratteristiche dei rispondenti proxy può servire ad identificare particolari subpopolazioni per le quali i dati raccolti sono *mediati* da altre unità; tale informazione è un indicatore della possibile esistenza di distorsioni.

Prospetto 3.6 - Indicatori relativi alle modalità d'intervista

INDICATORI	SIGNIFICATO	FONTE DI ERRORE		
		comuni	rilevatori	ISTAT
N_P^* / N_R^*	proxy	—	si	—
N_P / N_R	proxy	—	si	—
$1 - (N_P^*/N_R^*)$	dimensione campionaria reale	—	si	—
N_R / N	effettuazione interviste (questionari)	—	si	—
N_R^* / N^*	effettuazione interviste (individui)	—	si	—
$\Sigma D_i / N_R$	durata media intervista	—	si	si
G_a / G	giorni ammessi di intervista	si	si	si
G_b / G	giorni non ammessi di intervista	si	si	si
$N_R / (G_a + G_b)$	interviste giornaliere	si	si	si

N numerosità teorica (questionari)

N_R numero di unità rispondenti (questionari)

N_P^* numero di rispondenti proxy

N_P numero di questionari con almeno un rispondente proxy

N^* numerosità teorica (individui)

N_R^* numero di unità rispondenti (individui)

D_i durata dell'i-esima intervista

G periodo di riferimento (in giorni)

G_a numero di giorni in cui le interviste sono state effettuate (interni al periodo di riferimento)

G_b numero di giorni in cui le interviste sono state effettuate (esterni al periodo di riferimento)

I dati concernenti la situazione in cui si è svolta l'intervista (le relazioni tra unità presenti e rispondenti, quale ad esempio la relazione di parentela, la distribuzione per giorno della settimana e per ora il rispetto del calendario) possono essere analizzate allo scopo di individuare *tipologie d'intervista* da cui desumere utili indicazioni per le norme di rilevazione per la medesima indagine in tempi successivi o per indagini rivolte alla medesima popolazione.

Gli indicatori dell'identificazione delle unità

Nell'attuale organizzazione delle indagini, la responsabilità delle operazioni di trascrizione e di apposizione dei codici identificativi, ricade sui supervisori (gli uffici comunali).

Il calcolo degli indicatori di qualità di questo aspetto della rilevazione, dovrebbe essere, quindi, basato sull'analisi del materiale cartaceo, il solo effettivamente compilato dai suddetti soggetti.

Tuttavia, tale verifica è, generalmente, troppo dispendiosa, dal punto di vista organizzativo e della tempestività, cosicché si può ricorrere, per la determinazione degli errori di identificazione, al confronto tra l'identificatore della stessa unità presente nel file proveniente dalla registrazione con quello corretto nella fase di revisione quantitativa. Anche in questo caso, come per l'analisi delle mancate risposte parziali, l'abbinamento dei record presuppone un codice di accoppiamento esatto, indipendente dal sistema di identificazione adottato nell'indagine.

Il confronto può dar luogo a differenze (errori) o ad uguaglianze tra codici che saranno utilizzate per il calcolo degli indicatori del Prospetto 3.7, rapportando il numero di unità con codici errati al totale delle unità.

Tali indicatori, tuttavia, risentono delle modalità della registrazione; infatti, i codici identificativi vengono, generalmente, registrati in duplica e, quindi, un solo errore si ripercuote su tutte le unità cui è riferito l'identificatore. Per tali ragioni, non è possibile ipotizzare un errore di registrazione di modesta entità, distribuito casualmente ed uniformemente su tutte le unità rilevate, e, quindi, gli indicatori sovrastimano l'errore di identificazione dovuto alla fase di rilevazione. Ad esempio un errore di registrazione commesso sul codice identificativo del comune, comporta che, nel numeratore dell'indicatore, compaia il numero di tutti i questionari del comune in esame, pur essendo validi i codici riferiti al questionario.

Per calcolare un parametro, che approssimi l'errore commesso nella fase di rilevazione, si può supporre che la sistematicità sia collegata solo al codice comunale e calcolare al denominatore solamente le differenze riscontrate nell'identificatore di una delle unità subcomunali (questionario, area, rilevatore, individuo o evento).

I suddetti indicatori possono essere calcolati o per il concatenamento di tutti i codici subcomunali (ad esempio l'identificatore composto dai codici di rilevatore, famiglia, individuo), oppure per ogni singolo identificatore o combinazioni di identificatori; nel primo caso si valuta l'errore complessivo commesso sui codici identificativi, mentre, nel secondo, si può valutare il rischio di errore per ogni singolo o per una determinata combinazione di codici elementari.

La presenza di una doppia chiave di identificazione (come nell'indagine forze di lavoro dove esiste un codice di famiglia per l'aspetto trasversale ed uno per l'aspetto longitudinale dell'indagine) implica l'esistenza di una corrispondenza biunivoca tra i due identificatori; in questo caso le mancate relazioni, rapportate al numero di unità rilevate, danno luogo ad un specifico indicatore.

La numerazione progressiva dei questionari, relativi ad un dominio territoriale (usualmente il comune), comporta che il massimo dei progressivi od almeno quello relativo alle unità non sostituite, non può superare la numerosità campionaria assegnata. È possibile, quindi, calcolare un tasso di errore, mediante il rapporto tra il numero di progressivi maggiori della numerosità campionaria ed il totale delle unità rilevate.

I numeratori ed i denominatori degli indicatori possono essere calcolati facendo riferimento o al questionario (ad esempio il numero di questionari in cui è errato il codice di rilevatore o l'identificatore di famiglia), ovvero alle unità elementari di analisi (ad esempio il numero di individui o di eventi). Nel primo caso, si ha una misura dell'*errore commesso*, in quanto il questionario è il supporto cartaceo effettivamente compilato nella fase di rilevazione, mentre nel secondo si ottiene una misura dell'*impatto dell'errore* sui microdati dell'indagine.

Le caratteristiche strutturali

Le caratteristiche strutturali delle unità rispondenti, possono essere utilizzate per individuare eventuali distorsioni verificatesi nella fase di raccolta sul campo, ad esempio la sottovalutazione di particolari subpopolazioni.

Per le indagini sulla popolazione, si può far ricorso ad indicatori demografici che risultano stabili, se la base di calcolo è sufficientemente ampia, e sul cui livello si hanno informazioni a priori: il numero medio di componenti per famiglia, il rapporto di mascolinità, gli indici di dipendenza e di vecchiaia ecc..

Il confronto tra tali indicatori ed i corrispondenti desunti dal censimento, dalle risultanze anagrafiche o dalle previsioni, costituisce una verifica della *rappresentatività* dei risultati della rilevazione rispetto alla popolazione di riferimento.

Per le indagini campionarie è opportuno prendere in considerazione un livello di controllo sovracomunale per assicurare una numerosità sufficiente e garantire la stabilità dei rapporti.

Prospetto 3.7 - Indicatori relativi all'identificazione delle unità

INDICATORI	SIGNIFICATO	FONTE DI ERRORE		
		comuni	rilevatori	identificatori
	A) Errore commesso			
NQ_{cod} / NQ	identificatore completo	si	—	si
$NQ_{cod\ (i)}$ / NQ	singolo sub identificatore	si	—	si
NQ_{blu} / NQ	identificatori doppia chiave	si	—	si
NQ_{max} / NQ	identificatori progressivi	si	—	si
	B) Incidenza dell'errore sui microdati			
NU_{cod} / NU	identificatore completo	si	—	si
$NU_{cod\ (i)}$ / NU	singolo sub identificatore	si	—	si
NU_{blu} / NU	identificatori doppia chiave	si	—	si
NU_{max} / NU	identificatori progressivi	si	—	si

NQ^* = numero di questionari corrispondenti alle unità teoriche di rilevazione

NQ, NU = numero di questionari compilati e relative unità

NQ_{cod}, NU_{cod} = numero di questionari con almeno un identificatore subcomunale errato e relative unità

$N_{cod\ (i)}$, $NU_{cod\ (i)}$ = numero di questionari con l'i-esimo identificatore subcomunale errato e relative unità

NQ_{blu}, NU_{blu} = numero di questionari per i quali non è stata verificata la corrispondenza biunivoca nella doppia chiave di codici e relative unità

NQ_{max}, NU_{max} = numero di questionari con il codice progressivo $> NQ^*$ e relative unità

Gli indicatori succitati sono validi per la gran parte delle indagini sulla popolazione; per rilevazioni mirate a particolari subpopolazioni è necessario costruirne di specifici (ad esempio in una indagine sulla fertilità è opportuno indagare più approfonditamente sulle classi di età delle donne in età feconda).

5. La correzione degli errori

Del complesso degli errori derivanti dalla fase di rilevazione sul campo, solo le mancate risposte totali e quelle parziali (intese in senso lato, ovvero comprensive delle incongruenze logiche e dei valori fuori campo), possono essere riconosciute ed attribuite alla singola unità di analisi. Gli altri errori possono essere identificati e quantificati mediante indicatori indiretti, oppure stimati, mediante indagini di controllo, ma solo in riferimento al complesso dei dati.

La correzione può essere, quindi, apportata solo riguardo ai succitati errori, in quanto identificabili senza mutare le condizioni generali dell'indagine, in tempi e con costi contenuti; ciò implica che è possibile correggere solo una parte dell'errore totale.

La distinzione tra mancate risposte totali e parziali è funzionale ai metodi di correzione: le mancate risposte parziali sono corrette operando sui microdati nella fase di revisione del materiale raccolto, mentre si tenta di ridurre gli effetti delle mancate risposte totali o al momento della rilevazione, prevedendo le sostituzioni, o al momento delle stime finali, mediante appositi pesi correttivi.

Le mancate risposte parziali

Possiamo immaginare i risultati dell'indagine, dopo la fase di rilevazione, come una matrice unità/variabili, divisa in due sottoinsiemi: i dati relativi alle unità rispondenti e quelli dei non rispondenti.

Per l'analisi di tale matrice sono possibili tre strategie, tenendo presente, tuttavia, che, qualsiasi di esse venga adottata, si sconta una distorsione delle stime, se il meccanismo di generazione delle mancate risposte non è strettamente casuale:

— limitarsi all'insieme dei «dati completi» ovvero delle unità che hanno risposto a tutti i quesiti;
— includere, nell'analisi delle singole variabili, anche le unità che, per quelle caratteristiche, hanno fornito una risposta (dati disponibili);
— operare una qualche forma di correzione o di imputazione.

Nel primo caso, possiamo applicare le analisi statistiche standard e viene assicurata la comparabilità delle stime, poiché le statistiche sono calcolate sulla stessa base di dati; d'altro canto, tale scelta comporta una forte riduzione della numerosità campionaria, in funzione delle probabilità di mancata risposta sulle variabili e del numero delle stesse. Ad esempio, ipotizzando una probabilità di mancata risposta costante, casuale ed indipendente tra variabili, la riduzione del numero di unità è riportata nella tavola (3.1).

Tavola 3.1 - Riduzione della numerosità campionaria in funzione delle probabilità di mancata risposta

PROBABILITÀ DI MANCATA RISPOSTA	NUMERO DELLE VARIABILI RILEVATE		
	10	20	50
1%	90%	82%	60%
5%	60%	36%	8%

Nel caso dei *risultati disponibili*, si recupera tutta l'informazione contenuta nei dati, ma le statistiche univariate non sono immediatamente confrontabili, perché ottenute con numerosità diverse; inoltre, si deve far ricorso a procedure non standard per il calcolo di statistiche multivariate (ad es. la matrice di correlazione), a meno di non basarsi, in questi casi, solo sui *risultati completi*.

Mediante la correzione dei microdati, si eliminano gli inconvenienti dei primi due metodi, poiché tale tecnica fornisce una matrice di risultati completi per tutte le unità.

Data l'importanza dell'argomento, i procedimenti di determinazione e correzione dell'errore, qui classificato come mancata risposta parziale, sono oggetto di un apposito capitolo (Capitolo 5).

Le mancate risposte totali

Nel caso di indagini esaustive, in cui sono note le variabili strutturali dell'universo, la correzione può essere effettuata, stratificando le unità rispondenti secondo tali caratteristiche e *pesando* i risultati, mediante il rapporto tra numerosità teorica e numero di unità rispondenti. Tale tecnica equivale a sostituire le unità non rispondenti all'unità *media* di gruppo; in questo modo, però, le distribuzioni risultano appiattite su tali medie.

L'inconveniente viene superato, utilizzando una seconda tecnica, che consiste nel sostituire, a livello di micro dati, le unità

non rispondenti con *unità tipo*, determinate a priori, o da unità con le medesime caratteristiche, scelte a caso nello strato di appartenenza; il numero e l'omogeneità degli strati dipendono dalle caratteristiche delle unità, riportate nella lista. Tali tecniche possono essere considerate come un'estensione dell'applicazione dei piani di compatibilità e correzione al caso in cui tutte le variabili, tranne quelle di collegamento, sono mancanti (cfr. Capitolo 5).

Per le indagini campionarie, si può far riferimento a differenti metodologie.

Il primo metodo di correzione degli effetti delle mancate risposte totali è contemporaneo alla rilevazione sul campo: sostituire le unità non rispondenti con altre precedentemente selezionate in maniera casuale dalla medesima lista.

Tale tecnica ripristina la numerosità campionaria programmata e quindi la quota dei non rispondenti non influenza l'errore di campionamento. Tuttavia, possono permanere effetti distorsivi se la subpopolazione dei rispondenti, cui appartengono le unità sostitutive, presenta caratteristiche differenti da quella dei non rispondenti; continuano, perciò, a rimanere valide le considerazioni riportate nel paragrafo 2.

La seconda tecnica, più generalmente usata, consiste nella suddivisione in strati delle unità campionarie e nella correzione delle stime, mediante la modificazione delle probabilità di selezione in ciascuno strato.

In questo caso, ad esempio, lo stimatore diretto di Horwitz Thompson per la media,

$$\bar{y} = \Sigma_i \Sigma_j \pi_{ij}^{-1} y_{ij} / \Sigma_i \Sigma_j \pi_{ij}^{-1}$$

dove le π_{ij} rappresentano le probabilità di selezione della j-esima unità nell'i-esimo strato, verrà modificato in:

$$\bar{y} = \Sigma_i \Sigma_j (\pi_{ij}\, p_{ij})^{-1} y_{ij} / \Sigma_i \Sigma_j (\pi_{ij}\, p_{ij})^{-1} \qquad (3.4)$$

Le p_{ij} rappresentano le probabilità di risposta, usualmente stimate dalla proporzione di unità campionarie rispondenti nello strato n_{R_i} / n_i.

Nel caso in cui la probabilità di selezione sia uguale per tutte le unità, la (3.4) si trasforma nella:

$$\bar{y}' = \Sigma_i \bar{y}_{R_i} (n_{R_i} / n) \qquad (3.5)$$

dove le $\bar{y}_{R_i}$ sono le medie dei rispondenti nello strato.

Per stratificare sia le unità rispondenti che quelle non rispondenti secondo un unico criterio, è necessario che questo sia conosciuto a priori, indipendentemente dalle informazioni raccolte mediante l'indagine. Inoltre, la variabile di stratificazione non deve essere correlata con i fattori che determinano la mancata risposta, altrimenti gli strati rispecchierebbero ancora le popolazioni dei rispondenti e dei non rispondenti.

La riduzione della distorsione operata dal procedimento è funzione dell'omogeneità delle sub-popolazioni individuate a posteriori.

Se sono conosciute le numerosità degli strati nella popolazione, allora si può far ricorso allo stimatore post stratificato che, sotto le condizioni sopra enunciate, risulta non distorto:

$$\bar{y}' = \Sigma_i \bar{y}_{R_i} N_i / N$$

Un terzo metodo di correzione degli effetti delle mancate risposte, consiste nell'estrarre, dalle n_{NR} unità campionarie non rispondenti, un subcampione casuale semplice di n'_{NR} unità ed ottenerne l'intervista con successivi ritorni.

In questo caso, la stima può essere ottenuta come combinazione lineare delle due stime, quella dei rispondenti nell'indagine e quella ottenuta dal campione dei non rispondenti, con pesi pari ai rapporti delle rispettive numerosità con quella programmata.

Ad esempio nel caso dello stimatore media:

$$\bar{y}_t = (n_R / n)\, \bar{y}_R + (n_{NR} / n)\, \bar{y}_{NR}$$

dove $\bar{y}_R$ è calcolata dagli n_R rispondenti, mentre $\bar{y}_{NR}$ è calcolata sulle n'_{NR} unita campione selezionate dagli n_{NR} non rispondenti. Lo stimatore $\bar{y}_t$ sarà non distorto se tutte le unità del campione dei non rispondenti vengono intervistate.

Tale tecnica può risultare piuttosto costosa in termini economici ed organizzativi, per l'evidente difficoltà di reperire e/o intervistare unità, che nel corso dell'indagine principale non era stato possibile rilevare.

Gli errori di identificazione

La correzione degli errori di identificazione, si attua nella fase di revisione quantitativa (cfr. Capitolo 5), sulla base del confronto tra i documenti di rilevazione ed il file; essa consiste in operazioni di modificazione dei codici identificativi, di cancellazione od inserimento di record in detto file.

RIFERIMENTI BIBLIOGRAFICI

Lavori di carattere teorico

AA.VV. (1983), *Incomplete data in sample survey*, Volume 2, Academic Press, New York.

COCHRAN W. (1977), *Sampling techniques*, cap. 13, J. Wiley, New York.

GIOMMI A. (1986), *Sulla stima della probabilità di risposta nel campionamento da popolazioni finite*, in Atti della XXXII Riunione Scientifica della S.I.S., Sorrento.

GIOMMI A. (1986), *Procedimenti non parametrici nella stima delle probabilità individuali di risposta*, in Atti della XXXIII Riunione Scintifica della S.I.S., Bari.

HAASE K.W., WILSON R.W. (1972), *The study design of an experiment to measure the effects of using proxy reponses in the National Health Interview Survey*, Proceedings of the Social Statistics Section, A.S.A. pagg. 289-293.

KALTON G., KASPRZYK D. (1986), *The treatment of missing survey data*, in Survey Methodology, Volume 12 Number 1 June 1986, Statistics Canada.

KISH L. (1965), *Survey sampling*, cap. 13, J. Wiley, New York.

LITTLE R.J.A., RUBIN D.B. (1987), *Statistical analysis with missing data*, J. Wiley & Sons, New York.

MADOW W.G., OLKIN I., RUBIN D.B. (1983), *Incomplete data in sample surveys*, Academic Press, New York.

MARASINI D. e OLIVIERI D. (1983), *Il campionamento con risposte casualizzate come strumento per migliorare la qualità del dato statistico* in Atti del Convegno 1983 della S.I.S., Trieste.

NETER J., WAKSBERG J. (1964), *A study of response errors in expenditures data from household interview*, Journal of American Statistical Association, Volume 59, pagg 18-55.

PLATEK R., GRAY G.B. (1986), *On the definition of response rate*, in Survey Methodology, Volume 12 Number 1 June 1986, Statistics Canada.

SINGH D., CHAUDHARY F.S. (1986), *Theory and analysis of sample survey design*, John Wiley & Sons, New York.

SUDMAN S., BRADBUM N.M. (1973), *Effects of time and memory factors on response in surveys*, Journal of American Statistical Association, Volume 68, n. 344, pagg. 805-815.

U.N. (1982), *National household survey capability programme. Non-sampling errors in household surveys: sources, assesment and control*, New York.

WALTER K.M. (1983), *Coverage error models for census and survey data* in Proceedings of 47th session of I.S.I., Madrid.

ZARKOVICH (1967), *Sampling methods and censuses - Volume II - Quality of statistica data*, F.A.O.

Lavori di carattere applicativo in Italia

CORTESE A. (1983), *Indagini sul confronto censimento-anagrafe: scopi, modalità d'esecuzione, principali risultati*, Atti del Convegno S.I.S., Trieste.

FABBRIS L. (1981), *Metodi statistici per l'analisi della qualità dei dati sanitari*, in «Statistica e Ricerca epidemiologica», Cleup, Padova.

FABBRIS L. (1984), *Question wording e selezione delle alternative di risposta in una indagine postale*, in Atti della XXXII Riunione scientifica della S.I.S., Sorrento.

MANGANO S. (1984), *Analisi dell'influenza dei rilevatori sulla qualità dei dati raccolti nel terzo censimento generale dell'agricoltura, attraverso il metodo dell'analisi della varianza*, Atti della XXXII Riunione scientifica della S.I.S., Sorrento.

MASSELLI M. (1983), *Risultati dell'indagine di controllo sulla qualità dei dati del censimento 1981*, Atti del Convegno S.I.S., Trieste.

MASSELLI M., TERRA ABRAMI V. (1983), *L'indagine di controllo di copertura del censimento della popolazione*, Atti del Convegno S.I.S., Trieste.

MASSELLI M. (1988), *L'errore di identificazione delle unità ed il sistema di controllo di un'indagine statistica. Una applicazione all'indagine sulle forze di lavoro*, Atti della XXXIV Riunione Scientifica della SIS, Siena, Vol. II, Tomo I, pp. 169-176.

ROSSI F. (1983), *Il controllo dei dati nel censimento della popolazione del 1981*, in Statistica n. 4.

SCHIRINZI G. (1986), *Alcune prime annotazioni sulla ripartizione delle aziende agricole secondo la superficie*, Atti del Convegno della S.I.S. su «Statistica e risorse naturali», Messina.

ZANNELLA F., SABBADINI L.L., BURATTA V. (1986), *Analisi dell'effetto proxy in alcune recenti indagini sulle famiglie condotte dall'Istat: primi risultati*, documento interno ISTAT.

ZANNELLA F., SABBADINI L.L., BURATTA V. (1986), *Analisi dell'effetto proxy nell'indagine sulle forze di lavoro del luglio 1986 - Risultati preliminari*, documento interno ISTAT.

CAPITOLO 4 - LA REGISTRAZIONE

1. Introduzione

Nel processo di produzione del dato statistico, la registrazione costituisce l'anello di congiunzione tra il supporto cartaceo (questionario) e quello informatico; rende cioè elaborabili le informazioni raccolte.

In particolare, il piano di registrazione consente il trasferimento dei dati dal modello di rilevazione ad un record il cui tracciato è suddiviso in campi di uno o più byte, istituendo una corrispondenza biunivoca tra ciascuno dei suddetti campi e le variabili del modello originale.

I tipi di errore

I tipi principali di errore che si possono commettere durante il processo di registrazione riguardano il *valore* del dato e la sua *posizione* nel record finale. Il caso di errore sul valore si verifica quando un certo carattere (alfabetico o numerico) viene letto o interpretato male e quindi registrato in modo scorretto, così da eliminare la coincidenza fra quanto scritto sul modello e quanto risulta sul record. Ad esempio, se in alcune parti il modello viene codificato manualmente, può accadere che la lettura di certi dati risulti difficile e che taluni simboli (come i numeri 6 e 0) vengano confusi con altri. Il secondo tipo di errore accade quando un carattere viene letto e digitato correttamente rispetto al suo *valore*, ma in una posizione errata sul record. Per esempio può succedere che venga inserita la digitazione di uno o più byte, o al contrario si introducano valori blank non previsti, determinando uno slittamento (shift) indietro o in avanti di parte dei dati rispetto al tracciato record di riferimento.

Quando questi errori si verificano, le elaborazioni successive generano risultati affetti da errore: è necessario pertanto da un lato cercare di ridurre le fonti di errore, e dall'altro individuare metodi che forniscano una valutazione quantitativa dell'errore commesso.

2. La perdita di informazione dovuta all'errore di registrazione

Un aspetto fondamentale di cui tener conto nelle analisi sugli errori di registrazione è il contenuto informativo dei codici: l'influenza sui risultati finali di uno scambio di valori da zero a blank può essere quasi nulla se si tratta di dati quantitativi di cui interessa la somma (es. spese di consumo per determinati beni), ma può viceversa essere rilevante se al blank viene attribuito il significato di mancata risposta e se le mancate risposte sono oggetto di particolari elaborazioni.

Incidenza dell'errore

Alcuni tipi di errore possono incidere notevolmente sulla coerenza interna del questionario e richiedere un successivo intervento da parte dei piani di correzione, ciò dipende sostanzialmente dal piano di codifica predisposto.

Alcune volte è possibile definire una gerarchia degli errori legata a quella delle variabili-guida dei piani di compatibilità (cfr. Capitolo 5).

È chiaro che se l'errore di registrazione interessa proprio una variabile-guida, ne può risultare inficiata la sequenza di campi che da questa dipendono; particolare attenzione, quindi, deve essere rivolta ai campi del record che riguardano le variabili-guida.

Per converso certi errori possono non incidere sulle elaborazioni conclusive se non alterano il dato, portandolo oltre i valori-soglia di classificazione. Ad esempio nella rilevazione delle forze di lavoro un'età del rispondente superiore ai 14 anni determina la compilazione del foglio individuale, per cui è rilevante il caso in cui un minore di 14 anni diventi, a causa di un errore di registrazione, maggiore di tale età; è invece abbastanza indifferente che un valore al di sotto del valore-soglia, pur essendo digitato erroneamente, rimanga all'interno della classe (fra 0 e 13 anni).

L'effetto di un errore di registrazione va quindi valutato in termini del suo *contenuto informativo* nel contesto del questionario.

L'importanza relativa degli errori induce inoltre a guardare con particolare attenzione alle variabili che definiamo *strutturali*, cioè quelle che, essendo di intestazione alle tabelle finali, vengono incrociate con le altre variabili: ad esempio il *sesso* e le *classi di età*. Per queste variabili l'analisi dell'errore di registrazione va effettuata molto accuratamente tenendo anche conto dell'eventuale correlazione fra gli errori.

Errori sui codici identificativi

Infine bisogna sottolineare che l'errore di registrazione divine gravissimo quando intacca i codici identificativi dei record (e delle unità di rilevazione corrispondenti): quando si effettua il riporto all'universo delle elaborazioni sui dati campionari utilizzando un codice identificativo sbagliato vengono ad essere stravolti i risultati complessivi dell'indagine, in quanto si attribuiscono ad uno strato non pertinente i valori di variabili che spetterebbero ad un altro strato. L'effetto dell'errore viene tanto più amplificato quanto maggiori sono i coefficienti di riporto degli strati coinvolti.

I codici identificativi possono poi essere utilizzati per confronti e agganci longitudinali quando è prevista la reintervista della stessa unità di rilevazione in periodi diversi, o ancora possono servire come informazione di base in successive fasi di controllo dell'indagine: anche in questi casi la presenza di errore nel codice identificativo comporta l'impossibilità di effettuare l'accoppiamento fra record o porta ad accoppiamenti scorretti.

Sembra quindi opportuno ribadire la necessità di una verifica particolarmente attenta della registrazione per quanto riguarda i suddetti campi, che andrebbero pertanto controllati con una seconda digitazione di verifica creando la condizione di *assenza di errori sui codici identificativi*.

3. Prevenzione dell'errore di registrazione

Allo scopo di migliorare la qualità della registrazione è utile effettuare un'analisi preliminare per individuare eventuali possibilità di prevenzione degli errori. Cominciando dal modello o questionario si possono dare alcune indicazioni di massima.

Si dovrebbe preferire, ove possibile, la pre-codifica, ciò significa stabilire a priori i codici che ciascun quesito ammette come risposta, e stampare ciascun codice accanto alla corrispondente casellina da barrare; questa attenzione, di solito utilizzabile per variabili qualitative, evita il ricorso a codifiche manuali da parte del rilevatore e/o revisore, il quale potrebbe introdurre caratteri poco comprensibili.

ESEMPIO

CODIFICA MANUALE	PRE-CODIFICA
Titolo di studio ☐	Titolo di studio
1 Laurea	Laurea 1 ☐
2 Diploma superiore	Diploma superiore 2 ☐
3 Diploma inferiore	Diploma inferiore 3 ☐
4 Elementare	Elementare 4 ☐
5 Nessuno	Nessuno 5 ☐

Se la risposta è «Diploma superiore» nel primo caso il rilevatore segnerà un «2» nell'unica casellina, nel secondo caso bifferà la casellina accanto al numero 2.

Campi fissi e campi a serrare

Un'altra scelta importante è fra la registrazione di variabili a campi fissi o a serrare. Nel caso di campi fissi esiste una corrispondenza precisa fra la risposta ai quesiti e la posizione del rispettivo codice sul tracciato record, e quando si verifica una

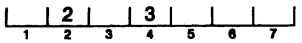
mancata risposta questo implica un blank in quella posizione. Coi campi a serrare invece il codice da registrare è univoco per ogni modalità di risposta, in modo che la posizione sul record non abbia rilevanza e sia possibile registrare questi codici di seguito, cioè senza inserire i valori blank.

ESEMPIO

CAMPI FISSI: Nel corso degli ultimi dodici mesi quante volte è stato fatto ricorso a uno dei seguenti medici specialistici o allo psicologo? Nel caso di visita indicare il tipo di servizio utilizzato

	Ricorso a servizio			
	pubblico	privato: per scelta	privato: per necessità	sia pubblico che privato
Dentista	1 ☐	2 ☐	3 ☐	4 ☐
Oculista	1 ☐	2 ☒	3 ☐	4 ☐
Cardiologo	1 ☐	2 ☐	3 ☐	4 ☐
Ortopedico	1 ☐	2 ☐	3 ☒	4 ☐
Endocrinologo	1 ☐	2 ☐	3 ☐	4 ☐
Psicologo	1 ☐	2 ☐	3 ☐	4 ☐
Altro	1 ☐	2 ☐	3 ☐	4 ☐

Il tracciato record corrispondente alle risposte barrate è il seguente:



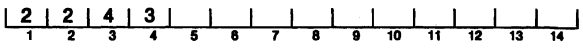
L'uso dei campi a serrare è vantaggioso quando ci sono domande complesse e/o multiple con lunghe sequenze di codici uguali che possono generare errori di registrazione, e consente di eliminare i valori blank intermedi che potrebbero dar luogo a slineamenti. I campi a serrare inoltre danno l'opportunità a chi registra di leggere i codici indifferentemente per riga o per colonna, quindi con maggior velocità.

D'altra parte comunque bisogna rilevare che i campi a serrare necessitano in genere di un numero di byte superiore rispetto ai campi fissi, dato che i codici previsti occupano due posizioni; questo può essere problematico se esistono limiti alla dimensione del record.

CAMPI A SERRARE: Nel corso degli ultimi dodici mesi quante volte è stato fatto ricorso a uno dei seguenti medici specialistici o allo psicologo? Nel caso di visita indicare il tipo di servizio utilizzato

	Ricorso a servizio			
	pubblico	privato: per scelta	privato: per necessità	sia pubblico che privato
Dentista	11 ☐	12 ☐	13 ☐	14 ☐
Oculista	21 ☐	22 ☒	23 ☐	24 ☐
Cardiologo	31 ☐	32 ☐	33 ☐	34 ☐
Ortopedico	41 ☐	42 ☐	43 ☒	44 ☐
Endocrinologo	51 ☐	52 ☐	53 ☐	54 ☐
Psicologo	61 ☐	62 ☐	63 ☐	64 ☐
Altro	71 ☐	72 ☐	73 ☐	74 ☐

Il tracciato record corrispondente alle risposte barrate è il seguente:



Altri suggerimenti che si possono aggiungere riguardano la predisposizione di un numero di byte adeguato e sufficiente a contenere tanto le risposte quantitative, nel caso in cui il valore massimo effettivo di una variabile superi quello ipotizzato, quanto i codici per variabili qualitative con risposte aperte.

Un ulteriore fattore importante è l'esatta definizione del piano di registrazione in cui alla descrizione dettagliata del tracciato record si affiancano indicazioni precise sulla compatibilità dei codici nei singoli campi.

4. Controllo amministrativo e statistico

Il controllo della qualità nella fase della registrazione attualmente effettuato dall'Istituto Nazionale di Statistica avviene in modo differenziato a seconda che quest'operazione venga eseguita all'interno o all'esterno (in service).

Nella registrazione interna i record sono sottoposti ad un controllo «leggero» interattivo e nel reparto stesso la registrazione viene controllata da *revisori*.

Diverso è il procedimento di controllo per la registrazione appaltata a ditte esterne: definita per contratto una soglia di errore (in percentuale sui byte digitati) al di sopra del quale l'Istituto ha facoltà di rifiutare lo stock di record registrati e di chiederne la ridigitazione, si prende in esame un campione di questionari e i relativi record. I modelli campionati vengono nuovamente registrati e verificata la coincidenza con quelli della prima registrazione della ditta: se la percentuale di byte errati supera la prefissata soglia di errore (5 per mille), l'intero stock viene rigettato, altrimenti esso viene considerato sufficientemente affidabile e quindi accettato.

L'errore totale

Questa procedura di verifica, che viene eseguita su richiesta del reparto responsabile dell'indagine, produce come risultato finale una stima dell'*errore totale*, espresso dal rapporto fra numero di byte errati e numero di byte utili (cioè quelli effettivamente utilizzati nel record), oltre che un certo numero di statistiche sul numero di errori per record, sui record saltati e duplicati ecc.

Pur essendo utilizzato a fini amministrativi, il dato sull'*errore totale* è poco indicativo dell'effettivo livello di *qualità* della registrazione e non dice nulla sulla tipologia degli errori commessi e sulla loro distribuzione all'interno e fra i record.

Si consideri a titolo di esempio il caso di 10 record lunghi 100 byte che, sottoposti a controllo, presentano un errore totale del 4 per mille, quindi al di sotto del valore-soglia; facendo l'ipotesi che la distribuzione degli errori sia tale da generare un solo byte errato per record si ottiene implicitamente un totale di quattro record errati ogni 10: in definitiva il 40% di record errati.

Naturalmente in realtà si verificano errori multipli sullo stesso record, cosicché la suddetta percentuale tende ad abbassarsi, ma è evidente come anche un valore abbastanza piccolo dell'errore totale calcolato sul numero delle battute possa incidere pesantemente sulla percentuale di record errati.

I record errati

Alcune verifiche empiriche (cfr. Zuchegna A.) su record del censimento della popolazione hanno riscontrato una percentuale di record errati del 13% in corrispondenza di un errore totale del 5 per mille.

L'esempio riportato vuole sottolineare lo scarso contenuto di informazione dell'errore totale e la necessità di elaborare altre informazioni disponibili o ricavabili dai dati provenienti dal controllo amministrativo, al fine di costruire indicatori specifici dell'entità e del tipo di errore, e quindi passare ad un *controllo statistico* che valuti sia i byte sia i campi errati: ad esempio il numero di record che contengono almeno un byte errato, il numero di byte errati per record, il numero di variabili (campi) errate, la distribuzione degli errori per record e per variabili, ecc..

5. Il controllo a campione

Per il controllo della qualità della registrazione si utilizza un campione di modelli, che vengono nuovamente digitati e confrontati con quelli provenienti dalla registrazione originale. L'obiettivo del controllo amministrativo è di pervenire ad una decisione circa il livello dell'errore totale e richiede quindi una verifica sul numero di byte utili errati: sono quindi i byte a costituire l'universo di riferimento da cui estrarre il campione. L'estrazione e la registrazione di singoli byte o di campi da ogni modello si presenta però molto laboriosa e di notevoli difficoltà organizzative, in quanto considerare un solo elemento per record costringerebbe a lavorare e a maneggiare un numero cospicuo di pacchi di modelli; pertanto, più agevolmente, una volta estratto un modello, vengono considerati nel campione tutti i byte in esso contenuti.

L'effetto «cluster»

Lo schema di campionamento dei byte è quindi di tipo cluster, cioè a grappoli o gruppi di byte che, appartenendo allo stesso modello o a un pacco di modelli sono in qualche modo omogenei fra loro (stesso operatore che li ha registrati, stesso Comune di rilevazione ecc.) e meno rappresentativi della variabilità dell'universo di riferimento. Il campionamento cluster implica un aumento della varianza delle stime rispetto al campionamento casuale semplice dei singoli byte e richiede, per mantenere la bontà delle stime a livello desiderato, che il campione di record abbia una numerosità adeguata. L'effetto cluster agisce anche quando volendo, effettuare controlli statistici su record individuali, si considera l'intero modello familiare, quindi un *grappolo* di individui.

Per esaminare l'effetto di un campionamento cluster consideriamo brevemente i risultati di una simulazione, effettuata sia nell'ipotesi che gli errori di registrazione fossero di tipo casuale, sia che fossero di tipo sistematico al fine del calcolo del «Deff», cioè della perdita di precisione delle stime del campionamento a grappoli rispetto al campionamento casuale semplice.

Se supponiamo di dividere la popolazione originaria (l'insieme di tutti i byte digitati) in S subpopolazioni (S questionari) che contengono M elementi, fissata la numerosità n del campione, non è indifferente procedere ad una estrazione casuale di n elementi o estrarre un certo numero K di subpopolazioni di M elementi da sondare, con il vincolo che $K*M >= n$. La differenza sta nella precisione delle stime ottenute con i due metodi, tendendo quella del secondo ad essere inferiore in funzione delle varianze interne alle singole subpopolazioni, che determinano il cosiddetto *effetto cluster*.

Per valutare l'entità di questo effetto sulla precisione delle stime del parametro «p» (percentuale di byte errati) e per corregere eventualmente la numerosità campionaria per tenere conto dell'ipotesi di estrazione di record invece di byte, quale indicatore di base di è utilizzato il «Deff», definito come il rapporto fra il valore della varianza della stima nell'ipotesi di campionamento cluster ed il valore della stessa varianza nell'ipotesi di campionamento casuale semplice, a parità di numerosità campionaria. Un valore del «Deff» prossimo all'unità indica assenza di *effetto cluster*, mentre valori via via superiori si rilevano in presenza di *effetto cluster* crescente, fino al punto che un «Deff» quasi uguale a M (numero di elementi per cluster) dovrebbe indurci ad estrarre K=n record di numerosità M per avere la stessa precisione del campione semplice di byte di numerosità n.

Se l'errore è casuale

Per verificare e valutare l'«effetto cluster» in assenza di errore sistematico un primo insieme di prove è stato eseguito simulando vari valori di «p», diverse lunghezze del record su K = 100 record, dando luogo ai risultati della Tavola 4.1.

Tavola 4.1 - Errori casuali

P	M	Repl.	Deff medio	Var. Deff	p stima
0.01	20	20	0.977	0.124	0.0094
0.005	20	100	0.991	0.119	0.0047
0.01	20	100	0.991	0.123	0.0096
0.01	100	100	0.992	0.129	0.0097

p = percentuale di byte errati; M = lunghezza del record; repl. = numero di replicazioni della procedura per il calcolo del deff medio; deff medio = deff medio sulle replicazioni; var deff = varianza del deff nelle replicazioni; p stima = valore stimato di p (in media).

Si osserva che, in assenza di errori sistematici, la precisione delle stime che si ottiene considerando ad esempio 100 record lunghi 20 al posto di 2000 singoli bytes non risente dell'*effetto cluster*. I due metodi sono sostanzialmente equivalenti, risultando quello per record più economico in termini di tempo e difficoltà di esecuzione. Quindi se si considerano K cluster (record) di numerosità M (numero di byte per record) per ottenere un errore campionario minore o uguale a quello che si otterrebbe estraendo n = K*M singoli byte è sufficiente (approssimativamente) utilizzare n/M = K cluster di M elementi.

Ad esempio dovendo estrarre, secondo un piano di campionamento casuale semplice, un campione di n=2000 byte in presenza di record lunghi M=80 potremmo considerare un campionamento di K record con K = 2000 / 80 = 25.

Queste considerazioni valgono in presenza di errori puramente casuali all'interno del record, mentre l'effetto «cluster» manifesta più problematici effetti in caso di errori sistematici, come dimostrano le prove successive.

Se l'errore è sistematico

Definiamo errore sistematico quello che accade al verificarsi di una condizione di errore su altri elementi, ad esempio un valore replicato identicamente su più byte contigui o su più record «vicini», oppure una certa «costanza» nell'interpretare certi codici. Abbiamo considerato il caso semplice di una procedura che genera un errore sul byte *x-esimo*, oltre che per evento casuale, ogni volta che si è determinato un errore sul byte *y-esimo*. Oltre all'errore sistematico singolo sono valutati gli effetti di errori multipli, cioè quelli che mettono in relazione coppie di byte, ed infine abbiamo considerato l'errore derivato dalla generazione di un record in cui i byte successivi al byte *x-esimo* sono tutti errati, ad esempio per uno slineamento.

Tavola 4.2 - Errori sistematici

Tipo errore	P	M	Repl.	Deff medio	Var. Deff	p stima
1 sist.	0.01	20	20	1.17	0.180	0.0091
2 sist.	0.005	20	100	1.08	0.176	0.0052
2 sist.	0.01	20	100	1.15	0.185	0.0108
5 sist.	0.01	20	100	1.41	0.263	0.0093
1 t.r.	0.005	20	100	6.98	6.980	0.0096
1 t.r.	0.005	80	100	26.64	28.150	0.0100
1 t.r.	0.005	100	100	27.45	34.144	0.0092

p = percentuale di byte errati; M = lunghezza del record; repl. = numero di replicazioni della procedura per il calcolo del deff medio; deff medio = deff medio sulle replicazioni; var deff = varianza del deff nelle replicazioni; p stima = valore stimato di p (in media); sist. = errore sistematico; t.r. = errore sistematico su tutto il record da un certo byte in poi.

