



Nota metodologica  
sulla strategia di  
campionamento

a cura di  
Stefano Falorsi



# Nota metodologica sulla strategia di campionamento del sistema nazionale di valutazione delle competenze per le classi seconda e quinta del primo ciclo della scuola primaria

Stefano Falorsi

## 1. Obiettivi

Il Sistema Nazionale di Valutazione (SNV) oggetto della presente nota è parte integrante del più ampio sistema di valutazione delle competenze alfabetiche e numeriche degli studenti della scuola primaria e secondaria condotto dall'INVALSI. Si tratta, a tutti gli effetti, di una rilevazione statistica sugli studenti della scuola primaria di primo grado, delle classi seconda e quinta, ai quali viene somministrato un insieme integrato di prove o *items*, di italiano e matematica, definiti in modo da essere comparabili tra tutti gli Stati appartenenti alla Comunità Europea. Nelle edizioni passate, svolte nei precedenti anni, le prove venivano somministrate direttamente dal personale docente delle scuole stesse, tuttavia per l'edizione di quest'anno si è ritenuto necessario rendere migliore la qualità dei dati raccolti standardizzando il più possibile le modalità di somministrazione delle prove. Ciò dovrebbe, tra l'altro, permettere una maggiore comparabilità dei risultati tra scuola e scuola ed, anche, tra differenti sottopopolazioni di tipo territoriale oppure relative alla tipologia di scuola. Si è deciso, pertanto, di far somministrare le prove da personale qualificato, esterno alla scuola, che avesse il compito di illustrare agli studenti le diverse prove effettuando, anche, la dovuta azione di controllo. I punteggi normalizzati riportati dagli studenti per i diversi items costituiscono, quindi, le *variabili di interesse* oggetto di rilevazione; in particolare la normalizzazione dei punteggi consiste nel dividere, per ciascuna prova, il punteggio assoluto assegnato allo studente per il valore massimo ottenibile per la suddetta prova in modo da arrivare ad un valore compreso tra zero e uno.

La *popolazione di interesse* – ossia l'insieme delle unità statistiche intorno alle quali si intende investigare – è costituita dagli studenti di seconda e quinta elementare iscritti nell'anno scolastico 2008-2009. Più precisamente si tratta di due distinte popolazioni di interesse costituite dagli alunni di seconda e quinta elementare rispettivamente di numerosità  $N_2$  e  $N_5$ . I test somministrati alle classi seconde, costituiscono, una valutazione sulla preparazione in entrata mentre quelli relativi agli alunni di quinta consentono una valutazione sulla preparazione raggiunta dagli stessi alla conclusione del primo ciclo di istruzione primaria.

Per ciascuna delle due popolazioni di interesse a partire dal punteggio assegnato a ciascuno studente, per ogni prova, si definiscono, infine, degli indici sintetici della distribuzione aggregando opportunamente i punteggi per tutte le unità della popolazione. Gli indici più importanti sono la media dei punteggi normalizzati e il numero di studenti che si trovano compresi tra i vari percentili

(generalmente i quartili) della distribuzione definiti a livello nazionale su tutta la popolazione di interesse. Tali quantità costituiscono quindi i *parametri di interesse* oggetto di stima.

I *domini di stima*, ossia le sottopopolazioni con riferimento alle quali si vogliono produrre le stime dei parametri di interesse, al fine di poter effettuare dei confronti tra le diverse sottopopolazioni, sono costituite dalle regioni geografiche, dalle ripartizioni geografiche e dall'intero territorio nazionale. A tale proposito occorre aggiungere che l'indagine ha, anche, la finalità secondaria di produrre stime - il più possibile attendibili anche se nel rispetto di limiti di spesa accettabili - dei parametri di interesse riferiti a ciascuna delle scuole campione.

## 2. Parametri di interesse

Si consideri la generica prova, denotata con l'indice  $b$  (per  $b=1, \dots, B$ ), costituita da un insieme di quesiti oggetto di valutazione finalizzati a valutare la preparazione dell'alunno su una determinata materia. Per ogni prova  $b$  e per ciascuna delle due popolazioni di riferimento (alunni di seconda e di quinta elementare) si indichi con:  $\lambda$  l'indice relativo al tipo di parametro oggetto di stima ( $\lambda = 1, \dots, \Lambda$ );  ${}_{b\lambda}y_{di}$  il valore della variabile di interesse  $y$  osservato sull'alunno  $i$  appartenente al dominio di stima  $d$  ( $d=1, \dots, D$ );  ${}_{b\lambda}\theta_d$  il parametro di interesse del tipo  $\lambda$  riferito al dominio  $d$  nell'ambito della prova  $b$ . I principali domini di stima considerati dall'indagine sono costituiti dalle regioni geografiche, pertanto, nel seguito a meno che non venga appositamente specificato per il generico dominio di stima  $d$  si intenderà la regione. Ciò premesso, di seguito si esplicita la natura statistica e la corrispondente forma funzionale dei parametri di interesse per ciascun tipo di parametro  $\lambda$  ( $\lambda = 1, \dots, \Lambda$ ). In particolare i parametri di interesse sono la *media* dei punteggi normalizzati e la *frequenza relativa* di alunni compresi tra due quantili della distribuzione dei punteggi. I diversi parametri considerati, per ciascuna prova  $b$  ( $b=1, \dots, B$ ) e per ogni dominio di stima  $d$  ( $d=1, \dots, D$ ), sono esprimibili mediante la seguente forma funzionale

$${}_{b\lambda}\theta_d = \frac{1}{N_d} \sum_{i=1}^{N_d} {}_{b\lambda}y_{di} \quad (1)$$

in cui  $i$  denota il generico alunno (di seconda o di quinta al variare della popolazione investigata),  $N_d$  è la numerosità della corrispondente popolazione di studenti ( $N_d = N_{II,d}$  per gli alunni di seconda e

$N_d = N_{V,d}$  per gli alunni di quinta) del dominio di stima  $d$  e  $y_{di}$  è il valore della variabile di interesse, riferita alla prova  $b$  e al tipo di parametro  $\lambda$ , osservata sulla  $i$  esima unità del dominio  $d$ . I diversi parametri considerati si differenziano in base alla definizione della variabile  $y_{di}$ . Nel caso della media, indicata con il simbolo  $\theta_d$ , la variabile di interesse  $y_{di} \equiv \theta_d$  è costituita dal punteggio normalizzato assegnato all' $i$ -esimo studente. Quando si considerano, invece, le frequenze relative, indicate con i simboli  $\theta_d, \dots, \theta_d$ , si hanno  $\Lambda - 1$  variabili strumentali, indicate come  $y_{di}, \dots, y_{di}$  essendo la generica di esse una variabile dicotomica pari a 1 se l'unità  $i$  è compresa tra due specifici quantili,  $(Q_{k-1} \text{ --| } Q_k)$  ( $k=0,1,\dots,K$ ) dove si è indicato con  $Q_{\min}$  e  $Q_{\max}$  rispettivamente il valore minimo e quello massimo della distribuzione dei punteggi. Il caso più frequente nell'indagine in esame è quello in cui i quantili presi in esame coincidono con i quartili e le relative classi sono quindi  $(Q_{\min} \text{ --| } Q_1)$ ,  $(Q_1 \text{ --| } Q_2)$ ,  $(Q_2 \text{ --| } Q_3)$  e  $(Q_3 \text{ --| } Q_{\max})$ . Pertanto  $y_{di}$  indica le unità comprese entro il primo quartile e .. così via fino a  $y_{di}$  che è la variabile indicatrice relativa alle unità che superano il terzo quartile della distribuzione. Quando si considerano i quartili il numero di parametri da stimare per ciascun dominio  $d$  e per l'intero territorio nazionale è pari a  $4 \cdot B$  mentre più in generale tale quantità è pari a  $\Lambda \cdot B$ .

### 3. Disegno di campionamento

#### 3.1. Premessa

Le rilevazioni condotte negli anni precedenti prevedevano la somministrazione delle prove da parte dei docenti delle scuole stesse; tale circostanza che determinava un costo molto basso della rilevazione, a fronte di dati non sempre comparabili e di buona qualità, consentiva di condurre un'indagine esaustiva su tutti gli alunni della popolazione di interesse.