Le simulazioni in gruppi di 100 record fanno riscontrare un aumento molto accentuato del «Deff», tanto al crescere del numero di campi coinvolti nell'errore sistematico quanto al crescere della lunghezza del record. Per record lunghi 80 byte ad esempio il «Deff» medio si assesta attorno al 26.6 nel totale delle replicazioni, comprendendo cioè sia i casi in cui effettivamente un errore sistematico sia stato generato sia i casi in cui questo non avviene.

In realtà quando la procedura di simulazione comporta l'effettiva generazione di un errore su tutto il record a causa di un

errore sul'x-esimo byte, il «Deff» assume valori che variano fra 47.9 e 69.3, implicando che per mantenere la precisione delle stime sarebbe necessario aumentare la numerosità del campione di record di almeno 70 volte, cioè secondo il caso ipotizzato sopra, estrarre (2000/80)*70 = 1850 record.

In pratica se c'è un errore sistematico il campionamento a grappoli è molto inefficiente, dando luogo anche a stime di «p» molto diverse dal vero valore: la numerosità campionaria pertanto andrebbe aumentata fino a far coincidere il numero K di gruppi (record) col numero originario di byte (elementi) da estrarre nel campionamento casuale semplice.

Per quanto riguarda il campionamento per attributi esposto nel Paragrafo 6 si conviene che, quando si abbia il forte sospetto della presenza di errori sistematici, nel caso si voglia utilizzare il controllo con l'approssimazione binomiale si adotterà una numerosità campionaria K di record pari al numero n di byte previsti per l'estrazione casuale semplice; se si vogliono adoperare le Tavole Military Standard si adotterà il campione *rinforzato*, con livello III (general inspection level) di numerosità campionaria.

6. Definizione degli standard di qualità

Per valutare se il materiale proveniente dalla registrazione è affidabile e quindi predisporre la procedura di controllo è fondamentale la definizione del livello di qualità che si ritiene accettabile o auspicabile, in modo da poter determinare un piano di campionamento che, con prefissata probabilità di errore, consenta di accertare se la percentuale di errore nel file registrato soddisfa o meno il prefissato standard.

Possiamo considerare diversi approcci per definire la quantità di errori riferendoci a:

1. numero di byte errati in totale sul numero di byte utili (errore totale), ad esempio 5 per mille;
2. numero medio di record errati sul totale dei record digitati, ad esempio 5 per cento;
3. esame complessivo dei due parametri precedenti.

Per ciascuno di questi casi si specifica una procedura di controllo.

Byte errati

CASO 1) Definiamo innanzitutto le battute utili del record come i *byte* effettivamente *occupati* da valori non sempre nulli sul tracciato record: se ad esempio un record fisico è lungo 80 colonne, ma ne vengono utilizzate soltanto 55, diremo che quest'ultimo è il numero di battute utili.

Dal punto di vista amministrativo, per la valutazione dei costi di registrazione in service, i quali sono proporzionali al numero di battute, vengono talvolta escluse dal conteggio le battute di più blank consecutivi o le battute in duplica, ma, relativamente al problema del controllo di qualità, è più conveniente considerare le battute utili come colonne del record occupate da variabili, anche se in queste risulteranno talvolta valori nulli.

Quando il questionario richiede la registrazione su più record fisici si effettueranno i passi della procedura di controllo campionario considerando l'insieme dei record dello stesso questionario come unità di estrazione, eseguendo i calcoli sul numero complessivo di battute.

Record errati

CASO 2) Alternativamente la qualità della registrazione può essere valutata basandosi sul numero di *record* errati (con almeno un errore): si applicherà allora una procedura di controllo in cui, pur rimanendo l'unità di rilevazione il questionario, la difettosità del lotto sarà riferita al record errato e non alla battuta errata.

Per quanto riguarda la determinazione del livello di qualità accettabile bisogna tener conto del fatto che una ridotta difettosità in termini di percentuale di battute errate può implicare un elevato numero di record errati. Se consideriamo che molto spesso le correzioni degli errori attraverso i piani di compatibilità utilizzano come riferimento i record «completi» (cioè quelli senza alcun errore) si capisce come sia importante cautelarsi rispetto alla possibilità che questi «scarseggino».

Tavola 4.3 - Percentuale di record errati al variare del numero di byte nel record, p = 5 per mille.

n numero byte nel record	$\lambda = np$	record senza errori (su 1000)	record con almeno un err. (su 1000)
50	0.25	779	221
60	0.30	741	259
70	0.35	705	295
80	0.40	670	330
90	0.45	638	362
100	0.50	607	393
110	0.55	577	423
120	0.60	549	451
130	0.65	522	478
150	0.75	472	528
200	1.00	368	632

Per fare un esempio abbiamo ipotizzato la sola presenza di errori casuali (situazione più sfavorevole rispetto alla distribuzione degli errori fra i record a parità di «p») e considerato una difettosità «p» del 5 per mille: si è utilizzata la distribuzione di Poisson per calcolare la percentuale teorica di record con almeno un errore al variare del numero di byte per record. La tavola 4.3 illustra i risultati ottenuti.

La probabilità che si riscontri un errore di registrazione all'interno di un record aumenta ovviamente all'aumentare della sua lunghezza, a parità di «p».

Byte e record errati

CASO 3) Una terza possibilità è rappresentata dalla considerazione congiunta degli errori sui *byte* e sui *record*. Si osserva allora che a parità di lunghezza del record la percentuale di errore calcolato sui record varia proporzionalmente al variare della percentuale «p» di errori sui byte. Nella tavola 4 abbiamo preso un record con un numero di byte prefissato pari a 100 per valutare la percentuale teorica di record errati in relazione a diversi valori di «p».

Tavola 4.4 - Percentuale di record errati al valore di «p», n = 100

p errore sui byte	$\lambda = np$	record senza errori (su 1000)	record con almeno un errore (su 1000)
0.0001	0.01	990	10
0.001	0.10	904	96
0.002	0.20	818	182
0.004	0.40	670	330
0.005	0.50	606	394
0.010	1.00	367	633
0.025	2.50	82	918
0.040	4.00	18	982
0.050	5.00	6	994
0.100	10.00	0	1000

È quindi necessario arrivare ad un compromesso fra percentuale «p» di errore sui byte e percentuale di record errati: nell'esempio, per ottenere il 9,6% di record errati bisogna stabilire un valore di «p» attorno all'1 per mille.

Sempre riguardo alla determinazione degli standard di qualità è necessario sottolineare che esistono due approcci al problema dai quali derivano due distinte metodologie.

Il primo è legato alla scelta di un unico parametro di qualità (p = percentuale di difettosi) che discrimina fra l'accettazione ed il rifiuto dei risultati della registrazione: il metodo da applicare in questo caso è quello statistico degli intervalli di confidenza (o verifica d'ipotesi) sul detto parametro.

Nel secondo approccio sono invece previste le due figure del fornitore (la ditta di registrazione) e dell'acquirente (l'Istat) a ciascuno dei quali viene attribuito un livello di qualità: LQA è il livello *buono* (accettabile) per il quale il fornitore è quasi certo dell'accettazione da parte dell'acquirente e al quale cerca di adeguarsi; LQT è il livello di qualità *cattivo* minimo, che il fornitore sa verrà rifiutato dalla controparte.

Questi due livelli di qualità servono alla costruzione della *curva operativa caratteristica* che sta alla base dei metodi di controllo della qualità industriale esaminati di seguito.

7. Piani di campionamento singolo per attributi

Se si assimila il processo di registrazione ad un processo produttivo in cui il pezzo prodotto è il singolo dato (o record) digitato è possibile applicare alla registrazione alcuni controlli utilizzando piani di campionamento ideati per i controlli industriali.

Nel caso della registrazione si tratta di verificare se il dato è digitato correttamente, cioè se è *buono*, o invece è *difettoso*: è opportuno allora utilizzare per il controllo statistico un piano di campionamento singolo per attributi, dove la caratteristica qualitativa da studiare è appunto la *difettosità*.

Il test di controllo viene effettuato mediante un piano di campionamento singolo (con una sola estrazione) per attributi (che discrimina fra pezzi difettosi e non). Dato un lotto di pezzi di numerosità N, un piano di campionamento singolo è definito da due parametri: n, la dimensione del campione e c, il numero di accettazione, cioè il numero di pezzi difettosi che si è disposti ad accettare nel campione senza che questo comporti la decisione di considerare inaccettabile la qualità complessiva della produzione, e di respingere pertanto il lotto in esame

Questi due valori vengono fissati sulla base di:

N dimensione del lotto;
LQA = p1 livello di qualità accettabile;
LQT = p2 livello di qualità tollerata (o rifiutabile);
$1 - \alpha$ probabilità di accettazione se p = p1, (α di solito uguale al 5%);
β probabilità di accettazione se p = p2, (β di solito uguale a 10%);
dove p indica la vera (e ignota) qualità del lotto espressa come percentuale di pezzi difettosi (a).

(a) Nel linguaggio tecnico proprio del controllo statistico industriale se p1 è il livello di qualità accettabile (AQL in inglese) e p2 il livello di qualità tollerata (LTPD in inglese) si usa dire che $1-\alpha$ è il rischio del fornitore (rischio che essendo la qualità *buona* il lotto venga rifiutato) e β è il rischio dell'acquirente (rischio che essendo la qualità *cattiva* il lotto venga accettato): questi valori definiscono due punti sulla curva operativa caratteristica (OC) che descrive al variare di p (% difettosi nel lotto) la probabilità di accettare il lotto e mostra la capacità discriminatoria del disegno campionario.

Se trattiamo un lotto di dimensione finita (N non eccessivamente grande) la distribuzione dell'errore nel lotto sarà una variabile casuale ipergeometrica:

$$p(d) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \qquad (4.1)$$

dove D è il numero di pezzi difettosi totali del lotto, N è la numerosità del lotto ed n la dimensione del campione, d il numero di difettosi nel campione.

Fissata la probabilità α (errore di prima specie) di rifiutare il lotto di record in presenza di una «buona» qualità (p1) dovremo determinare la numerosità campionaria n e il valore di accettazione c, tale per cui se il numero di errori riscontrati è superiore a c, il lotto non viene accettato. Porremo quindi:

$$p(c) \mid p1 > 1 - \alpha \qquad (4.2)$$

in cui la probabilità condizionata P(c) | p1 coincide con la (4.2) quando a «d» sostituiamo «c» e a «D» sostituiamo l'approssimazione D1 = p1*N.

Il vincolo (4.2) non è però da solo sufficiente a determinare entrambi i parametri (c ed n): si definisce allora una qualità *scadente* (p2) del lotto che vogliamo accettare con probabilità β molto bassa (di solito uguale a 0.10) cosicché:

$$P(c) \mid p2 > \beta \qquad (4.3)$$

che è la (4.2) con D = D2 = p2*N.

Risolvere simultaneamente per n e c le due equazioni (2 e 3) è però molto complicato e laborioso a causa dei calcoli richiesti dalla ipergeometrica. Si preferisce pertanto nella pratica determinare il piano di campionamento:

1. con l'approsimazione Binomiale, valida per N grande e per tassi di campionamento piccoli, n/N < 10%.
2. con le tavole Military Standard 105D per qualsiasi valore di N.

Metodo dell'approssimazione Binomiale

Per disegnare il piano di campionamento di accettazione (cioè per determinare n e c) è conveniente impiegare l'approssimazione Binomiale in luogo della complessa variabile ipergeometrica, fissati α, β, p1 e p2 le equazioni (4.2) e (4.3) divengono:

$$1 - \alpha = \sum_{d=o}^{c} \frac{n!}{d!\,(n-d)!} * p1^{d} * (1 - p1)^{n-d} \qquad (4.4)$$

$$\beta = \sum_{d=o}^{c} \frac{n!}{d!\,(n-d)!} * p2^{d} * (1 - p2)^{n-d} \qquad (4.5)$$

Le soluzioni delle due equazioni non lineari e simultanee sono ricavabili graficamente dal *nomografo* riportato nella Figura 4.1 seguendo una semplice procedura: si tracciano due rette che congiungono p1 a 1—α e p2 a β; l'intersezione delle due linee descrive una regione in cui giacciono varie possibili coppie di valori n e c; la scelta di una di queste coppie fornisce il piano di campionamento desiderato. Ad esempio se α = 0.05, p1 = 0.01, β = 0.10, p2 = 0.06 la procedura grafica definisce un'area in cui possiamo selezionare diverse coppie di valori n e c; n = 89 e c = 2 potrebbe essere un piano di campionamento appropriato. Oltre al procedimento grafico sono disponibili alcune tavole da cui ricavare n e c in funzione del livello di qualità desiderato (cfr. Duncan A.J., 1974).

Metodo delle Tavole Military Standard 105D

Queste tavole forniscono tipi di piani di campionamento standard a diversi livelli di ispezione:

- *normale* da utilizzarsi all'inizio dell'attività di controllo;
- *rinforzato* da usarsi quando la qualità del fornitore si è recentemente deteriorata;
- *ridotto* da usarsi quando la qualità del fornitore si è portata recentemente a livelli eccezionalmente buoni.

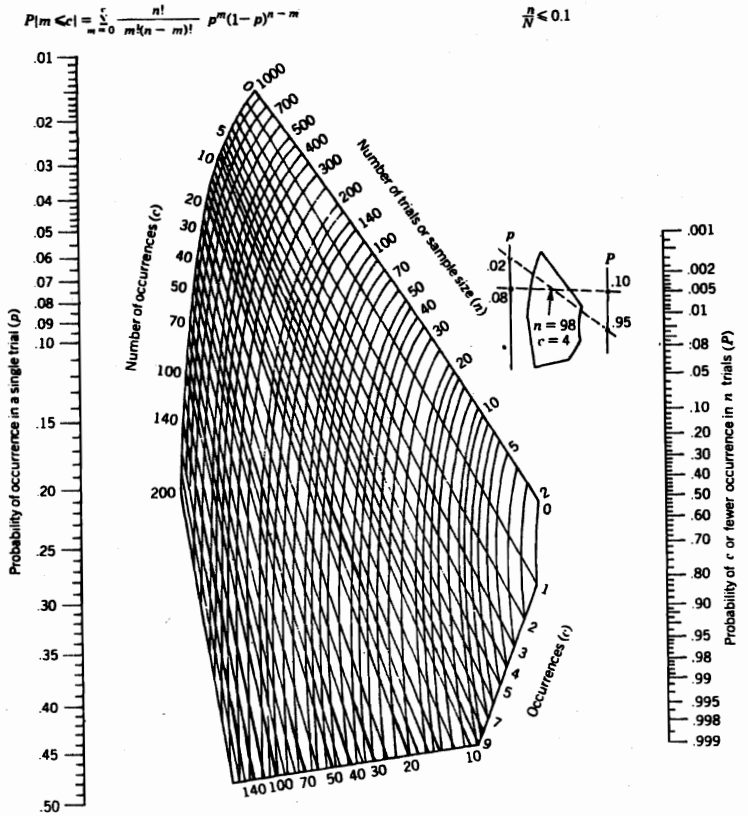
La procedura per un piano di campionamento singolo con le tavole MIL STD 105D è la seguente:

- si sceglie il livello di qualità accettabile (AQL, Acceptance Quality Level) espresso in percentuale;
- Si sceglie il livello generale di ispezione (relativo alla maggiore o minore numerosità campionaria (basso = I, medio = II, alto = III);
- si trova nelle Tavole I la «lettera-codice» (Sample Size Code Letter) corrispondente ai parametri (AQL ed n) sopra citati;
- si entra in una delle Tavole II, a seconda del livello di ispezione scelto (ridotto = reduced –> Tavola II-A; normale = Normal –> Tavola II-B; rinforzato = Tightened –> Tavola II-C) per trovare il piano di campionamento (n = sample size, c = Ac = Acceptance number).

Example:
Required: a sampling plan having

$$P = 0.95 \text{ at } p = 0.02$$
$$P = 0.10 \text{ at } p = 0.08$$

Solution: make alignments and read sample size (n) and acceptance number (c) as in diagram below:



Note:
If p is less than 0.01, set $k \times p$ on the p-scale and multiply the values on the n-scale by k, where $k = 0.01/p$ (taking k to the next higher integer).

Figura 4.1 - Nomografo binomiale

Lot or batch size	Special inspection levels					General inspection levels		
	S-1	S-2	S-3	S-4	I	II	III
2 to 8	A	A	A	A	A	A	B
9 to 15	A	A	A	A	A	B	C
16 to 25	A	A	B	B	B	C	D
26 to 50	A	B	B	C	C	D	E
51 to 90	B	B	C	C	C	E	F
91 to 150	B	B	C	D	D	F	G
151 to 280	B	C	D	E	E	G	H
281 to 500	B	C	D	E	F	H	J
501 to 1200	C	C	E	F	G	J	K
1201 to 3200	C	D	E	G	H	K	L
3201 to 10000	C	D	F	G	J	L	M
10001 to 35000	C	D	F	H	K	M	N
35001 to 150000	D	E	G	J	L	N	P
150001 to 500000	D	E	G	J	M	P	Q
500001 and over	D	E	H	K	N	Q	R

Figura 4.2

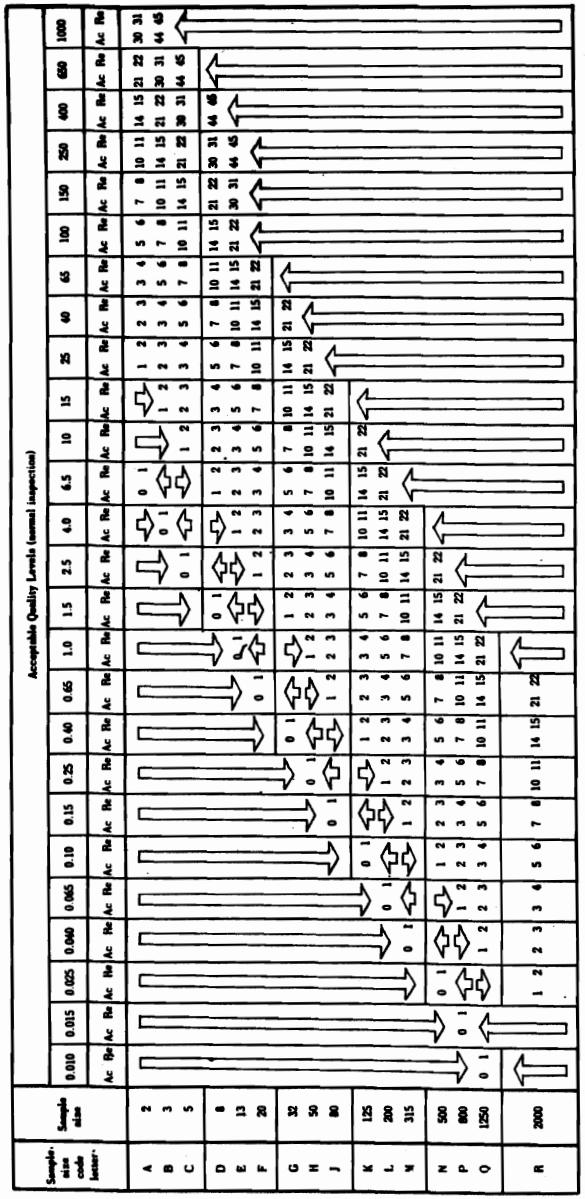


Figura 4.3

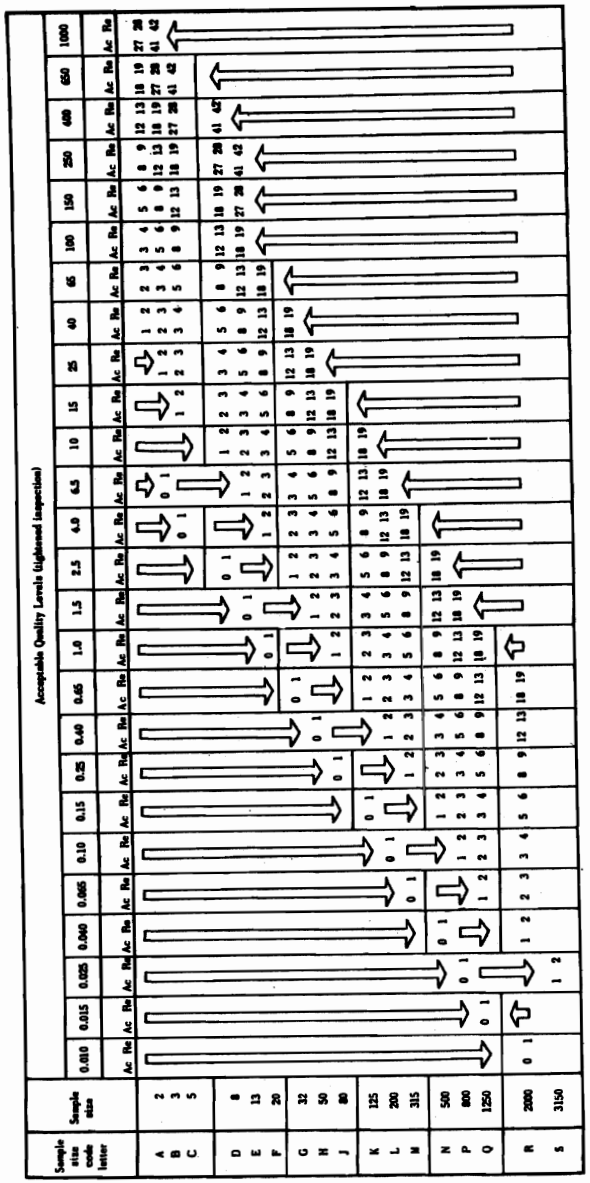


Figura 4.4

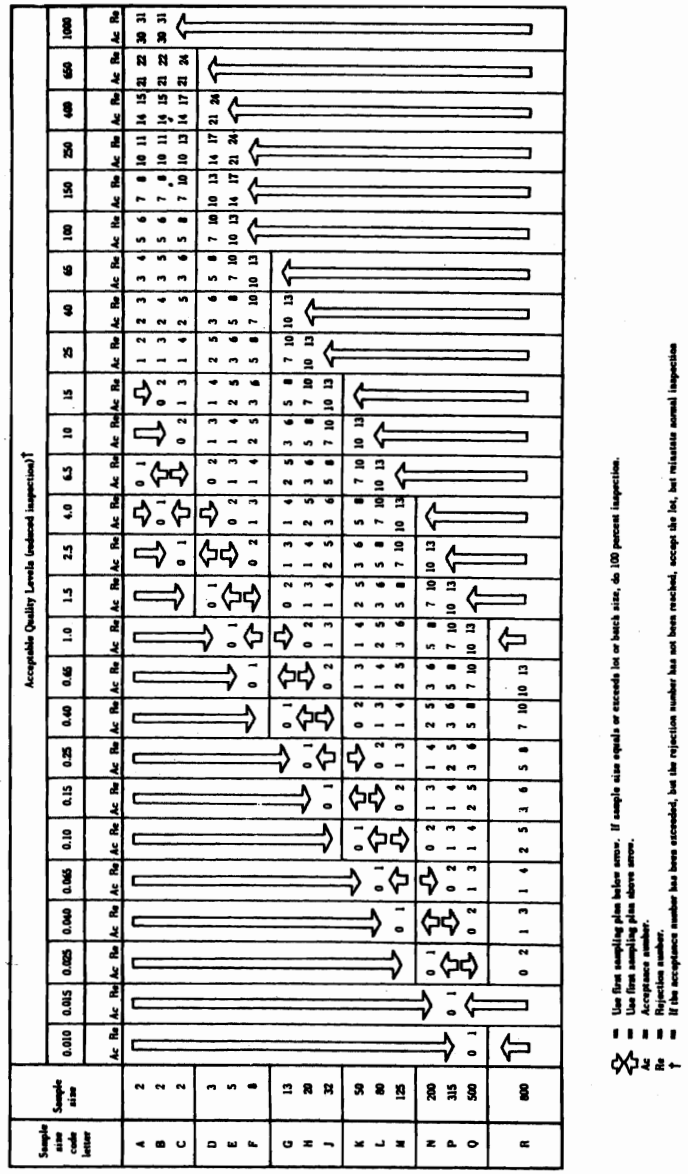


Figura 4.5

Estratto il campione casuale di numerosità n si confronta il numero di pezzi difettosi D con il numero di accettazione c:

se D > c il lotto verrà rifiutato;
se D < c il lotto verrà accettato.

Per un esempio di piani di campionamento semplice per attributi si confronti l'Appendice 3.

8. Test sequenziali

I metodi citati finora sono metodi a posteriori, che vengono applicati a registrazione ultimata. In alcuni casi può essere utile invece seguire il processo della registrazione nel suo svolgimento ed eseguire controlli in corso d'opera con test di tipo sequenziale. Questo approccio dei controlli di qualità consente di intervenire prontamente se si verificano situazioni anomale o quando il livello di errore sia superiore a quello stabilito. Pur richiedendo un certo sforzo organizzativo, se la registrazione viene eseguita all'interno oppure se c'è una certa regolarità nel flusso dei *ritorni* dalla registrazione esterna, il controllo sequenziale risulta sicuramente vantaggioso.

Nel campionamento sequenziale si considera una serie di campioni estratti successivamente da un lotto di pezzi prodotti (byte o record registrati). Il numero di campioni da estrarre è determinato dai risultati del processo di campionamento stesso: potrebbe teoricamente continuare all'infinito, ma in pratica si usa troncare dopo che il numero di pezzi ispezionati è circa pari a tre volte quello corrispondente nel piano di campionamento semplice. La dimensione del campione via via estratto può essere unitario (elemento per elemento) o maggiore di uno (per gruppi).

Esaminiamo il caso più frequente del campionamento sequenziale per singoli elementi (singoli record): si predispone un grafico, come nella Figura 4.6, in cui sull'ascissa si osserva il numero totale di pezzi estratti e sull'ordinata il numero di pezzi difettosi osservati.

Se nel processo di controllo i punti giacciono nella zona compresa fra la linea di accettazione XA e la linea di rifiuto XR, definite dalle formule date di seguito, si estrae un nuovo elemento. Quando un punto cade al di sopra della linea di rifiuto XR si rifiuta il lotto, quando cade al di sotto di quella di accettazione XA si accetta il lotto.

Le formule che consentono di tracciare le linee XA e XR si rcavano tenendo conto che:

n numero di unità estratte fino a quel momento;
p1 livello di qualità accettabile;

p2 livello di qualità tollerata (o rifiutabile);
$1 - \alpha$ probabilità di accettazione se p = p1;
β probabilità di accettazione se p = p2;

dove p è la percentuale di pezzi difettosi nel lotto.

$$XA = -h1 + s*n \quad (4.6)$$
$$XR = h2 + s*n \quad (4.7)$$

con:

$$h1 = (\log \frac{1-\alpha}{\beta}) / k \quad (4.8)$$

$$h2 = (\log \frac{1-\beta}{\alpha}) / k \quad (4.9)$$

$$k = \log \frac{p2*(1-p1)}{p1*(1-p2)} \quad (4.10)$$

$$s = (\log \frac{1-p1}{1-p2}) \quad (4.11)$$

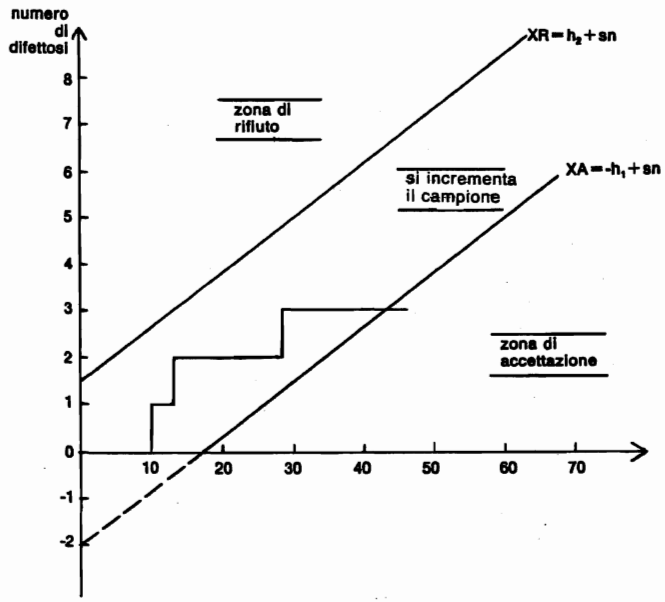


Figura 4.6

9. Analisi dei risultati campionari

Una volta che si disponga dei dati campionari si potrà valutare la quantità di errore a diversi livelli di analisi. Per il controllo amministrativo si calcolerà la percentuale di errore sul byte calcolando l'errore totale definito nei paragrafi precedenti (cfr. § 3). Per il controllo statistico l'errore potrà esser esaminato sia dal punto di vista dei record che dal punto di vista delle variabili.

Analisi statistica degli errori

Nel primo caso un'analisi a livello di record oltre a misurare la percentuale di record errati (con almeno un errore) sul totale dei record, potrà fornire la distribuzione dei record in funzione del numero di errori e l'errore medio (in termini di byte o di variabili) per record (cfr. Appendice 1).

Nel secondo caso si osserverà ad esempio la percentuale di errore sui *campi*, cioè sugli insiemi di byte contigui che definiscono il valore di una variabile: questo sia considerando il semplice rapporto fra il numero totale di campi errati e il numero di campi digitati, sia ricavando la distribuzione di frequenza degli errori per ogni singola variabile, tenendo conto che il peso degli errori andrà rapportato alla lunghezza del campo che definisce la variabile stessa e/o al numero di codici previsti come valore della variabile.

Di particolare interesse è la determinazione della casualità o meno dell'errore generato dalla fase di registrazione, quindi della presenza o meno di variabili o di posizioni più errate di altre.

Se gli errori di registrazione sono puramente casuali essi saranno distribuiti in maniera *neutrale* rispetto alle variabili successivamente elaborate, cioè tenderanno a compensarsi; se invece alcuni tipi di errore si verificano con maggior frequenza toccando particolari variabili si potranno determinare distorsioni nei risultati finali.

Test preliminari

La presenza di *errori sistematici* di registrazione può essere pertanto preliminarmente testata per eliminarne l'incidenza sui successivi passi di elaborazione, sia globalmente attraverso test di adattamento CHI QUADRATO, sulla distribuzione degli errori per modalità dalla variabile, oppure a livello di byte o di variabili, attraverso le *matrici di transizione*, a cui si applicano opportuni test.

Soprattutto per le variabili-guida dei piani di compatibilità (quelle da cui dipendono i valori accettabili di variabili gerarchicamente inferiori) sarà importante individuare la presenza di errori sistematici che potrebbero indurre correzioni improprie (e talvolta sistematiche) su altre variabili.

Inoltre, almeno per le variabili *strutturali*, quelle utilizzate come criterio di classificazione nelle tavole di pubblicazione, sarà necessario verificare la presenza di *errori correlati* fra variabili. Infatti può accadere che l'errore di registrazione si compensi all'interno della marginale (supponiamo ad esempio che la proporzione di maschi e femmine risultante dalla registrazione sia accettabile), ma che l'errore sulla variabile *sesso* sia correlato con qualche altra variabile (continuando l'esempio che all'errore *maschio* registrato come *femmina* si associ la variazione da *occupato* a *in cerca di occupazione*). Nella tabella che incrocia la variabile strutturale (sesso) con la variabile correlata (condizione) si otterrà una distribuzione delle frequenze sbilanciata verso alcune caselle (per es.: molte *femmine in cerca di occupazione*).

Vista la possibilità che l'errore sia rilevabile solo nell'interazione fra variabili è necessario considerare le correlazioni (sulle distribuzioni di frequenza doppie) degli errori, almeno sulle coppie di variabili strutturali più importanti.

10. Metodi per la ricerca degli errori sistematici

Per individuare la presenza di errori sistematici può essere utile, a partire dal campione di verifica, ricorrere alla costruzione di matrici di transizione *prima/dopo*, dove vengono riportate le frequenze con cui i valori registrati sul campione risultano identici a quelli originari o invece risultano diversi.

L'eventuale correlazione fra valori iniziali e finali illustrati in questo tipo di tabella può consentire di individuare la sistematicità dell'errore, sia a livello di carattere digitato (numerico o alfabetico) complessivamente, sia a livello di variabile.

Se non ci fosse nessuna differenza fra le due registrazioni allora le frequenze della matrice occuperebbero la sola diagonale principale, mentre valori con nulli al di fuori della diagonale principale paleserebbero il verificarsi di errori.

Talvolta la semplice ispezione della tabella è sufficiente per identificare l'errore sistematico, ma in generale conviene analizzare la tabella, considerando diversi aspetti dei legami tra errori, mediante specifici test di indipendenza, di simmetria, di omogeneità, illustrati nell'Appendice 2.

APPENDICE

1. Un metodo per la ricerca degli errori sistematici sui record

Un metodo per individuare la presenza di errori sistematici è quello di far riferimento alla distribuzione di una variabile casuale teorica e di valutare la bontà di adattamento degli errori osservati al modello.

Test sulla distribuzione degli errori per record

Nel caso di errori casuali si dovrebbe avere una distribuzione del numero di errori di registrazione per record che segue la legge ipergeometrica, a sua volta approssimabile — se la percentuale di errore è molto bassa e per campioni sufficientemente grandi come nel nostro caso — da una distribuzione di Poisson.

Riportiamo a titolo esemplificativo (Tavola 4.5) i risultati del già citato studio (§ 3) sugli errori di registrazione. Considerando uno dei tre tipi record sui quali venivano registrati i dati del questionario, la distribuzione del numero di errori per record, confrontata con la distribuzione di Poisson di parametro *lambda* uguale alla percentuale stimata di errore, indicava con chiarezza un basso livello di accostamento, facendo escludere che gli errori osservati fossero semplicemente di tipo casuale.

Tavola 4.5 - Alcuni risultati del controllo della registrazione del censimento della popolazione 1981

n. byte errati nel rk	n. rk errati val. ass. (a)	n. rk errati val. perc.	Poisson teorica	freq. assol. teoriche (b)	diff (a) — (b)
0	15.981	86.80	74.77	13.766	2.215
1	600	3.26	21.74	4.002	—3.402
2	1.295	7.03	3.16	582	713
3	350	1.90	0.31	57	293
4	185	1.00	0.02	4	181
totale	18.411	100.00	100.00	18.411	0

Partendo dai dati sopra illustrati è inoltre possibile eseguire il test CHI QUADRATO sull'adattamento alla distribuzione;

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \qquad (4A.1)$$

dove, essendo in questo caso $j = 1$, i valori n_{ij} sono le frequenze effettive (colonna (a) della tabella) e i valori E_{ij} sono quelle

teoriche della Poisson (colonna (b) della tabella). I gradi di libertà del X^2 sono pari al numero di modalità (5 nella tabella) meno una. Dal calcolo esemplificativo risulta

$X^2 = 13818.21$, $X^2\ (0.05)/4\ GL = 9.48$

Il test evidenzia che gli errori per record non sono distribuiti secondo la variabile casuale di Poisson e che verosimilmente si è in presenza di errori sistematici.

2. Test sulla matrice di transizione

Test sugli errori per variabile

Per la ricerca degli errori sistematici si possono costruire matrici di transizione. Esse vanno impostate in modo che l'intestazione di colonna indichi i valori-tipo digitati nella prima fase e l'intestazione di riga indichi i valori-tipo della seconda ed in modo che in corrispondenza dell'incrocio fra generica i-esima riga e generica j-esima colonna si legga il numero di volte che il valore di tipo «i» della prima registrazione è stato trovato uguale a un valore di tipo «j».

La forma generale della tabella che andiamo ad esaminare è:

Figura 4.7: Matrice di transizione tipo

	1	2	...	j	...	c	tot
1	n_{11}	n_{12}		n_{1j}		n_{1c}	$n_{1.}$
2	n_{21}	n_{22}		n_{2j}		n_{2c}	$n_{2.}$
...							
i	n_{i1}	n_{i2}		n_{ij}		n_{ic}	$n_{i.}$
...							
r	n_{r1}	n_{r2}		n_{rj}		n_{rc}	$n_{r.}$
tot	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.c}$	$n_{..}$

NOTA: n = frequenza assolute
i = indice di riga
j = indice di colonna
. = sommatoria fatta rispetto a quell'indice

La matrice di transizione può essere considerata come una tavola di contingenza di tipo quadrato e, sotto opportune ipotesi, ad essa è applicabile il test Chi Quadrato (X^2) sull'indipendenza delle variabili di riga e di colonna.

Le ipotesi che è necessario assumere sono:

- che la frequenza osservate seguano una distribuzione multinomiale, ovvero che il campione a cui esse si riferiscono sia casuale semplice;
- che le frequenze attese non siano troppo piccole (in ciascuna casella la frequenza non deve essere inferiore a 5).

Nella figura 4.7 i simboli hanno il seguente significato:

$n_{i.} = \Sigma_j\, n_{ij}$

$n_{.j} = \Sigma_i\, n_{ij}$

$n_{..} = \Sigma_i\, \Sigma_j\, n_{ij}$

Test di indipendenza

Se con p_{ij} indichiamo, in corrispondenza di ogni n_{ij}, la probabilità degli elementi della popolazione di appartenere alla i-esima modalità di riga ed alla j-esima modalità di colonna, nell'ipotesi nulla (H0) di indipendenza delle variabili di riga e di colonna, questa probabilità congiunta potrà esprimersi come:

$$H0:\ p_{ij} = p_{i.} * p_{.j} \qquad (4A.2)$$

e la corrispondente frequenza attesa come:

$$H0:\ F_{ij} = n_{..} * p_{ij} = n_{..} * p_{i.} * p_{.j} \qquad (4A.3)$$

Non conoscendo F_{ij} possiamo stimarla con i dati campionari della nostra tabella, stimando $p_{i.}$ e $p_{.j}$ con:

$$\widehat{p_{i.}} = n_{i.} / n_{..} \qquad e \qquad \widehat{p_{.j}} = n_{.j} / n_{..} \qquad (4A.4)$$

Allora, sostituendo le (4) nella (3)

$$E_{ij} = F_{ij} = (n_{i.} * n_{.j}) / n_{..} \qquad (4A.5)$$

Se le variabili sono indipendenti gli n_{ij} effettivi (frequenze effettive) saranno ben approssimati dalle stime E_{ij} (frequenze teoriche) e la statistica:

$$X^2 = \Sigma_i\, \Sigma_j\, \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \qquad (4A.6)$$

sempre nell'ipotesi nulla di indipendenza, seguirà una distribuzione Chi Quadrato e potrà essere utilizzata per il test: se l'indipendenza non è vera infatti la statistica X^2 assumerà valori più alti.

Fissato un livello di significatività α (del 5 o dell'1%), cioè una «bassa» probabilità di rifiutare H0 quando essa è vera, si confronta il valore del X^2, calcolato con la (6) con il valore della distribuzione Chi Quadrato ($X^2\ \alpha$) con gradi di libertà pari al numero di modalità di riga (numero uguale a quello di colonna dato che la matrice è quadrata) meno uno al quadrato.

$$G.L. = (r - 1)^2 \qquad (4A.7)$$

(dove con $X^2\ \alpha$ si intende quel valore che lascia alla sua destra un area pari ad α)

se X > X -> rifiuto H0 con significatività
se X < X -> accetto H0 con significatività.

Test di Quasi-Indipendenza

Nella tabella di transizione costruita sugli errori di registrazione ci si aspetta un elevato numero di zeri o di valori molto piccoli al di fuori della diagonale principale: questo fa cadere l'ipotesi 2) di frequenza non inferiore a 5 e non consente di utilizzare correttamente il precedente test.

Il problema dei valori nulli a priori è comunque risolvibile ricorrendo al cosiddetto «test di Quesi-Indipendenza», che si applica alla tabella di transizione modificata, ottenuta escludendo la diagonale principale e analizzando la sola parte relativa ai «flussi».

L'ipotesi nulla H0 diviene allora:

$$p_{ij} = 0 \text{ per } i = j$$

$$H0: \quad p_{ij} = \frac{p_{i.} * p_{.j}}{(1 - \sum_{i=1}^{r} p_{i.} * p_{.i})} \quad \text{per } i \neq j \qquad (4A.8)$$

con il vincolo: $\sum_{i=1}^{r} \sum_{J=1}^{r} p_{ij} = 1.$

È necessario quindi calcolare le frequenze teoriche che corrispondono alla tabella di transizione modificata. Tale calcolo richiede l'applicazione di una procedura iterativa che stima le frequenze E_{ij} teoriche in caso di indipendenza partire da valori iniziali $E_{ij}(0)$, riponderandoli ad ogni passo, in modo che venga soddisfatto una volta il vincolo dei totali di riga ed una volta quello dei totali di colonna, fino alla convergenza dei successivi E_{ij} ad un determinato valore.

In dettaglio:

$$E_{ij}(0) = \begin{cases} 1 \text{ per } i \neq j \\ 0 \text{ per } i = j \end{cases} \qquad (4A.9)$$

$$E_{ij}(1) = \frac{E_{ij}(0) * n_{i.}}{E_{i.}(0)} \qquad (4A.9a)$$

$$E_{ij}(2) = \frac{E_{ij}(1) * n_{.j}}{E_{.j}(1)} \qquad (4A.9b)$$

e così via, usando la formula (a) per successivi passi dispari e (b) per quelli pari, fino a che $E_{ij}(k + 1) - E_{ij}(k)$ è minore o uguale ad un prefissato valore piccolo (es: = 0,01).