Il passaggio all'utilizzazione di personale qualificato esterno alla scuola, se da un lato comporta un qualità più alta dei dati raccolti dall'altra introduce costi notevolmente maggiori, anche legati alla necessità di viaggiare per raggiungere le scuole interessate alla rilevazione. Per le suddette ragioni si è resa necessaria la selezione di due campioni casuali di alunni, rispettivamente appartenenti alle classi seconda e quinta elementare. Inoltre, per ridurre i costi di viaggio e di organizzazione della rilevazione sul campo – e quindi migliorare efficienza delle stime a parità di costi oppure ridurre i

costi a parità di efficienza campionaria - è stato necessario adottare un campionamento complesso a più stadi di selezione in cui al primo stadio viene estratto un campione di scuole. Per ovvi motivi di costi e praticabilità operativa, è stato selezionato il medesimo campione di scuole sia per la seconda che per la quinta elementare. Il disegno di campionamento adottato consente l'ottenimento di due vantaggi. Il primo consiste nella possibilità di ridurre e tenere sotto controllo il numero delle scuole coinvolte nella rilevazione, rispetto a quello ottenibile con un campione casuale semplice di pari numerosità in termini di alunni. Con quest'ultimo tipo di campionamento, infatti, potrebbero essere, anche, coinvolte tutte le scuole della popolazione risultando, inoltre, molto variabile il numero di alunni campione per ciascuna scuola estratta. Il secondo vantaggio riguarda la possibilità di selezionare il campione dalla lista aggiornata degli iscritti disponibile presso ciascuna scuola campione ove non fosse disponibile una lista centralizzata e aggiornata di tutti gli alunni della popolazione.

Altro elemento fondamentale di cui si è tenuto conto nella progettazione del disegno è stato quello relativo all'obiettivo di produrre stime regionali dei parametri di interesse al fine di poter effettuare confronti tra le performance medie rilevate nelle diverse regioni. I principali domini di stima sono costituiti dalle regioni essendo, quindi, indicata con  $d$  la generica di esse. Ciò ha comportato l'adozione di un disegno stratificato per regione in cui la numerosità campionaria regionale, in termini di alunni e scuole, è stata definita in modo da tenere sotto controllo gli errori di campionamento attesi delle stime dei parametri di interesse a livello regionale. Altre variabili di stratificazione utilizzate, non legate alla necessità di tenere conto dei domini di stima, che consentono di ridurre la variabilità campionaria delle stime a parità di numerosità sono: la tipologia di scuola, secondo la classificazione statale e non statale e l'ampiezza della scuola espressa in termini di numero di alunni iscritti.

In particolare, gli strati di base sono stati costruiti incrociando le 20 regioni geografiche con la tipologia di scuola, secondo la classificazione statale o non statale. Una volta suddivise le scuole negli strati di base sono state, poi, ulteriormente stratificate per *ampiezza* della scuola. Vale la pena osservare che, per quanto riguarda la variabile ampiezza, poiché l'indagine si riferisce sia agli alunni di seconda che a quelli di quinta elementare ed essendo necessario, per ovvi motivi di costo, selezionare un campione di scuole unico sul quale svolgere le prove sia per la seconda che per la quinta elementare, l'ampiezza di ciascuna scuola è stata definita come media degli alunni iscritti alla seconda e di quelli iscritti alla quinta elementare.

### 3.2. Caratteristiche generali del disegno

In questo paragrafo si illustra il disegno di campionamento riferito a ciascuno strato di base  $T_i$

( $l=1, \dots, L$ ). In proposito, occorre osservare che ogni strato di base  $T_l$  ( $l=1, \dots, L$ ) è comune a più domini di studio; ad esempio, lo strato di base formato dalle scuole elementari statali della regione Lazio fa parte sia della regione Lazio che dell'intero territorio nazionale. Conseguentemente, non è possibile adottare un distinto disegno di campionamento per ogni dominio, in quanto per ogni strato di base  $T_l$  si avrebbero tanti disegni di campionamento quanti sono i domini che lo contengono; è necessario pertanto che il disegno di campionamento venga definito per ogni strato di base  $T_l$ .

Le principali caratteristiche metodologiche del disegno campionario riferito a ciascuno strato di base  $T_l$  ( $l=1, \dots, L$ ) sono:

- stratificazione delle scuole in funzione della sola dimensione, espressa in termini di alunni iscritti;
- suddivisione delle scuole nei due insiemi: AR (scuole Auto Rappresentative), che include le scuole la cui popolazione è uguale o superiore ad una prefissata soglia  $v_l$ ; NAR (scuole Non Auto Rappresentative), che comprende le scuole la cui popolazione è inferiore alla suddetta soglia;
- ciascuna scuola AR costituisce strato a sé stante;
- le scuole dell'insieme NAR sono suddivise in strati di dimensione approssimativamente costante, in termini di ampiezza (media alunni di seconda e quinta), dopo essere state ordinate secondo una graduatoria decrescente in funzione dell'ampiezza delle stesse;
- il disegno di campionamento inerente all'insieme AR è del tipo ad uno stadio stratificato, in cui le unità primarie coincidono con le unità finali di campionamento, ossia gli alunni di seconda e quelli di quinta elementare;
- il disegno di campionamento relativo all'insieme NAR è del tipo a due stadi stratificato; le unità primarie sono le scuole, mentre le unità secondarie sono gli alunni;
- selezione, senza reimmissione, di un numero costante,  $m_{lh} = \bar{m}_l$ , di scuole campione in ogni strato  $h$  ( $h = 1, \dots, N_{AR} H_l$ ) dell'insieme NAR;
- assegnazione di un numero minimo,  $\bar{n}_l$ , di alunni da intervistare in ciascuna scuola campione;
- allocazione del campione di alunni tra gli strati in modo da rispettare la condizione di auto ponderazione;

Le ragioni che sottendono la scelta di questa particolare forma di disegno di campionamento sono determinate:

- dal desiderio di aumentare il livello di precisione delle stime, attraverso la suddivisione delle scuole in AR e NAR;
- dall'esigenza di conseguire vantaggi dal punto di vista organizzativo ed economico, attraverso la selezione di scuole (grappoli di alunni);
- la necessità operativa di selezionare un numero,  $\bar{n}_1$ , approssimativamente costante di alunni per le scuole medio piccole appartenenti all'insieme NAR e un numero di alunni maggiore uguale di  $\bar{n}_1$  per le scuole più grandi, appartenenti all'insieme AR.

### 3.3. Procedura di formazione del campione

La procedura di formazione del campione consta di due fasi. La *prima fase* riguarda la definizione delle numerosità campionarie ed in particolare delle quantità  $\bar{m}_l$ ,  $\bar{n}_1$  e del numero complessivo,  $n_1$ , di alunni da selezionare per ciascuno strato di base  $T_l$  ( $l=1, \dots, L$ ) attraverso criteri che tengono conto sia di aspetti economici ed operativi che di fattori legati all'efficienza attesa delle stime; nella *seconda fase*, avendo definito i valori delle suddette quantità è possibile attuare il disegno di campionamento, così come descritto nel precedente paragrafo, nell'ambito di ciascuna strato di base  $T_l$  ( $l=1, \dots, L$ ).

Occorre sottolineare che il processo che conduce alla definizione delle quantità  $\bar{m}_l$ ,  $\bar{n}_1$  e  $n_1$  è complesso in quanto la modifica di una di esse ha diverse implicazioni anche sulle altre. Per una migliore comprensione dell'intera procedura di formazione del campione si è ritenuto utile descrivere innanzitutto la seconda fase.

Prima di passare al successivo paragrafo occorre rilevare che nel disegno di campionamento adottato le quantità campionarie  $\bar{n}_1$  e  $n_1$  si riferiscono indifferentemente al campione di alunni della seconda elementare o a quello della quinta elementare.