Ottenuti in questo modo i valori E_{ij} teorici si effettua l'usuale test Chi Quadrato (cfr. (4A.6)).

Test di simmetria

Per verificare la non unidirezionalità degli errori, cioè per vedere se lo scambio fra due caratteri digitati avviene nei due sensi (1 diventa 2, ma anche 2 diventa 1) è utili effettuare un test di simmetria sulla tabella di transizione.

L'ipotesi nulla HO è così definita:

$$H0: p_{ij} = p_{ij} \text{ per } i \neq j \qquad (4A.10)$$

cioè la probabilità (e la corrispondente frequenza attesa) della casella ij-esima è uguale a quella della casella ji-esima.

La stima di massima verosimiglianza degli E_{ij} diviene:

$$E_{ij}(0) = \begin{cases} (1/2) * (n_{ij} + n_{ji}) & \text{per } i \neq j \\ n_{ii} & \text{per } i = j \end{cases} \qquad (4A.11)$$

da cui $X^2 = \Sigma i<j\ (n_{ij} - n_{ji}) / (n_{ij} + n_{ji})$ (4A.12)

con 1/2 * r * (r—1) gradi di libertà.

Test di omogeneità

Un'ipotesi più debole di quella di simmetria è quella sull'omogeneità, che postula l'eguaglianza delle probabilità marginali di riga con le corrispondenti probabilità di colonna.

$$HO: \quad p_{i.} = p_{.j} \text{ per } i = 1,2,... \qquad (4A.13)$$

si pone:

$$\begin{aligned} d_i &= n_{.i} - n_{.i} \\ v_{ii} &= n_{i.} + n_{.i} - 2 * n_{ii} \\ v_{ij} &= -(n_{ij} + n_{ji}) \end{aligned} \qquad (4A.14)$$

creando un vettore d (che contiene (r—1) differenze) ed una matrice V delle relative varianze e covarianze:

$$X^2 = d' * V^{-1} * d \qquad (4A.15)$$

con (r—1) gradi di libertà dove d' indica il vettore d trasposto e V^{-1} la matrice inversa di V.

L'analisi andrebbe condotta tanto sulla tabella di transizione iniziale, tanto su di una tabella «normalizzata», nella quale cioè si tiene conto del peso di ciascun carattere alfanumerico all'interno del record. Ad esempio i numeri «1» e «2» nei veri campi del record possono esserre più frequenti di altri — perché spesso sono previsti come codici di caratteri dicotomici del tipo «si»/«no» — e conseguentemente sarà probabile il riscontro di numerosi errori per questi valori: sarà necessario quindi ponderare la tabella di transizione attribuendo alle due cifre pesi proporzionali alla frequenza con cui essi compaiono come modalità all'interno del questionario, così da depurare i test dall'effetto suddetto.

3. Esempio sui piani di campionamento semplice per attributi

Consideriamo un esempio in ci si ipotizza di voler verificare la qualità della registrazione sui byte e sui record. Abbiamo visto (§ 5) che la percentuale di record con almeno un errore è strettamente legata alla percentuale di errore sui byte, secondo l'andamento illustrato dalle Tabelle 4.3 e 4.4, in cui si utilizzava la distribuzione di Poisson, avendo ipotizzato la casualità dell'errore. La casualità implica che gli errori di registrazione siano «dispersi» fra i vari record e che si presenti raramente il caso di errori multipli sul record. L'errore casuale quindi determina una più alta percentuale di record con almeno un errore, ma definisce una relazione fra errore sul byte ed errore sul record che rende pressocché indifferente effettuare controlli sugli uni o sugli altri.

Inoltre, per quel che riguarda il primo dei due aspetti, si è osservata la scarsa convenienza dell'estrazione di n singoli byte, per cui di solito si sceglie di ridigitare completamente i record estratti, eventualmente per un numero K di record inferiore ad n, ove si presuma l'assenza di errore sistematico.

In definitiva si preferisce predisporre un piano di campionamento per il controllo sui record, utilizzando poi i medesimi dati campionari per la verifica sui byte. Il procedimento apposto (prima i byte poi i record) è logicamente equivalente, ma per la determinazione del piano di campionamento, i calcoli sui *milioni* di byte risultano più complessi e/o richiedono l'estrapolazione delle Tavole Military Standard, che non prevedono numerosità così elevate.

Nel nostro esempio si suppone di avere un blocco di N = 32000 record di lunghezza pari a 100. Poniamo che la percentuale di errore sui byte («p-byte») che consideriamo adeguata sia pari a 0.001 e conseguentemente (Tabella 4) quella sui record («p-record») sia uguale a 0.096, approssimato a 0.10. (La definizione degli standard di qualità può essere ovviamente fatta fissando prima «p-record» e poi «p-byte»).

Effettuiamo quindi il primo *controlo sui record* con le *Tavole Military Standard*: per N = 32000 la Figura 5 con un *Inspection Level* = II, definisce la lettera-codice = «M». Adottando l'ispezione normale, si entra nella Tavola II-A per «M» ed LQA = «p-record» = 0.10, e si trova che la numerosità campionaria n ed il numero di accettazione c sono uguali a quelli dati nel caso di lettera-codice K (seguendo l'indicazione della freccia), ovvero n = 125 e c = 21. Pertanto, estratti 125 record da ridigitare, se il controllo definisce 22 o più record errati il lotto viene rifiutato, altrimenti accettato. Se si effettua il controllo rinforzato la Tavole II-B fornisce n = 125, c = 18.

Alternativamente sempre per il *controllo sui record* possiamo utilizzare la procedura grafica del *nomografo* aggiungendo le condizioni su: livello di qualità tollerata LQT = 0.15 (= p2)

α = errore di prima specie = 0.05
β = errore di seconda specie = 0.10.

Entrando nel nomografo della Figura 1 e ricordando che LQA = 0.10 (= p1), si tracciano le due rette che congiungono $1-\alpha$ (= 0.95 nella scala di destra) con p1 (nella scala di sinistra) e (a

destra) con p2 (a sinistra) ottenendo un'area di possibili piani di campionamento fra cui n = 250 e c = 45. Quindi se più di 45 record su 350 risultano errati si rifiuta il lotto.

Per i *controlli sui byte* possiamo utilizzare i dati campionari della registrazione eseguita per controllare i record. Supponiamo di avere il campione di n = 350 record e cioè di 35000 byte: possiamo effettuare la verifica di ipotesi su «p-byte» = 0.001, adoperando le tavole della Normale, quale approssimazione della Binomiale. Fissato il livello dell'errore di prima specie α (cioè la probabilità di accettare il lotto se la qualità è cattiva) uguale a 0.05, il valore $z\alpha$ per il test unidirezionale (interessa cautelasi solo contro valori di «p-byte» elevati) risulta pari a 1.64.

Il valore che discrimina la decisione di accettare o meno il lotto è definito da:

$$p = p + z\alpha * p*(1-p)/n$$

dove p indica il valore desiderato «p-byte».

Nell'esempio $p = 0.001 + 1.64 * (0.001*0.999)/35000 =$
$= 0.001 + 0.0000468 = 0.0010468$

Quindi se dal conteggio dei byte errati risulta una percentuale di errori superiore a p = 1.047 per mille si deve rifiutare il lotto.

RIFERIMENTI BIBLIOGRAFICI

BRAMBILLA F. (a cura di), *Trattato di statistica*, vol. II, pagg. 873-998, *Tecnica di controllo statistico* di MOLLER F., Unione Tipografica, Editrice Torinese, Torino, (1969).

EVERITT B.S. (1977), *The Analysis of Contingency Tables*, Chapman and Hall, London.

DUNCAN A.J., *Quality Control and Industrial Statistics*, IV ed., Irwin, Homewood, III, 1974.

IACOBINI A. (1978), *I metodi statistici nel controllo di qualità*, La Goliardica ed., Roma.

MONTGOMERY D.C. (1977), *Introduction to Statistical Quality Control*, John Wiley and Sons, New York.

PALAZZI A. (1964), *Metodi statistici nella ricerca industriale e nel controllo della produzione*, ETAS Kompass, Milano.

PANIZON F. (1988), *Il controllo statistico di qualità nella fase della registrazione dei dati*, Atti della SIS, Siena, vol. 2, tomo 1, pp. 185-192.

UNITES STATES DEPARTMENT OF DEFENSE (1963), *Sampling Procedures and Tables for Inspection by Attributes MIL STD 105D*, U.S. Government Printing Office, Washington D.C..

ZUCHEGNA A. (1984), *La digitazione dei dati ed il controllo statistico*, Tesi di laurea, Università di Roma.

CAPITOLO 5 · LA REVISIONE

1. La fase di revisione

La fase di revisione ha lo scopo di eliminare gli errori e le incongruenze presenti nel materiale di rilevazione, relativamente al numero delle unità statistiche, alle loro relazioni ed al contenuto delle informazioni raccolte.
Le operazioni di controllo e correzione possono essere effettuate con due metodi diversi:

I) esperti di settore che operano direttamente sui questionari;
II) procedure informatiche automatiche che elaborano il file proveniente dalla fase di registrazione.

I programmi informatici, pur scontando, rispetto agli esperti, una minore flessibilità, soprattutto in presenza di dati anomali e di errori sistematici, garantiscono una maggiore tempestività, un maggiore controllo sull'applicazione delle regole di identificazione e di correzione degli errori e l'uniformità del trattamento dell'informazione.

Ai medesimi criteri di uniformità e di controllo, deve essere ispirata l'organizzazione di eventuali operazioni di revisione manuale. A tal fine, devono essere fornite agli esperti le regole di coerenza e di correzione in forma di tabelle di decisione, ed un modello di riepilogo degli errori riscontrati e delle modificazioni apportate, se la procedura non permette di risalire in altro modo a tale informazione (sostanzialmente mediante l'archiviazione dei file ai vari passi del processo). Le informazioni desunte dal riepilogo devono essere analizzate per controllare la presenza di errori, in particolare di errori sistematici nel lavoro degli esperti.

Nella pratica, i due metodi, operazioni manuali e programmi informatici, sono spesso utilizzati in combinazione, in funzione della dimensione dell'indagine e del tipo di unità di rilevazione; ad esempio la revisione dei questionari delle grandi imprese industriali pone problemi diversi da quella effettuata sulle famiglie.

Una situazione abbastanza comune di mistura di metodi è quella in cui l'errore viene determinato mediante elaborazioni, mentre la correzione viene effettuata da esperti; di questo caso si riscontrano diverse varianti:

a) ricerca mediante programmi, correzione degli esperti, rielaborazione del file;
b) ricerca batch e correzione da video terminale;
c) ricerca e correzione da video terminale.

Nel caso (a), i programmi individuano l'errore e riportano il relativo record su supporto cartaceo, dove viene corretto dall'esperto; le correzioni, registrate su di un file di appoggio, vengono quindi rielaborate insieme al file principale, sostituendone gli errori, per da luogo ad un archivio *pulito*.

Nel caso (b), si listano solo i codici identificativi delle unità in cui i programmi hanno individuato gli errori; tali codici saranno utilizzati dall'esperto per richiamare e correggere i record errati.

Nel terzo caso, infine, un programma identifica l'errore ed il relativo record è richiamato automaticamente sul video, dove viene modificato dall'esperto.

Appare evidente, da quanto detto, che il ruolo e l'apporto dell'informatica, nella fase di revisione, è rilevante e, nel caso di indagini di medie — grandi dimensioni, insostituibile; tuttavia c'è da osservare che la predisposizione delle norme e le informazioni da derivare dalle operazioni di revisione e correzione, costituiscono una procedura di natura essenzialmente statistica ed in quanto tale di competenza del responsabile dell'indagine.

2. La procedura di controllo e correzione

La procedura di controllo e correzione è costituita da un insieme di operazioni interrelate, che agiscono sui dati registrati, raccolti in uno o più file: esse possono essere, riguardo agli scopi, suddivise in:

— controllo quantitativo del numero e dei legami tra unità;
— controllo qualitativo delle variabili;
— piani di compatabilità e correzione;
— controllo delle relazioni tra unità appartenenti ad uno stesso modello di rilevazione.

Obiettivo del controllo quantitativo è ricostruire la coerenza tra il numero di unità teoriche (previsto nel piano di rilevazione o risultante dai documenti di rilevazione), il numero di unità rilevate (riportate nei questionari) e quello delle unità presenti su supporto informatico. Tale controllo, inoltre, assicura l'uguaglianza tra il numero di unità rilevate e quelle registrate ed il ripristino dei collegamenti tra unità, mediante operazioni di inserimento e cancellazione di record o modificazioni dei codici identificativi.

Il secondo controllo è finalizzato ad una prima ricognizione qualitativa del materiale raccolto e alla determinazione di eventuali errori sistematici.

I piani di compatibilità e correzione agiscono a livello di singola unità, per identificare e correggere i valori fuori campo, le mancate risposte parziali e le incongruenze logiche tra variabili.

Infine il quarto controllo mira a ristabilire i legami tra le unità di ordine inferiore al modello, eventualmente modificati o non presi in considerazione nelle precedenti operazioni.

In Figura 5.1 è riportato il diagramma relativo alla sequenza dei controlli; quest'ultima, tuttavia, non è univocamente determinata. In particolare l'ordine tra i passi 3 e 4 può essere invertito nel caso in cui la coerenza tra le informazioni relative alla singola unità, siano ritenuti meno importanti dei legami tra queste ultime; ad esempio che la ricostruzione della famiglia sia prioritaria rispetto alle compatibilità tra le variabili del singolo individuo. Il problema verrà approfondito nel paragrafo 5.8.

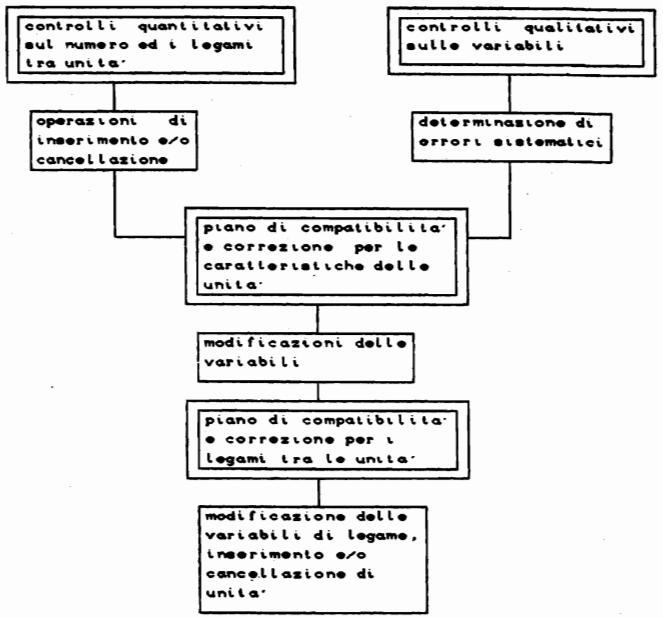


Figura 5.1 - La seguenza dei controlli nella fase di revisione

3. Le unità

La definizione di *unità*, nel contesto dei controlli quantitativi, si basa su considerazione di ordine pratico-organizzativo e sulla rappresentazione informatica del modello di rilevazione.

In particolare, poiché nelle indagini sulla popolazione, il nodo cruciale dell'organizzazione periferica è il comune (esso infatti costituisce la radice del sistema dei codici identificativi ed ad esso sono riferiti i documenti accessori di rilevazione), è conveniente considerare tale livello come unità di riferimento dei controlli quantitativi.

In generale, possiamo assumere che ciascun comune è suddiviso in *aree* (ad esempio le sezioni di censimento o le aree di circolazione dell'indagine forze di lavoro) e le interviste vengono condotte da uno o più rilevatori, ciascuno dei quali opera in una o più aree.

In ogni modello di rilevazione vengono raccolte informazioni riguardanti differenti unità di analisi, che possono essere identificate fisicamente (individuo, abitazione), istituzionalmente (famiglie, convivenze) o come eventi (nascite, morti, vacanze, spese, etc.).

Il modello viene rappresentato, su supporto informatico, mediante uno o più record (ad esempio un record famiglia, più record individui, più record eventi); ciascun record può contenere una o più unità di analisi presente nel modello di rilevazione.

Per *unità*, in questo contesto, si intende sia il modello di rilevazione, sia le unità di analisi, sia le istanze che sono coinvolte nell'organizzazione della raccolta e dell'elaborazione dei dati o sono rilevanti per essa.

Nel caso di indagini campionarie, in particolare, tra le unità verrà considerato anche lo strato, poiché rilevante ai fini della costruzione dei coefficienti di riporto all'universo.

In una generica indagine sulla popolazione, quindi, possiamo riconoscere quali unità nel senso sopra specificato:

- lo strato,
- il comune,
- l'area,
- il rilevatore,
- il modello di rilevazione,
- le unità di analisi.

4. I legami tra le unità

Nel modello di rilevazione, tra i differenti tipi di unità, vengono stabilite relazioni di *inclusione* o di *collegamento*; tali relazioni sono rappresentate, esplicitamente, mediante i codici identificativi, oppure, implicitamente, dal supporto fisico di rilevazione.

Ad esempio, la relazione tra famiglia ed individuo è implicita nel fatto che le interviste individuali compaiono sullo stesso questionario familiare, mentre la relazione tra modello e comune viene esplicitata dai codici identificativi riportati nel questionario.

Tra due differenti tipi di unità, viene stabilita una *relazione di inclusione*, se le prime possono essere considerate *grappoli* delle seconde, dei cui insieme costituiscono una partizione; ad esempio gli individui e le famiglie, le famiglie ed i comuni, le aree ed i comuni. Legami diversi dalla relazione di appartenenza, verranno definiti di *collegamento*; ad esempio il legame tra famiglia principale e coabitante, tra famiglia e abitazione, tra area e rilevatore.

Le due relazioni inducono un ordinamento tra le unità: possiamo definire quelle legate da una relazione di collegamento, come unità dello stesso ordine, mentre quelle legate da una relazione di inclusione, come di ordine superiore od inferiore, a seconda se includono o sono incluse.

Il modello di rilevazione contiene tutte le informazioni atte al riconoscimento delle diverse unità sia di ordine superiore (rilevatore, area, comune) sia di ordine inferiore (famiglia, abitazione, individuo, evento).

Il modello viene riportato su supporto informatico mediante un insieme di record collegati tra loro da un sistema di codici che permette di legare due o più unità diverse.

In particolare tale sistema deve assicurare:

- il riconoscimento dell'insieme di record corrispondente al modello;
- il collegamento tra le diverse unità di analisi appartenenti al modello;
- il collegamento tra l'insieme di record corrispondenti al modello e le unità di ordine superiore.

La struttura organizzativa delle indagini e la rappresentazione informatica del questionario determinano il sistema dei codici identificativi e le relazioni tra unità.

Il sistema dei codici identificativi gioca un ruolo centrale nell'analisi del materiale raccolto; infatti il controllo quantitativo dei file, poiché attuato sui record, non è altro che un insieme di operazioni su detti codici.

5. I controlli quantitativi

La prima operazione di revisione riguarda la dimensione quantitativa dell'indagine, ovvero il numero di unità e le loro relazioni, così come sono state definite nel paragrafo precedente. È necessario, infatti, che vi sia coerenza tra la programmazione dell'indagine, la sua effettuazione ed il risultato ottenuto, ovvero tra

— il piano teorico di rilevazione;
— il piano effettivo di rilevazione;
— l'insieme dei questionari;
— il file proveniente dalla registrazione.

L'obiettivo del controllo quantitativo consiste nel verificare, ed eventualmente ristabilire, l'uguaglianza tra il piano effettivo di rilevazione, l'insieme dei questionari ed il file, e nel ricostruire il bilancio tra piano teorico ed effettivo (dato dalla somma delle unità rilevate e le mancate risposte totali).

Tuttavia, per indagini di medie-grandi dimensioni, il ritorno al materiale cartaceo ed il confronto tra questo ed il file, sono operazioni estremamente dispendiose, in termini economici, organizzativi e di tempo; cosicché, è conveniente non considerare nei controlli l'insieme dei questionari, tranne ritornare a questi ultimi, nei casi in cui non è possibile risolvere altrimenti le incongruenze.

Le informazioni contenute nel piano teorico o nei documenti di rilevazione, possono essere disponibili sotto forma di liste (ad esempio l'elenco dei comuni per singolo strato, l'elenco dei rilevatori e delle aree per comune, gli elenchi delle assegnazioni etc.) o di riepiloghi di conteggi di unità (ad esempio il numero di famiglie intervistate e sostituite, il numero di rilevatori utilizzati etc.).

Nelle liste ciascuna unità è identificata mediante il medesimo codice che appare nel file; cosicché, se gli elenchi sono disponibili su supporto informatico, si può effettuare facilmente il controllo quantitativo mediante programmi di *linkage* sulle unità di tipo diverso, comuni al piano di rilevazione teorico, a quello effettivo ed al file.

Nella pratica, il piano teorico di rilevazione, per le indagini campionarie, è costituito dalla lista degli strati e dei comuni e dal numero di unità campione, essendo l'indicazione delle aree e dei rilevatori non essenziale, e spesso non determinabile, nel disegno campionario; nel piano dei censimenti è presente anche la lista delle aree, ma non quella dei rilevatori. I documenti di rilevazione, che variano da un'indagine all'altra, contengono, in genere, dei conteggi riepilogativi di unità e, laddove sono costituiti da liste (ad esempio le assegnazioni dei rilevatori), queste raramente sono registrate su supporto informatico.

Tuttavia la convenienza a strutturare in modo più analitico i documenti di rilevazione, così da avere a disposizione le liste di tutte le unità coinvolte nella rilevazione con i relativi riepiloghi, e a prevederne la trasposizione su supporto informatico, deve essere attentamente valutata in funzione dei seguenti aspetti: (i) il costo ed i tempi per la registrazione, (ii) la gestione di una consistente massa di informazioni, (iii) l'aggravio del lavoro di campo, (iv) l'errore di registrazione nelle liste e nei conteggi, che introdurrebbe elementi di incertezza nel controllo.

Nell'analisi che seguirà, si ipotizza che l'attuale organizzazione delle indagini renda disponibili, con modificazioni marginali delle attuali procedure, la lista di stratificazione ed i conteggi riassuntivi e desunti dai documenti di rilevazione (DR) e dal piano teorico (PT), secondo il Prospetto (5.1).

Prospetto 5.1 - Conoscenza a priori del numero di unità

	INDAGINE			
	CAMPIONARIA		TOTALE	
	P.T.	D.R.	P.T.	D.R.
Strato	si	si	no	no
Comune	si	si	si	si
Area	si	si	si	si
Rilevatore	si/no	no	no	no
Questionario	si	si	si	si
Famiglia	si	si	no	si
Individuo	no	no	no	no
Evento	no	no	no	no

La distinzione tra unità per cui si dispone di informazioni esterne al file e quelle per cui tali informazioni non sono disponibili, comporta un differente metodo di controllo quantitativo.

Nel primo caso, il controllo del file sarà sostanzialmente basato sul riscontro tra il file ed i documenti di rilevazione, nel secondo ci si avvarrà di un controllo induttivo, mediante l'analisi dei codici identificativi e di alcuni parametri statistici.

Gli strati, i comuni, le aree, i rilevatori e i modelli

Le informazioni sul numero di questionari e delle unità di ordine superiore e sulle reciproche relazioni, sono, generalmente disponibili, o facilmente ottenibili, dalle usuali procedure dell'indagine.

Mediante le notizie riportate sul piano teorico, sui documenti di rilevazione ed i conteggi effettuati sul file, verranno predisposte alcune *tavole di controllo*, per verificare le seguenti relazioni tra le suddette unità:

1) comuni-strati,
2) modelli-comuni,
3) rilevatori-comuni,
4) aree-comuni,
5) modelli-aree,
6) modelli-rilevatori,
7) aree-rilevatori.

La prima relazione viene controllata confrontando le liste di stratificazione provenienti dal piano teorico, dai documenti di rilevazione e dal file; per le ultime due deve essere verificata l'uguaglianza, mentre le prime due coincidono a meno delle mancate risposte totali.

Qualora le verifiche diano luogo ad incongruenze, devono essere effettuate le opportune correzioni, identificando gli errori (nella compilazione dei documenti di rilevazione o nei codici identificativi dei record) per mezzo delle tavole di controllo predisposte per la verifica delle relazioni (2) - (7).

Per quanto riguarda le relazioni (2), (3) e (4), il metodo consiste nell'accoppiare i codici identificativi comunali, estratti dal file, con quelli desumibili dal piano teorico e dai documenti di rilevazione, in una tavola di controllo contenente le seguenti informazioni per ogni codice comunale:

- dal piano di rilevazione teorico,
  numero di aree
  numero di rilevatori
  numero di modelli
- dai documenti di rilevazione,
  numero di aree
  numero di rilevatori
  numero di modelli
  numero di mancate risposte totali
  numero di sostituzioni
- dal file,
  numero di aree
  numero di rilevatori
  numero di modelli
  numero di sostituzioni

Lo schema suddetto è del tutto generale e deve essere adattato alle situazioni concrete, poiché, per determinate indagini,

alcune delle informazioni sopra riportate (ad esempio il numero di modelli del piano teorico di rilevazione per il censimento) non sono disponibili; in tal caso potrà essere utilizzata una stima od un valore ad un tempo precedente, che costituisca un punto di riferimento per validare i dati raccolti.

Una seconda tavola di controllo può aiutare nella verifica della relazione (2); per ogni codice comunale, nel file, si calcolano le interruzioni di sequenza nel numero d'ordine dei modelli, se essi sono, come è la regola, numerati progressivamente.

Una terza tavola di controllo, con la medesima struttura della prima, riportante, per ciascun comune e per ogni codice di area e di rilevatore, il numero dei relativi modelli, sarà utilizzata per il controllo delle relazioni (5) e (6); l'incrocio tra codice di area e rilevatore e relativo numero di questionari permetterà invece la verifica della relazione (7).

Mediante l'analisi delle tavole di controllo, è possibile non solo determinare l'esistenza di un errore, ma anche rintracciare eventuali blocchi di modelli con codici errati o duplicati nel file; in funzione del tipo di incongruenza verranno effettuate le seguenti operazioni:

— correzione del codice di comune, area, rilevatore;
— cancellazione dei record relativi ad uno o più modelli;
— inserimento dei record relativi ad uno o più modelli.

Le operazioni di cancellazioni vengono eseguite nel caso di duplicazione dei modelli; tale operazione e quella di correzione del codice saranno effettuate previo confronto a vista dei blocchi di modelli e di record *errati*.

A questo livello di controllo non è prevista altra operazione di inserimento, se non nel caso di un ritardo nell'acquisizione dei dati.

Con questa prima fase di controllo, si compie una riallocazione dei record mediante le operazioni di inserimento, cancellazione e correzione dei codici identificativi; tuttavia, alla fine del processo, possono rimanere alcune differenze tra unità rilevate e presenti nel file (ad esempio per un errore vengono aggregati con il medesimo codice due o più modelli), difficilmente rintracciabili per mezzo delle tavole di verifica di cui sopra.

La riallocazione definitiva dei modelli sarà, quindi, effettuata sulla base del controllo delle unità di analisi.

Le unità di analisi

Generalmente, per le unità di analisi presenti sul modello di rilevazione (famiglie, abitazioni, individui, eventi), non si dispone di numerosità desumibili né dal piano teorico di rilevazione

né da documenti aggiuntivi; spesso un conteggio di tali unità (ad esempio il numero di individui o di eventi rilevati) od un indicatore di presenza di altra unità (ad esempio l'abitazione), sono però presenti sul modello di rilevazione e riportati sui record.

Nel caso delle unità di analisi, quindi, il controllo quantitativo non ha riscontri esterni al file, ma si riduce a due verifiche complementari:

I) la verifica interna ad un gruppo di record individuati dallo stesso codice identificativo;
II) la verifica di plausibilità fondata sull'analisi delle distribuzioni di tali unità o di parametri statistici (medie, percentuali ecc.), calcolati sull'intero file o su domini territoriali.

In particolare, nel primo caso possiamo articolare il controllo:

- sul riscontro tra il valore di un campo di un record con il conteggio del numero di unità presenti sotto un medesimo codice identificativo;
- sull'accertamento della presenza di unità in base ad un indicatore contenuto in un record;
- sui legami tra codici identificativi di ordine inferiore.

Per quanto riguarda i controlli sub (I), si avranno quindi tre tavole di verifica.

La prima metterà in evidenza i modelli per i quali non sussiste l'eguaglianza tra gli eventuali conteggi riassuntivi contenuti nei record, il numero di unità di analisi presenti come codici identificativi diversi (ad es. il numero di individui riportato sui record famiglia ed il numero di record individuali), il contenuto del massimo codice progressivo e gli eventuali salti nei progressivi delle unità (se è prevista, per questa ultime, una numerazione progressiva).

La seconda evidenzierà i casi in cui ad un indicatore di presenza non corrisponde una unità di analisi, mentre la terza conterrà i modelli per i quali non sono stati verificati i legami tra codici all'intero del modello.

I controlli di cui al punto (II), vengono condotti calcolando alcuni parametri indicativi della *dimensione* delle unità da controllare in un determinato ambito territoriale (i comuni di una regione o le sezioni di censimento di un comune).

Ad esempio nel caso delle indagini sulle famiglie, il numero medio di componenti, l'indice di vecchiaia e di dipendenza, il rapporto di mascolinità, la percentuale di famiglie superiori ad una data dimensione, ecc.; nel caso di eventi quantitativi la media ed il coefficiente di variazione, il numero medio di eventi per modello, la distribuzione del numero di eventi, ecc.

L'analisi della distribuzione di un determinato parametro (o, simultaneamente, di più parametri) ha lo scopo di individuare eventuali valori *anomali* che potrebbero essere conseguenza di errori nei codici identificativi (ad esempio il numero di componenti eccezionalmente elevato di una famiglia, potrebbe essere dovuto ad un errore nei codici identificativi che ha comportato il raggruppamento di più famiglie).

Tali dati anomali possono essere individuati come valori esterni agli intervalli costruiti intorno ai valori medi (ad esempio, per la media, due volte lo scarto quadratico medio; per la mediana, lo scarto interquartile), oppure sulla base di più sofisticate tecniche di analisi multivariata (ad esempio l'analisi dei gruppi).

Il calcolo di indicatori

I controlli suddetti non hanno solo lo scopo di eliminare gli errori, ma anche di produrre informazioni per il controllo della rete e delle operazioni effettuate sul supporto cartaceo ed informatico.

Dai controlli quantitativi sarà quindi possibile calcolare indicatori relativi ai diversi tipi di errori, con riferimento al complesso del file e ai livelli di controllo appropriati, ovvero la fonte cui l'errore è imputabile, come riportato nel Capitolo 3.

6. I controlli qualitativi

Per controllo qualitativo si intende la verifica dei valori assunti dalle variabili nei dati, non ancora sottoposti alle procedure di controllo logico e di correzione; in questa fase si attua una prima analisi dell'efficienza complessiva dello strumento *rilevazione*.

In particolare, costituiscono obiettivi del controllo (I) la verifica delle informazioni raccolte e (II) l'individuazione di eventuali errori sistematici.

La verifica delle informazioni raccolte

L'analisi dell'informazione raccolta, oltre a costituire la *cartina di tornasole* della qualità dei dati, può fornire indicazioni per mettere a punto le elaborazioni successive: ad esempio, la revisione delle procedure di correzione, nel caso di una rilevante quota di questionari errati per particolari variabili, la revisione del piano di tabulazione, in funzione dell'attendibilità dei risultati, ecc.

Rispetto alle analisi da effettuare, possiamo dividere i risultati della rilevazione in due gruppi:

I) le variabili qualitative e quantitative intervallo;
II) le variabili quantitative, continue o discrete.

Nel primo caso verranno elaborate le distribuzioni di frequenza dei diversi codici presenti, per singola variabile e le più importanti distribuzioni congiunte; nel secondo, verranno calcolati i principali parametri statistici (ad esempio la media, il coefficiente di variazione, i valori minimo e massimo, la mediana ed i quartili, la distribuzione per classi), evidenziando in particolare la numerosità dei valori *blank*, *zero* ed alfanumerici riscontrati.

Per esprimere un giudizio sulla plausibilità dei dati raccolti, tali informazioni possono essere confrontate con fonti esterne (ovvero con i risultati di altre indagini o della medesima indagine in tempi precedenti) e tra domini di studio della stessa rilevazione, in particolare i domini territoriali rilevanti per l'organizzazione sul campo e per la diffusione dei dati.

Poiché la matrice dei parametri statistici per domini territoriali può essere di grandi dimensioni, rendendo difficoltosa l'analisi, ci si può limitare all'esame delle principali caratteristiche rilevate.

L'errore sistematico

Gli errori sistematici possono aver origine nella formulazione del questionario, nelle operazioni di rilevazione o nella fase di registrazione dei dati; essi possono manifestarsi come valori mancanti, incongruenze, valori fuori campo o come risposte coerenti ma accentrate sulle modalità di determinati quesiti.

La loro individuazione è propedeutica sia alla predisposizione di opportuni interventi correttivi sul materiale grezzo, sia alla distinzione, prima dell'applicazione della procedura di compatibilità e correzione, tra errori casuali e sistematici, prevista espressamente da alcune tecniche di imputazione di tipo stocastico.

La determinazione degli errori sistematici non è agevole; essi, infatti, possono essere definiti per negazione, come errori non casuali, ma difettano di una definizione operativa che ne permetta l'individuazione.

Una possibile specificazione del concetto di errore sistematico è definirlo come assenza di variabilità nei dati rilevati, in funzione del valore assunto da altre variabili e/o di particolari subpopolazioni, in cui la variabilità attesa è maggiore di zero; ovvero, in termini equivalenti, che la probabilità di una modalità, di essere rilevata in una data subpopolazione condizionata dal valore assunto da altre variabili, è pari ad uno. In ambedue i casi si potrebbe sottoporre a test, sulla base dei dati rilevati, l'ipotesi nulla; ma ciò comporterebbe il calcolo e l'analisi dei test in un numero elevato di domini, definiti dalle variabili correlate e dai livelli organizzativi coinvolti (i singoli comuni, supervisori, rilevatori, famiglie), con una affidabilità statistica, soprattutto nel caso di indagini campionarie, compromessa dalla ridotta numerosità delle informazioni disponibili per il singolo dominio.

In pratica, non facendo riferimento ad alcuna definizione, si ricorre ad alcuni indicatori indiretti; tuttavia, la loro significatività risente delle stesse difficoltà sopra accennate, riguardanti la dimensione della base di calcolo.

Il verificarsi di dati *anomali* nella distribuzione delle variabili, può essere considerato un indicatore indiretto dell'esistenza di un errore sistematico. In questo caso l'errore può essere individuato mediante l'analisi delle statistiche calcolate per il controllo qualitativo; dovendo, per le ragioni sopra esposte, stabilire un limite per le subpopolazioni da considerare, è conveniente scegliere il livello dei domini rilevanti per la diffusione dei risultati.

Quale indicatore per l'individuazione dell'errore sistematico, si assume, generalmente, il complemento a uno del tasso di *qualità del materiale raccolto*, riscontrato nelle singole variabili; tuttavia, se la determinazione dell'errore è finalizzata alla strategia di imputazione, è opportuno utilizzare il complemento del tasso di *qualità del materiale disponibile*, che include anche l'eventuale errore sistematico di registrazione.

Si ricorda (per maggiori ragguagli cfr. Capitolo 3, § 4) che, per ciascuna variabile, il primo tasso è calcolato come rapporto tra il numero di risposte dovute nette e la differenza tra il numero di unità rispondenti e quello dei valori fuori campo; il secondo, come rapporto tra il numero di risposte dovute nette e quello delle unità rispondenti.

Tali indicatori possono essere analizzati a livello aggregato ovvero per domini territoriali; nel primo caso, si determina la sistematicità dell'errore dovuta alle operazioni centralizzate che coinvolgono l'intera rilevazione (questionario, registrazione, norme, istruzioni, ecc.), mentre, nel secondo, viene individuata quella imputabile ai singoli organi della rete di rilevazione.

L'esistenza di un errore sistematico nella singola variabile, è generalmente derivata dal confronto dei livelli degli indicatori relativi alle altre variabili.

Indicando con q_j il complemento a uno del tasso di qualità del materiale disponibile, relativo alla j-esima delle k variabili presenti sul questionario, è possibile individuare l'errore sistematico utilizzando differenti metodi.

Un primo criterio di base sul confronto tra gli indicatori q_j ed una determinata soglia q^*, solitamente fissata dal 3% al 5%: la condizione

$$q_j > q^* \qquad (5.1)$$

indica il verificarsi dell'errore sistematico (NCBS Statistics Sweeden 1983).

Un secondo criterio si basa sulla considerazione che in presenza di un errore casuale (quindi non sistematico), si può ipotizzare che le probabilità di errore sulle k variabili siano uguali tra loro e pari a p; in questo caso, i tassi calcolati q_j rappresentano misure ripetute di p e si distribuiscono secondo una curva Gaussiana.

Allora, assimilando l'errore sistematico al dato anomalo, lo si può identificare come valore esterno all'estremo superiore dell'intervallo di confidenza

$$q_j > q + t_\alpha \hat{\sigma}_q \qquad (5.2)$$

dove:

$$E(q) = p$$

$$q = \Sigma_j q_j / k$$

$$\hat{\sigma}_q = \Sigma_j (q_j - q)^2 / (k-1)$$

e t_α è il livello corrispondente alla desiderata probabilità α.

Un terzo criterio per riconoscere il verificarsi di un errore di tipo sistematico, deriva dal definire la *sistematicità* in termini di *dipendenza* tra gli errori delle variabili del questionario. Nel singolo record è possibile riscontrare un errore in ciascuna delle k variabili, ovvero secondo una delle differenti disposizioni con ripetizione di due elementi (errato/non errato) presi a k a k.

Il numero teorico di tali disposizioni è pari a 2^k-1, ma, nel file, alcune di esse non si verificheranno mentre altre si presenteranno non frequenza n_g (ovvero il numero di record in cui la data combinazione g di errori si è manifestata).

Se gli errori delle variabili fossero indipendenti, la probabilità della g-esima combinazione di errore P_g sarebbe pari a:

$$P_g = \left[\prod_{j \epsilon S} p_j\right] * \left[\prod_{j \epsilon T} (1-p_j)\right]$$

dove p_j rappresenta la probabilità di errore della j-esima variabile, $j \epsilon S$ se errata mentre $j \epsilon T$ nel caso contrario.

La frequenza della g-esima combinazione di errore è allora data da:

$$N_g = P_g N$$

dove N rappresenta il numero di record.

Stimando P_g e N_g mediante le:

$$\hat{P}_g = \left[\prod_{j \epsilon S} q_j\right] * \left[\prod_{j \epsilon T} (1-q_j)\right]$$

$$\hat{N}_g = \hat{P}_g N$$

dove i q_j rappresentano, come sopra, i complementi a uno dei tassi di qualità del materiale disponibile, sarà possibile calcolare l'indice

$$X^2 = \sum^k (n_g - \hat{N}_g)^2 / \hat{N}_g \qquad (5.3)$$

e sottoporre a verifica l'ipotesi di conformità tra le distribuzione teorica, calcolata sotto l'ipotesi di indipendenza degli errori, e la distribuzione osservata.

Il test indica se, tra tutte le combinazioni di errore riscontrate nel file, si è verificato un errore sistematico; per individuare la particolare combinazione, occorre analizzare i relativi contributi all'indice.

Nell'ipotesi che l'errore sistematico venga generato non solo in funzione degli errori di altre variabili ma anche dei diversi livelli organizzativi dell'indagine (comune, rilevatore), sarebbe necessario calcolare l'indice di cui sopra con riferimento a ciascuno di tali livelli.

Nel Capitolo 3, si è data l'indicazione di calcolare i tassi q_j mediante i risultati dei piani di compatibilità e correzione; ovviamente, se la determinazione della sistematicità dell'errore è finalizzata alla scelta del tipo di imputazione, non è possibile utilizzare tali informazioni.

Ci si può allora basare su un tasso grezzo, avente al numeratore la somma dei valori fuori campo ed i rifiuti ed al denominatore il numero di unità rispondenti. Ma, mentre l'individuazione

dei valori fuori campo non presenta problemi, la determinazione dei *rifiuti*, in questa fase, è legata all'esistenza di un apposito codice sul questionario.

In mancanza di tale codice, si può calcolare il tasso includendo nel numeratore tutti i blank, ovvero sia le mancate risposte parziali che i blank significativi; tuttavia, tale operazione equivale a supporre che il livello del q_j sia indipendente dalla struttura del questionario e che l'errore sistematico non sia correlato a particolari subpopolazioni ma solo al valore di altre variabili.

In alternativa si possono elaborare dati grezzi mediante un programma di compatibilità, basato solo sulle principali regole *formali*, che discrimini i rifiuti dalle risposte non dovute.

Infine, è possibile iterare la procedura già predisposta di compatibilità e correzione, utilizzando i risultati della prima elaborazione per il calcolo degli indicatori.

7. I programmi di compatibilità e correzione

I programmi di compatibilità e correzione hanno la duplice funzione di determinare le incongruenze e di correggerle; tali funzioni sono logicamente distinte, anche se la maggior parte dei programmi le effettua simultaneamente.

L'identificazione dell'errore viene effettuata mediante un insieme di regole che, però, è in grado di determinare solo una parte

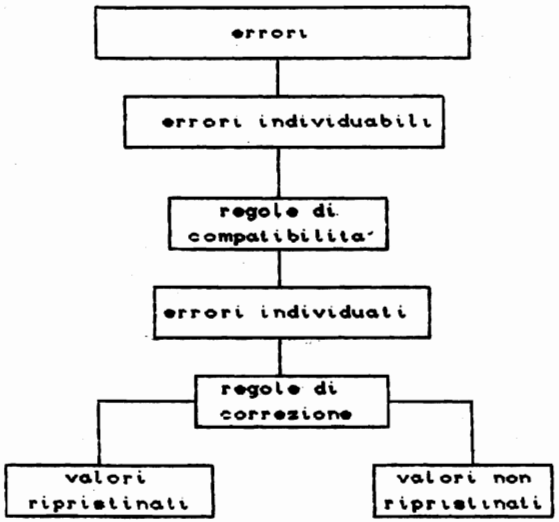


Figura 5.2 - Il processo di compatibilità e correzione

dell'errore totale: i valori fuori campo, le mancate risposte parziali e le incongruenze logiche tra variabili.

Sugli errori *determinabili* agiscono le regole di compatibilità; una loro insufficiente specificazione riduce tale insieme a quello degli errori *determinati*; su questi vengono applicate le regole di correzione che possono o meno ripristinare il valore *vero* originario. Il processo è schematizzato nella figura 5.2.

Nella costruzione di un programma di compatibilità e correzione, in cui confluiscono ed interagiscono aspetti e problematiche sia statistiche che informatiche, devono essere bilanciate esigenze diverse, a volte contraddittorie.

Dal punto di vista informatico le caratteristiche di un piano possono essere individuate:

I) nella possibilità di implementazione;
II) nella flessibilità, ovvero nella possibilità di adattamento a modificazioni dell'input e delle regole;
III) nella velocità di esecuzione.

Le proprietà statistiche cui far riferimento, possono essere invece sintetizzate:

I) nella *verosimiglianza* delle correzioni, ossia nel rendere coerenti le informazioni, imputando valori che trovano riscontro nella realtà indagata;
II) nel principio del minimo cambiamento, ovvero nel ridurre al minimo le modificazioni dell'informazione raccolta;
III) nell'efficienza e nella correttezza degli stimatori applicati ai dati *puliti*.

Le tecniche di correzione possono seguire, sostanzialmente, od una logica deterministica od una logica stocastica.

Il metodo deterministico consiste nell'imputare un solo valore predeterminato, oppure un valore casualmente scelto da una distribuzione predeterminata, in funzione o meno dei valori assunti da altra od altre variabili. Secondo il criterio stocastico, invece, i valori delle variabili da modificare vengono copiati da unità *simili* del file.

Il diverso modo di operare dei due metodi si riflette sulle fonti da cui vengono attinte le informazioni necessarie alla correzione; dati puramente interni all'indagine nel criterio stocastico, dati generalmente esterni in quello deterministico.

Nella correzione stocastica viene *salvaguardata* l'informazione raccolta, conservando le distribuzioni e le associazioni originarie riscontrate nell'insieme dei dati *completi*, mentre in quella deterministica si interviene sui dati dell'indagine mediante forzature decise *a priori*.

La differente logica dei programmi influenza anche le modalità della loro costruzione.

I piani deterministici implicano che vengano determinate contemporaneamente sia le regole di compatibilità che quelle di correzione, entrando quindi nel merito dei legami e delle modificazioni delle variabili della singola indagine. Al contrario, la costruzione di piani stocastici richiede la definizione di criteri formali di rappresentazione delle regole e dei metodi di correzione, senza entrare nello specifico dei dati rilevati.

Dal punto di vista della progettazione e della realizzazione informatica, i programmi deterministici sono molto meno complessi di quelli stocastici; essi, generalmente, sono costruiti ad hoc per la singola indagine, mentre i secondi, per la cui produzione sono necessarie risorse notevolmente maggiori che per i primi, sono programmi generalizzati, validi per più indagini differenti.

Le minori difficoltà di programmazione dei piani deterministici sono però bilanciate da inferiori *prestazioni* statistiche: la *verosimiglianza* delle forzature dipende dalle informazioni a priori o dalla soggettiva valutazione del responsabile dell'indagine e quindi possono essere introdotte distorsioni nelle stime; il principio del minimo cambiamento dei dati rilevati non può essere rispettato così come non vengono mantenute le distribuzioni e le associazioni presenti nell'insieme completo dei dati.

La correzione deterministica, tuttavia, è più adatta a trattare errori di tipo sistematico, al contrario dell'altra, che risulta più efficiente per quanto riguarda gli errori provenienti da un modello di generazione casuale. Essa, inoltre, supplisce ad alcune limitazioni di ordine informatico, che impediscono l'applicazione del criterio stocastico al trattamento delle variabili qualitative con un grande numero di modalità o delle variabili quantitative non riconducibili, ai fini delle elaborazioni, a variabili intervallo.

Per tali ragioni i due metodi sono spesso utilizzati in combinazione tra loro.

Per quanto riguarda il livello cui applicare i programmi, sarebbe teoricamente corretto costruire un unico piano di compatibilità che tratti contemporaneamente le variabili delle singole unità, di ordine inferiore al modello di rilevazione, e le reciproche relazioni (ad esempio famiglia/individui, famiglia/abitazione etc.), considerando quindi il questionario, come è in realtà, un unico insieme di informazioni omogenee.

Nella pratica, le difficoltà di ordine logico ed informatico, nel trattare contemporaneamente un grande numero di regole, consiglia l'adozione di una strategia in due tempi, distinguendo le correzioni dei dati relativi alle singole tipologie di unità (piani di compatibilità in senso stretto), da quelle apportate allo scopo di ristabilire i legami tra le unità del singolo questionario.

Le regole di compatibilità

Le *regole di compatibilità*, generalmente, sono asserzioni sulla *non ammissibilità* di codici per la singola variabile o di combinazioni di codici relativi a più variabili (che esprimono le reciproche relazioni logiche); esse, quindi, dovrebbero essere, più propriamente, chiamate *regole di incompatibilità*.

Assumere una logica di non ammissibilità ha il vantaggio di permettere un più stretto controllo nel processo di definizione delle regole che danno luogo ad errori (e, conseguentemente, a correzioni) poiché costringe a considerare analiticamente tutti i relativi casi invece di ottenerli come residuo.

Scopo delle regole è l'individuazione dell'errore che, in questo contesto, coincide con i valori fuori campo, le mancate risposte ai quesiti e alle incongruenze logiche tra variabili.

Gli errori possono riguardare una singola variabile o la relazione tra due o più variabili; essi sono logicamente della stessa natura e quindi devono essere trattati contemporaneamente e con i medesimi criteri. In particolare, nel caso esista una relazione logica tra variabili, i relativi controlli di campo possono essere *assorbiti* dalla regola di compatibilità che coinvolge tali variabili.

Sia per i controlli della singola variabile che per quelli di relazione, possiamo dividere le regole in *formali* o *sostanziali*; tale distinzione si riflette nella difficoltà ad esplicitare le regole e nel contenuto di soggettività delle medesime.

Sono regole del primo tipo quelle derivanti dalle norme di compilazione del questionario (ad esempio: se ha risposto SI al quesito 1, passare al quesito 3, altrimenti passare al quesito 2) e del piano di registrazione su supporto informatico; appartengono al secondo quelle derivanti da informazioni *a priori* sulla realtà indagata (ad esempio: se sesso è femmina non è possibile che la condizione sia militare di leva).

In particolare, per il controllo delle singole variabili, le relative regole formali derivano dal piano di registrazione (i codici non ammissibili); regole sostanziali possono, invece, essere considerate quelle riguardanti il campo di variazione *plausibile* per variabili quantitative (ad esempio l'intervallo di accettabilità del prezzo di acquisto di un determinato bene).

Le relazioni *formali* tra più variabili derivano dalle norme di compilazione e dalla struttura del questionario; per la loro individuazione, possono essere utilizzati due metodi, la diagrammazione del questionario e gli schemi della Progettazione Concettuale.

Nell'Appendice 1 del Capitolo 2 è stato riportato un esempio di diagrammazione, relativo al questionario individuale della rilevazione forze di lavoro, mediante il quale possono essere stabilite le relazioni formali tra i quesiti del questionario.

Lo stesso diagramma può essere utilizzato per evidenziare la funzione di «nodi» assunta da particolari quesiti (in questo caso il q. 10.1 ed il q. 14.1) e le incongruenze riscontrate nella redazione del questionario (cfr. il Capitolo 2); tale analisi può guidare nella definizione delle regole di compatibilità e di correzione che coinvolgono le rispettive variabili.

Il medesimo ruolo, ma in maniera più analitica ed esauriente, può essere svolto dagli elaborati del modello Entità-Relazioni; in questo caso, il lavoro verrebbe facilitato derivando le regole già al momento della predisposizione e dell'analisi del questionario nella fase di progettazione. Gli schemi E/R, inoltre, possono essere utilizzati come guida alla predisposizione delle regole sostanziali, analizzando i *percorsi* che definiscono le relazioni tra unità e variabili.

Sintetizzando le considerazioni sopra riportate, si può affermare che le regole di compatibilità sono funzione della struttura e delle disposizioni formali per la compilazione del questionario, del piano di registrazione su supporto informatico e delle relazioni tra variabili esistenti nella realtà in studio.

Per l'esplicitazione delle regole ci si potrà basare, per quanto riguarda i controlli dei singoli campi del record, sul piano di registrazione e su informazioni a priori sui limiti degli intervalli ammissibili per le variabili quantitative; sulla diagrammazione del questionario o sulla documentazione del modello Entità Relazioni, per quanto riguarda le incongruenze logiche tra variabili.

Una volta derivate dalle fonti suddette, le regole devono essere costituite in un *insieme* coerente, tale, cioè, da garantire:

— la non ridondanza, ovvero di non ripetere regole già poste in altra forma o derivabili da altre;
— la non contraddittorietà tra regole.

Le regole ridondanti e quelle contraddittorie inficiano le procedure basate sul principio del minimo cambiamento e l'intera operazione di correzione.

Per evitare tali inconvenienti, i programmi generalizzati di tipo stocastico fanno ricorso ad un algoritmo che garantisce, entro certi limiti, la costruzione di un insieme minimo, non ridondante e non contraddittorio, a partire dalle regole esplicitate (Fellegi - Holt, 1976).

Nel caso di piani ad hoc, che non possiedono il suddetto analizzatore di regole, è conveniente ricorrere alla diagrammazione del sistema di relazioni tra variabili, definito dalle regole; nella stesura dello schema è immediato determinare la ridondanza e/o la contraddittorietà di alcune di esse.

L'esplicitazione delle regole costituisce una parte rilevante dei metadati dell'indagine, in particolare per quanto riguarda la «trasparenza» del processo di formazione del dato statistico; di essa, quindi, deve essere mantenuta documentazione chiara ed esauriente.

I criteri di correzione

I criteri di correzione sono vari e non è agevole darne una classificazione esaustiva e precisa.

I metodi correntemente utilizzati per indagini di media grande dimensione possono comunque essere classificati in *deterministici, da donatore* e *mediante regressione*; essi sono spesso utilizzati in combinazione tra loro (misture).

Generalmente, i criteri deterministici sono applicati per la correzione sia di variabili quantitative che qualitative, quelli da donatore solo per l'imputazione di quest'ultime, mentre i metodi da regressione per correggere le caratteristiche quantitative; sotto determinate condizioni, alcuni di essi si equivalgono.

Una quarta tecnica, la *correzione multipla*, consiste, sostanzialmente, nel reiterare sullo stesso file un procedimento (o procedimenti diversi) di imputazione. La procedura dà luogo a più replicazioni dello stesso insieme di dati; la stima finale verà calcolata come media delle stime risultanti dalle singole replicazioni.

Il metodo è ancora sperimentale, data la sua dispendiosità in termini informatici ed organizzativi (si pensi ad esempio all'archiviazione delle diverse replicazioni per successive analisi).

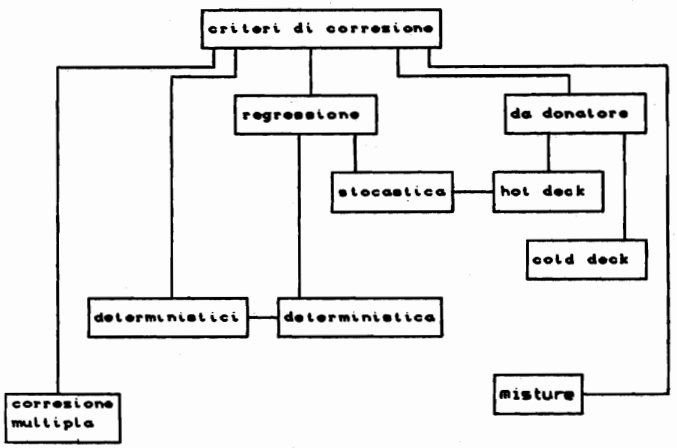


Figura 5.3 - I diversi criteri di correzione e le loro relazioni

I criteri deterministici

Gli algoritmi deterministici sono strettamente collegati alle regole di compatibilità e rispondono ad una logica «SE-ALLORA»: al verificarsi di una incompatibilità, data da una regola che coinvolge k variabili, di cui k^* già controllate ed eventualmente modificate in precedenza, si correggono le $k-k^*$ rimanenti, imponendo valori predeterminati o scelti a caso da una distribuzione definita a priori.

Le regole agiscono in maniera sequenziale e, quindi, il procedimento implica un ordinamento gerarchico tra esse.

La scelta della gerarchia influenza i risultati dell'algoritmo; la compatibilità e la correzione della i-esima variabile sono infatti funzione dei valori assunti o modificati delle precedenti.

Poiché, in questa logica, non è possibile applicare il principio del minimo cambiamento e la sequenza gioca un ruolo determinante nella procedura di correzione, la scelta del concatenamento delle regole, e quindi delle variabili, deve garantire dalla possibilità di errori indotti dalla procedura.

Possiamo assumere, quale *garanzia*, la minimizzazione della probabilità di modificare un valore *vero*; poiché essa è funzione delle probabilità a *priori* di errore sulle singole variabili e della probabilità condizionata di ripristinare un valore *vero sulla i-esima variabile, dato il valore assunto dalla j-esima, la sequenza dovrebbe essere ordinata in modo discendente secondo tali probabilità*.

Per fornire una base analitica a tali affermazioni, si può ricorrere ad un *modello* dell'operato della procedura che, pur se semplificato, ne rappresenta la logica di fondo.

Si abbia una procedura di compatibilità che coinvolge le k variabili X_i; in una logica deterministica si può assumere che le variabili vengono controllate e corrette in sequenza.

Sia stata scelta ad esempio la gerarchia $X_1, X_2 \ldots X_k$; essa dà origine alla sequenza di compatibilità e correzione:

$$C: X_1 \xrightarrow{R_1} X_1^* \xrightarrow{R_2} X_2 \xrightarrow{R_2} X_2^* \ldots\ldots X_{k-1}^* \xrightarrow{R_{k-1}} X_k \xrightarrow{R_k} X_k^*$$

Nella sequenza C sono stati tenuti distinti i valori *sporchi* X_i e quelli *puliti* X_i^* provenienti dall'applicazionie delle regole R_i; si è assunto, per semplicità espositiva, che ciascuna regola porta ad una modificazione della variabile originaria, che sarà reale, se viene riscontrata una incongruenza logica, od altrimenti fittizia (ovvero il valore grezzo e quello pulito coincidono).

Siano, inoltre, p_i le probabilità dell'errore, dovuto alle precedenti operazioni d'indagine, per la i-esima variabile e p_i^c le probabilità di errore della regola di correzione, ovvero la probabilità

dell'evento che il programma modifichi il valore della singola variabile in modo da generare un errore nel controllo con la successiva regola. Per errore, in questo contesto, si intende la presenza di un valore diverso da quello vero.

Indicando con E il verificarsi dell'errore, si ha per la generica X_i:

$$\text{Prob}\,(X_i = E) = p_i + (1 - p_i)\, p_{i-1}^* \qquad (5.4)$$

$$\text{Prob}\,(X_i^* = E) = p_i^* = p_i^c * \text{Prob}\,(X_i = E) \qquad (5.5)$$

in particolare, per $i = 1$ si ha:

$$\text{Prob}\,(X_1 = E) = p_1 \quad \text{e} \quad \text{Prob}\,(X_1^* = E) = p_1^* = p_1^c\, p_1$$

Come risultato della procedura si otterrà un record pulito in cui saranno presenti i valori X_i^*; sotto l'ipotesi di errori generati indipendentemente nelle variabili, possiamo, quindi, esprimere la probabilità che nel record si sia manifestato almeno un errore, come funzione delle probabilità p_i^*

$$\text{Prob}\,(E) = 1 - \text{Prob}\,(\bar{E}) = 1 - \prod_{i=1}^{k} (1 - p_i^*)$$

$$= 1 - \prod_{i=1}^{k} \left[1 - p_i^c (p_i + (1-p_i)\, p_{i-1}^*)\right] \qquad (5.6)$$

Tale probabilità dipende mediante la (5.6) dalla sequenza C utilizzata; al variare della sequenza, cambiano le p_i^c le p_i^* e quindi anche la Prob (E).

Le probabilità *a priori*, ovvero le probabilità dell'errore dovuto alle fasi precedenti, possono essere stimate sulla base delle mancate risposte parziali, degli errori di registrazione e dei valori fuori campo, mentre le probabilità condizionate mediante un indice di associazione asimmetrico ($\lambda_{A/B}$ e $\tau_{A/B}$ di Goodman e Kruskal, d di Somer), riscontrato in indagini precedenti per le medesime variabili.

Il procedimento di minimizzazione può riguardare il complesso delle caratteristiche rilevate oppure può essere limitato alle variabili più rilevanti.

In mancanza di tali informazioni, è conveniente che la sequenza delle regole di compatibilità e correzione preveda una gerarchia ordinata secondo l'importanza delle variabili, in modo da garantire, per le caratteristiche principali, il minimo di modificazioni.

Un caso particolare, ma che riveste notevole importanza, è rappresentato dal trattamento dei *salti* del questionario, ovvero dai quesiti in dipendenza dei quali vengono selezionate differenti

sequenze di domande. Se si accettasse, nelle regole di compatibilità e correzione, la gerarchia indotta dall'ordinamento dei quesiti del questionario, la possibilità di una modificazione di più variabili successive, verrebbe a dipendere dal valore assunto da una sola.

D'altro canto, cambiare il valore della domanda filtro può mutare radicalmente l'attribuzione di un particolare status al rispondente.

Per decidere quale sequenza accettare, è allora necessario considerare, contemporaneamente, tutti i quesiti coinvolti; è questo il metodo che, spesso implicitamente, viene seguito nella pratica. È, tuttavia, opportuno formalizzare tale criterio allo scopo di chiarirne i presupposti e rendere trasparenti ed esplicite le scelte assunte.

Sia X_f la variabile corrispondente al quesito filtro che può assumere i valori 1, 2, ...k; in corrispondenza di ciascun valore, le regole di compilazione del questionario impongono che sia presente una sola delle k sequenze ammesse:

sequenze	X_f	variabili			
S_1	1	$X_{1,1}$	$X_{1,2}$	...	$X_{1,n1}$
S_2	2	$X_{2,1}$	$X_{2,2}$	...	$X_{2,n2}$
:					
S_k	K	$X_{k,1}$	$X_{k,2}$	...	$X_{k,nk}$

Se i dati rilevati fossero esenti da errori, in ogni questionario sarebbe compilata con codici significativi una sola delle sequenze, mentre le rimanenti variabili, dipendenti dalla X_f, conterebbero un codice di «risposta non dovuta».

A causa degli errori derivanti dalle diverse fasi, invece, è possibile che nel singolo questionario siano presenti codici significativi nelle variabili $X_{i,j}$ che definiscono differenti sequenze S_i; inoltre, la stessa X_f potrebbe assumere un valore non significativo.

Il verificarsi di tali condizioni genera incertezza su quale delle k possibili risposte sia «vera».

In ogni questionario possono, quindi, corrispondere a k^* sequenze S_i, con $k^* \leq k$, altrettante sequenze osservate S_i^*, ciascuna delle quali ha la caratteristica che in almeno una delle variabili $X_{i,j}$ è presente un codice significativo.

Il problema consiste, allora, nello scegliere una sola S_i^* e, con le opportune regole di correzione, eventualmente completarne le informazioni mancanti.

In ciascuna sequenza osservata, possiamo discriminare tra le variabili che presentano un codice significativo e quelle che invece sono contrassegnate da blank (ovverso assenza di risposta o codice non ammissibile); se le $X_{i,j}$ sono contradditorie tra loro, all'interno della sequenza S_i, verranno considerate come blank.

Ad esempio, in corrispondenza di X_f le sequenze ammesse siano:

$$S_1 = (1;\ X_{1,1};\ X_{1,2};\ X_{1,3})$$
$$S_2 = (2;\ X_{2,1};\ X_{2,2})$$

Sul questionario invece sia presente la seguente configurazione:

X_f	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{2,1}$	$X_{2,2}$
blank	codice	codice	blank	codice	blank

che genera le due sequenze:

$$S_1^* = (X_f = \text{blank};\ X_{1,1} = \text{codice};\ X_{1,2} = \text{codice};\ X_{1,3} = \text{blank})$$
$$S_2^* = (X_f = \text{blank};\ X_{2,1} = \text{codice};\ X_{2,2} = \text{blank})$$

Occorrerà quindi decidere tra S_1^* e S_2^* e, come l'esempio suggerisce, completare la sequenza prescelta (imponendo almeno un codice alla X_f).

Quale criterio di decisione possiamo scegliere quella S_i^* per la quale sia massimo il rapporto tra la probabilità di aver osservato una data combinazione di codici, sotto l'ipotesi che la relativa sequenza sia «vera», e la probabilità dello stesso evento sotto la condizione che la sequenza sia «falsa»:

$$S_i^* \quad \text{tale che} \quad R_i = \max$$
$$R_i = \frac{\Pr\ (S_i^* / S_i = \text{vera})}{\Pr\ (S_i^* / S_i = \text{falsa})} \qquad (5.7)$$

La S_i^* che figura nella (5.7) è unica per il singolo questionario, ma può variare da un questionario all'altro; è perciò necessario che l'esperto consideri l'insieme di tutte le possibili combinazioni di codici (cui apparterranno le sequenze effettivamente osservate) ed assegni a ciascuna le probabilità condizionate.

Se la dimensione di tale insieme è ragguardevole, l'operazione può risultare impossibile. Una più semplice alternativa consiste nell'assegnare alle variabili X_f $X_{1,1}$ ...$X_{1,n1}$ $X_{2,1}$ ...$X_{2,n2}$ ...$X_{k,1}$ ...$X_{k,nk}$ le relative probabilità di errore e nel calcolare il numeratore ed il denominatore della (5.7) mediante:

$$\Pr\left[S_i^* / S_i = \text{vera}\right] = \left[p_f\,(1-I_f) + (1-p_f)\,I_f\right] \prod_{j\in A}(1-p_{i,j}) \prod_{j\in B} p_{i,j}$$

$$\Pr\left[S_i^* / S_i = \text{falsa}\right] = \left[p_f\,(1-I_f) + (1-p_f)\,I_f\right] \prod_{j\in B}(1-p_{i,j}) \prod_{j\in A} p_{i,j}$$

dove:

$$I_f = \begin{cases} 1 \text{ se } X_f = \text{codice} \\ 0 \text{ se } X_f = \text{blank} \end{cases}$$

A = insieme delle variabili con codici significativi
B = insieme delle variabili con codici blank
p_f, $p_{i,j}$ = probabilità di errore delle variabili X_f, $X_{i,j}$

Continuando nell'esempio sopra riportato si abbia:

p_f	$p_{1,1}$	$p_{1,2}$	$p_{1,3}$	$p_{2,1}$	$p_{2,2}$
0.05	0.07	0.09	0.10	0.09	0.04

In questo caso $I_f = 0$ e si ha:

	S_1^*	S_2^*
A	$X_{1,1}$; $X_{1,2}$	$X_{2,1}$
B	$X_{1,3}$	$X_{2,2}$
$\Pr\left[S_i^* / S_i = \text{vera}\right]$	0.00423	0.00182
$\Pr\left[S_i^* / S_i = \text{falsa}\right]$	0.000284	0.00432
R_i	14.89	0.431

Verrà quindi scelta la sequenza S_1^* che verrà resa completa correggendo $X_f = 1$ ed imponendo, un valore alla variabile $X_{1,3}$.

Le probabilità P_f, $P_{i,j}$ possono essere assegnate sulla base di analisi effettuate sulle medesime variabili in precedenti occasioni (ad esempio sulla base degli indicatori di mancata risposta esaminati nel capitolo 3), oppure da informazioni di tipo qualitativo; in mancanza di tali informazioni le probabilità possono essere interpretare come *pesi* assegnati alle diverse variabili in funzione della loro importanza.

Le considerazioni precedenti valgono, naturalmente, laddove non siano stati riscontrati errori sistematici; in caso contrario la formulazione delle regole di correzione, per eliminare la sistematicità, può non coincidere con il metodo consigliato.

Ad esempio, se si è riscontrato un errore sistematico in una variabile rilevante, è opportuno posporre, rispetto al piano programmato, la relativa regola di correzione e far dipendere il valore imputato per la suddetta variabile da altre caratteristiche, anche se meno importanti.

In conclusione, si possono indicare alcuni criteri per la predisposizione di un piano di compatibilità di tipo deterministico:

1) esplicitare le regole distinguendo chiaramente le condizioni di compatibilità (variabili SE) dall'azione correttiva (variabili ALLORA);
2) indicare l'esatta sequenza delle regole, in particolare in occasione di integrazioni del piano di compatibilità;
3) ordinare la sequenza, laddove è possibile, in funzione delle probabilità di errore, o di ripristino del valore vero, delle variabili coinvolte nelle diverse regole; oppure, in mancanza di tali informazioni, esplicitare la gerarchia delle variabili (e delle relative regole), in funzione della loro rilevanza a fini degli obiettivi conoscitivi dell'indagine;
4) in caso di quesiti di salto, esplicitare il sistema di pesi, o di probabilità, riportato nella (5.7).

I criteri da donatore

I programmi basati sui criteri *da donatore* utilizzano solo le informazioni desumibili dall'insieme completo e mirano, quindi, a salvaguardare le distribuzioni e le associazioni presenti in tale insieme.

In queste procedure, l'algoritmo determina, per ciascuna unità, l'esistenza di una incongruenza logica tra le variabili sulla base delle regole di compatibilità, decide quali di esse modificare, in funzione del principio del minimo cambiamento, e sceglie i valori da sostituire in modo tale da non attivare alcuna regola.

I valori delle variabili prescelte, vengono imputati ricorrendo a metodi diversi, ma tutti basati su un criterio casuale; essi possono essere generati da una distribuzione, semplice o congiun-

ta, desunta di dati puliti, oppure possono essere sostituiti con quelli presenti in una unità *donatrice*, in cui non è stato riscontrato alcun errore. In quest'ultimo caso, per le variabili qualitative si impone il medesimo valore, mentre per quelle quantitative è preferibile assegnare un valore *perturbato* (con un errore desunto dalla distribuzione dei valori *puliti* rispetto alla proprio media), per conservare la variabilità del fenomeno.

L'unità donatrice può essere individuata con due procedure: l'hot-deck ed il cold-deck.

Il metodo *cold-deck*, si differenzia da quello *hot-deck* in quanto nel primo, le unità vengono preventivamente suddivise in due insiemi (senza errori e con almeno un errore), mentre il secondo aggiorna dinamicamente un sottoinsieme di unità *pulite* da cui preleva il donatore; in funzione dell'ordinamento indotto nei file, vengono sfruttate, in questo modo, le eventuali correlazioni esistenti tra unità *vicine*.

Per *inizializzare* il sistema, ovvero nel caso che venga riscontrato un record errato prima del caricamento in memoria dell'insieme donatore, vengono previsti a priori alcuni record di default.

Nel criterio *hot-deck*, la numerosità dell'insieme donatore è stabilita in relazione alle performance informatiche dell'elaboratore. Tale insieme può essere costituito da tutte le unità del file *pulito* oppure, preferibilmente, da sottoinsiemi di unità stratificati mediante alcune caratteristiche (variabili di collegamento o matching) che non devono essere sottoposte a correzione e, possibilmente, devono risultare altamente correlate con le altre variabili del questionario; in genere assolvono tale compito i codici geografici e le variabili strutturali.

La stratificazione viene utilizzata non solo per soddisfare il principio di verosimiglianza delle forzature, ma anche per ridurre la distorsione imputabile all'eventuale non casualità delle mancate risposte parziali, tentando di identificare subpopolazioni omogenee per le quali sia ridotta la differenza tra le risposte fornite e quelle mancanti. Tale procedimento equivale a quello utilizzato per le mancate risposte totali e si basa sulle medesime considerazioni (cfr. Capitolo 3).

Una stratificazione delle unità abbastanza *fine*, e quindi con strati di ridotta numerosità, può comportare, però, un insieme donatore vuoto o un uso frequente del medesimo record donatore; per rimediare a tale inconveniente, si prevede la ricerca in altro strato simile (collassamento degli strati mediante la soppressione di una variabile di stratificazione).

Lo stimatore ottenuto con il metodo hot-deck può essere espresso come funzione dei valori degli m rispondenti; ad esempio nel caso della media

$$\bar{y}_{HD} = (\sum^{m} y_{i\cdot} + \sum^{m} t_i y_i) / n \quad \text{con} \quad \sum^{m} t_{i\cdot} = n-m$$
$$= (m \sum^{m} \bar{y}_R + \sum^{m} t_{i\cdot} y_i) / n \qquad (5.8)$$

dove le $t_{i\cdot}$ rappresentano il numero di volte che il valore $y_{i\cdot}$ viene utilizzato nella procedura di imputazione. La media e la varianza dello stimatore $\bar{y}_{HD}$ dipendono, mediante la 5.8, dai rispettivi parametri di $\bar{y}_R$ e dallo schema campionario utilizzato per generare i $t_{i\cdot}$.

Prospetto 5.2

A) Criteri di selezione dell'unità donatrice

I) mediante selezione casuale senza reimmissione per m > n/2 (criterio SR)
II) mediante selezione casuale con reimmissione per m < n/2 (criterio CR)
III) mediante selezione sequenziale con ordinamento casuale delle unità (criterio SEQ)
IV) mediante selezione da file ordinato (criterio ORD)

B) media e varianza dello stimatore $\bar{y}_{HD}$

stimatore	$E(\bar{y}_{HD})$	$Var(\bar{y}_{HD})$	
$\bar{y}_R$	$\bar{Y}_R$	$\frac{V}{n}\left[1 + \frac{n-m}{m}\right]$	
$\bar{y}_{HD}^{sr}$	$\bar{Y}_R$	$\frac{V}{n}\left[1 + \frac{2(n-m)}{n}\right]$	
$\bar{y}_{HD}^{cr}$	$\bar{Y}_R$	$\frac{V}{n}\left[1 + \left(\frac{n-m}{n}\right)\left(\frac{n+m-1}{n}\right)\right]$	
$\bar{y}_{HD}^{seq}$	$\bar{Y}_R$	$\frac{V}{n}\left[1 + \frac{2(n-m)}{m}\right]$	(a) (b)
$\bar{y}_{HD}^{ord}$	$\bar{Y}_R$	$\frac{V}{n}\left[1 + \frac{2(n-m)}{n} + 2\left(\frac{\varrho}{1-\varrho} - \frac{n-m}{n} \cdot \frac{2\varrho}{1-\varrho}\right)\right]$	(b)

dove: $V = Var(y)$
$\varrho = corr(y_i, y_j) \quad i, j = 1, 2, \ldots\ldots n$

(a) selezione sequenziale con il file ordinato casualmente
(b) varianze approssimate per $n \gg 0$

Nel caso in cui venga utilizzato un campione casuale semplice, selezionando l'unità donatrice con uno dei criteri casuali (i) — (iv) riportati nel Prospetto 5.2.A, lo stimatore è centrato sulla media dei rispondenti, ovvero $E(\bar{y}_R) = \bar{Y}_R$ e le relative varianze sono date dal prospetto 5.2.B.

Nel Prospetto 5.3 è riportato il confronto tra le varianze degli stimatori relativi ai criteri di selezione SR, CR, SEQ e $\bar{y}_R$.

Prospetto 5.3

	$Var(\bar{y}_R)$	$Var(\bar{y}^{sr}_{HD})$	$Var(\bar{y}^{cr}_{HD})$
$Var(\bar{y}^{sr}_{HD})$	>		
$Var(\bar{y}^{cr}_{HD})$	>	>	
$Var(\bar{y}^{seq}_{HD})$	>	>	>

Il confronto con $V(\bar{y}^{ord}_{HD})$ non è invece univoco, poiché esso dipende dal valore di ϱ e dalla relazione tra m ed n. Tale varianza risulta, in presenza di un $\varrho > 0$, sempre maggiore di $V(\bar{y}^{sr}_{HD})$, mentre è minore di $V(\bar{y}^{seq}_{HD})$ per $m > n/2$ e maggiore nel caso contrario; per un ϱ elevato ed un valore di m minore ma vicino a n/2, $V(\bar{y}^{ord}_{HD})$ è minore di $V(\bar{y}^{cr}_{HD})$.

Per quanto riguarda il caso di una stratificazione delle unità donatrici, con selezione operata mediante uno dei criteri (i) - (iv), non si dispone, attualmente, di una estensione delle suddette formule, che rimangono tuttavia valide all'interno di ogni strato.

Nel caso vengano utilizzate, quali variabili di collegamento, variabili quantitative è necessario definire, per identificare il record donatore, una funzione di distanza tra i record puliti Rk_p, e quello sporco, Rk_s:

$$D(Rk_s, Rk_p) = \left[\sum_i^K | y_{si} - y_{pi} |^r\right]^{1/r} \qquad (5.9)$$

dove le y sono le k variabili di collegamento ed r il valore che definisce la metrica utilizzata (r = 1 Manhattan, r = 2 euclidea, r = ∞ minimax). La selezione del donatore avviene allora mediante la condizione: $D(Rk_s, Rk_p) = \min$.

Per eliminare eventuali influenze dei valori di scala sulla (5.9), si trasformano le variabili y_i nelle:

$$Y_i^T = (y_i - a) / b$$

in cui a può essere la media, il minimo o la mediana e b lo scarto quadratico medio od il campo di variazione.

Per tener conto dei problemi connessi ad un uso frequente del medesimo record donatore, si modifica la funzione di distanza (5.9) facendola dipendere dal numero di replicazioni del donatore, d, e da una penalità, u:

$$D^*(Rk_s, Rk_p) = D(Rk_s, Rk_p)(1 + u\,d) \qquad (5.10)$$

I criteri «da donatore» sono teoricamente applicabili sia a caratteristiche qualitative che quantitative; in genere, però, per limitazioni di ordine informatico, si preferisce trattare queste ultime con programmi ad hoc, di tipo deterministico.

I criteri di regressione

Il criterio di correzione mediante regressione è generalmente applicato solo alle caratteristiche quantitative e consiste nella stima del valore correttivo mediante il modello di regressione lineare:

$$\hat{y}_i = \hat{b}_{R,0} + \sum_j^k \hat{b}_{R,j}\, x_{i,j} + \hat{e}_i \qquad (5.11)$$

Nella (5.11), $\hat{y}_i$ rappresenta il valore correttivo per l'unità i-esima, calcolato mediante il modello di regressione i cui coefficienti, $\hat{b}_{R,0}\hat{b}_{R,j}$ sono stati stimati sulla base delle informazioni fornite dai rispondenti; le X indicano le k variabili ausiliarie (che assumono per l'i-esima unità non rispondente i valori $x_{i,j}$) ed $\hat{e}_i$ è un residuo stocastico, con $E(\hat{e}_i) = 0$, da aggiungere al valore stimato $\hat{y}_i$.

Il modello di imputazione da regressione risulta deterministico se $\hat{e}_i$ viene posto uguale a zero, ovvero non si aggiunge nessun residuo generato da un modello casuale, mentre è stocastico nel caso contrario.

Operando sulle $x_{i,j}$ e sulle $\hat{e}_i$, questo metodo equivale ad una correzione apportata mediante altre tecniche.

Ad esempio, ponendo le variabili ausiliarie ed il residuo uguali a zero, esso coincide con l'imputazione di un valore pari alla media dei rispondenti; ponendo le variabili ausiliarie uguali a zero ed imputando un residuo desunto dalla distribuzione delle differenze tra i valori dei rispondenti e la loro media, il metodo equivale alla selezione casuale di un rispondente. La stratificazione delle unità equivale a considerare le x_j come variabili dummy, ovvero indicatrici di k strati; in questo caso si otterranno le medesime equivalenze di cui sopra a livello di singola classe. Se, oltre a stratificare, si ordinano le unità, il metodo di regressione stocastica è equivalente al criterio «hot-deck» sequenziale.

Nel prospetto 5.4 si riportano le principali caratteristiche di alcuni piani generalizzati di compatibilità e correzione, mentre informazioni più dettagliate su alcuni di essi, sono contenute in Appendice.

Prospetto 5.4: Principali caratteristiche di alcuni programmi generalizzati di compatibilità e correzione

nome	ente autore	tipo di correzione	elaborazione
AERO	CSO Ungheria	deterministiche stocastiche	batch
AERO Interactive Subsystem	FSO Yugoslavia	deterministiche stocastiche	interattivo
CANEDIT	Statistics Canada	stocastiche	batch
CONCOR	Bureau of the census U.S.A.	deterministiche stocastiche	batch
DIA	INE Spagna	deterministiche stocastiche	batch
EDIT 78	NCBS Svezia	deterministiche	batch interattiva
RESSAC	INSEE Francia	deterministiche	interattiva
SERIES IV	CSO Ungheria	deterministiche stocastiche	interattiva
TOSSVD	CISIS Bulgaria	deterministiche	interattiva
UNEDIT	Statistics Bureau ONU	deterministiche	batch

Il test sul piano di compatibilità

Le procedure di compatibilità e correzione, dopo la loro predisposizione e prima dell'applicazione ai dati dell'indagine, devono subire una validazione di tipo *formale* ed una di tipo *sostanziale*. L'obiettivo della prima consiste nella rispondenza del programma alle istruzioni impartite, mentre scopo della seconda è il controllo del funzionamento del piano in presenza delle due tipologie di errore, casuale e sistematico. I risultati di tali controlli comportano nel primo caso l'eventuale revisione delle istruzioni del programma, nel secondo la modifica delle regole di compatibilità e correzione.

I programmi generalizzati, al contrario di quelli prodotti ad hoc, non necessitano di verifiche formali, perché garantiti dal produttore, ma è opportuno approfondire le loro prestazioni statistiche, che dipendono, in ultima analisi, dalle regole di compatibilità esplicitate.

Per rendere possibile ed agevole il controllo e, successivamente, anche l'analisi dei risultati delle elaborazioni sui dati «reali» della rilevazione, è opportuno prevedere i seguenti accorgimenti:

- assegnare ad ogni regola di compatibilità un codice alfanumerico che dovrà essere riportato per ciascun record errato su uno degli output del programma;
- in mancanza di un codice *esatto* di individuazione del record (ad esempio, il caso in cui il codice che identifica l'individuo viene sottoposto a modificazioni da una regola di correzione), prevedere una rinumerazione sequenziale, dopo la fase di controllo quantitativo, di tutti i record; eventuali inserimenti nella successiva fase di controllo dei legami tra unità saranno caratterizzati da un valore blank, nell'apposito campo del record.

Per la verifica formale si compilerà, per ogni regola di compatibilità esplicitata, uno o più questionari errati, in dipendenza delle possibili correzioni previste; ad esempio, se una regola di correzione impone un valore da una distribuzione di frequenza, dovrà essere compilato un numero di questionari sufficiente a verificare la casualità della correzione. Realizzato l'accoppiamento tra file pulito e sporco, mediante il codice *esatto*, si verificherà l'esistenza dell'incompatibilità, il risultato della correzione e la rispondenza tra la regola ed il relativo codice di errore apposto dal programma.

La verifica sostanziale del piano di compatibilità si effettua sostituendo valori *errati* in un insieme *pulito* di dati ed elaborando questi ultimi mediante i programmi di compatibilità e correzione. I valori da imputare, ammissibili e non ammissibili, saranno generati da un modello casuale o da uno sistematico (adottando, in questo caso, una delle definizioni riportate nel sottoparagrafo «l'errore sistematico») o da misture, a seconda dell'aspetto che si vuole sottoporre a controllo.

Da tale procedura sono disponibili tre file (origine, sporco, pulito), sui cui è possibile operare tre confronti:

i) tra il file origine e quello sporco;
ii) tra il file sporco e quello *pulito*;
iii) tra il file origine e quello *pulito*.