### 3.4. Formazione del campione per ciascuno strato di base

Definite le quantità  $\bar{m}_l$ ,  $\bar{n}_l$  e  $n_l$  (sulla base della procedura che verrà descritta nel successivo paragrafo 3.5) con riferimento a ciascuno degli strati di base  $T_l$  ( $l=1, \dots, L$ ), si passa all'attuazione del disegno, in ciascuno strato di base, mediante le seguenti fasi operative:

(a) calcolo della soglia  $v_l$  ( $l=1, \dots, L$ ) sulla base della relazione

$$v_l = \frac{\bar{n}_l}{f_l}, \quad (2)$$

in cui  $f_l = n_l / N_l$ , essendo  $N_l$  il numero di alunni appartenenti allo strato di base  $T_l$ . Si ritiene utile precisare che l'utilizzo della (2) per il calcolo della soglia garantisce, applicando la condizione di auto ponderazione formalizzata nel successivo punto (e), per ciascuna scuola dell'insieme AR una numerosità campionaria, in termini di alunni, uguale o superiore ad  $\bar{n}_l$ ;

(b) suddivisione delle  $M_l$  scuole di  $T_l$  ( $l=1, \dots, L$ ) negli insiemi AR e NAR;

(c) ordinamento delle scuole NAR secondo una graduatoria decrescente basata sull'ampiezza delle stesse;

(d) suddivisione delle scuole NAR in strati aventi ampiezza<sup>1</sup>  $N_{lh} = \bar{N}_l$  ( $h=1, \dots, {}_{nar}H_l$ )

(approssimativamente) costante in termini di alunni iscritti, essendo

$$\bar{N}_l = \bar{m}_l v_l;$$

(e) attribuzione a ciascuna scuola AR del numero,  $n_{lh}$ , di alunni campione espresso da

$$n_{lh} = N_{lh} f_l \quad (h=1, \dots, {}_{ar}H_l),$$

essendo  ${}_{ar}H_l$  il numero complessivo di strati dell'insieme AR ed  $N_{lh}$  il numero di alunni nello strato  $h$ . Si ritiene utile sottolineare che nel caso delle scuole AR, in cui ciascuna scuola  $c$  fa strato a se stante, si ha  $n_{lh} \equiv n_{lhc}$ ,  $N_{lh} \equiv N_{lhc}$  e la precedente relazione può essere, anche, riscritta come

---

<sup>1</sup> Data la necessità di selezionare un unico campione di scuole sia per le prove di seconda che per quelle di quinta le quantità che definiscono l'ampiezza di scuola e di strato sono ottenute come media del numero degli alunni iscritti alla seconda e alla quinta elementare.

$$n_{lhc} = N_{lhc} f_l$$

- (f) selezione, senza reimmissione<sup>2</sup>, da ciascuno strato  $h$  ( $h = 1, \dots, {}_{nar}H_l$ ) di  $\bar{m}_l$  scuole campione, attribuendo alla  $c$ -esima ( $c = 1, \dots, M_{lh}$ ) scuola dello strato una probabilità di selezione variabile, espressa da  $Z_{lhc} = N_{lhc} / N_{lh}$  ;

- (g) attribuzione a ciascuna scuola campione NAR del numero  $n_{lhc}$  di alunni da selezionare, dato da

$$n_{lhc} = \frac{N_{lh} f_l}{\bar{m}_l} \quad (c = 1, \dots, \bar{m}_l; h = 1, \dots, {}_{nar}H_l).$$

Tenendo conto della procedura di formazione degli strati, sopra descritta, si rileva che per gli strati (e quindi scuole) AR si ha  $n_{lhc} \geq \bar{n}_l$ . La precedente disuguaglianza è direttamente legata alla disuguaglianza tra le corrispondenti quantità di popolazione:  $N_{lhc} \geq \bar{N}_l$ ; per gli strati NAR si ha, invece,  $n_{lhc} \cong \bar{n}_l$ ;

- (h) estrazione del campione di alunni, dalle scuole dell'insieme AR e da quelle dell'insieme NAR, la cui numerosità è stata definita nelle fasi precedenti. Per il campione in oggetto si è utilizzato il metodo di selezione Bernoulli sequenziale che consiste nel generare, per ciascuna unità della popolazione, un numero casuale dalla distribuzione uniforme nell'intervallo (0,1); formata poi una graduatoria crescente o decrescente delle unità della popolazione in base ai suddetti numeri si selezionano le prime  $n_{lhc}$  unità della graduatoria. Questo metodo assegna a ciascun alunno della generica scuola campione  $c$  dello strato  $h$  appartenente allo strato di base  $l$  probabilità di selezione uguali e pari a  $n_{lhc} / N_{lhc}$  ;

### 3.5. Definizione delle quantità necessarie alla formazione del campione per ciascuno strato di base

Fissati gli errori di campionamento ammessi  $\delta_{(b_\lambda \theta_d)}$  delle stime dei parametri  $b_\lambda \theta_d$  ( $b=1, \dots, B$ ;

$\lambda = 1, \dots, \Lambda$  ;  $d=1, \dots, D$ ), per la formazione del campione a livello degli  $L$  strati di base è necessario

<sup>2</sup> Per l'estrazione delle scuole si ricorre alla procedura di selezione sistematica, suggerita da Madow (1949) e Murthy (1967), che presenta le seguenti caratteristiche: (i) assegna una probabilità di inclusione espressa nella forma  $\pi_{lhc} = \bar{m}_l Z_{lhc}$  ; (ii) la sua implementazione è estremamente semplice; (iii) conduce all'ottenimento di stime generalmente più efficienti rispetto a quelle ottenibili con altre procedure di selezione (Cicchitelli *et al.*, 1997; Fabbris, 1991).

definire le quantità  $\bar{m}_l$ ,  $\bar{n}_l$  e  $n_l$  ( $l=1,\dots,L$ ), in modo tale che gli errori di campionamento attesi soddisfino i vincoli  $\delta(b_{\lambda} \theta_d)$  prefissati. Il processo per conseguire tale risultato parte dalla determinazione delle quantità  $\bar{m}_l$ ,  $\bar{n}_l$  sia in base a considerazioni di tipo operativo che di efficienza campionaria. Una volta che sono stati fissati i valori delle quantità  $\bar{m}_l$ ,  $\bar{n}_l$  per la definizione delle numerosità campionarie  $n_l$  ( $l=1,\dots,L$ ) si adotta una procedura iterativa basata su una metodologia di allocazione multivariata.

La scelta di  $\bar{n}_l$  si basa sia su criteri di tipo operativo che legati all'efficienza campionaria delle stime. Esiste, infatti, un *trade off* tra le necessità di tipo operativo, per le quali il valore di  $\bar{n}_l$  dovrebbe essere il più elevato possibile, ed esigenze, legate all'efficienza delle stime, che spingono verso la scelta di un valore il più basso possibile di tale quantità. Ovviamente viene prescelto un valore intermedio definito in base ad un compromesso tra le opposte esigenze. Più precisamente, al diminuire di  $\bar{n}_l$  cresce il numero di unità primarie (le scuole) campione, anche se aumenta, generalmente, l'efficienza campionaria delle stime. Per contro al crescere di  $\bar{n}_l$  si riduce, generalmente, l'efficienza campionaria delle stime anche se diminuisce il numero di unità primarie coinvolte nella rilevazione. La concentrazione della rilevazione in poche unità primarie, comporta, ovviamente, notevoli vantaggi dal punto di vista operativo e dei costi anche se un numero troppo elevato di alunni campione in ciascuna scuola potrebbe determinare una difficoltà di organizzazione della rilevazione nelle scuole.

Per la definizione del numero di unità primarie campione per strato,  $\bar{m}_l$ , si adotta, in genere, il criterio di scegliere tale quantità più piccola possibile. Infatti, tenendo conto di quanto descritto al precedente punto (d), la scelta di  $\bar{m}_l=1$  è la più opportuna dal punto di vista dell'efficienza delle stime, poiché porta alla costruzione di un maggior numero di strati e quindi ad un migliore effetto della stratificazione di primo stadio. Tuttavia in tale caso occorre collassare gli strati (generalmente a coppie) nella fase di stima della varianza campionaria, poiché non è possibile calcolare la varianza in base ad un'unica unità primaria campione.

Una volta prefissate le numerosità  $\bar{m}_l$ ,  $\bar{n}_l$  si utilizza una procedura iterativa che arriva alla determinazione delle numerosità  $n_l$  ( $l=1,\dots,L$ ), mediante l'applicazione della metodologia di allocazione multivariata, *multi dominio, per disegni complessi* proposta nel lavoro di Falorsi S. e Russo A (2003). Questa tecnica generalizza - al caso dei disegni di campionamento a più stadi e di più domini di stima - il metodo suggerito da Bethel (1989) finalizzato alla determinazione della dimensione ottimale in un'ottica multivariata, relativamente al caso di un disegno ad uno stadio stratificato e di un solo dominio di studio.