Ciascuna comparazione presenta una differente valenza informativa: mediante la prima si conoscono analiticamente i ri-

sultati del processo di generazione degli errori, che, essendo di natura stocastica, sono *ex ante* fuori controllo; il secondo confronto simula la situazione reale dell'indagine ma, rispetto a questa, presenta il vantaggio di conservare la *storia* del singolo record e quindi permette di verificare il funzionamento delle regole di compatibilità e di correzione; mediante il terzo confronto, istituito tra i valori *puliti* e quelli *veri*, si può, infine, valutare distintamente l'effetto, dovuto al tipo di errore indotto nei dati e al procedimento di correzione, sulle stime finali.

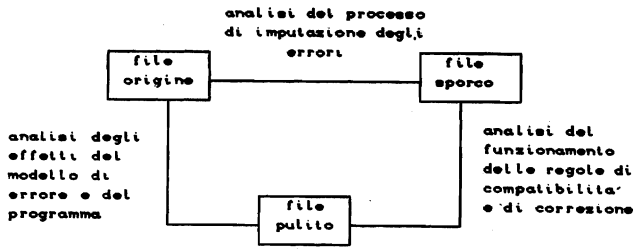


Figura 5.4 - Il processo di simulazione per il controllo dei programmi di compatibilità e correzione

Poiché i risultati del procedimento dipendono dal campione di errori generato, è necessario simulare più replicazioni indipendenti del file *sporco* e riferire l'analisi alla *media* delle prestazioni.

Sia y_{rj}^p il valore della generica variabile y, dopo l'elaborazione della procedura di correzione nell'r-esima replicazione, della j-esima unità ed y_j il relativo valore *vero* del file originario (che rimane costante al variare delle replicazioni); sia inoltre $R \geq 2$ il numero delle replicazioni ed n il numero delle unità.

Nel caso delle caratteristiche qualitative, la y rappresenta la variabile dicotomica (presenza / assenza) relativa alla modalità esaminata; si assume, quindi, che le matrici dei dati originarie vengano trasformate nelle rispettive matrici disgiuntive complete.

Un primo aspetto delle prestazioni della procedura è rappresentato dalla «distanza» tra il file pulito e quello originario, che possiamo misurare come media delle differenze assolute tra i rispettivi valori:

$$\sum_r \sum_j |y_{rj}^p - y_j| \,/\, Rn \qquad (5.12)$$

Nel caso delle variabili dicotomiche, la (5.12) assume il significato di un tasso atteso di modificazioni.

Un secondo aspetto da valutare è il comportamento della procedura rispetto agli stimatori che verranno utilizzati nell'analisi dei dati finali; a titolo di esempio, ci si riferirà alla funzione *media* (che coincide con la frequenza relativa nel caso delle variabili dicotomiche).

Siano, quindi, per la generica variabile y,

$$\bar{y}_r^p = \sum_j y_{rj}^p / n \quad ; \quad \bar{y} = \sum_j y_j / n \quad ; \quad \bar{y}^p = \sum_r \bar{y}_r^p / R$$

rispettivamente, la media nell'r-esima replicazione, la media *vera* nel file origine e la media nelle R replicazioni.

Un primo indicatore è rappresentato dalla stima della distorsione della media, calcolata dopo la procedura di correzione, rispetto al medesimo parametro nei dati originari. La distorsione può essere misurata nel duplice aspetto di distorsione attesa (in intensità e direzione)

$$\hat{B} = \sum_r (\bar{y}_r^p - \bar{y}) / R = \bar{y}^p - \bar{y} \qquad (5.13)$$

e di distanza dal valore vero

$$|\hat{B}| = \sum_r |\bar{y}_r^p - \bar{y}| / R \qquad (5.14)$$

Poiché possiamo considerare le stime $\bar{y}_r^p$ come provenienti da *campioni di errori*, la variabilità campionaria dei risultati ottenibili dalla procedura, stimata in termini assoluti o relativi (rispetto alla media *vera*), costituisce un secondo indicatore della bontà del metodo di correzione:

$$\hat{SE}(\bar{y}) = \left[\sum_r (\bar{y}_r^p - \bar{y}^p)^2 / R(R-1)\right]^{1/2} \qquad (5.15)$$

$$\hat{CV}(\bar{y}) = \hat{SE}(\bar{y}) / \bar{y} \qquad (5.16)$$

Infine, mediante la (5.17) possiamo unificare l'errore variabile e la distorsione in un solo indicatore di errore totale:

$$\begin{aligned} \hat{MSE}(\bar{y}) &= E_r(\bar{y}_r^p - \bar{y}^p + \bar{y}^p - \bar{y})^2 \\ &= (\hat{SE}(\bar{y}_r^p))^2 + B^2 \end{aligned} \qquad (5.17)$$

Per confrontare gli indicatori relativi a differenti variabili, è necessario svincolare le (5.13), (5.14) e (5.17) dall'unità di misura, rapportandole alla media $\bar{y}$:

$$\hat{B}' = \hat{B} / \bar{y}$$

$$|\hat{B}|' = |\hat{B}| / \bar{y} \qquad (5.18)$$

$$M\hat{S}E' = M\hat{S}E / \bar{y}$$

Infine, dalle R replicazioni, se in numero sufficientemente elevato, possono essere calcolate le distribuzioni di due indicatori, relativi alla distorsione della stima dell'errore campionario e all'influenza della procedura sulle correlazioni:

$$I_r^v = \hat{V}_r(\bar{y}) / \hat{V}(\bar{y})$$

$$I_r^c = corr_r(y^p, x) / corr(y, x) \qquad (5.19)$$

dove:

$$\hat{V}_r(\bar{y}) = \Sigma_j (y_{rj}^p - \bar{y}_r^p)^2 / n(n-1)$$

$$\hat{V}(\bar{y}) = \Sigma_j (y_j - \bar{y})^2 / n(n-1)$$

Data la gran mole di calcoli e di risultati derivanti dalle (5.12) - (5.19), è opportuno limitare la simulazione e l'analisi alle più rilevanti caratteristiche oggetto d'indagine.

Le informazioni desumibili dall'elaborazione del piano di compatibilità

Le informazioni che è possibile ottenere dai programmi di compatibilità e correzione, riguardano da un lato l'esame ex post della stessa procedura, ovvero delle regole utilizzate e delle correzioni effettuate, e dall'altro la valutazione di alcuni aspetti della *qualità* dei dati e delle prestazioni della rete di rilevazione, utilizzando gli indicatori riportati nel Prospetto 3.5 e nel Prospetto 5.5.

Gli indicatori desumibili dalla procedura di compatibilità e correzione, anche se indiretti e grezzi, hanno il pregio però di essere economici, poiché facilmente ottenibili dalle procedure esistenti, ed analitici, in quanto possono essere calcolate per qualsiasi livello di controllo.

Prospetto 5.5

INDICATORI	SIGNIFICATO	FONTE DI ERRORE		
		comuni	rilevatori	ISTAT
	A) REGOLE			
$\Sigma_r T_r / NR$	regole utilizzate	—	—	si
$\Sigma_j T_{rj} / NRK$	utilizzazioni della regola	—	—	si
$\sum_{r} \Sigma_j M_{ij} / NRK$	variabili corrette per regola	—	—	si
$\sum_{r} \Sigma_j M_{ij} / V_r \cdot NRK$	variabili corrette per regola	—	—	si
	B) CORREZIONI			
$\Sigma_j M_{rj} / NRK$	correzioni per variabile	—	si	si
$\Sigma_i \Sigma_j M_{ij} / NRK$	correzioni per record	—	si	si
$\Sigma_j RK_j / NRK$	record corretti	—	si	si
$\sum_{s} \Sigma_j M_{ij} / V_s \cdot NRK$	correzioni per gruppo di variabili	—	si	si

NRK	numero di record
NR	numero di regole
V_r	numero di variabili coinvolte nella regola r
V_s	numero di variabili nell'insieme S
T_{rj}	indicatore di utilizzazione della regola r per il record j; si = 1 no = 0
T_r	indicatore di utilizzazione della regola r; $T_r = 1$ per $\Sigma_j T_{rj} > 0$; $T_r = 0$ per $\Sigma_j T_{rj} = 0$
M_{ij}	indicatore di correzione per la variabile i nel il record j; si = 1; no = 0
RK_j	indicatore di correzione del record j; $Rk_j = 1$ per $\Sigma_j M_{rj} > 0$; $Rk_j = 0$ per $\Sigma_j M_{rj} = 0$

La valutazione delle regole potrà basarsi sull'esame degli indicatori riportati nel Prospetto 5.5.A. La percentuale di regole utilizzate ed il numero medio di utilizzazioni, misurano l'estensione e l'intensità dell'operato del sistema di regole; mediante i due rimanenti indicatori è possibile analizzare e confrontare le prestazioni per la singola regola.

L'analisi delle correzioni apportate può essere effettuata a diversi gradi di approfondimento: rispetto alle modalità della singola variabile, a ciascuna od a gruppi rilevanti di variabili, a livello record.

È conveniente, per evitare la ridondanza di informazioni, limitare a livello totale o di domini di studio, le analisi relative alle modalità (per ciascuna caratteristica la distribuzione incrociata tra i valori assunti prima e dopo del piano di compatibilità), ai record (la distribuzione per numero di correzioni) e al tasso di modificazioni per le singole variabili.

Gli indicatori, riportati nel prospetto 5.5.B, assumono un diverso significato a seconda del livello cui sono riferiti: a livello aggregato, misurano l'effetto del piano di compatibilità e confluiscono, quindi, nell'insieme degli indicatori della qualità dei dati; se calcolati per comune e per rilevatore costituiscono degli indicatori delle prestazioni della rete periferica.

L'esame delle regole e delle correzioni può essere utilizzato quale ulteriore controllo dell'esistenza di errori sistematici, prima dell'elaborazione definitiva dei dati. L'analisi condotta a livello del complesso dei dati può individuare l'errore derivante dal questionario, dalla registrazione e dalle regole di correzione; ad esempio, una eccessiva frequenza nell'uso di determinate regole di compatibilità, od un numero anomalo di correzioni per una data modalità, possono evidenziare un quesito od una sequenza di quesiti mal formulati o concatenati.

Un modello per l'identificazione dell'errore sistematico sulla base dei risultati del piano di compatibilità è riportato in Marchetti & Masselli, 1984 ed in Marchetti 1986.

L'analisi degli indicatori a livelli più disaggregati può invece evidenziare errori sistematici commessi dai singoli organi periferici.

Studi più analitici, infine, possono essere condotti, mettendo in relazione i suddetti indicatori con le caratteristiche socio demografiche dei rispondenti e dei rilevatori.

8. Il controllo dei legami tra unità

Come anticipato nel paragrafo 2, la fase di controllo e correzione delle variabili che rappresentano legami tra unità può essere effettuata sia prima che dopo quella di imputazione delle

caratteristiche individuali. La scelta dell'una o dell'altra sequenza deve essere effettuata in funzione dell'importanza attribuita alle informazioni attinenti alle singole unità oppure al contenuto informativo dei loro legami; essa determina tipologie diverse di controllo e correzione dei legami.

Ad esempio, in una generica indagine sulla popolazione, in cui sono presenti le unità «famiglia» ed «individuo», la variabile «relazione di parentela» può essere considerata sia come legame tra unità sia come caratteristica degli individui. In questo caso la variabile sarà in relazione con altre, ad esempio l'età e mediante questa con la condizione professionale.

Possiamo allora scegliere due diverse «strategie di imputazione»:

i) come primo passo, considerare la «relazione di parentela» come una caratteristica dell'individuo e quindi controllarla ed eventualmente correggerla sulla base delle relazioni intercorrenti tra le variabili individuali; come secondo passo controllare e correggere la stessa variabile, quale legame tra gli individui appartenenti alla stessa famiglia.

ii) ribaltare la sequenza di cui sopra, controllando e corregendo dapprima la relazione di parentela come legame tra individui, e quindi operando sulle relazioni tra le caratteristiche individuali, considerando immodificabile la relazione di parentela.

Nella prima strategia, la correzione dei legami interni all'unità famiglia assume un ruolo residuale rispetto a quella apportata alle relazioni tra le informazioni individuali; cosicché essa verrà prescelta se si vogliono salvaguardare i contenuti informativi del singolo individuo rispetto alla struttura della famiglia.

Le due strategie implicano metodi di imputazione diversi: nella prima è opportuno che le correzioni della «relazione di parentela» vengano apportate in maniera interattiva da esperti di settore per evitare cicli successivi di modificazioni individuo/famiglia, mentre nella seconda, possono essere effettuate mediante piani automatici.

APPENDICE

1. Programmi generalizzati di compatibilità e correzione automatica

Presso l'Istituto sono stati sperimentati due programmi generalizzati per la compatibilità e correzione automatica: l'AERO ed il CONCOR mentre di un terzo programma, il DIA, si stanno attualmente verificando le performance; del CANEDIT si è in possesso di numerosi elementi di valutazione.

Quella che segue è una sintetica panoramica su questi quattro prodotti.

- IL CONCOR

Il CONCOR (CONsistency and CORrection) è un programma prodotto dall'International Statistical Program Center del Bureau of Census U.S.A. in collaborazione con l'United Nations Demographic Center for Latin America, cui si deve la prima versione; è datato 1979 ed è scritto in COBOL con parti in Assembler.

È applicabile solo a variabili qualitative e può identificare errori nella struttura delle risposte di un questionario (structural check), nel range dei valori delle risposte (range check) e la congruità tra risposte a differenti quesiti correlati tra loro, sia che siano fisicamente sullo stesso records, sia in record diversi (consistency check); genera automaticamente un programma COBOL e dei files per analisi e tabulazione.

L'ultima versione del CONCOR, contrariamente alle prime, per le quali veniva utilizzato un criterio deterministico provvede alle correzioni automatiche con la tecnica dell'*hot-deck*; a mano a mano che i record vengono sottoposti al controllo, quelli giudicati *puliti* aggiornano delle tabelle in memoria (che rappresentano in pratica dei record medi donatori) le quali, in caso di erroneità dei dati, forniranno i valori da sostituire.

Il sistema CONCOR è composto da 4 sottosistemi:

- l'analizzatore del linguaggio di comando (tra cui le *regole*);
- il generatore del programma COBOL;
- l'Editor, per la corretta imputazione;
- il generatore dei report.

Il linguaggio di comando consiste di istruzioni di definizione, di istruzioni eseguibili (per l'organizzazione ed il controllo del programma, per assegnazioni o trasformazioni di valori, per il test

di condizioni, per l'imputazione e la creazione di output), di istruzioni di controllo (stabiliscono il livello ed il tipo delle statistiche) e di istruzioni di commento.

La sintassi è tipo-COBOL (tipica la struttura del programma in Divisioni, Sezioni e Paragrafi) e presuppone l'intervento di un informatico.

Il CONCOR può accettare fino a 50 tipi record diversi, di lunghezza fissa, con il tipo record sempre alle stesse posizioni.

Come output sono previsti: il file esatto, report statistici con notizie sui test effettuati e sugli errori riscontrati, nonché file opzionali derivanti dai diversi passi della procedura, il cui formato è stabilito dall'utente; è prevista inoltre la possibilità di effettuare confronti tra i dati originali e quelli puliti

Nelle sperimentazioni Istat è stato evidenziato un limite nelle performance informatiche; a tutt'oggi è impiegato nella versione deterministica nel controllo dei dati dell'indagine sulle Forze di lavoro.

- L'AERO

L'AERO è un sistema di controllo e correzione prodotto dall'Ufficio di statistica ungherese e utilizzato, non solo in Ungheria, per il Censimento della Popolazione 1981.

È composto da due sottoinsiemi. Il primo, sottosistema di specificazione, crea il dizionario e definisce le regole per la correzione mentre il secondo, sottosistema di generazione, provvede, in base ai parametri forniti dall'utente, a controlli di range, di relazione e consistenza, a correzioni automatiche, a report statistici e a liste diagnostiche di vario tipo.

È possibile specificare tre tipi di regole:

— regole Y, ovvero condizioni di errore o di rigetto che provocano la bipartizione del file originario in errati ed esatti e l'utilizzo di questi ultimi per le correzioni automatiche;
— regole X le quali, in una fase successiva, assicurano che particolari condizioni sui campi del record siano soddisfatte (si tratta in pratica di imputazioni deterministiche);
— regole Z che permettono di scrivere più sinteticamente le regole Y e X.

Anche l'AERO effettua le correzioni secondo il metodo hot-deck, identificando il record donatore nel file pulito che viene creato a mano a mano che si procede nella lettura e nel controllo.

La correzione dei campi errati non viene effettuata simultaneamente, ma avviene per sottoinsiemi di campi: è l'utente che

decide le variabili per ogni gruppo e stabilisce l'ordine con il quale i campi debbono essere corretti.

L'AERO può gestire solo variabili qualitative con un ridotto numero di modalità. È programmato in PL/1 con parti in Assembler e, a differenza del CONCOR che origina dapprima un sorgente COBOL e poi lo compila automaticamente, l'AERO produce dei moduli già eseguibili.

La sperimentazione condotta ha evidenziato che è possibile utilizzare AERO solo con insiemi di dati di ridotte dimensioni provenienti da questionari non troppo complessi.

- IL CANEDIT/SPIDER

Il CANEDIT è un sistema di controllo e correzione prodotto da Statistics Canada (dove è più noto con il nome originario di GEISHA: Generalized Edit and Imputation System using the Hot deck Approach).

È stato utilizzato nel Censimento della Popolazione 1981 insieme al prodotto SPIDER, capace di gestire variabili continue o discrete con un gran numero di valori che il CANEDIT non è capace di supportare.

Esso utilizza una tecnica hot deck ed è composto da un analizzatore logico e da un sottosistema di controllo ed imputazione. Il primo identifica le regole ridondanti e contradditorie mentre il secondo ricerca i valori corretti da sostituire copiandoli da record donatori che soddisfano le condizioni volute. Nel caso in cui i donatori sono più di uno, viene effettuata una scelta casuale; se non vengono identificati donatori allora ogni campo corretto viene copiato da un donatore diverso.

SPIDER è un sistema collaterale (System for Processing Instructions from Directly Entered Requirements) che recepisce le regole dell'utente sotto forma di tavole di decisione e le organizza in un programma PL/1.

Il sistema è capace di riconoscere 4 diversi tipi di unità: persona, abitazione, famiglia censuaria e famiglia economica.

Anche SPIDER è basato sulla tecnica hot-deck e di un sistema di pesi e di una stratificazione dei record possibili donatori; nel caso in cui, però, non riesca a rintracciare il donatore, utilizza una imputazione deterministica (in pratica applicando dei valori di default).

La lunghezza della ricerca può essere controllata sia indicando il numero massimo di record da esaminare sia restringendola a determinati livelli geografici; se la ricerca nel «serbatoio» nell'ambito di uno strato non ha avuto esito positivo, il donatore viene selezionato a caso tra quelli più vicino allo strato.

Il binomio CANEDIT/SPIDER si muove in ambiente DB (RAPID) con programmi PL/1.

- IL NEIS

Il NEIS è un sistema generalizzato per il controllo e l'imputazione di variabili quantitative prodotto recentemente da Statistics Canada.

Esso trae origine da un prototipo del 1976 scritto in Fortran; la versione attuale, che incorpora il prototipo, è programmata in C ed utilizza l'SQL.

Più che un sistema integrato, il prodotto è un insieme di programmi ognuno dei quali esegue funzioni di controllo e di imputazioni diverse, sulla base del principio del minimo cambiamento e del metodo hot-deck.

Può essere considerato un prodotto sperimentale per la necessità di ricorrere a programmi ad hoc e ad interventi manuali.

- IL DIA

Il DIA (sistema de Deteccion y Imputacion Automatica) è un sistema per il controllo e l'imputazione per variabili qualitative sviluppato dal 1981 al 1984 dall'Istituto de Estadistica de Espana.

È stato utilizzato in varie indagini: l'inchiesta sulla fecondità 1985, il Censimento della popolazione mussulmana di Ceuta e Melilla 1986, il Censimento della Popolazione 1986 (questionario ridotto) e l'inchiesta sulla popolazione attiva 1987; verrà applicato all'elaborazione del Censimento della Popolazione 1991.

Scritto in PL/1, per elaboratori IBM o IBM-compatibili, la prima versione completa è del luglio 1987.

Pur essendo conosciuto come il CANEDIT spagnolo, è in realtà un prodotto completamente diverso, che ha in comune con il programma canadese la generazione dell'insieme minimo, completo e non contraddittorio di regole secondo la metodologia di Fellegi e Holt.

L'imputazione non avviene mediante record donatori, ma generando i valori da imputare, con un algoritmo stocastico, dalle distribuzioni ottenute dai record puliti.

L'integrazione tra criteri stocastici e deterministici, del resto già presente in CANEDIT/SPIDER, viene raggiunta risolvendo le inconsistenze ed i conflitti tra regole di incompatibilità e regole di imputazione.

La scrittura delle «regole» per il DIA è estremamente semplice e concisa, tale da consentire, ad esempio, di scrivere le regole per i controlli di range delle variabili di un'indagine in pochi minuti.

Il DIA necessità di 4 input:

— l'elenco delle variabili del record con i codici assumibili;
— l'elenco delle variabili del record con le specifiche di posizione e lunghezza;
— le regole di compatibilità;
— le regole di imputazione deterministica

Il limite del DIA è che esso non tratta le variabili quantitative o qualitative con un gran numero di modalità dovendo ricorrere per il trattamento di tali caratteristiche a programmi di controllo ed imputazione ad hoc.

2. Analisi delle prestazioni di un programma di compatibilità

Nell'esempio che segue, si riporta l'analisi delle prestazioni di un programma di compatibilità di tipo deterministico, tratto da M. Masselli in *La qualità dei dati nell'indagine Istat sulla salute 1983*, in Atti del Convegno *Salute e ricorso ai servizi nel Veneto*, Regione Veneto, novembre 1987.

La procedura di compatibilità e correzione dell'indagine Istat sulla salute degli italiani 1983 è basata su 174 regole la cui funzione è diversa: controllare che il valore della singola variabile sia all'interno del campo di variazione ammesso o porre in relazione variabili diverse. Se la regola viene contraddetta, il programma modifica il contenuto di una caratteristica secondo una determinata gerarchia; tale piano è quindi di tipo deterministico.

Sui dati relativi alla regione Veneto, la procedura ha utilizzato 98 delle 174 regole, (pari al 56.3%), previste dal piano di compatibilità; tali regole, che denomineremo *fallite*, hanno dato luogo, quindi, ad almeno una modificazione dei valori originari. Analizzando tale insieme di regole è stato possibile risalire ad alcune delle cause delle correzioni (Tavola 5.A.1). In essa, particolarmente interessanti sono i dati relativi alla «risposta non dovuta» e alle «relazioni quantitative non verificate», che possiamo ritenere imputabili al mancato rispetto delle norme formali per la compilazione del modello di rilevazione da parte del rilevatore; la loro somma è superiore al 20% del totale delle correzioni.

Nella Tavola 5.A.2, sono stati costruiti alcuni indicatori riguar-

Tavola 5.A.1 - Correzioni apportate dalla procedura secondo la causa

CAUSA CORREZIONI	NUMERO CORREZIONI	
	val ass.	val. %
— mancata risposta	291	11.47
— risposta non dovuta	231	9.10
— registrazione su supporto informatico	27	1.06
— relazioni quantitative non verificate	317	12.50
— altre	1.670	65.85
Totale	2.536	100.00

danti l'operato della procedura, in relazione alle diverse sezioni in cui si articola il modello di rilevazione. C'è da notare, innanzitutto, l'assenza di correzioni nella sezione «malattie presenti», probabilmente dovuta alla semplicità delle risposte previste (di tipo dicotomico) e, soprattutto, all'impossibilità di istituire controlli incrociati con altre variabili del questionario: le relative regole, quindi, sono solo di controllo di campo.

Il totale delle modificazioni subite dai dati grezzi è stato di 2536 a fronte dei 4470 questionari, con un numero medio di correzioni per modello pari a 0.57.

La terza colonna scompone il totale delle correzioni per sezioni del modello, ed indica che la maggior parte delle modificazioni si sono concentrate nella sezione dei «dati generali» ed «accertamenti diagnostici». Tale parametro, però, non tiene conto della struttura del questionario ed è inadeguato a rappresentare una gerarchia di qualità tra i diversi tipi di informazioni; a tal fine, nella quinta colonna, è riportato un indicatore calcolato come rapporto percentuale tra correzioni e variabili soggette a correzione in ciascuna sezione. Le conclusioni che è possibile trarre, cambiano di segno: sono le risposte fornite alle sezioni «rispondente» ed «attività fisica» quelle più soggette ad incongruenze, seguite da «visite mediche» e «dati generali». Nell'ultima colonna, infine, è riportato un indicatore che potremmo definire di *efficienza* delle regole: esso misura la quantità *filtrata* dalle stesse rapportando il numero di correzioni al numero di regole fallite. Valori di tale parametro particolarmente alti o bassi possono indicare una non adeguata formulazione delle regole o del questionario. Nel caso in esame i valori relativi alle sezioni «malattie presenti» e «attività fisica» sono da considerarsi sospetti.

Tavola 5.A.2 - Principali parametri del piano di correzione

SEZIONI MODELLO	regole	regole fallite	CORREZIONI %	variabili soggette a correzioni	NUMERO MEDIO CORREZIONI per variabile	NUMERO MEDIO CORREZIONI per regola fallita
— dati generali	41	19	22.2	13	50.5	29.68
— stato di salute	25	15	8.7	8	27.5	14.67
— malattie presenti	20	—	—	20	—	—
— invalidità	9	3	3.4	8	10.8	28.67
— visite mediche	9	8	14.8	7	53.4	46.75
— diagnostici	28	24	16.9	10	42.8	17.83
— consumo farmaci	13	4	3.4	11	7.8	21.50
— consumo tabacco	17	13	10.7	7	38.7	20.85
— consumo bevande	6	6	8.2	5	41.8	34.83
— attività fisica	2	2	6.0	2	75.5	75.50
— rispondente	4	4	5.8	2	73.5	36.75
— Totale	174	98	100.0	93	27.3	25.88

Una ulteriore analisi delle prestazioni delle regole è riportata nella Tavola 5.A.3, in cui il numero di correzioni per singola regola è stato ridotto a classi significative (gli estremi superiori corrispondono allo 0.5%, al 1%, all'1.5%, al 2% ed al 2.4% dei 4470 questionari). Da tale tavola risulta che delle 98 regole fallite, 5 hanno pesato sulla correzione, avendo modificato da 101 a 106 questionari, mentre la gran parte di esse, 56, è responsabile di modifiche su un massimo di 22 modelli di rilevazione.

Tavola 5.A.3 - Classi di correzioni per regole fallite

CLASSI DI CORREZIONI PER SINGOLA REGOLA	numero regole utilizzate
1 - 22	56
23 - 44	20
45 - 90	17
91 - 106	5
Totale	98

3. Analisi degli effetti di un piano di compatibilità

Si riporta un esempio di analisi degli effetti di un piano di compatibilità tratto da M. Masselli in *La procedura di controllo degli effetti del piano di compatibilità dell'indagine sulle forze di lavoro*. Istat, documento interno.

Mediante le informazioni derivate dall'abbinamento dei record *grezzi* e di quelli sottoposti a correzione, sono stati studiati gli effetti prodotti dai piani di compatibilità, prendendo in considerazione tre diversi livelli di controllo:

- i record
- le variabili
- le modalità delle singole variabili.

Analisi per record

La distribuzione dei record secondo il numero di correzioni è un indicatore di quanto ha inciso la procedura a livello di unità di analisi e, nel contempo, dello stato del materiale grezzo.

Tavola 5.A.4 - Percentuale di record corretti per numero di correzioni

NUMERO DI CORREZIONI	% di record	NUMERO DI CORREZIONI	% di record
0	68.6	6	0.6
1	19.0	7	0.6
2	4.8	8	0.6
3	2.5	9	0.9
4	1.1	10	0.3
5	0.7	10-18	0.4

Dalla Tavola 5.A.4 risulta che il 31% delle unità ha subito almeno una modificazione e che il numero massimo di modificazioni apportate su di un record è pari a 18; la gran parte di essi (l'80%) è stato comunque soggetto, al massimo, a 2 correzioni.

Poiché le variabili del questionario non hanno tutte la stessa importanza, è opportuno analizzare tale distribuzione in funzio-

ne di blocchi omogenei di informazione, raggruppando opportunamente le variabili:

A. variabili per la definizione degli aggregati delle forze di lavoro;
B. variabili demografiche;
C. identificatori individuali;
D. variabili relative alla condizione lavorativa;
E. variabili relative alla ricerca di lavoro;
F. variabili relativi al lavoro precedente;
G. variabili relative al corso professionale.

Tavola 5.A.5 - Percentuale di record corretti per numero di correzioni e per gruppi di quesiti

NUMERO CORREZIONI	GRUPPI DI VARIABILI						
	A	B	C	D	E	F	G
0	90.5	90.0	98.4	96.2	95.2	88.0	97.3
1	5.6	9.5	1.5	2.8	2.7	8.2	2.6
2	1.9	0.4	..	0.5	0.4	1.7	..
3	1.6	0.1		0.1	1.4	0.4	
4	0.3			0.4	0.1	0.5	
5	0.1				0.2	0.4	
6	..					0.8	
7						..	
Numero variabili	9	5	3	5	5	7	2

La massima percentuale di record modificati si riscontra nel gruppo di variabili relative al *lavoro precedente*, seguito da quelle utilizzate per la definizione degli aggregati delle forze di lavoro.

Poiché le percentuali di record modificati variano notevolmente da gruppo a gruppo e non sembra esistere una relazione tra tali tassi ed il numero di variabili coinvolte, si può dedurre che gli errori derivanti dalla fase di rilevazione non sono uniformemente distribuiti e/o che la procedura privilegia la correzione delle variabili di tipo F,A,B.

Analisi per variabile

La percentuale delle modificazioni subite dalle singole variabili costituisce una misura dell'impatto dei piani di compatibilità sui risultati finali; nella Tavola 5.A.6, le variabili sono state ordinate secondo i tassi di modificazioni subite.

Tavola 5.A.6 - Tassi di modificazioni per variabile

VARIABILE	QUESITO	% MODIFICAZIONI	
Mese di nascita		—	
Sesso		0.1	
Anno di nascita		0.1	
Residenza		0.1	*
Età		0.1	
Età < 14 anni		0.1	*
Ore attività secondaria	12	0.2	
Tipo corso	15.2	0.2	
Posizione della professione	11.3	0.7	
Mesi di ricerca lavoro	13.2	0.7	
Orario di lavoro	11.5	0.8	*
Ore effettive	11.1	0.9	*
Carattere perm. occupazione	11.6	1.0	
Motivo ore eff. >ore abit.	11.2	1.1	*
Attività economica	11.4	1.2	
Disponibilità a lavorare	14.5	1.3	*
Proxy	16	1.4	
Ore abituali	11.1	1.7	*
Condizione momento ricerca	14.3	1.7	
Parentela		1.8	
Ex posizione	13.4.1	1.8	
Ex ramo	13.4.2	1.8	
Stato civile	9.	1.9	
Motivo abbandono lavoro	13.3	2.0	
Mesi ricerca occupazione	14.4	2.0	
Aggregati forze di lavoro		2.1	
Tipo occupazione	14.2	2.3	
Attività secondaria	12	2.4	
Obiettivo corso	15.3	2.5	
Numero azioni ricerca	14.6	2.7	
Ore lavoro si/no	10.2	2.8	*
Lavoro precedente si/no	13.1	2.9	
Corso professionale	15.1	3.3	
Condizione	10.1	3.4	*
Ufficio collocamento		3.8	
Motivo non ricerca	14.8	4.2	
Quando ultima azione	14.7	4.8	
Cerca attivamente lavoro	14.1	5.5	*
Istruzione	8.	6.7	

Nella stessa tavola sono segnalate con un asterisco le caratteristiche utilizzate per la definizione dell'occupazione «stimata».

Di particolare interesse, nella Tavola 5.A.6, sono i risultati relativi a tale variabile, indicata come *aggregati*, che rappresenta la stima più importante fornita dall'indagine. Essa risulta modificata nel 2.3% dei casi; questo valore deriva dalle correzioni subite dalle caratteristiche rilevate ed usate nella definizione operativa, i cui tassi di modificazioni si presentano molto differenziati: da un minimo dello 0.1% dell'età ad un massimo del 5.5% della ricerca attiva di lavoro.

Analisi per modalità

I tassi lordi di modificazioni della Tavola 5.A.6 possono essere studiati più analiticamente mediante matrici di flusso che conteggiano, a livello di modalità della singola variabile il numero di correzioni apportate dal programma; a titolo di esempio si riportano i dati relativi alla variabile AGGREGATI.

Ad un primo esame, le distribuzioni marginali, derivate dai dati grezzi e da quelli puliti, non presentano rilevanti differenze (Tavola 5.A.7, colonna 1 e 2); ma già il loro rapporto (colonna 3) indica un valore anomalo in corrispondenza della modalità *disoccupati*.

Tavola 5.A.7 - Distribuzioni marginali della variabile AGGREGATI

AGGREGATI	Dati grezzi (1)	Dati elaborati (2)	(2)/(1)*100
Non fdl	60.31	59.92	99.35
Occupati	28.96	29.30	101.17
Sottoccupati	1.99	2.19	110.05
Disoccupati	1.77	1.16	65.54
Prima occupazione	4.53	4.56	100.66
Altri in cerca	7.36	6.87	99.34
Totale	100.00	100.00	

Le percentuali per riga della Tavola 5.A.8, che possono essere letti come flussi prima/dopo, confermano la peculiarità del trattamento subito dalla sub-popolazione *disoccupati*.

Il 36% dei disoccupati originari è stato, infatti, riclassificato in un altro aggregato alla fine del processo di *pulizia*. La medesima tavola, inoltre, mette in evidenza l'intensità e la direzione dei passaggi, tra i quali i più rilevanti risultano

«in cerca prima occupazione» → «non fdl»
«altri in cerca» → «non fdl»

Tavola 5.A.8 - Spostamenti dai dati grezzi a quelli puliti

DATI GREZZI	DATI PULITI						
	non fdl	occupati	sottoccupati	disoccupati	prima occup.	altri in cerca	Totale
Non fdl	98.59	0.89	0.11	0.04	0.13	0.24	100.0
Occupati	0.27	99.29	0.35	—	0.02	0.05	100.0
Sottoccupati	—	—	99.83	—	—	0.17	100.0
Disoccupati	1.86	0.56	1.12	64.13	4.46	27.88	100.0
Prima occupaz.	2.91	0.15	0.07		96.87	—	100.0
Altri in cerca	8.96	—	0.27	—	—	91.03	100.0

La Tavola 5.A.9 completa il quadro, mostrando la provenienza degli aggregati definitivi dalle subpopolazioni originarie.

Tavola 5.A.9 - Variabile AGGREGATI - Derivazione degli aggregati del file «pulito»

DATI GREZZI	DATI PULITI					
	non fdl	occupati	sottoccupati	disoccupati	prima occup.	altri in cerca
Non fdl	99.24	1.82	3.01	1.99	1.74	5.16
Occupati	0.13	98.11	4.66	—	0.14	0.57
Sottoccupati	—	—	90.98	—	—	0.11
Disoccupati	0.06	0.03	0.90	98.01	1.74	17.22
Prima occupaz.	0.22	0.02	0.15	—	96.38	—
Altri in cerca	0.35	0.30	—	—	—	76.92
Totale	100.0	100.0	100.0	100.0	100.0	100.0

L'analisi delle matrici di flusso può dare indicazioni sull'*efficienza* del questionario; infatti, l'accumularsi di correzioni in

particolari quesiti od in particolari modalità, può essere indice di inadeguatezza dello strumento di misura.

4. Schemi di tavole di controllo per la fase di revisione

A) *Controllo quantitativo delle unità strati, comuni, aree, rilevatori e questionari.*

Tavola 1
- controllo della stratificazione teorica e quella presente sul file;
- per ciascuno strato riportare i codici dei singoli comuni ed il loro numero totale desunto dal piano teorico e dal file.

Tavola 2
- controllo dei codici di questionario presenti nel file;
- per ciascun comune riportare il numero dei codici doppi, il numero progressivo minimo e massimo, gli eventuali salti di numerazione nei codici progressivi.

Tavola 3
- controllo del numero di modelli, numero di aree e numero di rilevatori presenti nel file, nel piano teorico e nei documenti di rilevazione;
- per ciascun comune riportare il numero complessivo delle aree, dei rilevatori e dei questionari distintamente per il piano teorico, i documenti di rilevazione ed il file; inoltre, dai documenti di rilevazione riportare il numero delle mancate interviste e delle sostituzioni e dal file il numero delle sostituzioni.

Tavola 4
- controllo dei codici di area e dei rilevatori e dei relativi modelli presenti nel file;
- per ciascun comune riportare per ciascuna area il codice ed i relativi modelli, per ciascun rilevatore il codice ed i relativi modelli.

Tavola 5
- controllo, nel file, dell'abbinamento codice di area e di rilevatore e dei relativi modelli;
- per ciascun comune riportare il codice di area, i codici dei rilevatori ed i relativi modelli.

Tavola 6
- controllo dei tipi-record non previsti nel file;
- per ciascun comune riportare il codice del modello ed il relativo numero del record con il tipo-record non previsto.

B) *Controllo quantitativo delle unità di analisi*

Tavola 7
- controllo del numero dei tipi-record nel file;
- per ciascun comune riportare il numero dei record riscontrati per ciascun tipo-record ed il loro totale.

Tavola 8
- controllo dell'eguaglianza tra le variabili di conteggio ed il numero di unità di analisi presenti nel file;
- per ciascun comune riportare il codice del modello errato, il contenuto della variabile di conto, il relativo numero dei record riscontrati, il valore massimo assunto nella loro numerazione progressiva ed il numero degli eventuali salti e/o doppioni in tale numerazione.

Tavola 9
- controllo dei casi in cui ad un indicatore di presenza non corrisponde una unità di analisi;
- per ciascun comune riportare il codice del modello errato ed il numero dei record non presenti, per ogni tipo-record.

C) *Controllo qualitativo*

Tavola 10
- variabili quantitative;
- riportare per ciascuna variabile la media, il coefficiente di variazione, il minimo, il massimo (calcolati sia includendo che escludendo lo zero dal calcolo) e le percentuali relative ai valori non numerici ed agli zeri riscontrati;.

D) *Controllo dei piani di compatibilità e correzione*

Tavola 12
- controllo del numero di imputazioni e del numero dei record imputati;
- per la prescelta unità di analisi riportare il totale delle correzioni (a), il totale dei record con almeno una correzione (b), il totale dei record (c) ed i rapporti (b)/(c) e (a)/(b).

Tavola 13
- — controllo della distribuzione dei record secondo il numero di imputazioni;
- — per la prescelta (in funzione degli obiettivi) unità di analisi riportare la distribuzione delle imputazioni subite dai record ed il numero totale degli stessi.

Tavola 14
- — controllo del numero di imputazioni e dei record imputati per tipo di regola;
- — per la prescelta unità di analisi riportare per ciascuna regola il numero delle correzioni dovute a ciascuna di esse ed il totale dei record.

Tavola 15
- — controllo del numero di imputazioni per variabile;
- — per la prescelta unità di analisi riportare per ciascuna delle variabili il numero di correzioni subite.

RIFERIMENTI BIBLIOGRAFICI

Lavori di carattere teorico

ABBATE C., BOVE G., CRESCENZI F. (1990), *Metodi statistici multivariati per la ricostruzione dell'informazione mancante*, Relazione al convegno «Avanzamenti metodologici e statistiche ufficiali», Roma 13-14 dicembre 1990 - ISTAT.

FELLEGI I.P., HOLT D. (1976), *A systematic approach to automatic editing & imputation*, J.A.S.A.

GARCIA RUBIO, GOMEZ ALFONSO, VILLAN (1983), *Desarollo de un sistema de detection y imputation automatica basado en la metodologia de Fellegi-Holt ampliada* Atti I.S.I.

GRANQUIST L. (1987), *On the need for generalized numeric and imputation system*, U.N. - CES, documento CES/SEM. 23/R. 10, Seminar on statistical methodology.

KREWSKI D., PLATEK R., RAO J.N.K. (1981), *Current topics in survey sampling*, Academic Press, New York.

MARCHETTI E. (1986), *Large sample models for editing response errors*, documento interno ISTAT.

MASSELLI M., MARCHETTI E. (1984), *I piani di compatibilità ed il controllo dell'attendibilità del dato*, Atti della XXXII Riunione Scientifica della S.I.S., Sorrento.

MASSELLI M. (1990), *Un modello per l'individuazione della sequenza di regole e variabili in un piano di compatibilità di tipo deterministico*, documento interno ISTAT.

NATIONAL CENTRAL BUREAU OF STATISTICS (NCBS Sweden) (1983), *On generalized Editing Programs and the solution of the data quality problems*, manoscritto non pubblicato.

PULLUM T.W., HARPHAM T., OZSEVER N. (1986) *The machine editing of large sample surveys: the experience of the World Fertility Survey*, International Statistical Review 54, 3, pp 311-326.

U.N.D.P. - CES Statistical Computing Project (1984), *Description and feature analyses of the selected data editing software systems*, documento SCP/DA/WP. 76.

Sperimentazioni sulle indagini Istat

CARIANI G. (1983), *I controlli ED del censimento demografico* in Atti del Convegno della S.I.S., Trieste.

MASSELLI M. (1986), *Valutazione dei piani di compatibilità e correzione automatici. Una sperimentazione*, Atti della XXXIV Riunione Scientifica della S.I.S., Bari.