Più precisamente la metodologia adottata parte dal definire le numerosità campionarie  $n_l$  ( $l=1,\dots,L$ ) per ciascuno strato di base  $T_l$  sotto l'ipotesi, non realistica, di aver adottato un campionamento casuale semplice di alunni in ciascuno strato di base  $T_l$ . A partire da questa soluzione iniziale le scuole appartenenti a ciascuno strato di base vengono ripartite negli insiemi AR e NAR e si calcola una stima *dell'effetto del disegno di campionamento*, o *deff*, per ciascuna stima presa in considerazione nel processo di allocazione. A tale proposito vale la pena ricordare che il *deff* è una statistica frequentemente utilizzata nelle indagini campionarie su larga scala sia *ex-ante*, nella fase di progettazione del disegno di campionamento (per tenere conto, nel calcolo degli errori attesi e nel processo di allocazione della numerosità campionaria, dell'inflazione o deflazione della varianza dovuta al particolare disegno di campionamento considerato), sia *ex-post*, nella fase di analisi di efficienza del disegno adottato (per effettuare valutazioni sull'efficienza del disegno di un campionamento adottato ed operare eventualmente delle correzioni in successive edizioni dell'indagine). Tale statistica, come è noto, è misurata dal rapporto tra la varianza del disegno di campionamento complesso utilizzato rispetto alla varianza di un ipotetico disegno di campionamento casuale semplice, di pari numerosità in termini di unità finali, a quella adottata nel disegno complesso. Pertanto se il *deff* è minore di uno, il piano campionario adottato produce un miglioramento di efficienza rispetto a quella ottenibile mediante un disegno di campionamento casuale semplice; ciò generalmente avviene per i disegni di campionamento stratificati, nella misura in cui le variabili di

stratificazione sono legate alla variabile di interesse. Se, invece, il  $deff$  è maggior di uno ciò indica un peggioramento di efficienza del disegno complesso rispetto al disegno di riferimento; ciò avviene molto spesso per i disegni clusterizzati, anche in presenza di una stratificazione delle unità primarie. Per il disegno sopra descritto, applicato alle scuole, in cui sono stati calcolati i  $deff$  per i diversi strati di base necessari per effettuare l'allocazione del campione si sono osservati, infatti, relativamente ai diversi strati di base  $deff$  minori di uno nell'insieme AR e  $deff$  maggiori di uno per l'insieme NAR. I  $deff$  complessivi ottenuti considerando contemporaneamente parte AR e NAR sono risultati complessivamente maggiori di uno per la maggior parte degli strati di base.

Da quanto detto si evince che nella prima iterazione del processo di allocazione il  $deff$  è posto pari ad uno, poiché è effettuata sotto l'ipotesi di campionamento casuale semplice. Dopo la prima iterazione, le varianze delle stime vengono moltiplicate per i corrispondenti  $deff$ , ed il processo di allocazione viene iterato un certo numero di volte fino a quando non si raggiunge una stabilità nelle numerosità campionarie  $n_l$  ( $l=1, \dots, L$ ) assegnate ai diversi strati di base.

Il processo precedente può essere schematizzato per passi come di seguito illustrato.

- (1) determinazione delle numerosità  $n_l$  ( $l=1, \dots, L$ ), mediante l'applicazione della metodologia di allocazione multivariata, *multi dominio, per disegni complessi*;
- (2) applicazione della fase (a) del par. 3.4;
- (3) applicazione della fase (b) del par. 3.4 e nuovo calcolo dei  $deff$  delle stime considerate nel processo di allocazione;
- (4) ricalcolo delle numerosità  $n_l$  ( $l=1, \dots, L$ ), mediante l'applicazione della metodologia suddetta;
- (5) reiterazione del passo (2), sulla base dei risultati ottenuti in (4);
- (6) reiterazione del passo (3) sulla base dei risultati conseguiti in (5);
- (7) reiterazione dei passi (4), (5) e (6) fino a quando le numerosità  $n_l$  ( $l=1, \dots, L$ ) non si stabilizzano tra un'iterazione e la successiva.

Una volta definite le numerosità  $n_l$  ( $l=1,\dots,L$ ), si può procedere alla formazione del campione nell'ambito di ciascuna area territoriale  $T_l$ , attraverso le fasi già descritte nel par. 3.4.

### 3.6. Disegno realizzato

Nel disegno di campionamento realizzato per l'indagine si è definito per tutte le regioni un valore di  $\bar{n}_l$  pari a 23 per le scuole statali e pari a 8 per le scuole non statali. Preliminarmente è stato necessario eliminare dalla lista di campionamento, contenente tutte le scuole della popolazione, le scuole la cui ampiezza (calcolata come media degli alunni di seconda e di quinta) fosse inferiore ai valori prefissati per  $\bar{n}_l$ ; l'eventuale selezione nel campione di tali scuole avrebbe, infatti, determinato una caduta della numerosità campionaria prefissata in termini di alunni. La lista di campionamento ottenuta dopo tale operazione mantiene, comunque, una buona copertura rispetto alla popolazione oggetto di indagine; si hanno, infatti, coperture del 99.8% e 99.06% rispettivamente per gli alunni di seconda e di quinta elementare.

Per la definizione del numero di scuole campione per strato si è adottata la scelta  $\bar{m}_l=2$  che è quella che conduce alla stratificazione più fine possibile pur mantenendo la possibilità di stimare la varianza campionaria senza la necessità di dover collassare gli strati.

Per determinare le numerosità campionarie  $n_l$  mediante il processo di allocazione multivariata sopra descritto, sono state prese in considerazione le stime dei parametri  ${}_{b_2}\theta_d, \dots, {}_{b_\lambda}\theta_d$ , che esprimono le frequenze relative percentuali di alunni compresi tra due quantili della distribuzione dei punteggi. Non sono state considerate, invece, le stime dei parametri  ${}_{b_1}\theta_d$ , punteggi medi, in quanto esse sono caratterizzate da errori di campionamento molto bassi sia a livello nazionale che regionale. In particolare sono stati considerati due valori tipici delle stime di frequenza relativa percentuale, pari all'5% a livello nazionale e al 10% a livello regionale. Per la stima a livello nazionale è stato fissato un vincolo sull'errore relativo percentuale del 5,2% mentre per le stime regionali tale vincolo è stato posto pari al 9.6%.

Oltre agli errori campionari delle stime che hanno guidato il processo di allocazione, sono stati valutati gli errori di campionamento relativi percentuali attesi per un insieme di stime del tipo  $b_2 \theta_d, \dots, b_\lambda \theta_d$  desunte in base ad informazioni raccolte in una precedente indagine esaustiva dello stesso tipo rivolta agli alunni delle seconde e quarte elementari. Le variabili considerate sono riassunte nel prospetto di seguito riportato.

Prospetto - Schema delle variabili considerate per il calcolo degli errori attesi

Variabili	Quantili	Classe	Anno	Prova
Y1-Y3	< 25°, 50°, >75°	II	04-05	Italiano
Y4-Y8	<10°, 25°, 50°, 75°, >75°	II	04-05	Matematica
Y9-Y14	<10°, 25°, 50°, 75°, 90°, >90°	IV	04-05	Italiano
Y15-Y20	<10°, 25°, 50°, 75°, 90°, >90°	IV	04-05	Matematica
Y21-Y25	<10°, 25°, 50°, 75°, >75°	II	05-06	Italiano
Y26-Y30	<10°, 25°, 50°, 75°, >75°	II	05-06	Matematica
Y31-Y35	<10°, 25°, 50°, 75°, >75°	IV	05-06	Italiano
Y36-Y41	<10°, 25°, 50°, 75°, 90°, >90°	IV	05-06	Matematica

Alla precedente lista si devono aggiungere le variabili Y42, Y43 e Y44. Le variabili Y43 e Y44 si riferiscono alle percentuali che hanno guidato il processo di allocazione multivariata. A Y43 corrisponde una frequenza relativa percentuale del 5% ed un massimo errore relativo percentuale ammesso a livello nazionale inferiore al 5,2%, invece ad Y44 corrisponde una frequenza relativa percentuale del 10% ed un massimo errore relativo percentuale ammesso a livello regionale inferiore al 9.6%. Infine a Y42 corrisponde una frequenza relativa percentuale del 1% senza vincoli sugli errori ammessi come nel caso delle variabili Y1-Y41 desunte dai dati reali. Nella tabella 1 sono riportati i valori minimi, medi e massimi della distribuzione degli errori relativi percentuali ottenuta per ciascuna variabile Y1-Y44 e per ciascun tipo di dominio: Intero territorio nazionale (DOM1); Regioni geografiche (DOM2); Tipologia di scuola: statale /non statale (DOM3); Regione incrociata con la

tipologia di scuola (DOM4) . Come si evince dalla tabella, in relazione alla variabile Y44 gli errori relativi percentuali per il dominio regionale (DOM2) sono compresi tra un minimo del 7.45130 % ed un massimo del 9.5840% con un valore medio del 9.4241% in conformità al vincolo del 9.6% ammesso sull'errore relativo percentuale. Per quanto riguarda, invece la variabile Y43 il valore minimo, medio e massimo dell'errore a livello nazionale (DOM1) ovviamente coincidono e sono pari a 5.199 anche in questo caso in conformità al vincolo imposto nella fase di allocazione. Dall'analisi generale della tabella si osserva che le variabili Y1-Y41 presentano errori più che accettabili sia a livello nazionale che a livello regionale ed anche a livello di tipologia di scuola. Invece le stime al livello dell'incrocio regione per la tipologia di scuola presentano errori estremamente elevati, pertanto, una stima attendibile a tale livello di disaggregazione potrebbe essere ottenuta solamente mediante stimatori indiretti per piccole aree che *prendono forza* dalle aree circostanti. Invece per i domini DOM1, DOM2 e DOM3 si possono utilizzare tranquillamente i più robusti stimatori diretti.

Tabella 1 - Valori minimo, medio, e massimo, degli errori relativi percentuali per variabile e tipo di dominio:  
 DOM1 intero territorio nazionale;  
 DOM2 regioni geografiche;  
 DOM3 tipologia di scuola: statale /non statale;  
 DOM4 regione x tipologia di scuola.

<b>Dominio</b>	<b>Min.</b>	<b>Med.</b>	<b>Max.</b>
<b><u>Variabile Y1</u></b>			
DOM1	2.99236	2.9924	2.992
DOM2	6.52012	10.9635	15.342
DOM3	3.14615	5.9744	8.803
DOM4	6.78426	29.2124	107.593
<b><u>Variabile Y2</u></b>			
DOM1	2.25019	2.2502	2.2502
DOM2	5.14771	8.5192	12.0182
DOM3	2.37998	4.4580	6.5360
DOM4	5.42459	20.9533	55.7634

<b>Dominio</b>	<b>Min.</b>	<b>Med.</b>	<b>Max.</b>
<b><u>Variabile Y3</u></b>			
DOM1	1.87183	1.8718	1.8718
DOM2	4.56116	6.7243	9.8905
DOM3	2.02451	3.1607	4.2970
DOM4	4.89910	12.9061	32.3262
<b><u>Variabile Y4</u></b>			
DOM1	3.55448	3.5545	3.554
DOM2	7.24191	13.8843	20.097
DOM3	3.72305	7.3418	10.960
DOM4	7.55109	40.7857	163.990
<b><u>Variabile Y5</u></b>			
DOM1	2.66329	2.6633	2.6633
DOM2	5.50195	10.5084	15.6921
DOM3	2.80724	5.3685	7.9298
DOM4	5.77715	26.9788	79.4884
<b><u>Variabile Y6</u></b>			
DOM1	2.13068	2.1307	2.1307
DOM2	4.91514	8.0467	11.0713
DOM3	2.26347	4.0982	5.9329
DOM4	5.25279	18.9094	45.9168
<b><u>Variabile Y7</u></b>			
DOM1	2.48765	2.4877	2.4877
DOM2	7.02601	8.6401	10.3294
DOM3	2.66966	4.6133	6.5569
DOM4	7.47344	19.3703	52.4916
<b><u>Variabile Y8</u></b>			
DOM1	4.5861	4.5861	4.5861
DOM2	12.6152	15.3740	21.8367
DOM3	4.9970	7.8687	10.7405
DOM4	12.7889	33.7165	90.3778
<b><u>Variabile Y9</u></b>			
DOM1	3.61163	3.6116	3.612
DOM2	7.66512	13.1050	19.007
DOM3	3.79590	7.1003	10.405
DOM4	7.98285	37.3892	170.746
<b><u>Variabile Y10</u></b>			
DOM1	2.77988	2.7799	2.7799
DOM2	6.06133	10.3704	14.9171
DOM3	2.92315	5.7092	8.4952
DOM4	6.33130	27.4710	70.5973
<b><u>Variabile Y11</u></b>			
DOM1	2.13747	2.1375	2.1375
DOM2	4.71330	8.1060	11.5928
DOM3	2.26486	4.2136	6.1624
DOM4	4.99014	19.5019	55.2438
<b><u>Variabile Y12</u></b>			
DOM1	2.19731	2.1973	2.1973
DOM2	5.34940	7.9641	10.6852
DOM3	2.34369	4.2032	6.0627
DOM4	5.70158	18.9445	60.9756

<b>Dominio</b>	<b>Min.</b>	<b>Med.</b>	<b>Max.</b>
<b><u>Variabile Y13</u></b>			
DOM1	3.34145	3.3415	3.3415
DOM2	8.56758	11.2435	16.2877
DOM3	3.60669	6.0289	8.4511
DOM4	9.14892	24.4378	56.6790
<b><u>Variabile Y14</u></b>			
DOM1	6.8860	6.8860	6.886
DOM2	18.1562	21.8684	30.268
DOM3	7.4639	12.4009	17.338
DOM4	18.1562	47.3962	123.623
<b><u>Variabile Y15</u></b>			
DOM1	3.97876	3.9788	3.979
DOM2	8.17555	15.0595	21.397
DOM3	4.17396	8.1895	12.205
DOM4	8.46749	41.8342	113.226
<b><u>Variabile Y16</u></b>			
DOM1	2.64967	2.6497	2.6497
DOM2	5.51039	10.4150	15.3376
DOM3	2.78911	5.3989	8.0086
DOM4	5.76679	28.7225	76.2652
<b><u>Variabile Y17</u></b>			
DOM1	2.19213	2.1921	2.1921
DOM2	4.71228	8.4511	12.7245
DOM3	2.32402	4.2713	6.2187
DOM4	4.98578	20.5416	57.7260
<b><u>Variabile Y18</u></b>			
DOM1	2.55161	2.5516	2.5516
DOM2	6.34295	9.0791	12.2953
DOM3	2.72348	4.8023	6.8812
DOM4	6.71684	21.2353	58.6480
<b><u>Variabile Y19</u></b>			
DOM1	3.22757	3.2276	3.2276
DOM2	8.89319	10.9252	14.5344
DOM3	3.49933	5.6406	7.7818
DOM4	9.29008	23.6751	64.1712
<b><u>Variabile Y20</u></b>			
DOM1	7.1849	7.1849	7.185
DOM2	18.9794	22.4981	29.419
DOM3	7.7966	12.8838	17.971
DOM4	19.2275	54.1663	175.751
<b><u>Variabile Y21</u></b>			
DOM1	3.99114	3.9911	3.991
DOM2	7.51411	14.8330	21.186
DOM3	4.31107	6.5680	8.825
DOM4	7.51411	46.3393	177.409

<b>Dominio</b>	<b>Min.</b>	<b>Med.</b>	<b>Max.</b>
<b><u>Variabile Y22</u></b>			
DOM1	2.40076	2.4008	2.4008
DOM2	5.03995	9.0269	12.8082
DOM3	2.53918	4.6456	6.7520
DOM4	5.26031	25.6134	93.3420
<b><u>Variabile Y23</u></b>			
DOM1	2.13889	2.1389	2.1389
DOM2	5.04472	8.0736	11.1372
DOM3	2.25194	4.4525	6.6531
DOM4	5.34271	20.2143	55.4753
<b><u>Variabile Y24</u></b>			
DOM1	2.57611	2.5761	2.5761
DOM2	6.52464	9.2982	10.9762
DOM3	2.72615	5.2718	7.8175
DOM4	6.92458	21.8953	53.2826
<b><u>Variabile Y25</u></b>			
DOM1	3.36492	3.3649	3.3649
DOM2	8.02057	11.6139	16.7716
DOM3	3.64033	6.0314	8.4224
DOM4	8.49987	23.5272	59.7684
<b><u>Variabile Y26</u></b>			
DOM1	3.35631	3.3563	3.356
DOM2	7.33521	12.7233	18.672
DOM3	3.55189	6.3991	9.246
DOM4	7.64227	37.3854	114.128
<b><u>Variabile Y27</u></b>			
DOM1	2.46027	2.4603	2.4603
DOM2	5.41782	9.5635	14.1764
DOM3	2.62949	4.5093	6.3891
DOM4	5.68375	26.6550	89.9224
<b><u>Variabile Y28</u></b>			
DOM1	2.14606	2.1461	2.1461
DOM2	5.06096	7.9737	10.5162
DOM3	2.27563	4.1587	6.0417
DOM4	5.34846	19.7047	55.6087
<b><u>Variabile Y29</u></b>			
DOM1	2.51982	2.5198	2.5198
DOM2	6.74141	9.1913	13.3580
DOM3	2.65468	5.3058	7.9569
DOM4	7.15104	22.5004	64.2032
<b><u>Variabile Y30</u></b>			
DOM1	3.49139	3.4914	3.4914
DOM2	9.47202	12.1719	18.7525
DOM3	3.76598	6.3105	8.8551
DOM4	9.97994	25.9420	74.0247
<b><u>Variabile Y31</u></b>			
DOM1	3.71868	3.7187	3.719
DOM2	6.87469	12.9527	17.981
DOM3	4.01376	6.0298	8.046
DOM4	6.87469	38.6591	156.454