MASSELLI M. (1987), *La procedura di controllo degli effetti del piano di compatibilità dell'indagine forze di lavoro*, documento interno Istat.

MASSELLI M. (1987), *La qualità dei dati nell'indagine Istat sulla salute 1983*, in Atti del Convegno «Salute e ricorso ai servizi nel Veneto», Padova novembre 1987.

PANIZON F., SIGNORE M. (1987), *Analisi dell'effetto dei piani di compatibilità dell'indagine forze di lavoro con accoppiamento statistico dei record*, documento interno Istat.

CAPITOLO 6 - L'ELABORAZIONE FINALE E L'ANALISI DEI RISULTATI

1. I controlli nella fase di elaborazione e di validazione dei risultati

I dati elementari, corretti nella fase di revisione quantitativa e qualitativa, vengono infine elaborati, in funzione degli obiettivi prefissati; le elaborazioni, quindi, possono risultare di vario tipo e di differente complessità: tavole, indicatori, archivi, campioni per gli utenti, analisi di secondo livello ecc..

Poiché, attualmente, il prodotto-tipo di una indagine è rappresentato dal *piano di tabulazione dei risultati*, che consiste nell'elaborazione di tabelle semplici e a più entrate contenenti frequenze relative ed assolute, totali, medie ed altri indici descrittivi, tale prodotto standard sarà l'oggetto delle considerazioni che seguono.

Il controllo da effettuare nella fase di elaborazione e validazione dei risultati può essere distinto in un controllo formale, relativo ai possibili errori generati nella specificazione analitica del piano di tabulazione (ovvero al *come* vengono prodotti i risultati), ed in un controllo sostanziale, relativo alla rilevanza dell'informazione fornita (ovvero a *quali* risultati vengono prodotti).

I controlli di tipo formale hanno l'obiettivo di prevenire e di identificare eventuali errori di «quadratura» all'interno della singola tavola o tra tavole diverse; scopo dei controlli di tipo sostanziale è, invece, la verifica dell'uso dell'informazione rilevata, la valutazione dei risultati sotto il profilo della *plausibilità* rispetto alla realtà esaminata, l'eventuale integrazione del piano di tabulazione con tavole relative ad aspetti non previsti o non considerati ex ante.

Riassumendo, nella fase di elaborazione e validazione dei risultati, è necessaria:

la verifica formale:

— della specificazione dei parametri necessari per la costruzione delle singole tavole;
— della coerenza tra i dati relativi al medesimo fenomeno contenuti in tavole diverse;
— del dizionario nel caso di programmi generalizzati.

e la verifica sostanziale:

— della selezione e della complessità delle tavole;
— della validità dei risultati.

I controlli di quadratura delle tavole

Il controllo necessario per garantire la quadratura della singola tavola o tra tavole diverse, dipende dalla procedura utilizzata per la predisposizione di un piano di tabulazione; questi, infatti, può essere generato, in funzione degli strumenti informatici disponibili:

a) definendo dapprima un *dizionario* contenente tutte le variabili coinvolte nel piano (sia quelle originarie che quelle derivate) ed i relativi riferimenti ai record di elaborazione, ed utilizzando poi le variabili così definite nel processo di tabulazione (tale tecnica implica la disponibilità di programmi generalizzati).
b) definendo di volta in volta, per ciascuna tavola, gli elementi necessari per la sua elaborazione.

Nel primo caso la possibilità di generare errori è confinata alla predisposizione del dizionario; un controllo accurato delle variabili ivi contenute (definizione e corrispondenza con il piano di registrazione), garantisce l'impossibilità di errori nel processo di tabulazione e rende superflue altre verifiche.

Nel secondo caso, invece, ogni tavola, per vizio logico o per banali sviste, può essere fonte di errore, essendo stata materialmente definita in modo indipendente dalle altre; è quindi necessario sottoporre a verifica sia la singola tavola che il singolo aggregato presente in tavole diverse.

Il controllo della singola tavola

Elaborare una tavola statistica equivale a raggruppare le unità di analisi in subpopolazioni, caratterizzate dalle modalità delle variabili indicate in testata ed in fiancata, ed a calcolare l'indicatore statistico d'interesse per ciascuna delle suddette subpopolazioni.

Gli elementi che caratterizzano la definizione di una tavola statistica sono pertanto le *variabili di classificazione*, il *parametro statistico di analisi* e le *corrispondenze* tra le variabili coinvolte nel calcolo ed i campi del record del file sottoposto ad elaborazione. Ad esempio per elaborare la tavola riportante la spesa media per famiglia, per regione e classi di età del capofamiglia è necessario definire l'indicatore statistico d'interesse (la media), le variabili di classificazione (la regione, la classe d'età e la relazione di parentela) ed i riferimenti sul record delle variabili coinvolte (spesa, età, regione, relazione di parentela).

Come caratteristiche di classificazione possono essere utilizzate sia le variabili originarie (nell'esempio il codice regionale presente sul record) sia le variabili da queste derivate (le classi

di età del capofamiglia provenienti dall'informazione anno di nascita e relazione di parentela).

Possiamo considerare quali elementi costitutivi del parametro di analisi, la *funzione* dei dati elementari da utilizzare (in genere totali, medie, frequenze relative ed assolute) e l'*argomento* della detta funzione, ovvero le variabili coinvolte nel calcolo; nell'esempio precedente, la funzione media ponderata (con i coefficienti di riporto dell'universo) della variabile spesa per consumi.

Nel predisporre le tavole, gli errori possono essere generati da una insufficiente od errata specificazione dei suddetti elementi, in particolare dalle definizioni di variabili derivate, spesso complesse per l'uso congiunto di condizioni logiche AND, OR e NOT.

Il controllo può essere attuato preventivamente mediante l'analisi logica dei vari elementi delle specifiche (seguendo ad esempio la distinzione sopra riportata) e a posteriori mediante la verifica della quadratura delle singole tavole.

Il controllo tra tavole

Generalmente, in un piano di tabulazione, un parametro relativo ad una data subpopolazione (ad esempio il numero di occupati maschi di una determinata regione) compare più volte o può essere ricalcolata da tavole diverse. Se il piano di tabulazione non è generato mediante un dizionario, errori di definizione in una o più tavole comportano valori differenti del medesimo parametro; il controllo, allora, consiste nella coincidenza di tali valori nelle diverse tavole in cui sono contenuti.

A tale scopo è opportuno predisporre, insieme al piano di tabulazione, un elenco analitico delle diverse subpopolazioni con le tavole in cui compaiono i relativi dati e le relazioni tra tali tavole; ad esempio la consistenza della popolazione attiva nelle ripartizioni territoriali contenuta in tavole diverse, deve sommare al complesso Italia presente in altra tavola. L'elenco costituirà una guida per il confronto dei dati contenuti nelle varie tavole di spoglio: a titolo di esempio, anche se costruito con diversi intendimenti, si cita il «Riepilogo delle tavole di spoglio» nelle pubblicazioni del Censimento della popolazione 1981.

Se la fase di elaborazione prevede degli archivi di *riepilogo*, è possibile informatizzare la procedura di controllo, traducendo le relazioni tra subpopolazioni, in operazioni sugli elementi del file.

La selezione delle tavole

Il piano di tabulazione rappresenta lo strumento principale mediante il quale i risultati della rilevazione vengono messi a disposizione dell'utente; la scelta delle tavole, quindi, deriva dallo schema concettuale che ha guidato la definizione degli obiettivi

dell'indagine e dovrebbe essere effettuata nella fase di programmazione.

Come guida alla selezione, possono essere utilizzati gli schemi Entità Relazioni, oppure la specificazione per aree d'interesse degli obiettivi della rilevazione (cfr. Capitolo 2). Un criterio che può risultare utile in tale operazione, è la distinzione tra tavole che hanno lo scopo di dare informazioni *di livello* (analisi primaria) da quelle il cui obiettivo sono le *relazioni* tra variabili (analisi secondaria); in genere, le prime risultano definite dalle stime principali dell'indagine e dai domini di studio, stabiliti in sede di programmazione.

Tuttavia, il piano così programmato risponde ad uno schema concettuale basato sulle conoscenze a priori della realtà e può quindi accadere che sia inappropriato a descrivere i fattori rilevati od emergenti del fenomeno, a causa di lacune nello schema o di cambiamenti effettivamente sopravvenuti.

Il rispetto del piano programmato potrebbe, perciò, portare ad una sottoutilizzazione del contenuto informativo od, anche, ad una rappresentanzione distorta della realtà, ovvero ad errori di *rilevanza*. È quindi opportuno analizzare i dati per individuare le eventuali tavole aggiuntive, o sostitutive, a quelle programmate, che rappresentano aspetti emergenti o non considerati ex ante.

Tale ricerca può essere basata sull'analisi di un insieme di tavole più vasto di quello programmato per la pubblicazione (come usualmente accade); tuttavia tale metodo implica una notevole mole di lavoro e un elevato livello di soggettività. Un criterio più oggettivo e meno dispendioso consiste nell'utilizzare modelli statistici e tecniche multivariate di analisi per esplorare l'insieme dei dati ed individuare le relazioni tra le variabili maggiormente esplicative.

La definizione del piano di tavole è, infine, vincolato da due fattori: da un lato la comparabilità nel tempo e nello spazio (ad esempio nel caso di indagini ripetute o di confronti internazionali) e dall'altro l'opportunità di integrare i risultati prodotti con quelli del sistema informativo cui l'indagine appartiene.

La validazione dei risultati

La prima verifica della plausibilità dei risultati ottenuti, rispetto alla realtà esaminata, è costituita dalle informazioni derivanti dal sistema di controllo. Come è stato già ricordato, gli errori riscontrati nelle diverse fasi non rappresentano solo fattori di disturbo, ma costituiscono anche «informazione» sui fenomeni indagati. Essi, pertanto, possono aiutare ad individuare i limiti dell'analisi effettuata, gli eventuali fenomeni emergenti, nuove o diverse interpretazioni della realtà.

A tale scopo dovrebbero essere utilizzate le analisi che sono state suggerite nei capitoli precedenti come approfondimenti per fase o per tipologia di errore. Ad esempio l'analisi delle mancate risposte totali, può segnalare particolari subpopolazioni sfuggite alla rilevazione o suggerire la presenza di effetti distorsivi nei risultati finali; dalle mancate risposte parziali e dalle correzioni effettuate, possiamo invece desumere fenomeni emergenti o inadeguatezze nello schema concettuale utilizzato per l'indagine.

Oltre agli indicatori derivanti dai controlli di qualità, la validazione dei risultati deve essere basata su fonti esterne alla singola rilevazione:

I) la serie storica dei dati dell'indagine;
II) le informazioni provenienti da altre indagini od altre fonti.

Anche in questo caso è opportuno formalizzare il controllo, inquadrando l'indagine nel contesto più ampio del sistema informativo di pertinenza e precostituendo un archivio delle fonti disponibili, dei relativi risultati e modalità di rilevazione.

La descrizione sintetica del piano di tabulazione

La descrizione della struttura del piano di tabulazione, mediante alcuni parametri quantitativi, ha l'obiettivo di controllare se il dettaglio ed il diverso peso assunto dalle differenti variabili ed unità di analisi, corrisponde effettivamente alle priorità programmate.

Gli indicatori dell'utilizzazione dell'informazione rilevata che verranno consigliati, non devono essere interpretati come indicatori di errore, ma costituiscono dei parametri sintetici che possono contribuire a verificare la completezza, l'organicità e l'equilibrio dei risultati pubblicati.

I parametri descrittivi del piano di tabulazione fanno riferimento (I) all'utilizzazione delle variabili e delle unità rilevate e (II) al numero ed al tipo delle tavole pubblicate.

Le variabili oggetto di pubblicazione possono non coincidere con quelle rilevate; alcune caratteristiche, infatti, possono comparire nel questionario al solo scopo di fornire gli elementi per il calcolo di variabili *derivate*, non direttamente rilevate o rilevabili. Come caratteristiche rilevate si considerano, inoltre, i codici identificativi significativi (ad esempio i codici geografici). La distinzione, secondo l'utilizzazione, delle variabili complessivamente presenti nel questionario e nelle tavole, costituisce un primo insieme di parametri (Prospetto 6.1).

Il numero e le dimensioni delle tavole pubblicate rappresentano il grado di analiticità dei dati forniti e, quindi, costituisco-

Prospetto 6.1 - Parametri descrittivi di un piano di tavole

PARAMETRI	
numero della variabili presenti nel questionario V_q	K_1
numero delle variabili V_q utilizzate nelle tavole	K_2
numero delle variabili V_q utilizzate per derivarne altre	K_9
numero delle variabili derivate	K_4
numero delle variabili V_q non utilizzate	K_4

no un indicatore, anche se indiretto e approssimato, dell'informazione a disposizione degli utenti.

Indicando, inoltre, con V_i^m il numero delle ricorrenze della variabile i-esima nell'insieme di tavole a m dimensioni e con $V_i = \Sigma_m V_i^m$ il numero totale delle ricorrenze, si possono ottenere due tassi di utilizzazione della singola variabile, il primo rispetto al totale delle tavole, il secondo, più analitico, relativo al numero di tavole di dimensione m:

$$V_i / \Sigma_i V_i$$

$$V_i^m / \Sigma_i V_i^m \qquad (6.1)$$

Se i risultati pubblicati si riferiscono a più unità statistiche (ad esempio famiglie ed individui), il piano di tabulazione può essere esaminato rispetto al peso relativo assegnato a ciascuna di esse mediante i rapporti:

$$T_u / T \quad ; \quad T_u^m / T^m \qquad (6.2)$$

dove T rappresenta il numero di tavole prodotte, u la generica unità statistica e m la dimensione delle tavole.

Un ulteriore parametro descrittivo può essere ottenuto, calcolando il rapporto tra numero di tavole a m dimensioni effettivamente pubblicate ed il massimo numero di tavole producibili dalle k variabili, presenti nel piano di tabulazione:

$$T^m / \binom{k}{m} \qquad (6.3)$$

Se si assume che uno degli obiettivi di un piano di tavole è fornire un riassunto sintetico del contenuto informativo del file, un rapporto (6.3) molto elevato può indicare una selezione non particolarmente accurata delle tavole.

Infine, è possibile che in un insieme di tavole particolarmente complesso, alcune tavole risultino ridondanti, perché l'informazine da esse contenute è presente in tabelle di dimensione superiore; ad esempio la distribuzione marginale di una variabile compare necessariamente in tavole di ordine superiore che coinvolgono la detta variabile.

Le tavole ridondanti rappresentano una perdita di efficienza del piano di tabulazione; tuttavia, a volte, è preferibile la loro presenza per renderne più semplice la consultazione.

RIFERIMENTI BIBLIOGRAFICI

Di Ciaccio A., Sabbadini L.L. (1990), *Presentazione del contenuto informativo di un'indagine complessa: selezione di tabelle di contingenza in un approccio multivariato*, Relazione al convegno «Avanzamenti metodologici e statistiche ufficiali» Roma, 13-14 dicembre 1990 - Istat.

Volle M. (1985), *Analyse des donnes*, 3_{eme} edition ESA Parigi 1985.

CAPITOLO 7 - LA STIMA DELL'ERRORE GLOBALE DI MISURA

1. Descrizione dell'errore di misura

Con l'espressione *errori di misura* o *errori di risposta* si indicano tutti gli errori che sorgono nella fase della raccolta dei dati e per effetto dei quali si osserva un valore diverso da quello che si intendeva misurare.

A differenza degli errori di campionamento, che sono dovuti al fatto che viene rilevata solo una parte della popolazione oggetto di studio, gli errori di risposta, derivando dal processo di misurazione adottato, possono verificarsi anche quando si osservano tutte le unità della popolazione. Di conseguenza il problema degli errori di misura riguarda sia le indagini campionarie sia quelle totali.

Con riferimento ad indagini che hanno come unità di analisi la famiglia o l'individuo, gli errori di misura possono, in linea generale, essere dovuti:

alla predisposizione del questionario, ovvero:

- alla formulazione delle domande;
- alla sequenza delle domande;
- alla lunghezza del questionario;
- alla scelta delle classificazioni;

al rispondente, in particolare:

- a problemi di memoria;
- alla mancanza di informazione;
- alla scarsa motivazione a rispondere attentamente;
- al fraintendimento di alcune domande;
- ad errori accidentali;
- ad errori volontari;
- a problemi di condizionamento (dovuti alla presenza di altre persone);
- all'effetto «proxy» (quando cioè l'intervistato risponde per altre persone);

al criterio di raccolta adottato, cioè:

- autocompilazione del questionario;
- intervista diretta;
- intervista telefonica;

al rilevatore (ogni volta che è presente), quindi:

- al grado di preparazione sul questionario (conoscenza e comprensione dei quesiti);
- al grado di preparazione sulla conduzione dell'intervista (comunicazione, partecipazione e non influenza del rispondente);
- ad errori di compilazione del questionario;

alla codifica, quindi:

- alla completezza del sistema di codifica;
- ad errori accidentali;
- all'inosservanza delle norme;

al supervisore, ovvero:

- a carenze nelle istruzioni e nel controllo degli intervistatori.

Gli errori di risposta quindi possono sorgere in maniera accidentale o sistematica, essere introdotti volontariamente o derivare da una mancanza di informazione.

È chiaro che le possibili fonti d'errore elencate possono interagire e combinarsi in maniera diversa tra loro e che non sempre è possibile tenere separati i diversi effetti in fase di analisi e di valutazione. In effetti il questionario, il rispondente, il criterio di raccolta, il rilevatore e le loro interazioni costituiscono il processo di misurazione stesso e di conseguenza contribuiscono nel loro insieme a determinare *l'errore di risposta globale*. Inoltre l'errore di misura stimato utilizzando i dati finali risulta influenzato anche dagli errori di registrazione e dall'effetto dei piani di compatibilità. Pertanto la distinzione operata è più logica che effettiva ed è stata adottata a scopo esemplificativo.

Tuttavia, come le sperimentazioni condotte in diversi Paesi hanno mostrato, i contributi più determinanti all'errore di misura complessivo, provengono dal rilevatore e dal rispondente e quindi su di essi sarà focalizzata l'attenzione nei prossimi paragrafi.

Gli effetti degli errori di risposta sono:

- l'introduzione di una *distorsione* nelle stime finali;
- l'aumento della *variabilità* delle stime finali.

La distorsione e la variabilità dovute agli errori di misura saranno esplicitate formalmente nel paragrafo 3 mentre i metodi di stima di tali effetti saranno descritti nei paragrafi 5 e 6.

In particolare si dimostra che l'effetto della distorsione è costante rispetto al numero delle osservazioni effettuate; quindi un censimento presenta la stessa distorsione di un'indagine campionaria, se svolto nelle stesse condizioni *essenziali*.

Viceversa la varianza di risposta è inversamente correlata con la numerosità campionaria e può quindi essere diminuita aumentando la dimensione del campione, così come è possibile ridurre la variabilità dovuta all'intervistatore impiegandone un numero maggiore. Questi risultati sono validi per indagini svolte nelle stesse condizioni generali; è però presumibile che ad un incremento del numero di osservazioni possano corrispondere maggiori difficoltà, soprattutto organizzative, per il controllo delle varie fasi della rilevazione, con conseguente perdita di accuratezza dei risultati.

Per quanto concerne la stima dei suddetti effetti, notiamo che per misurare la distorsione è necessario disporre di dati da una fonte esterna all'indagine, mentre le componenti della varianza di risposta possono essere stimate a partire dalle osservazioni campionarie.

Le conseguenze degli errori di risposta, come è stato ampiamente dimostrato, possono risultare superiori a quelle prodotte dagli errori di campionamento e comunque non sono mai di entità trascurabile.

Ne deriva la necessità di sperimentare e di mettere a regime alcune procedure di controllo degli errori di misura. Questi controlli devono essere sia preventivi, nel senso di ricercare quelle tecniche (disegno campionario, criterio di raccolta, ecc.) che minimizzano gli errori di risposta, sia di stima a posteriori allo scopo di quantificarne gli effetti. Infatti l'informazione sugli errori di misura è indispensabile sia al produttore di dati che ne voglia migliorare la qualità, sia all'utilizzatore che deve conoscere il livello di precisione delle stime fornitegli.

Nell'affrontare la problematica connessa allo studio degli errori di misura Cochran ha individuato quattro aspetti principali (Tenenbein, 1984):

- i tipi di modelli matematici usati per rappresentare gli errori di misura;
- la misura in cui gli errori di risposta sono automaticamente presi in considerazione dalle tecniche standard di analisi e la misura in cui questi metodi diventano fuorvianti se certi tipi di errore sono presenti;
- il danno provocato dagli errori di misura nel produrre distorsioni o nel diminuire la precisione delle stime e le procedure disponibili per ridurre tali conseguenze indesiderate;
- le tecniche per lo studio degli errori di misura.

2. Quadro concettuale di riferimento

Valore vero individuale

Alla base del modello presentato vi è la definizione di *valore vero individuale*.

Questo significa che ad ogni individuo della popolazione corrisponde un *valore vero* di una variabile oggetto di studio: una media o un aggregato di tali valori veri individuali costituisce il *valore da stimare mediante l'indagine*.

Il valore vero individuale viene concepito come una caratteristica propria dell'unità di analisi ed è quindi indipendente dalle condizioni in cui si effettua l'indagine che invece influenzano la risposta individuale.

Consideriamo, come esempio, la variabile età. Solitamente l'età viene definita come l'intervallo di tempo che intercorre tra due eventi. In base a questa definizione risulta chiaro che ad ogni individuo corrisponde un'età *vera* e che tale valore individuale non dipende dal criterio scelto per determinarlo. Tuttavia ciò non assicura che la risposta individuale che si ottiene, ad esempio chiedendo ad una persona l'età, sia il valore vero di tale variabile così come è stata definita. Infatti un individuo può non conoscere la propria età, può mentire oppure essersi confuso per problemi di memoria.

La definizione del valore vero individuale può essere in alcuni casi piuttosto complicata. Si pensi ad esempio all'intelligenza: come definire l'intelligenza vera di una persona?

Hansen, Hurwitz e Madow (1953) hanno affrontato per primi il problema e hanno indicato tre criteri per la definizione del valore vero:

- il valore vero deve essere univocamente definito;
- il valore vero deve essere definito in maniera tale da soddisfare gli obiettivi dell'indagine;
- quando non è in contrasto con i primi due criteri, il valore vero deve essere definito in termini di operazioni che possono essere effettivamente eseguite (anche se ciò può risultare difficile o costoso).

Nelle situazioni pratiche può accadere che i criteri esposti siano in conflitto tra loro e richiedano una scelta o un compromesso tra i tre. Occorre però tenere presente che i primi due criteri sono essenziali, mentre il terzo, pur essendo utile, non lo è.

La definizione del valore vero in termini di operazioni eseguibili permette di eliminare o di rendere trascurabili gli errori di misura. Tuttavia basarsi solo su criteri operazionali può allontanare dagli obiettivi della ricerca e può non portare ad una definizione unica.

Consideriamo il seguente esempio in cui il luogo di nascita di una persona viene definito come la risposta trascritta dall'intervistatore alla domanda: «In quale città o paese è nato?».

Questa è una definizione in termini operazionali che però non può essere accettata come valore vero. Infatti si è interessati a conoscere dove una persona è effettivamente nata e non la risposta che è stata data ad una domanda e che può risultare alterata per effetto di un insieme di fattori descritti in precedenza.

In ogni caso trascurare il terzo criterio può far aumentare sensibilmente la distorsione causata dagli errori di risposta.

Anche quando il valore vero viene definito con precisione possono, però, sorgere notevoli difficoltà per determinarlo. Tali difficoltà sono strettamente connesse al tipo di variabile che si intende misurare. Infatti è presumibile che un'indagine riesca a cogliere il valore vero per una larga proporzione di individui per variabili quali l'età o il sesso, mentre per altre, come ad esempio il reddito, ciò sarà possibile solo in misura molto minore.

Valore di risposta atteso

Quando non è possibile definire il valore vero in modo tale da soddisfare i tre criteri sopra esposti, si può dare una definizione che soddisfi i primi due requisiti e definire una operazione il cui *valore atteso* approssimi in maniera soddisfacente il valore vero.

Questo porta al concetto di un *campione* di risposte da un insieme di possibili misurazioni.

Infatti l'indagine viene considerata *concettualmente ripetibile* e le ripetizioni dell'indagine *indipendenti* tra di loro. Questo significa che le ripetizioni si considerano riferite allo stesso istante o intervallo di tempo e che l'esecuzione dell'operazione non influenza i risultati successivi. I risultati particolari osservati in una indagine sono considerati come i risultati di una prova.

In particolare si suppone di intervistare ogni individuo un gran numero di volte sotto le *stesse* condizioni. Questa operazione genera una popolazione di risposte per tutti gli individui. Allora si può pensare di estrarre un campione di individui e quindi un campione costituito da una delle risposte possibili per ciascun individuo. Sotto queste ipotesi il valore atteso di una stima ottenuta da un campione di possibili risposte può essere considerato come una approssimazione del valore vero.

In questo modo la risposta individuale viene considerata come una variabile aleatoria e l'insieme dei valori che ognuna di esse può assumere costituisce l'universo delle risposte individuali. Tale universo deve essere caratterizzato con maggiore pre-

cisione: in realtà si farà riferimento all'insieme delle risposte ottenibili sotto certe condizioni che chiameremo *essenziali*.

Condizioni essenziali di un'indagine

Queste condizioni sono specificate dal disegno dell'indagine e quindi vengono determinate nella fase di progettazione dell'indagine stessa quando si stabiliscono, ad esempio, l'oggetto di analisi e il criterio di raccolta delle informazioni.

Dal punto di vista del controllo degli errori di misura, si è interessati alle condizioni sotto le quali si svolge l'indagine in quanto esse caratterizzano la situazione nella quale si ottiene la risposta individuale e quindi esercitano un'influenza su di essa. Più in particolare, la distorsione e la varianza degli errori di misura possono considerarsi determinate da tali condizioni. Quindi è possibile ridurre la variabilità delle risposte individuali attraverso le specificazioni dell'indagine anche se è impossibile eliminarla completamente.

Allora si definiscono condizioni *essenziali* di un'indagine quelle variabili che si cerca di mantenere costanti per tutti i casi in esame, cioè le condizioni che si cerca di tenere *sotto controllo* attraverso l'introduzione di regole uniformi e di opportune procedure.

Errore di misura individuale

Si definisce l'*errore di misura individuale* come la differenza tra il valore osservato in una particolare indagine e il valore vero dell'individuo.

L'errore di misura individuale, come il valore osservato, viene concepito come una variabile aleatoria con una sua distribuzione di probabilità. Quindi l'errore di risposta di un particolare individuo in una particolare indagine avrà un valore atteso, che costituisce la *distorsione di risposta individuale* e una componente variabile intorno a questo valore, denominata *deviazione di risposta individuale*.

La deviazione di risposta misura la differenza tra il valore atteso individuale e il valore vero, e di conseguenza l'influenza delle condizioni essenziali, che caratterizzano l'indagine, sui risultati osservati. Essa è quindi funzione, ad esempio, del criterio di raccolta dei dati adottato e del tipo di intervistatori scelti.

Le fluttuazioni dell'errore intorno al suo valor medio, invece, sono imputabili alle condizioni particolari in cui si è effettuata l'osservazione, ad esempio, ai singoli intervistatori scelti.

Allo stesso modo, una media o un aggregato di un insieme di risposte di individui differenti saranno affetti da una *distorsione di risposta* e da una *varianza di risposta* determinate dalle distorsioni e dalle varianze individuali.

Oltre agli effetti sopra menzionati, può esistere una *correlazione* tra gli errori di risposta individuali relativi a persone diverse. Questo accade, ad esempio, nel caso di indagini con intervista diretta, nelle quali il rilevatore può influenzare le risposte degli individui da lui intervistati.

3. Un modello matematico per lo studio degli errori di misura

La formalizzazione del problema mediante un modello matematico permette di evidenziare le conseguenze della presenza di errori di misura sulle stime finali e di mettere a punto delle procedure specifiche per quantificarne gli effetti. Come già richiamato nel paragrafo 1, il modello è utile per stimare l'errore globale di risposta e per evidenziare l'influenza di alcune fonti specifiche poste sotto controllo quali, ad esempio, il rilevatore e il rispondente, ma non consente di scindere l'errore totale nelle singole componenti che lo hanno generato.

Il modello che viene descritto è il più noto ed utilizzato per lo studio degli errori di misura; è stato introdotto da Hansen, Hurwitz e Bershad (1961) ed è stato poi ripreso ed applicato da differenti autori, tra i quali Cochran (1977) e Fellegi (1963), (1964) e (1974), e dal U.S. Bureau of the Census che lo ha adottato per la valutazione degli errori di misura nel censimento della popolazione (cfr. U.S. Bureau of the Census, 1969) e per studi specifici sull'effetto rilevatore.

Nel descrivere il modello si farà riferimento ad una indagine campionaria, tuttavia i risultati ottenuti possono essere facilmente estesi al caso di indagini totali.

Allo scopo di non appesantire eccessivamente la trattazione, si rimanda alla bibliografia per le dimostrazioni dei risultati presentati.

Componenti dell'errore di misura individuale

Le ipotesi che stanno alla base di questa formulazione matematica sono state dettagliatamente discusse nel paragrafo precedente comunque è necessario richiamarle brevemente per introdurre un'adeguata simbologia.

Ipotesi del modello:

- l'indagine è ripetibile sotto le stesse condizioni essenziali;
- le replicazioni del processo di misurazione sono tra loro indipendenti;
- esiste un valore vero individuale che indicheremo con μ_{ij};
- esiste un valore osservato per l'i-esimo individuo nella t-esima replicazione che indicheremo con y_{it};

- l'errore di misura per l'i-esimo individuo nella t-esima replicazione sarà indicato con e_{it}.

Sotto queste assunzioni, si ha:

$$y_{it} = \mu_i + e_{it} \qquad (7.1)$$

ovvero il valore osservato di una variabile per l'i-esimo individuo nella t-esima replicazione è composto dal *valore vero individuale* μ_i e da un *errore di misura individuale* e_{it}. Il valore vero individuale, (variabile da individuo a individuo secondo l'indice i), rimane costante nelle diverse replicazioni, mentre l'errore, oltre a variare da individuo a individuo, è variabile al variare della replicazione del processo di misurazione (indice t).

Di conseguenza, sotto misurazioni ripetute sulla stessa unità i, gli errori e_{it} seguiranno una certa distribuzione di frequenza. In particolare esisterà un valore medio e una variabilità intorno a quest'ultimo, espressa dalla varianza, che indichiamo rispettivamente con:

$$b_i = E(e_{it} \mid i) \qquad (7.2)$$

$$\sigma_i^2 = V(e_{it} \mid i) \qquad (7.3)$$

dove con $E(e_{it} \mid i)$ si è indicato il valore atteso al variare della replicazione del processo di misurazione sullo stesso individuo, cioè il valore atteso condizionato all'i-esima unità.

Allora b_i rappresenta la *distorsione di risposta individuale*, ovvero l'errore sistematico relativo alla i-esima unità, mentre σ_i^2 misura la dispersione intorno al valore medio individuale.

L'entità di b_i e di σ_i^2 dipende principalmente dal tipo di variabile considerata e dal processo di misurazione adottato, ma può essere influenzata da numerosi altri fattori, quali ad esempio il grado di sensibilizzazione dei rispondenti nei confronti dell'indagine.

La differenza tra l'errore di misura individuale e il suo valor medio al variare della replicazione, costituisce la componente variabile dell'errore o *deviazione di risposta individuale*, che indichiamo con il simbolo d_{it}, ovvero:

$$d_{it} = e_{it} - b_i \qquad (7.4)$$

In base alle ipotesi fatte segue che:

$$E(d_{it} \mid i) = 0 \qquad (7.5)$$

$$V(d_{it} \mid i) = E(d_{it}^2 \mid i) = \sigma_i^2 \qquad (7.6)$$

Finora è stato considerato l'errore di misura individuale, cioè relativo ad una singola unità. Occorre, però, esplicitare la relazione che intercorre tra gli errori di misura di due unità distinte.

In effetti ci può essere una correlazione tra i valori dell'errore e_{it}, ovvero tra le deviazioni di risposta d_{it}, per differenti unità appartenenti allo stesso campione. Il caso più semplice di correlazione degli errori è quello in cui esiste una distorsione dovuta all'intervistatore che si riflette su tutte le unità a lui assegnate. Di conseguenza non si può assumere l'indipendenza tra errori relativi ad individui facenti parte della medesima assegnazione, anche se può ipotizzarsi l'indipendenza tra assegnazioni differenti.

Nel seguito saranno analizzati separatamente gli effetti di errori di misura incorrelati e di errori correlati, nel senso sopra esposto, sulla stima della media di una popolazione.

A completamento del modello si assume l'esistenza di una distorsione costante, indicata con B, che agisce su tutte le unità della popolazione. Si ipotizza, quindi, che l'errore sistematico b_i non si annulli in media, cioè che non esista una compensazione degli errori di misura relativi a tutte le unità della popolazione. Da un punto di vista formale, quanto detto equivale ad assumere che:

$$E(b_i) = B$$

Dall'espressione precedente deriva l'esistenza di una componente variabile della distorsione, rappresentata dalla differenza $b_i - B$. Questa componente ha media zero, ma può risultare correlata con il valore vero μ_i; ad esempio il processo di misurazione può essere tale da sottostimare valori grandi di μ_i e da sovrastimare valori piccoli.

La conseguenza delle assunzioni fatte sull'errore di misura individuale è che, pur replicando le misurazioni sull'individuo, il valore atteso individuale non coincide con il valore vero per effetto della distorsione b_i. Infatti dalla (7.1) e dalla (7.2) si ottiene:

$$m_i = E(y_{it} \mid i) = \mu_i + b_i \qquad (7.7)$$

La quantità m_i è concettualmente il *valore atteso di risposta* relativo all'i-esimo individuo, calcolato sulle possibili ripetizioni del processo di misurazione.

Come si vede dalla (7.7), il valore atteso m_i differisce, per effetto degli errori di misura, dal valore vero μ_i di una quantità pari all'errore medio b_i.

Inoltre, sostituendo nell'espressione (7.4) le formule (7.1) e (7.7), si può esprimere la deviazione di risposta individuale, ovvero la componente variabile dell'errore, come la differenza tra il valore osservato e valore atteso:

$$d_{it} = y_{it} - m_i \qquad (7.8)$$

Nel Prospetto 7.1 si riporta uno schema riassuntivo delle componenti dell'errore di misura individuale introdotte con i relativi simboli, valori attesi e varianze.

Prospetto 7.1 - Errore di misura individuale e sue componenti

Simbolo	Natura della componente	Valore atteso e varianza
$e_{it} = y_{it} - \mu_i$	Errore di misura individuale	$E(e_{it} \mid i) = b_i$ $V(e_{it} \mid i) = \sigma_i^2$ può essere: $Cov(e_{it}, e_{jt}) \neq 0$
$d_{it} = e_{it} - b_i$ $= y_{it} - m_i$	Deviazione di risposta individuale o componente variabile dell'errore	$E(d_{it} \mid i) = 0$ $V(d_{it} \mid i) = \sigma_i^2$ può essere: $E(d_{it}\, d_{jt}) \neq 0$
$b_i = m_i - \mu_i$	Distorsione di risposta individuale	$E(b_i) = B$
B	Distorsione costante su tutte le unità	
$b_i - B$	Componente variabile della distorsione	$E(b_i - B) = 0$ può essere: $Cov(b_i - B, \mu_i) \neq 0$

Allo scopo di esaminare situazioni più complesse (ad es. l'influenza di supervisori, codificatori, ecc.), sono stati sviluppati da

Fellegi (1964 e 1974) alcuni modelli matematici a cui si rimanda per eventuali approfondimenti.

Tuttavia è opportuno segnalare che i tipi di correlazione ritenuti più comuni, in base agli studi effettuati, sono rappresentati da questo modello o possono esserlo mediante lievi modificazioni (cfr. Cochran, 1977).

Effetti degli errori di misura sulla stima della media di una popolazione

Allo scopo di esplicitare il modello e di evidenziare gli effetti degli errori di misura sulle stime finali di un'indagine, supponiamo di voler stimare la media μ di una popolazione di N elementi mediante un campione (casuale semplice) di n elementi, dove:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \mu_i$$

In base alle ipotesi finora fatte, il valore osservato per l'i-esimo individuo nella t-esima replicazione può scriversi, per la (7.8):

$$y_{it} = d_{it} + m_{it} \qquad (7.9)$$

La differenza tra il valore osservato y_{it} e la media della popolazione può essere espressa nel seguente modo:

$$y_{it} - \mu = d_{it} + (m_i - M) + (M - \mu) \qquad (7.10)$$

dove M è la media, calcolata nella popolazione, dei valori attesi individuali m_i, ovvero:

$$M = \frac{1}{N} \sum_{i=1}^{N} m_i \qquad (7.11)$$

La (7.10) esplicita la differenza tra il valore osservato in una particolare indagine per un dato individuo e la media della popolazione, nella somma di tre componenti:

- la deviazione di risposta individuale, che rappresenta la differenza tra il valore osservato e il valore atteso per l'individuo i;
- la differenza tra il valore atteso individuale e il suo valor medio calcolato su tutte le unità della popolazione;
- la differenza, dovuta alla distorsione, tra il valore medio delle risposte attese e la media della popolazione.

Se si considera di aver osservato un campione casuale di n unità, allora si avrà un'espressione analoga alla (7.10) per ciascuna delle unità del campione e facendone la media campionaria si ottiene:

$$\bar{y} - \mu = \bar{d}_t + (\bar{m} - M) + (M - \mu) \qquad (7.12)$$

dove:

$$\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_{it}$$

$$\bar{d}_t = \frac{1}{n}\sum_{i=1}^{n} d_{it}$$

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$$

Nell'espressione (7.12), l'*errore totale della stima*, ovvero, la differenza tra la media campionaria e la media della popolazione è attribuibile, rispettivamente:

- all'errore di misura variabile, espresso da $\bar{d}_t$;
- all'errore campionario, misurato dalla differenza $(\bar{m} - M)$;
- e alla distorsione o errore sistematico, espresso da $(M - \mu)$.

La formula (7.12) costituisce la base per calcolare l'errore quadratico medio della stima della media, il cui quadrato sarà indicato con il simbolo $MSE(\bar{y}_t)$:

$$MSE(\bar{y}_t) = E(\bar{y}_t - \mu)^2$$

dove il valore medio è calcolato rispetto a tutti i possibili campioni e replicazioni del processo di misurazione sotto le stesse condizioni essenziali. Sviluppando il quadrato della (7.12) e calcolando il valor medio si ottengono le seguenti quantità non nulle:

$$MSE(\bar{y}_t) = \sigma^2_{\bar{d}_t} + \sigma^2_{\bar{m}} + B^2 + 2\,Cov(\bar{d}_t, \bar{m}) \qquad (7.13)$$

le quali rappresentano rispettivamente:

- la *varianza di risposta*, dovuta alla variabilità nelle risposte al variare delle replicazioni, data da:

$$\sigma^2_{\bar{d}_t} = E(\bar{d}_t^2) = E(\bar{y}_t - \bar{m})^2 \qquad (7.14)$$

- la *varianza campionaria*, espressa da:

$$\sigma^2_{\bar{m}} = E(\bar{m} - M)^2 \qquad (7.15)$$

- il quadrato della *distorsione complessiva*, uguale a:

$$B^2 = (M - \mu)^2 \qquad (7.16)$$

- il doppio della *covarianza tra la deviazione di risposta e il valore atteso di risposta*, dovuta all'interazione tra errori campionari ed errori di risposta, ovvero:

$$2Cov(\bar{d}_t, \bar{m}) = 2E((\bar{y}_t - \bar{m})(\bar{m} - M)) \qquad (7.17)$$

Poiché il *valor medio* della stima campionaria $\bar{y}_t$, calcolato al variare delle replicazioni e dei campioni, è uguale a:

$$E(\bar{y}_t) = M = \mu + B \qquad (7.18)$$

allora l'MSE di $\bar{y}_t$ può essere scomposto nel seguente modo:

$$MSE(\bar{y}_t) = E(\bar{y}_t - M)^2 + (M - \mu)^2 = \sigma^2_{\bar{y}_t} + B^2 \qquad (7.19)$$

dove con $\sigma^2_{\bar{y}_t}$ si è indicata la *varianza totale* della stima della media che, per la (7.13), risulta uguale a:

$$\sigma^2_{\bar{y}_t} = \sigma^2_{\bar{d}_t} + \sigma^2_{\bar{m}} + 2Cov(\bar{d}_t, \bar{m}) \qquad (7.20)$$

Analizzando il valor medio e la varianza di $\bar{y}_t$, possiamo riassumere gli effetti degli errori di misura sulla stima della media di una popolazione nei seguenti due:

- *lo stimatore è distorto*: la sua distorsione è pari a B;
- *la precisione dello stimatore è minore* poiché la sua varianza risulta aumentata della parte relativa alla varianza degli errori di risposta.

Dalla (7.18) risulta che la distorsione è indipendente dalla numerosità campionaria; essa quindi rimarrebbe inalterata anche se si effettuasse una indagine totale sotto le stesse condizioni essenziali.