<b>Dominio</b>	<b>Min.</b>	<b>Med.</b>	<b>Max.</b>
----------------	-------------	-------------	-------------

**Variabile Y32**

DOM1	2.55195	2.5519	2.5519
DOM2	5.42151	9.5639	14.1477
DOM3	2.71874	4.6661	6.6135
DOM4	5.67032	26.0480	93.3420

**Variabile Y33**

DOM1	2.28801	2.2880	2.2880
DOM2	5.09580	8.6493	12.2759
DOM3	2.43087	4.3887	6.3465
DOM4	5.36822	23.0200	65.8236

**Variabile Y34**

DOM1	2.21960	2.2196	2.2196
DOM2	5.19689	8.2263	11.1963
DOM3	2.34556	4.5266	6.7076
DOM4	5.53830	18.6586	48.9834

**Variabile Y35**

DOM1	4.0395	4.0395	4.039
DOM2	10.1306	13.6409	17.656
DOM3	4.4370	6.7130	8.989
DOM4	10.8266	30.6986	121.482

**Variabile Y36**

DOM1	3.57880	3.5788	3.579
DOM2	7.35925	13.4071	21.204
DOM3	3.85173	5.9621	8.073
DOM4	7.35925	39.1606	134.165

**Variabile Y37**

DOM1	2.77378	2.7738	2.7738
DOM2	5.86365	10.7924	17.5260
DOM3	2.97896	4.8651	6.7512
DOM4	5.86365	31.0926	97.8940

**Variabile Y38**

DOM1	2.44751	2.4475	2.448
DOM2	5.32143	9.5111	14.947
DOM3	2.59409	4.7585	6.923
DOM4	5.60298	27.3127	125.236

**Variabile Y39**

DOM1	2.32124	2.3212	2.3212
DOM2	5.63370	8.6876	15.0687
DOM3	2.44425	4.8377	7.2311
DOM4	5.95854	20.5379	73.4519

**Variabile Y40**

DOM1	4.3754	4.3754	4.375
DOM2	11.9235	16.7439	53.661
DOM3	4.6984	8.1714	11.644
DOM4	12.2098	33.9882	101.986

<b>Dominio</b>	<b>Min.</b>	<b>Med.</b>	<b>Max.</b>
<b><u>Variabile Y41</u></b>			
DOM1	13.0632	13.0632	13.063
DOM2	33.7441	50.1303	72.999
DOM3	14.5393	20.8322	27.125
DOM4	35.8502	99.8723	281.642
<b><u>Variabile Y42</u></b>			
DOM1	10.0707	10.071	10.071
DOM2	29.0655	36.761	37.385
DOM3	10.6068	21.277	31.947
DOM4	30.4168	103.179	297.252
<b><u>Variabile Y43</u></b>			
DOM1	5.1988	5.1988	5.199
DOM2	15.0044	18.9770	19.299
DOM3	5.4755	10.9836	16.492
DOM4	15.7020	53.2640	153.450
<b><u>Variabile Y44</u></b>			
DOM1	2.58175	2.5818	2.5818
DOM2	7.45130	9.4241	9.5840
DOM3	2.71918	5.4545	8.1899
DOM4	7.79770	26.4513	76.2042

Il processo di allocazione multivariata ha portato alla definizione delle numerosità campionarie in termini di scuole e di alunni illustrate nella tabella 2. Come è possibile osservare dalla suddetta tabella si tratta di un compromesso tra un'allocazione uguale tra le regioni e una proporzionale alla popolazione di alunni iscritti in ciascuna regione. Infatti un'allocazione approssimativamente proporzionale è quella che si avvicina all'ottimo per gli errori campionari delle stime a livello nazionale mentre l'allocazione uguale è quella che si avvicina all'ottimo per le stime regionali.

Tabella 2 - Numero di alunni e di scuole campione per regione e totale Italia

Regione	Campione	
	Scuole	Alunni
Piemonte	1200	58
Valle D.	563	52
Lombardia	2003	102
Trentino A. A.	1107	137
Veneto	1207	58
Friuli V. G.	1103	53
Liguria	1125	58
Emilia Romagna	1199	58
Toscana	1191	58
Umbria	1061	48
Marche	1136	52
Lazio	1216	64
Abruzzo	1116	52
Molise	833	34
Campania	1484	80
Puglia	1205	56
Basilicata	1002	45
Calabria	1163	54
Sicilia	1219	59
Sardegna	1134	53
<b>Italia</b>	<b>1261</b>	<b>23267</b>

Si ritiene utile riportare, infine, nella tabella 3 i valori medi, calcolati su tutte le regioni, delle stime delle frequenze relative percentuali e dei corrispondenti deft (ottenuti alla fine del processo di allocazione per ciascuna delle variabili reali considerate) ordinate per valori crescenti delle percentuali considerate. Come è possibile notare dalla tabella il deft cresce in misura proporzionale all'incremento delle percentuali considerate; quindi a valori alti delle percentuali corrispondono valori relativamente più alti dei deft. Tenendo conto di tale relazione, nel processo di allocazione multivariata, i deft delle percentuali relative alle variabili Y42, Y43 (1% e 5%) sono stati posti pari a 1.3 mentre il deft della percentuale relativa alla variabile Y44 (10%) è stato posto pari a 1.4.

Tabella 3 - Percentuali medie ordinate  
per valori crescenti e rispettivi deft per  
ognuna delle variabili considerate

<b>Variabile</b>	<b>perc.</b>	<b>deft</b>
<b>Y21</b>	5.8	1.3
<b>Y41</b>	5.9	1.3
<b>Y15</b>	6.9	1.3
<b>Y20</b>	7.3	1.4
<b>Y14</b>	7.6	1.4
<b>Y31</b>	8.0	1.4
<b>Y9</b>	9.0	1.4
<b>Y4</b>	9.1	1.4
<b>Y36</b>	10.2	1.4
<b>Y26</b>	10.7	1.4
<b>Y10</b>	12.6	1.5
<b>Y1</b>	14.3	1.5
<b>Y37</b>	15	1.5
<b>Y16</b>	16.5	1.5
<b>Y5</b>	16.8	1.5
<b>Y8</b>	17.1	1.5
<b>Y24</b>	18	1.5
<b>Y13</b>	18.1	1.5
<b>Y40</b>	18.1	1.5
<b>Y32</b>	19.2	1.5
<b>Y38</b>	19.3	1.5
<b>Y27</b>	19.3	1.5
<b>Y29</b>	19.6	1.5
<b>Y18</b>	19.9	1.5
<b>Y33</b>	21.1	1.5
<b>Y2</b>	22.5	1.6
<b>Y19</b>	23.2	1.6
<b>Y35</b>	24.1	1.6
<b>Y22</b>	24.5	1.6
<b>Y23</b>	24.8	1.7
<b>Y30</b>	24.9	1.8
<b>Y28</b>	25.1	1.8
<b>Y12</b>	25.8	1.8
<b>Y17</b>	25.9	1.9
<b>Y11</b>	26.5	2.0
<b>Y25</b>	26.7	2.0
<b>Y7</b>	27.2	2.1
<b>Y34</b>	27.3	2.1
<b>Y6</b>	29.5	2.3
<b>Y39</b>	31.2	2.6
<b>Y3</b>	63.1	3.3

## **4. Stimatore**

### 4.1. Premessa

Nei precedenti paragrafi sono stati descritti i principi metodologici che hanno guidato nella scelta del disegno di campionamento adottato nonché i criteri che hanno condotto alla definizione delle numerosità campionarie complessive, ai vari stadi di selezione, e la loro allocazione tra i domini pianificati di stima. Qui di seguito si illustrano, invece, le principali caratteristiche metodologiche della procedura di stima in base al quale sono costruiti i coefficienti di riporto all'universo e le modalità di calcolo degli errori di campionamento assoluti, relativi e dei corrispondenti intervalli di confidenza.

#### 4.2. Procedura di stima

Con riferimento al generico dominio pianificato  $d$  ( $d=1,\dots,D$ ), formato da  $L_d$  strati di base  $T_l$  ( $l=1,\dots,L_d$ ), è utile riformulare il parametro (1), utilizzando la notazione già introdotta per descrivere formalmente il piano campionario adottato. Si ha pertanto

$$b_\lambda \theta_d = \frac{1}{N_d} b_\lambda Y_d \quad (3)$$

essendo

$$b_\lambda Y_d = \sum_{h=1}^{AR} \sum_{i=1}^{H_l} b_\lambda y_{dlhi} + \sum_{h=1}^{NAR} \sum_{c=1}^{M_{lh}} \sum_{i=1}^{N_{lhc}} b_\lambda y_{dlhci} \quad (4)$$

in cui per lo strato di base  $T_l$  incluso nel dominio di stima  $d$ , per l'insieme  $NAR$  si è indicato con:

- $b_\lambda y_{dlhci}$ , la variabile di interesse osservata sull'alunno  $i$  della scuola  $c$  appartenente allo strato  $h$ ;
- $M_{lh}$ , il numero di scuole dello strato  $h$ ;
- $N_{lhc}$ , il numero di alunni - di seconda o di quinta elementare a seconda del parametro considerato - della scuola  $c$  dello strato  $h$ ;

le corrispondenti quantità dell'insieme  $AR$  - ricordando che in  $AR$ , per definizione, si ha  $M_{lh}=1$  e quindi non è necessario utilizzare l'indice  $c$  di scuola ma è sufficiente l'indice  $h$  di strato - sono indicate come:

- $b_\lambda y_{dlhi}$ , la variabile di interesse osservata sull'alunno  $i$  della scuola  $AR$  che costituisce lo strato  $h$ ;
- $N_{lh}$ , il numero di alunni - di seconda o di quinta elementare a seconda del parametro considerato - della scuola  $AR$  che forma lo strato  $h$ .

Lo stimatore di Horvitz-Thompson del parametro (3) è dato da

$$b_\lambda \hat{\theta}_d^* = \frac{1}{N_d} b_\lambda \hat{Y}_d^* \quad (5)$$

$$b_\lambda \hat{Y}_d^* = \sum_{h=1}^{AR} \sum_{i=1}^{\tilde{n}_{lh}} b_\lambda y_{dlhi} \pi_{dlhi}^{-1} + \sum_{h=1}^{NAR} \sum_{c=1}^{\bar{m}_l} \sum_{i=1}^{\tilde{n}_{lhc}} b_\lambda y_{dlhci} \pi_{dlhci}^{-1} \quad (6)$$

in cui per lo strato di base  $T_l$  incluso nel dominio di stima  $d$ , per l'insieme  $NAR$  si è indicato con:

- $\pi_{dlhci}$ , la probabilità di inclusione nel campione dell'alunno  $i$  della scuola  $c$  appartenente allo strato  $h$ ;
- $\bar{m}_l = m_{lh} = 2$ , il numero di scuole campione dello strato  $h$ ;
- $\tilde{n}_{lhc}$ , il numero di alunni campione rispondenti - di seconda o di quinta elementare a seconda del parametro considerato - della scuola  $c$  dello strato  $h$ ;

Le corrispondenti quantità dell'insieme  $AR$ , ricordando che in  $AR$ , per definizione, si ha  $M_{lh} = m_{lh} = 1$ , sono indicate come:

- $\pi_{dlhi}$ , la probabilità di inclusione nel campione dell'alunno  $i$  della scuola  $c$  appartenente allo strato  $h$ ;
- $\tilde{n}_{lh}$ , il numero di alunni rispondenti - di seconda o di quinta elementare a seconda del parametro considerato - della scuola  $AR$  che forma lo strato  $h$ .

E' utile rilevare che sotto l'ipotesi di assenza della mancata risposta totale, il numero di alunni selezionati nelle scuole campione  $AR$  e  $NAR$ , rispettivamente  $n_{lh}$  e  $n_{lhc}$ , coincide con il corrispondente numero di alunni effettivamente rispondenti, rispettivamente indicati come  $\tilde{n}_{lh}$  e  $\tilde{n}_{lhc}$ .

Tenendo presente la procedura di stratificazione ed il meccanismo di selezione degli alunni negli insiemi  $AR$  e  $NAR$  si ha:

$$\pi_{dlhi} = \frac{n_{lh}}{N_{lh}} \quad , \quad \pi_{dlhci} = \bar{m}_l \frac{N_{lhc}^*}{N_{lh}^*} \frac{n_{lhc}}{N_{lhc}} \quad ,$$

Nell'indicare il significato delle quantità  $N_{lhc}^*$  e  $N_{lh}^*$  è necessario ricordare che:

- (a) data la necessità di selezionare un unico campione di scuole sia per le prove di seconda che per quelle di quinta, le quantità che definiscono l'ampiezza di scuola  $N_{lhc}^*$  e di strato  $N_{lh}^*$  sono ottenute come media del numero degli alunni iscritti alla seconda e alla quinta elementare;

- (b) poiché sono state eliminate le scuole più piccole di una certa soglia dimensionale dal processo di stratificazione (cfr. par. 3.6) le quantità  $N_{lh}^*$  ( $h = 1, \dots, ARH_l; h = 1, \dots, NARH_l$ ) non rappresentano tutti gli alunni della popolazione investigata ma solamente quelli appartenenti alle scuole sopra la soglia;
- (c) da quanto detto nei precedenti punti si desume che le quantità  $N_l^*$  e  $N_d^*$  sono differenti dalle corrispondenti quantità  $N_l$  e  $N_d$  ( $l = 1, \dots, L_d; d = 1, \dots, D$ ) essendo:

$$N_l = \sum_{h=1}^{ARH_l} N_{lh} + \sum_{h=1}^{NARH_l} N_{lh} \quad \text{e} \quad N_l^* = \sum_{h=1}^{ARH_l} N_{lh}^* + \sum_{h=1}^{NARH_l} N_{lh}^*$$

$$N_d = \sum_{l=1}^{L_d} N_l \quad \text{e} \quad N_d^* = \sum_{l=1}^{L_d} N_l^* .$$

E' utile riscrivere le quantità  $\pi_{dlhci}^{-1}$  e  $\pi_{dlhi}^{-1}$ , che indicheremo nel seguito come *pesi base*, con i corrispondenti simboli  $D_{dlhi}$  e  $D_{dlhci}$ .

Un certo grado di mancata risposta totale è fisiologico in tutte le indagini campionarie. Come è noto è necessario contenere il più possibile tale fenomeno al fine di evitare effetti discorsivi sulle stime prodotte. Per l'indagine in oggetto la mancata risposta totale degli alunni è risultata molto contenuta, essendo legata essenzialmente all'impossibilità di partecipare al test, per diverse ragioni legate alle assenze. In ogni caso si è operata una correzione per mancata risposta totale moltiplicando i pesi base per l'inverso del tasso di risposta nella scuola. I nuovi pesi sono detti *pesi corretti per mancata risposta totale* e sono espressi in formule come:

$$\tilde{D}_{dlhi} = D_{dlhi} \frac{n_{lh}}{\tilde{n}_{lh}} = \frac{N_{lh}^*}{\tilde{n}_{lh}} \quad , \quad \tilde{D}_{dlhci} = D_{dlhci} \frac{n_{lhc}}{\tilde{n}_{lhc}} = \frac{1}{\bar{m}_l} \frac{N_{lh}^*}{N_{lhc}^*} \frac{N_{lhc}}{\tilde{n}_{lhc}} .$$

Per quanto riguarda la mancata risposta delle scuole campione si è operata la sostituzione delle scuole cadute pertanto, in questo caso, le numerosità progettate coincidono con quelle realizzate. Anche per quanto riguarda la caduta delle scuole si sottolinea il buon profilo di qualità dell'indagine in quanto il numero di scuole cadute è risultato molto ridotto.