Notiamo che nel caso di indagini totali, il valore $\bar{m}$ coincide con il valore M, di conseguenza la varianza campionaria (7.15) e la covarianza tra errori di misura e di campionamento (7.17) si annullano. Tale covarianza è nulla anche se si considerano replicazioni della indagine su un campione fissato, in quanto si è assunto che $E(d_{it} \mid i) = 0$. Questo termine sarà trascurato negli sviluppi successivi per motivi di semplicità.

Prima di esplicitare la varianza di risposta sotto l'ipotesi di errori incorrelati e di errori correlati all'interno del campione riportiamo nel Prospetto 7.2 uno schema riassuntivo delle componenti del MSE.

Prospetto 7.2 - MSE della media campionaria e sue componenti

Simbolo	Nome della componente
$MSE(\bar{y}_t) = \sigma^2_{\bar{y}_t} + B^2$	Errore quadratico medio
$B^2 = (M - \mu)^2$	Quadrato della distorsione complessiva
$\sigma^2_{\bar{y}_t} = \sigma^2_{\bar{d}_t} + \sigma^2_{\bar{m}} + 2\,Cov(\bar{d}_t, \bar{m})$	Varianza totale
$\sigma^2_{\bar{d}_t} = E(\bar{y}_t - \bar{m})^2$	Varianza di risposta
$\sigma^2_{\bar{m}} = E(\bar{m} - M)^2$	Varianza campionaria
$Cov(\bar{d}_t, \bar{m}) = E((\bar{y}_t - \bar{m})(\bar{m} - M))$	Covarianza tra deviazione di risposta e valore atteso

Effetti di una distorsione costante

Gli effetti di una distorsione costante che agisce su tutte le unità non sono misurabili mediante i dati campionari.

Infatti se le osservazioni y_i sono tutte distorte di una quantità costante e incognita B, tale risulta anche la media campionaria $\bar{y}$. Ne consegue che l'errore campionario di $\bar{y}$ calcolato a partire dai risultati osservati mediante la nota formula per campioni casuali semplici senza reimmissione, ovvero:

$$S^2_{\bar{y}} = \frac{(N - n)}{nN} \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n - 1}$$

non tiene conto della distorsione B poiché essa si elide nelle differenze $(y_i - \bar{y})^2$.

Lo stesso accade nel caso di campioni stratificati, cfr. Cochran (1977).

Nei paragrafi 4 e 5 vedremo in quali casi è possibile stimare la distorsione complessiva B, data dalla (7.18), cui è soggetto lo stimatore $\bar{y}_t$.

Effetti di errori di misura incorrelati

Esplicitiamo la varianza di risposta, che misura l'aumento di variabilità di uno stimatore dovuto alla presenza di errori di risposta, nella situazione più semplice in cui tali errori sono incorrelati all'interno del campione.

Questa situazione si verifica, ad esempio, in indagini postali o con questionari autocompilati, cioè in quelle indagini dove gli individui non si consultano tra di loro e dove non c'è l'influenza di un intervistatore comune a più persone. Di conseguenza si può ritenere che non esista alcun legame tra gli errori di risposta relativi a due distinte unità campionarie.

Anche in questo tipo di indagini, tuttavia, si può verificare una correlazione intracampionaria degli errori dovuta, ad esempio, alla codifica e alla registrazione dei dati, per le unità assegnate ad una stessa persona. Tali errori, pur non essendo propriamente classificabili come errori di misura, in base alla definizione data nel paragrafo 1, possono però contribuire alla varianza di risposta.

Infatti, i dati finali utilizzati per stimare la varianza di risposta risentono di tutti gli errori che si sono accumulati nelle varie fasi del processo di produzione del dato con le relative correlazioni, ovvero dell'*errore globale di misura*.

Sotto l'ipotesi di errori di misura incorrelati, la stima usuale della varianza campionaria di uno stimatore riflette sia gli errori

di risposta che quelli campionari, a condizione che la frazione di campionamento f = n/N sia trascurabile.

Se gli errori di misura, relativi a due distinte unità appartenenti allo stesso campione, sono incorrelati, si ha:

$$\mathrm{Cov}(e_{it}, e_{jt}) = E(d_{it}\, d_{jt}) = 0 \qquad i \neq j \qquad (7.21)$$

Sotto questa assunzione, si dimostra che la varianza di risposta, definita dalla (7.14), assume la seguente espressione:

$$\sigma^2_{\bar{d}_t} = E(\bar{d}_t^2) = \frac{1}{n}\sigma^2_d \qquad (7.22)$$

dove σ^2_d è il valor medio nella popolazione delle varianze di risposta individuale σ^2_i, date dalla (7.6), ovvero:

$$\sigma^2_d = \frac{1}{N}\sum_{i=1}^{N}\sigma^2_i \qquad (7.23)$$

L'espressione (7.22) è calcolata effettuando prima il valore atteso al variare delle replicazioni e quindi il valore atteso al variare dei possibili campioni.

Se gli errori di misura sono incorrelati, la (7.22) rappresenta il contributo della varianza degli errori di misura alla varianza totale dello stimatore $\bar{y}_t$.

La varianza di risposta si riduce, all'aumentare della numerosità campionaria, in proporzione al fattore 1/n, a differenza della distorsione di risposta che, come abbiamo visto in precedenza, è indipendente dal numero di osservazioni effettuate.

Notiamo inoltre che l'espressione (7.22) rimane valida anche per indagini totali, posto n=N.

Come è noto, la varianza campionaria dello stimatore $\bar{y}_t$, definita dalla (7.15), è uguale a:

$$\sigma^2_{\bar{m}} = E(\bar{m} - M)^2 = \frac{N-n}{N-n}\frac{1}{n}\sigma^2_m \qquad (7.24)$$

poiché stiamo considerando un campione casuale semplice senza reimmissione. Con σ^2_m si è indicata la varianza nella popolazione dei valori attesi di risposta m_i, definiti dalla (7.7), cioè:

$$\sigma^2_m = \frac{1}{N}\sum_{i=1}^{N}(m_i - M)^2 \qquad (7.25)$$

La *varianza totale* della media campionaria è quindi data da:

$$\sigma^2_{\bar{y}_t} = \frac{1}{n}\sigma^2_d + \frac{N-n}{N-n}\frac{1}{n}\sigma^2_m \qquad (7.26)$$

come si verifica facilmente, sostituendo la (7.22) e la (7.24) nella espressione (7.20) e trascurando la covarianza tra deviazioni di risposta e valori attesi di risposta.

Consideriamo ora il problema di stimare la varianza totale di $\bar{y}_t$, sulla base dei dati campionari. Lo stimatore usuale, in assenza di errori di misura, è:

$$S^2_{\bar{y}_t} = \frac{(1-f)}{n}\sum_{i=1}^{n}\frac{(y_{it} - \bar{y}_t)^2}{n-1} \qquad (7.27)$$

Se le osservazioni sono affette da errori di risposta, dalla (7.9) si ha:

$$y_{it} - \bar{y}_t = (d_{it} - \bar{d}_t) + (m_i - \bar{m})$$

e sostituendo queste espressioni nella (7.27) si dimostra che:

$$E(S^2_{\bar{y}_t}) = \frac{(1-f)}{n}\sigma^2_d + \frac{N-n}{N-1}\frac{1}{n}\sigma^2_m \qquad (7.28)$$

Dal confronto tra la (7.28) e la (7.26) si evince che $S^2_{\bar{y}_t}$ è uno stimatore *corretto* della varianza totale della media se la frazione di campionamento f è trascurabile; questo risultato è estendibile anche a campioni stratificati, (cfr. Giusti, 1969).

Possiamo concludere che, per popolazioni sufficientemente grandi e per campioni sia semplici che complessi, le formule usuali per il calcolo degli errori campionari delle stime rispecchiano anche l'effetto degli errori di misura, se tali errori sono incorrelati.

I risultati appena illustrati sono riportati, a scopo riassuntivo, nel Prospetto 7.3.

Prospetto 7.3 - Varianza totale della media, sue componenti e stimatore nell'ipotesi di errori di misura incorrelati

Simbolo	Nome della componente
$\sigma^2_{\bar{y}_t} = \frac{1}{n}\sigma^2_d + \frac{N-n}{N-1}\frac{1}{n}\sigma^2_m$	Varianza totale
$\sigma^2_{\bar{d}_t} = \frac{1}{n}\sigma^2_d$	Varianza di risposta
$\sigma^2_{\bar{m}} = \frac{N-n}{N-1}\frac{1}{n}\sigma^2_m$	Varianza campionaria
$S^2_{\bar{y}_t} = \frac{(1-f)}{n}\sum_{i=1}^{n}\frac{(y_{it}-\bar{y}_t)^2}{n-1}$	Stimatore usuale della varianza totale
$E(S^2_{\bar{y}_t}) = \sigma^2_{\bar{y}_t}$ se $f \simeq 0$	Stimatore corretto se la frazione di campionamento è trascurabile

Effetti di errori di misura correlati

La situazione che si verifica più frequentemente nella pratica è quella in cui esiste una correlazione tra gli errori di misura relativi a unità campionarie distinte.

Tale correlazione può sorgere per diversi motivi, tuttavia nelle indagini per intervista diretta, la causa principale è la presenza di un intervistatore comune a più individui. Infatti l'interpretazione di una domanda, la non comprensione o la non osservanza di alcune istruzioni, la tendenza ad accettare le non-risposte sono solo alcune delle caratteristiche proprie del rilevatore che influenzano i valori osservati, introducendo una correlazione tra gli errori relativi ad individui facenti parte della stessa assegnazione.

La conseguenza diretta dell'esistenza di una correlazione intracampionaria degli errori è un aumento della variabilità delle stime. Inoltre non è più possibile tenere conto dell'effetto degli errori di misura con le formule abituali per il calcolo della varianza campionaria, come accade invece sotto l'ipotesi di errori incorrelati. Si dimostra, infatti, che la varianza campionaria stimata dai dati osservati nel modo tradizionale, costituisce una sottostima della variabilità complessiva di uno stimatore. Si rende quindi necessario disporre di metodi di stima appropriati per valutare l'aumento di variabilità delle stime causato dalla presenza di errori di misura correlati; i principali metodi saranno discussi nei paragrafi seguenti.

Se gli errori sono correlati, la varianza di risposta assume la seguente espressione:

$$\sigma^2_{\bar{d}_t} = \frac{1}{n}\sigma^2_d + \frac{(n-1)}{n}\varrho\,\sigma^2_d \qquad (7.29)$$

dove σ^2_d è definito dalla (7.23), e:

$$\varrho = \frac{E(d_{it}\,d_{jt})}{\sigma^2_d} \qquad i \neq j \qquad (7.30)$$

è il coefficiente di correlazione interna tra le deviazioni di risposta individuali in una data indagine o in una data prova.

La *varianza di risposta totale*, (7.29), quindi è la somma di due componenti denominate, rispettivamente, *varianza di risposta semplice* e *componente correlata* della varianza di risposta totale.

La varianza di risposta semplice rappresenta la varianza degli errori di misura individuali e diminuisce al crescere della numerosità campionaria.

La componente correlata, invece, misura l'effetto della correlazione tra gli errori di risposta relativi a individui diversi e non dipende dal numero di osservazioni effettuate. Quando la correlazione interna è dovuta principalmente all'effetto rilevatore, allora la componente correlata è funzione della assegnazione degli intervistatori e può essere ridotta diminuendo il numero di individui intervistati da una stessa persona, a parità di dimensione campionaria. Infatti se a ciascuno dei k intervistatori viene assegnato un numero n' di rispondenti, in modo che sia $n = n'k$, la varianza di risposta totale diventa:

$$\sigma^2_{\bar{d}_t} = \frac{1}{n}\sigma^2_d + \frac{(n'-1)}{n}\varrho\,\sigma^2_d \qquad (7.31)$$

La (7.31) è equivalente alla (7.29) nel caso in cui c'è un solo intervistatore ($k = 1$, $n' = n$); mentre all'altro estremo, la varianza di risposta è minima quando a ciascun intervistatore viene assegnata una sola persona ($k = n$, $n' = 1$): in questo caso infatti la correlazione intracampionaria è nulla e la (7.31) si riduce all'espressione (7.22) relativa a errori incorrelati.

Espressioni analoghe alla (7.31) possono essere utilizzate per analizzare anche altri tipi di correlazione intracampionaria tra er-

rori di risposta, ad esempio la correlazione attribuibile ai supervisori o alla registrazione.

È opportuno segnalare l'importanza del coefficiente di correlazione interna ϱ, definito dalla (7.30); infatti anche valori relativamente piccoli di ϱ possono avere considerevoli effetti sulla varianza di risposta totale. Se, ad esempio, $\varrho = 0.01$ e $n = 2000$, allora la varianza di risposta risulta aumentata di circa 20 volte, ovvero del 2000%, per effetto della correlazione interna, rispetto al caso in cui $\varrho = 0$. Quindi anche se la varianza di risposta semplice non è molto elevata, la varianza di risposta totale può risultare molto grande a causa della correlazione interna degli errori.

La *varianza totale* della media campionaria $\bar{y}_t$ è uguale a:

$$\sigma^2_{\bar{y}_t} = \frac{1}{n}\sigma^2_d(1 + (n-1)\varrho) + \frac{N-n}{N-1}\frac{1}{n}\sigma^2_m \qquad (7.32)$$

come si verifica facilmente, sostituendo la (7.29) e la (7.24) nella espressione (7.20) e trascurando la covarianza tra deviazioni di risposta e valori attesi di risposta.

Nel caso di indagini totali, la varianza campinaria, (7.24), è nulla e la varianza dello stimatore si riduce alla varianza di risposta, (7.29), posto $n = N$.

Come abbiamo accennato in precedenza, la formula usuale per stimare la varianza della media campionaria non riflette l'effetto degli errori di misura nelle osservazioni. Infatti lo stimatore $S^2_{\bar{y}_t}$, dato dalla (7.27), ha il seguente atteso:

$$E(S^2_{\bar{y}_t}) = \frac{(1-f)}{n}(\sigma^2_d(1-\varrho)) + \frac{N-n}{N-1}\frac{1}{n}\sigma^2_m \qquad (7.33)$$

Confrontando questo valore atteso con la varianza totale di $\bar{y}_t$, data dalla (7.32), si vede che lo stimatore $S^2_{\bar{y}_t}$ è *distorto* e la sua *distorsione* è pari a:

$$E(S^2_{\bar{y}_t}) - \sigma^2_{\bar{y}_t} = -\varrho\,\sigma^2_d \qquad (7.34)$$

se la frazione di campionamento f è trascurabile.

Poiché è verosimile supporre che la correlazione tra gli errori sia positiva, allora la varianza campionaria stimata con la formula usuale è una sottostima della varianza complessiva di uno stimatore. Di conseguenza le stime campionarie vengono considerate più precise di quanto non siano effettivamente, a meno di non utilizzare metodi di stima adeguati a tenere conto della correlazione interna degli errori.

Nel Prospetto 7.4 sono schematizzati gli effetti, sulla stima della media, di errori di misura correlati all'interno del campione.

Prospetto 7.4 - Varianza totale della media, sue componenti e stimatore nell'ipotesi di errori di misura correlati

Simbolo	Nome della componente
$\sigma^2_{\bar{y}_t} = \frac{1}{n}\sigma^2_d(1 + (n-1)\varrho) + \frac{N-n}{(N-1)n}\sigma^2_m$	Varianza totale
$\sigma^2_{d_t} = \frac{1}{n}\sigma^2_d + \frac{(n-1)}{n}\varrho\,\sigma^2_d$	Varianza di risposta
$S^2_{\bar{m}} = \frac{N-n}{N-1}\frac{1}{n}\sigma^2_m$	Varianza campionaria
$\varrho = \frac{E(d_{it}\,d_{jt})}{\sigma^2_d} \qquad i \neq j$	Coefficiente di correlazione interna tra deviazioni di risposta
$S^2_{\bar{y}_t} = \frac{(1-f)}{n}\sum_{i=1}^{n}\frac{(y_{it} - \bar{y}_t)^2}{n-1}$	Stimatore usuale della varianza totale
$E(S^2_{\bar{y}_t}) - \sigma^2_{\bar{y}_t} = -\varrho\,\sigma^2_d$	Distorsione dello stimatore

4. Metodi di stima degli errori di misura

Nel paragrafo 3 è stato descritto il modello matematico che permette di evidenziare l'effetto degli errori di misura su una stima campionaria quale la media aritmetica.

Come abbiamo visto le principali conseguenze della presenza degli errori di misura sono:

- l'introduzione di una distorsione, denominata *distorsione di risposta*, nelle stime campionarie;

- l'aumento di variabilità espresso dalla *varianza di risposta totale*, delle stime campionarie.

Inoltre è stato dimostrato che se esiste una distorsione costante B che agisce su tutte le unità allora anche la media risulta distorta di una quantità B la quale non è stimabile a partire dai dati campionari.

La varianza di risposta, invece, è stimabile dalle osservazioni campionarie solo nel caso in cui è nulla la correlazione tra errori di misura di unità appartenenti allo stesso campione. In questo caso, infatti, lo stimatore usuale della varianza della media riflette sia gli errori di campionamento sia quelli di risposta, (cfr. la 7.28).

Questa situazione, tuttavia, è piuttosto teorica, poiché nella pratica tale correlazione interna può essere causata dalla presenza di intervistatori comuni a un gruppo di unità campione nelle indagini per intervista diretta, oppure da altri *agenti* che eseguono le operazioni di registrazione o di codifica su uno stesso insieme di questionari.

Se gli errori di misura sono correlati allora la varianza di risposta totale è la somma di due addendi: la varianza di risposta semplice, che esprime la variabilità dovuta agli errori di misura, e la componente correlata che misura l'effetto della correlazione interna degli errori, (cfr. la 7.29). Inoltre non è più possibile stimare correttamente l'aumento di variabilità causato dagli errori di risposta con i dati del campione. La varianza della media stimata con la formula tradizionale, infatti, sottostima quella totale poiché non tiene conto, in maniera esatta, della correlazione interna, come la (7.34) dimostra.

È necessario, quindi, predisporre delle tecniche che consentano di tenere conto anche degli errori di misura se si vogliono fornire delle indicazioni esatte sulla precisione dei dati forniti agli utenti.

La conoscenza dell'entità degli errori di risposta e l'analisi delle fonti che li hanno generati permette, inoltre, di intervenire nel processo di produzione dei dati allo scopo di migliorarne la qualità.

Occorre, tuttavia, sottolineare che gli errori di risposta dipendono sia dal processo di misurazione adottato (questionario, tipo di indagine scelto, *agenti* impiegati) sia dal tipo di variabile oggetto di studio; di conseguenza raramente i risultati sugli errori di misura di una indagine possono essere applicati ad un'altra. Il confronto tra i risultati ottenuti per rilevazioni differenti può, comunque, risultare molto utile per analizzare le fonti di errore ed individuare i miglioramenti più adatti da apportare.

Nei prossimi paragrafi saranno analizzati e confrontati i due principali metodi di stima degli errori di misura: la *reintervista* e la *compenetrazione del campione*. Con il primo è possibile stimare la distorsione o, in alternativa, la varianza di risposta, mentre il secondo consente di stimare solo la variabilità di risposta. La compenetrazione del campione, però, non altera il costo di una indagine in quanto si risolve in fase di predisposizione del campione, mentre la reintervista che consiste nella replicazione dell'indagine o di una parte di essa, può risultare molto costosa e anche piuttosto lunga.

Saranno, inoltre, sottolineati i problemi organizzativi che ciascun metodo comporta e le condizioni che devono essere rispettate per una corretta applicazione e, di conseguenza, per una corretta utilizzazione delle due tecniche esaminate.

A titolo illustrativo della teoria esposta, si riportano, in appendice, i risultati relativi alla applicazione della tecnica della compenetrazione del campione all'indagine Istat sugli sport e sulle vacanze del 1985.

5. Il metodo della reintervista

Il metodo della *reintervista* consiste nel replicare l'indagine o parte di essa sotto le stesse condizioni generali, ma variando le condizioni particolari di cui si vuole studiare l'influenza sulla qualità dei dati rilevati.

Con questo metodo è possibile stimare:

- la *distorsione di risposta* di uno stimatore quale la media campionaria;
- la *varianza di risposta totale* della media e le sue *componenti* e valutare il contributo relativo degli errori di misura alla varianza campionaria dello stimatore.

I due obiettivi non sono però conciliabili; per stimare la distorsione, infatti, è necessario adottare un processo di misurazione più preciso dell'indagine originaria, mentre per ottenere una stima della varianza di risposta è necessaria una replicazione indipendente dell'indagine sotto le stesse condizioni generali.

Se l'obiettivo è la stima della *variabilità di risposta* dovuta, ad esempio all'impiego degli intervistatori, ovvero la stima dell'*effetto intervistatore* allora la reintervista deve essere condotta da persone diverse da quelle dell'indagine originaria, ma della stessa abilità, esperienza e con il medesimo addestramento,

lasciando inalterati tutti gli altri aspetti quali il questionario, la codifica, la registrazione, i controlli automatici di correzione e così via. In questo modo si ottengono per ciascun individuo due misurazioni indipendenti ed *equivalenti*, in quanto rilevate sotto le stesse condizioni generali; la differenza tra i due valori osservati consente di valutare l'influenza delle mutate condizioni particolari quali gli intervistatori. In maniera analoga possono essere analizzati, ad esempio, gli effetti degli errori di codifica o di registrazione, a parità delle altre condizioni.

Se, invece, l'obiettivo è la stima della *distorsione di risposta* allora la reintervista deve essere predisposta allo scopo di individuare il valore *vero*; sono, quindi, necessari degli accorgimenti specifici i quali, mutando le condizioni generali dell'intervista, non consentono di utilizzare i risultati per stimare la variabilità di risposta.

A tal fine si può replicare l'indagine con *riconciliazione* delle risposte: con questo metodo il rilevatore è fornito, durante la reintervista, delle risposte originarie e in caso di discordanza cerca di appurare con l'aiuto del rispondente quale sia la risposta *vera*. Con questa tecnica è possibile controllare la rete di rilevazione e valutare la parte di distorsione dovuta all'intervistatore se, durante la riconciliazione, si tenta l'attribuzione al rilevatore o al rispondente delle differenze riscontrate, separando, così, le due possibili cause di errore. Se, inoltre, sono previsti dei quesiti sul motivo di tali differenze e sulla conduzione dell'intervista originaria, si possono evidenziare alcune fonti di errore, quali ad esempio carenze nel questionario o nelle istruzioni fornite agli intervistatori.

Altri metodi per ottenere, nella reintervista, una misurazione accurata da poter assumere come valore *vero* possono, ad esempio, essere l'utilizzazione di un questionario più dettagliato con domande di controllo, l'impiego di intervistatori più esperti che abbiano ricevuto un addestramento migliore ed istruzioni più particolareggiate o una mistura di questi accorgimenti.

È interessante riportare, a titolo di esempio, il programma di reinterviste per l'indagine sulle forze di lavoro svolto da Statistics Canada con il duplice scopo di quantificare l'effetto degli errori di misura e di controllo della rete di rilevazione e dell'aggiornamento delle liste (delle abitazioni). La reintervista è condotta da intervistatori esperti (senior interviewers) seguendo la stessa procedura della rilevazione delle forze di lavoro e riguarda un sottocampione dell'indagine. In due terzi di esso la reintervista è effettuata con riconciliazione delle risposte, mentre nel rimanente terzo vengono rilevate le risposte fornite senza alcun processo di riconciliazione. Assumendo che dal sottocampione con riconciliazione si ottengano i valori *veri* e che le reinterviste senza riconciliazione siano una replicazione indipendente delle interviste originarie, allora due terzi delle reinterviste forniscono una stima della distorsione e un terzo una stima della variabilità di risposta. Il sottocampione con riconciliazione viene utilizzato anche per il controllo della rete di rilevazione; durante il processo di riconciliazione, infatti, l'intervistatore esperto cerca di appurare se le differenze riscontrate devono essere attribuite al rilevatore o al rispondente e il motivo della discrepanza (ad esempio errori nella procedura o non comprensione dei quesiti). Queste informazioni vengono riportate su un apposito modello, (discusso insieme all'intervistatore), che consiste di quattro parti principali: la prima riporta il grado di aggiornamento delle liste delle abitazioni; la seconda gli errori dell'intervistatore emersi durante la riconciliazione e i risultati delle consultazioni; la terza il giudizio complessivo dell'intervistatore esperto sulla conduzione dell'intervista da parte del rilevatore sottoposto a controllo; la quarta il giudizio e le raccomandazioni del supervisore che segue l'intervistatore esperto.

I risultati ottenibili con il metodo della reintervista sono influenzati dalla scelta dell'intervallo temporale che separa le due indagini. Infatti se sono troppo ravvicinate, i risultati della reintervista sono condizionati da quelli originari in quanto l'intervistato ricorda le risposte date in precedenza e tende a ripeterle nella reintervista anche se incorrette. Un intervallo di tempo troppo lungo, tuttavia, crea notevoli difficoltà poiché risulta difficile fornire delle risposte precise con riferimento a situazioni lontane nel tempo; inoltre l'incidenza dell'intervallo temporale è collegata al tipo di variabile considerata. Citiamo a titolo di esempio alcune esperienze effettuate negli Stati Uniti, (cfr. Ballar, 1968), dalle quali risulta un intervallo di tempo *ottimale*, per alcune variabili, di circa tre mesi.

Per le indagini correnti, essendo troppo lungo e complesso dal punto di vista organizzativo, replicare totalmente la rilevazione, si ricorre alla reintervista di un *sottocampione* di dimensione n' (con $n' < n$) dell'indagine originaria; gli stimatori che così si otterranno possono essere facilmente ricondotti al caso in cui $n' = n$.

In questo modo per ciascuna delle n' unità reintervistate si dispone di due valori osservati che indichiamo con:

$$y_{i1} \quad e \quad y_{i2} \qquad (i = 1,\ldots,n')$$

dove il sottoscritto 1 si riferisce all'indagine originaria e il sottoscritto 2 alla replicazione; inoltre siano:

$$\bar{y}_1 \quad e \quad \bar{y}_2$$

le medie campionarie nelle due rilevazioni, relative alle n' unità in comune nelle due indagini.

Stima della distorsione

Se la reintervista è stata condotta con riconciliazione o se, comunque, si può assumere che sia più accurata dell'indagine originaria allora una stima corretta della distorsione B, definita dalla (7.18), è data semplicemente da:

$$\hat{B} = \bar{y}_1 - \bar{y}_2$$

ovvero dalla differenza tra le due medie campionarie.

Stima della varianza di risposta totale

Se la reintervista è stata effettuata sotto le stesse condizioni generali e se ogni individuo del secondo campione è stato intervistato da un rilevatore diverso da quello che aveva condotto la prima indagine allora la differenza tra le medie campionarie nelle due indagini costituisce la base per stimare la varianza di risposta totale, definita dalla (7.29). A tale scopo indichiamo con C la seguente espressione:

$$C = \frac{1}{2}(\bar{y}_1 - \bar{y}_2)^2 \qquad (7.35)$$

C misura la variabilità tra gli intervistatori poiché corrisponde alla varianza tra le medie delle due indagini ed ha un solo grado di libertà. La (7.35) è un caso particolare, con m = 2, dello stimatore relativo a m replicazioni indipendenti dell'indagine:

$$C = \frac{1}{m-1}\sum_{t=1}^{m}(\bar{y}_t - \bar{y})^2$$

con (m — 1) gradi di libertà e con $\bar{y}$ uguale alla media generale.

Si dimostra che, si veda Hansen, Hurwitz e Bershad (1961), il valore atteso di C è pari a:

$$E(C) = \frac{1}{n'}\sigma_d^2\,(1 + (n' - 1)\,\varrho) \qquad (7.36)$$

se sono soddisfatte le seguenti ipotesi:

$$_1m_i = {}_2m_i$$

$$\sigma^2_{d_1} = \sigma^2_{d_2}$$

$$\varrho_1 = \varrho_2$$

$$E(d_{i1}\,d_{i2}) = 0 \qquad (7.37)$$

ovvero se rispettivamente:

- i valori attesi di risposta, cfr. la (7.7), relativi allo stesso individuo sono uguali in entrambe le indagini;
- la varianza di risposta semplice è uguale nelle due indagini;
- il coefficiente di correlazione interna tra deviazioni di risposta, dato dalla (7.30), è uguale due indagini;
- la covarianza tra deviazioni di risposta delle due indagini è nulla.

Notiamo che le prime tre ipotesi sono soddisfatte se le due rilevazioni sono state condotte sotto le stesse condizioni generali, mentre per la quarta è necessaria l'indipendenza tra le due prove. In realtà è presumibile che ci sia una correlazione positiva tra gli errori di misura poiché l'individuo, ricordando le risposte fornite in precedenza, tende a ripeterle, anche se errate. Questo effetto può essere ridotto aumentando l'intervallo di tempo che separa le due indagini; in questo modo, tuttavia, si può produrre un aumento della distorsione delle stime e un aumento della variabilità di risposta nella reintervista, causati dall'insorgere di problemi di memoria da parte del rispondente. Le due prove possono, invece, considerarsi indipendenti se la replicazione è finalizzata, ad esempio, a valutare l'effetto di errori di codifica o di registrazione e si impiegano *agenti* della stessa abilità che non hanno accesso al lavoro svolto precedentemente da altri e non possono esserne influenzati.

Dal confronto tra la (7.36) e la (7.29) segue che:

$$C' = \frac{n'}{n} C \qquad (7.38)$$

fornisce una sottostima della varianza di risposta totale della media dell'indagine originaria di una quantità pari a $((n'/n) - 1)\ \varrho\ \sigma_d^2$. La distorsione, quindi, si riduce all'aumentare di n', fino ad annullarsi per $n' = n$, cioè quando si replica completamente l'indagine e può considerarsi trascurabile per n' sufficientemente grande. In ogni caso C, (7.35), stima correttamente la varianza di risposta totale della media del sottocampione. Se, ad esempio, si reintervistano tutte le unità appartenenti ad una data area geografica (comune, provincia o regione) incluse nel campione di una indagine condotta su scala nazionale, allora C fornisce una *stima corretta* della varianza di risposta totale della media di quella data area, mentre (n'/n)C sottostima la corrispondente varianza della media del campione nazionale.

Più gravi conseguenze ha, invece, la *caduta* della quarta delle ipotesi (7.37). Se, infatti, la covarianza tra deviazioni di risposta è positiva allora la (7.35) sottostima la varianza di risposta totale con riferimento al sottocampione. La sottostima sarà tanto maggiore quanto maggiore è la correlazione tra le due indagini inficiando la possibilità di utilizzare i risultati della reintervista per quantificare l'effetto degli errori di misura sulle stime campionarie.

Stima della varianza di risposta semplice

Uno stimatore corretto della varianza di risposta semplice può essere costruito a partire dalle differenze tra i valori individuali rilevati nelle due indagini. Sia D la media dei quadrati di tali differenze:

$$D = \frac{1}{n'} \sum_{i=1}^{n'} (y_{i1} - y_{i2})^2 \qquad (7.39)$$

Sotto le ipotesi (7.37), si dimostra che, (cfr. Hansen, Hurwitz e Pritzker, 1964), il valore atteso di D è:

$$E(D) = 2\ \sigma_d^2 \qquad (7.40)$$

Dalla (7.40) segue che è possibile stimare correttamente la varianza di risposta semplice della media sia a livello di sottocampione sia a livello del campione originario. Infatti la quantità:

$$\frac{1}{2n'} D \qquad (7.41)$$

fornisce una *stima corretta* della varianza di risposta semplice del sottocampione reintervistato, mentre:

$$\frac{1}{2n} D \qquad (7.42)$$

stima correttamente la corrispondente varianza della media relativa al campione complessivo.

Anche in questo caso sono valide le considerazioni fatte precedentemente sulla quarta delle condizioni (7.37), infatti una correlazione positiva tra gli errori di misura nelle due indagini riduce il valore atteso (7.40) di una quantità pari a tale correlazione; di conseguenza gli stimatori (1/2n)D e (1/2n')D forniranno delle sottostime della varianza di risposta semplice.

Il rapporto tra la (7.41) e la (7.35) stima correttamente il contributo relativo della varianza di risposta semplice alla varianza di risposta totale della media del sottocampione, mentre per il campione complessivo occorre rapportare la (7.42) alla (7.38).

Stima della componente correla

La differenza tra i due stimatori C e D permette di valutare l'effetto della componente correlata della varianza di risposta totale. Indicando con F tale differenza:

$$F = C - \frac{1}{2n'} D \qquad (7.43)$$

si ha:

$$E(F) = \frac{n' - 1}{n'}\ \varrho\ \sigma_d^2 \qquad (7.44)$$

Quindi F è uno *stimatore corretto* della componente correlata per il sottocampione, mentre:

$$F' = \frac{n'}{n' - 1} \frac{n - 1}{n} F \qquad (7.45)$$

fornisce una stima corretta della componente correlata relativa al campione complessivo.

I rapporti F/C e F'/C' stimano il contributo relativo della componente correlata alla varianza di risposta totale della media del sottocampione e della media del campione originario.

Stima della varianza campionaria

Indichiamo con G la varianza interna alle osservazioni nelle due replicazioni:

$$G = \sum_{t=1}^{2} \sum_{i=1}^{n'} \frac{(y_{it} - \bar{y}_t)^2}{2(n' - 1)} \qquad (7.46)$$

Se sono soddisfatte le ipotesi (7.37), G ha il seguente valore atteso, (si veda Cochran, 1977):

$$E(G) = \sigma_d^2 (1 - \varrho) + \frac{N}{N - 1} \sigma_m^2 \qquad (7.47)$$

Confrontando la (7.47) con la (7.33), si vede che (1/n)G ha lo stesso valor medio di $S^2_{\bar{y}_t}$, trascurando la frazione di campionamento, cioè dello stimatore usuale della varianza campionaria. Ne consegue che il rapporto tra varianza di risposta e varianza di campionamento stimate:

$$\frac{C}{G}$$

fornisce una misura dell'influenza, in termini di variabilità, degli errori di risposta rispetto a quelli campionari.

Stima dell'indice di inconsistenza

Il rapporto

$$I = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_m^2} \qquad (7.48)$$

è noto nella letteratura con il nome di indice di inconsistenza e misura la parte di varianza totale *elementare* (n = 1) dovuta alla variabilità di risposta. In base a quanto visto precedentemente, l'indice I può essere stimato mediante il seguente rapporto:

$$\hat{I} = \frac{D/2}{G} \qquad (7.49)$$

Nel Prospetto 7.5 si riporta uno schema riassuntivo degli stimatori che si possono ottenere mediante la replicazione parziale dell'indagine e la loro estensione all'indagine complessiva.

Prospetto 7.5 - Schema riassuntivo degli stimatori ottenibili con la reintervista

Varianze stimate	Stimatore per il sottocampione	Stimatore per il campione
Varianza di risposta totale	C	$\frac{n'}{n}$ C
Varianza di risposta semplice	$\frac{1}{2n'}$ D	$\frac{1}{2n}$ D
Componente correlata	F	$\frac{n'(n-1)}{(n'-1)n}$ F
Varianza campionaria	$\frac{1}{n'}$ G	$\frac{1}{n}$ G

dove:

$$C = \frac{1}{2}(\bar{y}_1 - \bar{y}_2)^2$$

$$D = \frac{1}{n'} \sum_{i=1}^{n'} (y_{i1} - y_{i2})^2$$

$$F = C - \frac{1}{2n'} D$$

$$G = \sum_{t=1}^{2} \sum_{i=1}^{n'} \frac{(y_{it} - \bar{y}_t)^2}{2(n' - 1)}$$

roblemi operativi

Analizziamo i problemi organizzativi e le condizioni che devono essere rispettate per una corretta applicazione del metodo della reintervista.

La reintervista può essere predisposta per valutare l'effetto degli errori di misura introdotti da vari *agenti*; i contributi maggiori all'errore globale, tuttavia, sono dovuti ai rilevatori.

Se la reintervista è finalizzata a quantificare l'aumento di variabilità delle stime (o in maniera analoga la distorsione), causato dagli intervistatori, allora ogni individuo del campione originario deve essere reintervistato da un rilevatore differente. Come è stato già sottolineato si rende necessario, per ragioni di costo, replicare parzialmente una indagine su larga scala rinunciando ad alcune proprietà degli stimatori. Occorre, quindi, una particolare attenzione nella scelta della numerosità del sottocampione e nell'estensione dei risultati ottenuti all'indagine originaria.

Nel caso di replicazione parziale della rilevazione deve essere possibile associare ad ogni unità intervistata un codice di rilevatore ed estrarre per ogni codice un sottocampione di unità da assegnare ad un intervistatore diverso ma della stessa capacità ed esperienza. Inoltre poiché lo stimatore della varianza di risposta semplice D, definito dalla (7.39), si basa sul confronto tra i valori individuali rilevati nelle due indagini, il sistema di identificazione delle unità deve essere tale da consentire l'aggancio tra i due codici relativi allo stesso individuo. Quanto detto presuppone:

- l'esistenza di un elenco base di rilevatori;
- l'esistenza di un codice unico per ciascun rilevatore;
- l'aggancio individuo-rilevatore nell'indagine originaria e nella replicazione;
- l'aggancio tra i codici individuali nelle due rilevazioni.

Se l'unità di rilevazione è la famiglia e l'unità di analisi è l'individuo, allora l'assegnazione delle reinterviste deve essere eseguita rispetto alle famiglie; di conseguenza le condizioni sopra esposte devono riferirsi alla famiglia così come gli stimatori illustrati nel sottoparagrafo precedente devono riguardare la varianza o la distorsione della media calcolata per famiglia (ad esempio numero medio di componenti per famiglia).

Una particolare attenzione deve essere rivolta per assicurare che il rispondente sia lo stesso individuo in entrambe le rilevazioni soprattutto se l'indagine prevede la possibilità di risposte «proxy», anche nel caso in cui non sia specificato chi deve fornire le notizie di carattere generale, per non introdurre una ulteriore distorsione e variabilità nei dati dovute al cambiamento del rispondente.

6. Il metodo della compenetrazione del campione

Il metodo della compenetrazione del campione è una tecnica che permette di stimare, dagli stessi dati campionari, sia la *varianza totale* di uno stimatore (ovvero varianza campionaria e varianza di risposta) sia la *componente correlata* della varianza di risposta; inoltre tale tecnica non implica costi aggiuntivi ma solo una maggiore attenzione nell'organizzazione della rilevazione sul campo.

Il metodo della compenetrazione del campione è stato introdotto da Mahalanobis (1946) e ripreso da numerosi autori e istituti ufficiali di statistica che lo hanno adattato alle caratteristiche particolari delle indagini oggetto di studio.

Nella sua formulazione standard (cfr. Cochran, 1977) tale tecnica consiste nel suddividere *a caso* un campione casuale di n unità in k sottocampioni di uguale numerosità $n' = n/k$, ognuno dei quali costituisce un campione rappresentativo della popolazione di origine.

I sottocampioni così ottenuti non risultano statisticamente indipendenti (per approfondimenti teorici sulla tecnica di campionamento si vedano Koop (1960) e Deming (1964)); quindi l'organizzazione sul campo dell'indagine deve essere pianificata in modo da eliminare la correlazione tra errori di misura di unità appartenenti a sottocampioni differenti, dovuta all'impiego dei medesimi intervistatori, revisori e supervisori.

Nell'ipotesi semplificatrice in cui la correlazione tra deviazioni di risposta è attribuibile esclusivamente agli intervistatori è sufficiente assegnare *casualmente* ciascun sottocampione ad un intervistatore diverso per ottenere una corretta applicazione del metodo. Sotto queste assunzioni, infatti, si può supporre che la correlazione tra gli errori di misura relativi ad unità appartenenti a campioni differenti sia nulla.

La componente correlata della varianza di risposta che in questo caso misura l'*effetto intervistatore*, può essere stimata partendo dal confronto tra la varianza tra le assegnazioni degli intervistatori (che misura la variabilità tra le medie di ogni sottocampione e la media generale) e la varianza interna alle assegnazioni degli intervistatori (che misura la variabilità all'interno di ogni sottocampione). Inoltre la varianza esterna permette di stimare anche la varianza totale della media, ovvero di calcolare la precisione dello stimatore tenendo conto sia degli errori di misura sia di quelli campionari.

Nel seguito faremo riferimento alla situazione appena descritta e al modello matematico illustrato nel paragrafo 3; per gli sviluppi successivi è però conveniente riferire l'indice i ($i = 1, 2, \ldots, k$)

ai sottocampioni o equivalentemente agli intervistatori e l'indice j (j = 1, 2,...,n') alle unità all'interno di ciascun sottocampione.

Dalla (7.9) il valore osservato per la j-esima unità assegnata all'i-esimo intervistatore nella t-esima replicazione può scriversi:

$$y_{ijt} = m_{ij} + d_{ijt} \tag{7.50}$$

Indichiamo con:

$$\bar{y}_{it} = \frac{1}{n'} \sum_{j=1}^{n'} y_{ijt} = \bar{m}_i + \bar{d}_{it} \tag{7.51}$$

la media dei valori osservati dall'i-esimo intervistatore e con:

$$\bar{y}_t = \frac{1}{n} \sum_{j=1}^{n'} \sum_{i=1}^{k} y_{ijt} = \frac{1}{k} \sum_{i=1}^{k} \bar{y}_{it} \tag{7.52}$$

la media dei valori osservati con riferimento al campione complessivo.