Da quanto detto nei precedenti punti (a), (b) e (c) risulta chiaro che la somma dei pesi  $\tilde{D}_{dlhci}$  e  $\tilde{D}_{dlhci}$  non riproduce il totale della popolazione di alunni, di seconda o di quinta, investigata. Per tale ragione il sistema dei pesi relativi al campione degli alunni di seconda e quello relativo al campione di quinta sono stati ulteriormente corretti al fine di produrre due sistemi di pesi, detti *pesi finali*, uno per il campione delle seconde e l'altro per il campione delle quinte, che riproducessero i totali delle rispettive popolazioni investigate a livello regionale. Con riferimento alla generica regione  $d$  ( $d = 1, \dots, D$ ) i pesi finali sono ottenuti come

$$W_{dlhi} = \tilde{D}_{dlhi} \frac{N_d}{\hat{N}_d^*} \quad , \quad W_{dlhci} = \tilde{D}_{dlhci} \frac{N_d}{\hat{N}_d^*}$$

dove

$$\hat{N}_d^* = \sum_{h=1}^{AR} H_1 \sum_{i=1}^{\tilde{n}_{lh}} \tilde{D}_{dlhi} + \sum_{h=1}^{NAR} H_1 \sum_{c=1}^{\bar{m}_1} \sum_{i=1}^{\tilde{n}_{lhc}} \tilde{D}_{dlhici}$$

E' facile verificare che la somma dei pesi finali in ciascun dominio riproduce il totale della popolazione del dominio stesso essendo

$$\hat{N}_d = \sum_{h=1}^{AR} H_1 \sum_{i=1}^{\tilde{n}_{lh}} W_{dlhi} + \sum_{h=1}^{NAR} H_1 \sum_{c=1}^{\bar{m}_1} \sum_{i=1}^{\tilde{n}_{lhc}} W_{dlhici} = N_d$$

Sulla base dei pesi finali è possibile definire lo stimatore finale del parametro di interesse espresso come

$$b_\lambda \tilde{\theta}_d = \frac{1}{N_d} b_\lambda \tilde{Y}_d \quad (7)$$

Dove

$$b_\lambda \tilde{Y}_d = \sum_{h=1}^{AR} H_1 \sum_{i=1}^{\tilde{n}_{lh}} b_\lambda y_{dlhi} W_{dlhi} + \sum_{h=1}^{NAR} H_1 \sum_{c=1}^{\bar{m}_1} \sum_{i=1}^{\tilde{n}_{lhc}} b_\lambda y_{dlhici} W_{dlhici} \quad (8)$$

Lo stimatore ottenuto è uno stimatore del rapporto che può essere espresso nella forma equivalente

$$b_\lambda \tilde{\theta}_d = \frac{b_\lambda \hat{Y}_d}{\hat{N}_d^*} \quad (9)$$

### 4.3. Stime a livello di scuola

L'indagine in oggetto ha, anche, la finalità secondaria di produrre stime a livello di singola scuola, sia per quanto riguarda le scuole selezionate nel campione progettato che per quanto concerne le scuole che si sono aggiunte volontariamente che sono dette *scuole volontarie*. A tal fine è utile modificare la notazione sopra adottata indicando con  $j$  ( $j=1, \dots, J$ ) la generica scuola campione e con  $J$  il numero complessivo delle scuole  $AR$ , di quelle  $NAR$  e delle scuole volontarie. Il parametro di interesse riferito alla scuola  $j$  ( $j=1, \dots, J$ ) è

$$b_\lambda \theta_j = \frac{1}{N_j} \sum_{i=1}^{N_j} b_\lambda y_{ji} \quad (10)$$

in cui  $N_j$  è il numero di alunni, di seconda o di quinta elementare, della scuola  $j$  e  $b_\lambda y_{ji}$  è il valore della variabile di interesse, riferita alla prova  $b$  e al tipo di parametro  $\lambda$ , osservata sulla  $i$  esima unità. Poiché nell'ambito di ciascuna scuola campione è stato utilizzato un campionamento casuale semplice senza reimmissione lo stimatore corretto del parametro (10) è

$$b_\lambda \hat{\theta}_j = \frac{1}{\tilde{n}_j} \sum_{i=1}^{\tilde{n}_j} b_\lambda y_{ji} \quad (11)$$

in cui  $\tilde{n}_j$  indica il numero di alunni rispondenti della scuola  $j$  appartenenti alla seconda o alla quinta elementare.

## 5. Indicatori di qualità

### 5.1. Errori e intervalli di confidenza delle stime riferite ai diversi domini

Al fine di valutare la qualità delle stime prodotte da un'indagine campionaria è necessario produrre indici relativi alla variabilità campionaria delle stesse che consentono di calcolare, anche, gli intervalli di confidenza delle stime.

Lo stimatore della varianza campionaria, sulla base del disegno di campionamento e dello stimatore utilizzato, è espresso da

$$\hat{Var}(b_\lambda \tilde{\theta}_d) = \frac{1}{N_d^{*2}} \sum_{l=1}^{L_d} \left[ \sum_{h=1}^{AR H_l} N_{lh}^2 \frac{(N_{lh} - \tilde{n}_{lh})}{N_{lh}} \frac{s_{e,hl}^2}{\tilde{n}_{lh}} + \sum_{h=1}^{NAR H_l} \frac{\bar{m}_l}{\bar{m}_l - 1} \sum_{c=1}^{\bar{m}_l} (\hat{E}_{lhc} - \hat{E}_{lh})^2 \right]. \quad (12)$$

Il primo addendo della precedente formula, costituisce la varianza della componente AR dello stimatore, in cui

$$s_{e,hl}^2 = \frac{1}{n_{lh} - 1} \sum_{i=1}^{\tilde{n}_{lh}} (e_{dlhi} - \bar{e}_{dlh})^2$$

essendo

$$e_{dlhi} = b_\lambda y_{dlhi} - \frac{b_\lambda \hat{Y}_d}{\hat{N}_d^*}, \quad \bar{e}_{dlh} = \sum_{i=1}^{\tilde{n}_{lh}} e_{dlhi}.$$

Per il secondo addendo della (12), che costituisce la varianza della componente NAR dello stimatore, si è adottata l'usuale approssimazione della varianza con reimmissione che costituisce uno stimatore conservativo della varianza, in cui

$$\hat{E}_{lhc} = \sum_{i=1}^{\tilde{n}_{lhc}} e_{dlhci} W_{dlhci}, \quad \hat{E}_{lh} = \frac{1}{\bar{m}_l} \sum_{c=1}^{\bar{m}_l} \hat{E}_{lhc}.$$

Per ciascuna stima prodotta con l'indagine, a partire dalla varianza di campionamento, è possibile costruire il corrispondente errore assoluto e l'errore relativo percentuale espressi in formule come

$$\hat{\sigma}_{(b_\lambda \tilde{\theta}_d)} = \sqrt{\hat{V}ar(b_\lambda \tilde{\theta}_d)}$$

$$CV_{(b_\lambda \tilde{\theta}_d)} = \frac{\sqrt{\hat{V}ar(b_\lambda \tilde{\theta}_d)}}{\tilde{\theta}_d} \cdot 100.$$

Infine gli estremi inferiore e superiore dell'intervallo di confidenza al 95% sono dati da  $\tilde{\theta}_d - 1.96 \hat{\sigma}_{(b_\lambda \tilde{\theta}_d)}$  e  $\tilde{\theta}_d + 1.96 \hat{\sigma}_{(b_\lambda \tilde{\theta}_d)}$ .

## 5.2. Errori e intervalli di confidenza delle stime a livello di scuola

La varianza di campionamento dello stimatore  $b_\lambda \hat{\theta}_j$ , espresso dalla (11) riferito alla generica scuola campione  $j$  è espressa come

$$Var(b_\lambda \hat{\theta}_j) = N_j^2 \frac{(N_j - \tilde{n}_j)}{N_j} \frac{s_j^2}{\tilde{n}_j},$$

in cui

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{\tilde{n}_j} (y_{ji} - \bar{y}_j)^2.$$

Il corrispondente errore assoluto e errore relativo percentuale espressi in formule sono

$$\hat{\sigma}_{(b_\lambda \hat{\theta}_j)} = \sqrt{\hat{V}ar(b_\lambda \hat{\theta}_j)}$$

$$CV_{(b_\lambda \hat{\theta}_j)} = \frac{\sqrt{\hat{V}ar(b_\lambda \hat{\theta}_j)}}{b_\lambda \hat{\theta}_j} \cdot 100.$$

Infine gli estremi inferiore e superiore dell'intervallo di confidenza al 95% sono dati da  $b_{\lambda} \hat{\theta}_j - 1.96 \hat{\sigma} (b_{\lambda} \hat{\theta}_j)$  e  $\tilde{\theta}_d + 1.96 \hat{\sigma} (b_{\lambda} \hat{\theta}_j)$ .

### **Riferimenti bibliografici**

Bethel J. (1989), "Sample Allocation in Multivariate Surveys", *Survey Methodology*, Vol. 15, 47-57.

Cicchitelli G., Herzel A. e Montanari G. E. (1992) "*Il campionamento statistico*", Il Mulino, Bologna.

Fabbris L. (1991), "Campioni di numerosità due o tre per strato selezionati con probabilità variabili: valutazione empirica di alcune stime di frequenze assolute", in *Atti della giornata di studio sul campionamento statistico*. Annali di statistica ,Serie IX, ISTAT.

Falorsi S., Russo A. (2003) "*Il disegno di rilevazione per le indagini panel sulle famiglie*" , Rivista di statistica ufficiale, n. 1/2003, Franco Angeli.

Madow W. G. (1949), "On theory of systematic sampling", *Annals of Mathematical Statistics*, 20.

Murthy M.N. (1967), "Sampling theory and methods", Calcutta.