È facile verificare che il valore atteso di $\bar{y}_{it}$ e di $\bar{y}_t$ è:

$$E(\bar{y}_{it}) = E(\bar{y}_t) = M \tag{7.53}$$

dove M è la media nella popolazione dei valori attesi individuali m_{ij}; quindi sia la media generale che le medie di ciascun sottocampione sono distorte, per effetto degli errori di misura, come nel caso di un campione casuale semplice, cfr. (7.18).

Il metodo dei campioni interpenetranti non fornisce elementi che consentano di stimare la distorsione B, infatti a tale scopo è necessario conoscere il valore *vero* o almeno una stima più precisa con cui confrontare il valore osservato.

Per quanto concerne la stima della varianza totale della media campionaria, ricordiamo che, se gli errori di misura sono correlati e se a ciascuno dei k intervistatori sono assegnate n' unità, la varianza di risposta di $\bar{y}_t$ assume l'espressione (7.31), mentre la varianza campionaria è sempre data dalla (7.24). Ne consegue che la varianza totale della media campionaria $\bar{y}_t$ è:

$$\sigma^2_{\bar{y}_t} = \frac{N}{(N-1)n} \sigma^2_m + \frac{1}{n} \sigma^2_d (1 + (n' - 1)\, \varrho) \tag{7.54}$$

Nella (7.54) si è supposto che la frazione di campionamento f sia trascurabile e che siano nulle tutte le correlazioni relative ad unità assegnate ad intervistatori diversi.

Notiamo che la componente correlata della varianza di risposta, in questa situazione, assume la seguente espressione:

$$\frac{n' - 1}{n} \varrho\, \sigma^2_d \tag{7.55}$$

ed è quindi funzione decrescente della numerosità campionaria complessiva n e funzione diretta del numero n' di unità assegnate a ciascun intervistatore, ovvero funzione inversa del numero di intervistatori utilizzati.

Come abbiamo detto precedentemente la stima della varianza totale di $\bar{y}_t$ e della componente correlata si ottengono dal confronto tra la varianza esterna e la varianza interna ai sottocampioni e quindi direttamente dai dati rilevati.

Indichiamo con S^2_b la *varianza esterna* o *varianza tra le assegnazioni degli intervistatori*, definita come la somma dei quadrati degli scostamenti tra le medie di ciascun intervistatore e la media generale, divisa per i gradi di libertà, pari a k−1, ovvero:

$$S^2_b = n' \sum_{i=1}^{k} \frac{(\bar{y}_{it} - \bar{y}_t)^2}{k - 1} \tag{7.56}$$

Con S^2_w indichiamo la *varianza interna alle assegnazioni degli intervistatori*, data dalla somma delle deviazioni al quadrato tra i valori osservati nell'i-esimo sottocampione e la relativa media, diviso per k(n'−1) gradi di libertà, cioè:

$$S^2_w = \sum_{i=1}^{k} \sum_{j=1}^{n'} \frac{(y_{ijt} - \bar{y}_{it})^2}{k(n' - 1)} \tag{7.57}$$

Consideriamo i valori attesi della (7.56) e della (7.57) al variare delle assegnazioni, del campione e delle replicazioni; nell'ipotesi di assenza di correlazione tra errori di risposta di unità

assegnate a differenti rilevatori e trascurando i fattori di correzione per popolazioni finite, si dimostra che (cfr. Cochran, 1977):

$$E(S_b^2) = \frac{N}{N-1}\sigma_m^2 + \sigma_d^2 (1 + (n' - 1)\varrho) \qquad (7.58)$$

$$E(S_w^2) = \frac{N}{N-1}\sigma_m^2 + \sigma_d^2 (1 - \varrho) \qquad (7.59)$$

Stima della varianza totale

Confrontando tra la (7.54) e la (7.58) si evince che:

$$\frac{1}{n} S_b^2 \qquad (7.60)$$

è uno *stimatore corretto* della *varianza totale* della media. La (7.60) è, quindi, l'espressione *corretta* da utilizzare per il calcolo della variabilità delle stime in quanto tiene conto anche dell'effetto degli errori di risposta.

Stima della varianza campionaria

Il confronto tra la (7.59) e la (7.33) mostra che:

$$\frac{1}{n} S_w^2 \qquad (7.61)$$

può essere utilizzato per stimare la *varianza campionaria*.

Stima della componente correlata

Inoltre il confronto tra la varianza esterna e la varianza interna permette di stimare l'*effetto intervistatore*, espresso dalla (7.55). Infatti indicando con:

$$S_d^2 = \frac{n' - 1}{n'} (S_b^2 - S_w^2) \qquad (7.62)$$

si ha:

$$E(S_d^2) = (n' - 1)\,\varrho\,\sigma_d^2 \qquad (7.63)$$

Ne consegue che:

$$\frac{1}{n} S_d^2 \qquad (7.64)$$

è lo stimatore cercato in quanto fornisce una *stima corretta*, della *componente correlata* della varianza di risposta.

Notiamo infine che il rapporto:

$$\frac{S_d^2}{S_b^2} \qquad (7.65)$$

misura il contributo relativo della componente correlata alla varianza totale della stima. Se la componente correlata può considerarsi molto maggiore della varianza di risposta semplice, allora la (7.65) misura il contributo relativo della varianza di risposta alla varianza totale della media, ovvero la percentuale della variabilità totale dovuta esclusivamente agli errori di misura. In maniera analoga è possibile valutare il peso degli errori di risposta rispetto a quelli campionari considerando il rapporto:

$$\frac{S_d^2}{S_w^2} \qquad (7.66)$$

Riportiamo nel Prospetto 7.6 uno schema riassuntivo degli stimatori appena descritti.

Prospetto 7.6 - Schema riassuntivo degli stimatori ottenibili con la compenetrazione del campione

Simbolo	Varianze e stimatori
$S_b^2 = n' \sum_{l=1}^{k} \frac{(\bar{y}_{lt} - \bar{y}_t)^2}{k-1}$	Varianza esterna o tra le assegnazioni
$S_w^2 = \sum_{l=1}^{k} \sum_{j=1}^{n'} \frac{(y_{ljt} - \bar{y}_{lt})}{k(n'-1)}$	Varianza interna alle assegnazioni
$\frac{1}{n} S_d^2$	Stimatore corretto della varianza totale
$\frac{1}{n} S_w^2$	Stimatore della varianza campionaria

Prospetto 7.6 segue - Schema riassuntivo degli stimatori ottenibili con la compenetrazione del campione

Simbolo	Varianze e stimatori
$S_d^2 = \frac{n' - 1}{n'} (S_b^2 - S_w^2)$	Stimatore corretto della componente correlata della varianza di risposta
$\frac{S_d^2}{S_b^2}$	Contributo relativo della componente correlata alla varianza totale
$\frac{S_d^2}{S_w^2}$	Contributo relativo della componente correlata alla varianza campionaria

Problemi operativi

Ricordiamo che la tecnica della compenetrazione del campione non è solo potente dal punto di vista dei risultati conseguibili, ma presenta anche l'ulteriore vantaggio dell'economicità in quanto, a differenza della reintervista, non incide sul *budget* di una indagine. Essa, infatti, si risolve in fase di predisposizione dell'indagine poiché riguarda esclusivamente il disegno campionario; è però necessario un maggior controllo del rispetto delle norme nella fase di rilevazione sul campo.

Indipendentemente dal piano di campionamento scelto, che può essere semplice, o complesso come nelle indagini correnti dell'Istituto, ricordiamo che i sottocampioni estratti devono risultare tali da eliminare la correlazione tra errori di risposta di unità assegnate ad intervistatori diversi. Il modo standard per soddisfare questo requisito consiste, come abbiamo detto, nell'estrarre *a caso* i sottocampioni e nell'assegnarli *casualmente* a ciascun rilevatore.

A questo proposito occorre distinguere tra piccoli e grandi comuni. Per i primi il vincolo è costituito dalla numerosità campionaria che può non essere sufficiente, se sdoppiata, a garantire un guadagno minimo richiesto dal rilevatore. Per i grandi comuni la casualizzazione delle assegnazioni degli intervistatori può comportare un notevole dispendio di tempo e risultare antieconomica in quanto i rilevatori sarebbero costretti a coprire un'area troppo vasta. In questo caso si preferisce procedere in maniera alternativa accorpando due aree di rilevazione contigue e casualizzando le assegnazioni tra due intervistatori. Di conseguenza ciascun rilevatore lavora su un'area doppia, in termini di distanze, rispetto a quella originaria, ma viene salvaguardata la compenetrazione del campione (anche se occorre esplicitare l'*effetto cluster*, cioè la correlazione tra deviazioni campionarie di unità appartenenti alla stessa area).

Per stimare l'effetto degli errori di risposta è sufficiente che le assegnazioni degli intervistatori siano state effettuate correttamente e che siano rispettate le condizioni elencate di seguito:

- esistenza di un elenco base di rilevatori;
- esistenza di un codice unico per ciascun rilevatore;
- selezione casuale dei rilevatori dell'elenco base;
- controllo della casualità delle assegnazioni di ciascun rilevatore;
- possibilità di agganciare il rilevatore alla famiglia e a ciascun componente.

APPENDICE

1. Applicazione del metodo della compenetrazione del campione all'indagine Istat sugli sport e sulle vacanze

Il disegno campionario dell'indagine Istat sugli sport e sulle vacanze del 1985 prevedeva la compenetrazione delle assegnazioni degli intervistatori nei comuni campione di Milano e Firenze, quindi è stato possibile misurare l'influenza dell'intervistatore sulla qualità dei dati rilevati (cfr. Signore, 1988b).

Trattandosi di grandi comuni, le famiglie campione non sono state assegnate casualmente agli intervistatori in quanto questo avrebbe comportato spostamenti su di un territorio troppo vasto. Si è invece proceduto nel modo seguente. Ciascuna città è stata suddivisa in aree, per semplicità ci si è riferiti alle circoscrizioni, e in ciascuna di esse le famiglie estratte sono state assegnate in maniera casuale a due rilevatori; in particolare ogni rilevatore ha intervistato n' = 12 famiglie a Milano e n' = 9 famiglie a Firenze.

Di conseguenza in ciascuna area si possono calcolare la varianza esterna, tra le assegnazioni dei due rilevatori, e la varianza interna alle assegnazioni e, quindi, stimare la varianza totale e la componente correlata. Per k = 2, la (7.56) e la (7.57) assumono rispettivamente le seguenti espressioni:

$$ {}_hS_b^2 = \frac{n'}{2} ({}_h\bar{y}_1 - {}_h\bar{y}_2)^2 \qquad (7A.67) $$

$$ {}_hS_w^2 = \frac{1}{2} ({}_hS_1^2 + {}_hS_2^2) \qquad (7A.68) $$

dove l'indice h si riferisce all'area e, per semplicità, si è omesso l'indice t relativo alla replicazione; nella (7A.68) si è indicata con ${}_hS_i^2$ la varianza campionaria corretta relativa al sottocampione assegnata all'i-esimo intervistatore. Infine sia:

$$ {}_hS_d^2 = \frac{n' - 1}{n'} ({}_hS_b^2 - {}_hS_w^2) \qquad (7A.69) $$

Si ottengono, quindi, i seguenti stimatori per ciascuna area:

stimatore della varianza totale:

$$ \frac{1}{2n'} {}_hS_b^2 \qquad (7A.70) $$

stimatore della varianza campionaria:

$$ \frac{1}{2n'} {}_hS_w^2 \qquad (7A.71) $$

stimatore della componente correlata:

$$ \frac{1}{2n'} {}_hS_d^2 \qquad (7A.72) $$

Per sintetizzare le stime a livello di città e per aumentarne l'affidabilità, (si noti che la (7A.67) ha un solo grado di libertà), si è provveduto a calcolare una media, rispettivamente, delle espressioni (7A.67), (7A.68) e (7A.69), nel modo seguente:

$$ S_b^2 = \frac{\sum_{h=1}^{L} N_h^2 \, {}_hS_b^2}{(\sum_{h=1}^{L} N_h)^2} \qquad (7A.73) $$

$$ S_w^2 = \frac{\sum_{h=1}^{L} N_h^2 \, {}_hS_w^2}{(\sum_{h=1}^{L} N_h)^2} \qquad (7A.74) $$

$$ S_d^2 = \frac{\sum_{h=1}^{L} N_h^2 \, {}_hS_d^2}{(\sum_{h=1}^{L} N_h)^2} \qquad (7A.75) $$

dove N_h indica il numero di famiglie residenti nella h-esima circoscrizione (L = 20 a Milano e L = 14 a Firenze). Le espressioni (7A.73), (7A.74) e (7A.75) consentono di stimare rispettivamente la varianza totale, la varianza campionaria e la componente correlata con riferimento ad una circoscrizione di dimensione media (cfr. U.S. Bureau of the Census, 1968).

Nella tavola 7A.1 sono riportati i valori stimati del rapporto tra componente correlata e varianza campionaria della media o percentuale per famiglia delle 22 variabili riportate in nota (a).

Tavola 7A.1 - Valori del rapporto S_d^2/S_w^2

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
MI	.27	.28	1.36	1.09	.72	1.18	.45	.39	.57	.26	1.87
FI	.81	.31	.76	.54	.55	.51	.39	.44	.28	.29	.47

	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22
MI	.05	1.35	.29	.05	1.27	.87	.00	2.14	.79	1.04	.86
FI	.53	1.03	.86	.04	.17	.37	.33	.54	.39	1.96	.32

Dall'analisi della tavola 7A.1 si vede che esiste una grande variabilità nei valori del rapporto S_d^2/S_w^2 sia per differenti variabili nella stessa città, sia per una stessa variabile nei due comuni. Ad esempio variabili *simili* come V15 (spese sostenute per viaggio tutto compreso) e V21 (altre spese sostenute) assumono rispettivamente valori molto bassi (.05 e .04) e valori molto elevati (1.04 e 1.96) in entrambe le città. Viceversa una stessa variabile può assumere valori molto diversi nelle due città, si vedano ad esempio V1 (numero di componenti della famiglia) e V16 (spese sostenute per mezzi di trasporto).

In generale risulta esserci un considerevole *effetto intervistatore*, in quanto si riscontrano valori piuttosto elevati del rapporto tra componente correlata e varianza campionaria, anche se Firenze presenta valori più bassi di Milano. In media questo rapporto è pari a .54 per Firenze e .78 per Milano; questo significa che le varianze campionarie, in media, sottostimano notevolmente la variabilità totale. Infatti per ottenere una stima corretta della varianza totale si dovrebbero moltiplicare le varianze campionarie, in media, per 1.54 a Firenze e per 1.78 a Milano.

(a) V1 = numero di componenti della famiglia; V2 = numero di componenti che hanno effettuato almeno un periodo di vacanza; V3 = numero di periodi di vacanza; V4 = numero di periodi di vacanza effettuati in Italia; V5 = numero di periodi di vacanza effettuati all'estero; V6 = numero di periodi di vacanza iniziati da giugno a settembre; V7 = numero di periodi di vacanza iniziati da ottobre a maggio; V8 = numero di periodi di vacanza effettuati in auto, camper o moto; V9 = numero di periodi di vacanza effettuati con altro mezzo di trasporto; V10 = numero di periodi di vacanza effettuati in alloggio di proprietà di un componente, di un parente o di amici; V11 = numero di periodi di vacanza effettuati in alloggio non di proprietà; V12 = numero di periodi di vacanza effettuati con viaggi organizzati; V13 = numero di periodi di vacanza effettuati con viaggi non organizzati; V14 = durata complessiva delle vacanze (in giorni); V15 = spese sostenute per viaggio tutto compreso (in migliaia di lire); V16 = spese sostenute per mezzi di trasporto (in migliaia di lire); V17 = spese sostenute per pensione completa (in migliaia di lire); V18 = spese sostenute per mezza pensione (in migliaia di lire); V19 = spese sostenute per vitto (in migliaia di lire); V20 = spese sostenute per alloggio (in migliaia di lire); V21 = altre spese sostenute (in migliaia di lire); V22 = spesa complessiva sostenuta (in migliaia di lire).

RIFERIMENTI BIBLIOGRAFICI

BAILAR B.A. (1968), *Recent Research in Reinterview Procedures*, «Jour. Amer. Stat. Ass.», Vol. 63, pp. 41-63.

BAILAR B.A. (1987), *Nonsampling errors*, «Jour. of Official Statistics», Vol. 3, n. 4, pp. 323-325.

BAILAR B.A., L. BAILEY e J. STEVENS (1977), *Measures of Interviewer Bias and Variance*, «Jour. of Marketing Research», 14, pp. 337-343.

BAILAR B.A. e P.P. BIEMER (1984), *Some Methods for Evaluating Nonsampling Error in Household Census and Surveys*, in Rao P.S. e J. Sedransk (eds.), «W.G. Cochran's impact on Statistics, J. Wiley, New York.

BAILAR B.A. e T. DALENIUS (1970), *Estimating the Response Variance Components of the U.S. Bureau of the Census Survey Model*, Ser. B, pp. 341-360.

BAILEY L., T.F. MOORE e B.A. BAILAR (1978), *An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample*, «Jour. Amer. Stat. Assoc.», pp. 16-23.

COCHRAN W. (1977), *Sampling Techniques*, J. Wiley, New York.

DALENIUS T. (1983), *Error and Other Limitation of Surveys*, in Wright T. (ed.), «Statistical Methods and the Improvement of Data Quality», Academic Press, New York.

DEMING W.E. (1960), *Sample design in buisness research*, J. Wiley, New York.

DEMING W.E. (1964), *On Some of the Contribution of Interpenetrating Networks of Samples*, in Rao (ed.), «Contribution to Statistics presented to Prof. Mahalanobis on the Occasion of His 70th Birthday», Pergamon Press, Calcutta, pp. 57-66.

ECLER A.R. e W.N. HURWITZ (1958), *Response Variance and Biases in Censuses and Surveys*, «Bull. Int. Stat. Inst.», n. 36, pp. 12-35.

FABBRIS L. (1981), *Metodi statistici per l'analisi e il controllo della qualità dei dati sanitari*, in: Bellini P., S. Rigatti Luchini e F. Vian (eds.), «Statistica e Ricerca Epidemologica», CLEUP, Padova, pp. 67-74.

FABBRIS L. (1983), *Una esperienza di stima dell'errore non campionario mediante reintervista e compenetrazione dell'assegnazione degli intervistatori*, «Atti del Convegno SIS», Trieste, vol. I, pp. 515-531.

FABBRIS L. (1991), *Abbinamento tra fonti di errore nella formazione dei dati e misure dell'effetto degli errori sulle stime*, in Bollettino SIS, n. 22, pp. 19-54.

FELLEGI I.P. (1963), *An Analysis of Response Variance*, «Bull. Int. Stat. Inst.», n. 40, pp. 758-759.

FELLEGI I.P. (1964), *Response Variance and its Estimation*, «Jour. Amer. Stat. Assoc.», n. 59, pp. 1016-1041.

FELLEGI I.P. (1974), *An Improved Method of Estimating the Correlated Response Variance*, «Jour. Amer. Stat. Assoc.», pp. 496-501.

GIUSTI F. (1969), *Su gli errori di osservazione nei censimenti e nelle rilevazioni campionarie*, «Atti del Convegno SIS», Firenze, pp. 345-367.

HANSEN M.H. et al. (1951), *Response Errors in Surveys*, «Jour. Amer. Stat. Assoc.», n. 46, pp. 147-190.

HANSEN M.H., W.N. HURWITZ e M.A. BERSHAD (1961), *Measurement Errors in Censuses and Surveys*, «Bull. Int. Stat. Inst.», n. 38, Vol. II, pp. 359-374.

HANSEN M.H., W.N. HURWITZ e W.G. MADOW (1953), *Sample Survey Methods and Theory*, Vol. I e Vol. II, J. Wiley, New York.

HANSEN M.H., W.N. HURWITZ e L. PRITZKER (1964), *The Estimation and the Interpretation of Gross Differences and the Simple Response Variance*, in Rao (ed.), «Contribution to Statistics presented to Prof. Mahalanobis on the Occasion of His 70th Birthday», Pergamon Press, Calcutta, pp. 111-136.

HANSEN M.H. e J. WAKSBERG (1970), *Research on Non-Sampling Errors in Censuses and Surveys*, «Rev. Int. Stat. Inst.», n. 38, pp. 318-332.

HANSON R.H. e E.S. MARKS (1958), *Influence of the Interviewer on the Accuracy of Survey Results*, «Amer. Stat. Assoc.», n. 53, pp. 635-655.

IACHAN R. (1983), *Nonsampling Errors in Surveys: A review*, «Comun. Statist. Theor. Meth.», 12 (19), pp. 2273-2287.

JONES H.W. (1955), *Investigating the Properties of a Sample Mean by Employing Random Subsample Means*, «Jour. Amer. Stat. Assoc.», n. 51, pp. 54-83.

KISH L. (1962), *Studies of Interviewer Variability for Attitudinal Variables*, «Jour. Amer. Stat. Assoc.», n. 57, pp. 92-115.

KISH L. (1965), *Survey Sampling*, J. Wiley, New York.

KOCH G. (1973), *An Alternative Approach to Multivariate Response Error Models for Sample Survey Data With Applications to Estimators Involving Subclass Means*, «Jour. Amer. Stat. Ass.», n. 68, pp. 906-913.

KOOP J.C. (1960), *On Theoretical Questions Underlying the Technique of Replicated or Interpenetrating Samples*, «Proc. Soc. Stat. Amer. Stat. Assoc.», pp. 196-205.

LESSLER J.T. e R.A. KULKA (1983), *Reducing the Cost of Studing Survey Measurement Error: Is a Laboratory Approach the Answer?*, in T. Wright (ed.), «Statistical Methods and the Improvement of Data Quality», Academic Press, New York.

MAHALANOBIS P.C. (1946), *Recent Experiments in Statistical Sampling in the Indian Statistical Institute*, «Jour. Roy. Stat. Soc.», n. 109, pp. 326-370.

O'MUIRCHEARTAIG C.A. (1977), *Response Errors*, in C.A. O'Muircheartaig and C. Payne (eds.), «The Analysis of Survey Data», Vol. 2, J. Wiley, New York.

O'MUIRCHEARTAIG C.A. (1983), *Statistical Methods of Assessing the Quality of Survey Data*, «Atti del Convegno SIS», Trieste, Vol. I, pp. 79-102.

PRITZKER L. e R. HANSON (1962), *Measurement Errors in the 1960 Census of Population*, «Proc. of the Soc. Stat. Sect. A.S.A.».

SIGNORE M. (1988a), *Stima dell'errore di misura: alcune riflessioni sui problemi teorici e pratici per l'applicazione ad indagini su larga scala*, «Atti della XXXIV Riunione Scientifica della SIS», Siena, Vol. 2, Tomo 1, pp. 193-200.

SIGNORE M. (1988b), *Evaluation of the Interviewer's Influence on the Quality of the 1985 Sports and Holidays Survey Data*, «Proceedings of the First Conference of the International Association for Official Statistics», Roma, pp. 252-255.

SIGNORE M. (1989), *Stima dell'effetto intervistatore: estensione del metodo della compenetrazione del campione alle indagini ISTAT sulle famiglie*, ISTAT, documento interno.

SIGNORE M. (1990), *Valutazione dell'effetto intervistatore nell'indagine sulle forze di lavoro*, ISTAT, documento interno.

STATISTICS CANADA (1978), *A Compendium of Methods of Error Evaluation in Censuses and Surveys*, Catalogue 13-564 E occasional, a cura di J.F. Gosselin, B.N. Chinnappa, P.D. Ghangurde e J. Tourgny. pp. 80-89.

TENENBEIN A. (1984), *Cochran's Contributions to Errors of Measurement in Statistics*, in Rao P.S. e J. Sedransk (eds.), «W.G. Cochran's Impact on Statistics», J. Wiley, New York.

UNITED NATIONS (1982), *National Household Survey Capability Programme. Non-sampling Errors in Household Surveys: Sources, Assesment and Control: Preliminary version*, DP/UN/INT-81-041/2, New York.

U.S. Bureau of the Census (1968), *Evalution and Research Program of the U.S. Censuses of Population and Housing, 1960: Effects of Interviewers and Crew Leaders*, Series ER 60, n. 7, Washington D.C.

CAPITOLO 8 - L'ARCHIVIO DI QUALITÀ

1. Il patrimonio informativo dell'indagine

Nel volume si è assunta, per il «controllo di qualità», un'ottica «globale» e «di processo»; globale poiché il controllo deve essere esteso a tutti gli aspetti dell'indagine (dalla programmazione alla diffusione dei risultati) e di processo in quanto il controllo deve isolare gli errori peculiari alle singole fasi ma nel contesto più generale delle loro interazioni.

In questa logica, le informazioni necessarie alla validazione dei risultati coincidono con il «patrimonio informativo» dell'indagine.

Quest'ultimo è costituito da molteplici fonti e da informazioni di tipo quantitativo o qualitativo: la documentazione sulla progettazione dell'indagine (definizioni, classificazioni, schemi E/R, i diversi piani di campionamento, di registrazione, di elaborazione, i manuali e le norme di istruzione, il calendario, il questionario ecc.), i risultati della rilevazione, i risultati del controllo delle fasi del processo di produzione (registrazione, revisione, ecc.), i documenti accessori di rilevazione, le relazioni degli ispettori e degli uffici regionali ed infine la documentazione amministrativa.

Allo scopo di gestire in maniera efficiente tale massa di dati e per effettuare rapidamente le necessarie analisi statistiche è conveniente riportare su supporto informatico il massimo possibile delle informazioni di cui sopra, eventualmente trasformando in quantitative quelle qualitative (ad esempio il contenuto delle relazioni degli ispettori e degli uffici regionali Istat).

Da tali fonti sono stati prodotti nelle diverse fasi dell'indagine (cfr. i precedenti capitoli), indicatori ed analisi specifiche della qualità dei dati.

L'insieme di questi elaborati costituisce l'archivio di qualità dell'indagine, che può essere considerato come la base informativa per la determinazione del profilo dell'errore dell'intera rilevazione, per il controllo della procedura nelle differenti fasi e la programmazione di indagini future.

Il profilo dell'errore è dato dal complesso delle analisi per fase ed è sinteticamente rappresentato dalle informazioni sull'errore delle variabili di studio; il controllo delle singole fasi si attua mediante le analisi specifiche e/o gli indicatori sintetici, riportati nei diversi capitoli del volume. Mediante tali verifiche si determinano e si quantificano le carenze dell'indagine e possono essere prodotti i metadati di qualità ad uso degli utenti finali.

Le informazioni desumibili dall'archivio sono rilevanti per il miglioramento e la revisione delle procedure della medesima indagine in tempi successivi o per la programmazione del disegno di una nuova indagine; in particolare per la ripartizione del bilancio tra rilevazione principale ed indagini di controllo, per la progettazione del questionario, per la programmazione dell'attività di istruzione e di supervisione della rete di rilevazione.

L'archivio di qualità in senso stretto, cui d'ora in avanti faremo riferimento, è invece costituito da un sottoinsieme di tutte le informazioni disponibili, più precisamente dal complesso degli indicatori, individuato nei capitoli precedenti, accessibili su supporto informatico e suscettibili di ulteriori elaborazioni di tipo statistico; essi possono essere considerati un riassunto, una sintesi, della misurazione dei differenti aspetti della qualità dei dati.

2. L'Archivio di qualità

È stato già osservato nel Capitolo 1 che il contenuto dell'archivio deve essere stabilito ex ante, nel piano dei controlli e che la sua efficienza e completezza, rispetto a quanto programmato, dipende dal sistema di codici identificativi che legano le informazioni disponibili, per le medesime unità, nelle differenti fonti. Gli obiettivi e l'analiticità dei dati provenienti dalle fonti informative, determinano la struttura dell'archivio, ovvero gli indicatori ed i livelli di controllo.

In generale, l'archivio può essere concettualmente distinto, in tre parti: indicatori relativi alle variabili, alle fasi ed alla rete di rilevazione.

L'archivio delle variabili

Per le indagini campionarie, l'archivio delle variabili, oltre agli indicatori relativi all'errore non campionario, conterrà anche quelli derivanti dal disegno di campionamento; in particolare per ciascuna variabile saranno presenti nell'archivio:

- la stima
- l'errore campionario
- gli indicatori dell'efficienza del disegno campionario, «Deff» e «roh»
- l'errore minimo di registrazione e quello derivante dal controllo campionario della registrazione
- gli indicatori di mancata risposta parziale
- il tasso di risposta proxy per la singola variabile, se disponibile sul questionario
- il tasso di correzioni per variabile apportate dal piano di compatibilità

- l'eventuale distorsione e la varianza correlata di risposta stimata mediante indagini di controllo o compenetrazione del campione.

La gran parte dei suddetti indicatori possono essere calcolati a regime, mentre altri derivano da controlli particolarmente accurati svolti nel corso della rilevazione (l'indicatore di risposta proxy sulla singola variabile) o dalle indagini di controllo.

Poiché le utilizzazioni possibili dell'archivio «variabili» riguardano sia la conoscenza dell'affidabilità dei dati, sia la programmazione di future indagini, i livelli di controllo più opportuni sono il complesso della rilevazione ed i principali domini territoriali, organizzativi e di diffusione dei dati.

L'archivio della rete

L'esistenza di un archivio centralizzato della rete periferica, alimentato dagli archivi delle singole rilevazioni, non esime i responsabili di queste ultime dall'analisi di tale aspetto della qualità dei dati. Come è stato più volte richiamato nel corso del volume, infatti, la validazione dei risultati non può prescindere dall'analisi dei dati sulla qualità di questa fase insieme con quelli relativi a tutte le fasi del processo di produzione. Inoltre, per assicurare la gestibilità dell'archivio centralizzato, gli indicatori che ad esso vengono ceduti sono meno numerosi ed analitici di quelli disponibili per la singola rilevazione, su cui è conveniente, quindi, condurre un più approfondito esame.

La scelta di un livello minimo di controllo, scaturisce da un compromesso tra l'analiticità dell'informazione e la capacità di gestire la medesima; in ogni caso è conveniente predisporre archivi con dati relativi a livelli più disaggregati (ad esempio, per le indagini campionarie, il rilevatore), anche se l'analisi viene condotta a livelli superiori. Deve comunque essere possibile risalire, mediante gli appropriati identificatori ed una razionale organizzazione delle informazioni elementari, ai livelli di controllo più disaggregati.

Per ciascun livello di controllo (ad esempio il rilevatore, il comune, la regione), l'archivio della rete conterrà gli indicatori, diretti od indiretti, dei differenti aspetti della fase di rilevazione: quelli che è possibile costruire sulla base delle relazioni degli ispettori o degli uffici regionali Istat e quelli relativi alle mancate risposte totale e parziali, all'intervista ed ai risultati del piano di compatibilità, già presentati nel Capitolo 3 e 5.

In dettaglio:

A) Per le mancate risposte totali, gli indicatori:
   — dell'errore di lista
   — di rifiuto dell'intervista
   — di mancato contatto
   — di mancata risposta totale
   — della dimensione media delle famiglie sostituite e sostitutive.

B) Per le mancate risposte parziali, gli indicatori:
   — di qualità del materiale raccolto
   — delle incongruenze
   — di rifiuto
   — di efficacia dell'intervista e della raccolta dati.

C) Per l'intervista, gli indicatori:
   — di risposta proxy
   — di effettuazione dell'intervista
   — di durata media e dei giorni in cui sono state condotte le interviste
   — di errore nei codici identificativi
   — di rispetto dei tempi del calendario (dai documenti amministrativi) per i comuni.

D) Quali indicatori indiretti:
   — il numero di correzioni per record
   — il numero di record corretti
   — il numero medio di correzioni per gruppi significativi di variabili

E) Le caratteristiche strutturali della popolazione rilevata, per domini territoriali significativi:
   — l'indice di mascolinità
   — il numero medio di componenti per famiglia
   — l'indice di vecchiaia
   — l'indice di dipendenza
   — le percentuali di individui per classi significative di età.

F) A tali indicatori, và aggiunta, laddove sia disponibile, una sintesi dell'errore di risposta delle variabili del questionario; in genere, tale indicatore è disponibile al livello minimo di comune.

L'archivio delle fasi

Gli indicatori di cui sopra, considerati a livello aggregato, costituiscono la base informativa per l'analisi delle fasi in cui è stata distinta la rilevazione statistica; vanno altresì aggiunti gli indicatori che non sono presenti nell'elenco, ma sono stati indicati nei capitoli precedenti, ovvero:

I) gli indicatori riguardanti le regole di compatibilità e correzione;
II) l'indicatore di qualità del materiale disponibile, l'errore minimo di registrazione e l'indicatore di efficacia dell'indagine;
III) il tasso di campionamento effettivo;
IV) gli indicatori di errore di codifica;
V) gli indicatori del piano di tabulazione.

L'Analisi dell'archivio

I tre archivi di qualità costituiscono matrici di dati «osservazioni / variabili» che possono essere analizzate con le consuete tecniche di analisi univariata o multivariata, in funzione degli obiettivi prefissati.

In particolare, l'esame dell'archivio della rete di rilevazione, può risultare utile per la programmazione di iniziative (quali i controlli e le ispezioni, particolari corsi di istruzione) *mirate* a singole realtà, mediante l'individuazione di osservazioni *anomale* e di gruppi omogenei di unità. L'analisi della serie storica degli indicatori per le medesime unità, inoltre, può costituire una misura dell'efficacia degli interventi effettuati.

APPENDICE

1. Il sistema di controllo dell'indagine sulla salute 1983

L'esempio che segue è relativo all'integrazione di fonti diverse per il controllo di un'indagine, realizzato però *a posteriori*, ovvero non programmato nella fase di progettazione. Esso è tratto da M. Masselli *La qualità dei dati nell'indagine Istat sulla salute 1983*, in Atti del Convegno «Salute e ricorso ai servizi nel Veneto», Regione Veneto, novembre 1987.

Il sistema di controllo dell'indagine Istat sulla salute degli italiani, anno 1983, si avvale della possibilità di integrare le informazioni provenienti da fonti diverse e di riferirle ai differenti livelli di controllo (individuo, famiglia, rilevatore, USL, comune ecc.) mediante un adeguato sistema di identificazione di tali unità. Le informazioni sono desumibili dalle seguenti fonti:

- A) norme e documenti amministrativi derivanti dal piano di rilevazione
- B) questionario base
- C) procedura di controllo e correzione
- D) foglio notizie del rilevatore
- E) questionario per la reintervista di un campione di capifamiglia.

In particolare il questionario base può essere suddiviso in blocchi di informazione:

- B1) codici identificativi
- B2) composizione della famiglia intervistata
- B3) variabili individuali oggetto di studio
- B4) individuo rispondente per ciascun foglio individuale
- B5) composizione della famiglia sostituita
- B6) ragioni della sostituzione
- B7) durata, data ed ora di inizio dell'intervista
- B8) difficoltà delle domande per ogni capitolo del questionario

Le informazioni desunte dal foglio notizie del rilevatore possono invece essere scomposte in:

- D1) caratteristiche socio-demografiche del rilevatore
- D2) numero di interviste effettuate

mentre il questionario utilizzato per la reintervista:

- E1) avvenuta intervista (si/no) e modalità della stessa

- E2) alcune delle variabili individuali oggetto di studio.

C'è da osservare che mentre le informazioni desumibili dalle fonti da A a D sono disponibili per tutte le unità di tutti i livelli di controllo considerati nei codici di identificazione, quelle della fonte E, proveniendo da un sub campione, non sono riferibili a tutte le unità dei livelli di controllo desiderati.

Nel Prospetto 8.1 vengono riportati i principali tipi di controllo che è possibile effettuare per l'indagine in esame e le relative fonti di informazione.

Prospetto 8.1 - Il sistema di controllo dell'indagine

CONTROLLO/ANALISI	FONTI
— Quantitativo (consistenza del materiale raccolto)	A B1 D2
— Sistema di identificazione	A B1 B2 D2
— Distorsione nel numero e nella tipologia delle famiglie e degli individui intervistati rispetto a quelli selezionati	B2 B5
— Caratteristiche della rete di rilevazione	D1
— Operato dei singoli rilevatori	B2 B4 B5 B6 B7 C
— Relazione tra caratteristiche dei rilevatori e loro prestazioni	B2 B4 B5 B7 C D1
— Schema dell'intervista	B4 B7
— Liste di estrazione	B6
— Adeguatezza del questionario	B8 C
— Stima del numero di interviste non effettuate e delle modalità di quelle effettuate	E1
— Stima della distorsione dei risultati	B3 E2

RIFERIMENTI BIBLIOGRAFICI

De Marchis M.A. (1988), *Interviewer file of Istat household surveys*, Atti della I Conferenza I.A.O.S., Roma.

Manicardi D., Venturi M. (1988), *Analisi integrata di dati e funzioni nei sistemi informativi statistici*, documento interno Istat.

Masselli M. (1987), *La qualità dei dati nell'indagine Istat sulla salute 1983*, in Atti del Convegno «Salute e ricorso ai servizi nel Veneto», Padova novembre 1987.

PUBBLICAZIONI ISTAT

BOLLETTINO MENSILE DI STATISTICA
La più completa ed autorevole raccolta di dati congiunturali concernenti l'evoluzione dei fenomeni demografici, sociali, economici e finanziari
Abbonamento annuo L. 122.000 (Estero L. 147.000) Ogni fascicolo L. 16.000

INDICATORI MENSILI
Forniscono dati riassuntivi e tempestivi sull'andamento mensile dei principali fenomeni interessanti la vita nazionale
Abbonamento annuo L. 31.000 (Estero L. 37.000) Ogni fascicolo L. 4.000

NOTIZIARI ISTAT
È attualmente in corso una radicale trasformazione della struttura dei "Notiziario ISTAT" per cui, pur essendo stato fissato il prezzo di un singolo fascicolo (L. 1.700) valido per alcuni numeri eccezionali che potranno essere ancora pubblicati, non è previsto un canone di abbonamento.
Le informazioni sul sistema di diffusione sostitutivo dell'abbonamento saranno diffuse quanto prima.

INDICATORI TRIMESTRALI
Conti economici trimestrali
Abbonamento annuo L. 12.000 (Estero L. 14.000) Ogni fascicolo L. 4.000

STATISTICA DEL COMMERCIO CON L'ESTERO
Documentazione statistica ufficiale, a periodicità trimestrale, sul commercio dell'Italia con l'estero; fornisce, per tutte le merci comprese nella classificazione merceologica della tariffa dei dazi doganali, l'andamento delle importazioni e delle esportazioni da e per i principali Paesi
Abbonamento annuo L. 105.000 (Estero L. 119.000) Ogni fascicolo L. 33.000

Abbonamento annuo cumulativo a tutti i periodici, compresa la "Statistica del commercio con l'estero": L. 243.000 (Estero L. 288.000); esclusa la "Statistica del commercio con l'estero" L. 148.000 (Estero L. 180.000)
Gli abbonamenti decorrono dal 1° gennaio anche se sottoscritti nel corso dell'anno. In tal caso l'abbonato riceverà i numeri dell'annata già pubblicati. L'abbonato ai periodici ISTAT ha diritto a ricevere gratuitamente i fascicoli non pervenutigli soltanto se ne segnalerà il mancato arrivo entro 10 giorni dal ricevimento del fascicolo successivo. Decorso tale termine, si spediscono solo contro rimessa dell'importo. Le variazioni di indirizzo devono essere segnalate dall'abbonato per iscritto. Nel sottoscrivere l'abbonamento cumulativo, gli interessati possono chiedere che l'ISTAT provveda, senza ulteriori richieste, all'invio di tutte le pubblicazioni non periodiche non appena liberate dalle stampe, contro assegno o con emissione di fattura, con lo sconto del 30%. Le singole pubblicazioni possono essere richieste direttamente all'Istituto nazionale di statistica (Via Cesare Balbo, 16 - 00100 Roma) versando il relativo importo, maggiorato del 10% per spese di spedizione, sul c/c postale n. 619007.
Tutti i prezzi sono riferiti all'anno 1992.

ANNUARIO STATISTICO ITALIANO - Edizione 1991 - L. 49.000
Sintetizza in semplici tabelle numeriche di facile lettura ed attraverso appropriate note illustrative e rappresentazioni grafiche, i dati fondamentali della vita economica, demografica e sociale e fornisce un quadro panoramico della corrispondente situazione degli altri principali Paesi del mondo.

COMPENDIO STATISTICO ITALIANO - Edizione 1991 - L. 24.000
Sintetizza i risultati delle rilevazioni ed elaborazioni statistiche di maggior interesse nazionale.

ITALIAN STATISTICAL ABSTRACT - Edition 1992 - L. 25.000 *(in corso di stampa)*
Fornisce i principali risultati delle rilevazioni ed elaborazioni statistiche concernenti la situazione sociale ed economica italiana - Edizione in lingua inglese.

I CONTI DEGLI ITALIANI - Vol. 25, edizione 1991 - L. 17.000
Illustra in forma divulgativa i principali aspetti quantitativi dell'economia italiana.

LE REGIONI IN CIFRE - Edizione 1991 - Distribuzione gratuita
Fornisce i dati delle singole regioni e delle due grandi ripartizioni geografiche: Nord-Centro e Mezzogiorno.