La strategia di campionamento[1]

1. Descrizione del disegno di campionamento

Nelle pagine che seguono si illustrano gli obiettivi conoscitivi e gli aspetti più significativi della strategia di campionamento dell'indagine sugli alunni con disabilità nelle scuole primarie e secondarie di I grado dell'anno scolastico 2012/2013.

La *popolazione di interesse* dell'indagine in oggetto, ossia l'insieme delle unità statistiche intorno alle quali si intende investigare, è costituita dagli studenti con disabilità presenti nelle scuole nell'anno scolastico 2012/2013. I *domini* di riferimento delle stime sono:

- l'intero territorio nazionale;
- tre ripartizioni geografiche (Nord, Centro e Meridione);
- due ordini scolastici, primarie e secondarie di I grado;
- le modalità ottenute dall'incrocio tra la ripartizione e l'ordine scolastico.

Il disegno di campionamento è a due stadi di selezione con stratificazione delle unità di primo stadio. Le unità di primo stadio sono le scuole, stratificate per regione geografica e ordine scolastico. Le unità di secondo stadio sono gli alunni con disabilità.

La numerosità campionaria di primo e di secondo stadio è stata definita tenendo conto sia di esigenze organizzative e di costo, sia degli errori di campionamento attesi delle principali stime di interesse a livello dei domini di stima sopra menzionati. La dimensione complessiva del campione di scuole è stata fissata a 3002 unità, mentre la dimensione del campione di alunni da intervistare è stata fissata a 13.743.

L'archivio di selezione per l'indagine è costituito dalla lista delle scuole primarie e secondarie di I grado in cui è presente almeno un alunno con disabilità; tale archivio è stato fornito dal Ministero dell'Istruzione e contiene per ogni scuola il numero di alunni con disabilità.

Le scuole sono state stratificate nei domini ottenuti come incrocio della regione e dell'ordine scolastico.

La dimensione complessiva del campione di scuole è stata distribuita tra gli strati ottenuti dall'incrocio delle variabili ordine scolastico e regione in modo da garantire che gli errori di campionamento attesi delle principali stime riferite ai diversi domini di interesse non superassero prefissati livelli.

Da ciascuno strato è stato selezionato un campione di scuole mediante selezione casuale con probabilità proporzionale alla dimensione espressa in termini di alunni con disabilità.

Per ciascuna scuola inclusa nel campione sono stati selezionati, a cura della stessa scuola, 5 alunni con disabilità; qualora nella scuola ne fossero presenti meno di 5, sono stati intervistati tutti gli alunni con disabilità presenti.

Nella fase di rilevazione si sono verificate numerose cadute di scuole campione, portando il campione realizzato da 3.002 a 2.435, per un totale di 10.489 alunni intervistati.

2. Procedimento per il calcolo delle stime

Le stime prodotte dall'indagine sono principalmente stime di frequenze assolute e relative.

Il principio su cui è basato ogni metodo di stima campionaria è che le unità appartenenti al campione rappresentino anche le unità della popolazione che non sono incluse nel campione.

Questo principio viene realizzato attribuendo ad ogni unità campionaria un peso che denota il numero di unità della popolazione rappresentate dalla unità medesima. Se, ad esempio, ad una unità campionaria viene attribuito un peso pari a 30, vuol dire che questa unità rappresenta se stessa ed altre 29 unità della popolazione che non sono state incluse nel campione.

[1] A cura di Claudia De Vitiis e Monica Russo

Al fine di rendere più chiara la successiva esposizione, introduciamo le seguenti notazioni simboliche. Sia:

- d indice di dominio di stima;
- i indice di scuola;
- j indice di alunno con disabilità;
- h indice di strato (regione geografica per ordine scolastico);
- M_h numero totale di alunni con disabilità nello strato h;
- M_{hi} numero totale di alunni con disabilità nella scuola i appartenente allo strato h;
- m_{hi} numero di alunni con disabilità campione nella scuola i appartenente allo strato h;
- N_h numero totale di scuole nello strato h;
- n_h numero di scuole campione nello strato h;
- H_d numero totale di strati nel dominio d;
- x generica variabile oggetto di indagine;
- X_{hij} valore osservato della variabile x sul j-mo alunno della scuola i appartenente allo strato h.

Ipotizziamo di voler stimare, con riferimento ad un generico dominio d, il totale di popolazione espresso dalla seguente relazione:

$$X_d = \sum_{h=1}^{H_d} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} X_{hij} \quad . \tag{1}$$

La stima del totale (1) si ottiene in generale mediante la seguente formula:

$$\hat{X}_d = \sum_{h=1}^{H_d} \sum_{i=1}^{{}_r n_h} \sum_{j=i}^{{}_r m_{hi}} X_{hij} W_{hij} \quad , \tag{2}$$

dove: W_{hij} è il *peso finale* assegnato all'individuo j, ${}_r n_h$ ed ${}_r m_h$ sono rispettivamente il numero di scuole ed il numero di alunni con disabilità campione rispondenti appartenenti allo strato h.

I pesi finali da attribuire agli individui campione sono stati calcolati in base ad uno stimatore post-stratificato, che utilizza la conoscenza di totali noti di popolazione disponibili da fonti esterne all'indagine. Tali totali sono il numero di alunni con disabilità a livello di strato, ottenuto dal concatenamento delle modalità delle variabili regione geografica e ordine scolastico, e sono stati desunti dall'archivio aggiornato fornito dal Ministero dell'Istruzione. La post-stratificazione garantisce che sussista l'uguaglianza tra tali totali noti e le corrispondenti stime campionarie.

La procedura di costruzione dei pesi è stata così articolata:

1) si è determinato un *peso base* (o *peso diretto*), D_{hij}, uguale per tutti gli individui j appartenenti alla medesima scuola i dello strato h. Tale peso è ottenuto dal prodotto del peso di riporto all'universo di primo stadio – dato dall'inverso della probabilità di inclusione della scuola i – moltiplicato per il peso di riporto all'universo di secondo stadio – ottenuto dal reciproco della probabilità di inclusione dell'individuo j condizionata all'inclusione nel campione della scuola i a cui l'individuo j appartiene. In simboli:

$$D_{hij} = \pi_{hij}^{-1} = \pi_{hi}^{-1} \cdot \pi_{hj|i}^{-1} = \left(n_h \frac{M_{hi}}{M_h} \right)^{-1} \left(\frac{m_{hi}}{M_{hi}} \right)^{-1} \quad ,$$

in cui: π_{hij} è la probabilità di inclusione dell'individuo j, π_{hi} è la probabilità di inclusione nel campione della scuola i e $\pi_{hj|i}$ è la probabilità di inclusione dell'individuo j condizionata al fatto che la scuola i è stata inclusa nel campione;

2) si è definito il *fattore correttivo della mancata risposta totale*[2], R_{hi}, anch'esso uguale per tutti gli individui j appartenenti alla medesima scuola i dello strato h. Tale fattore è definito come reciproco della probabilità di risposta dell'individuo j della scuola i nello strato h, ottenuta dal prodotto della probabilità di rispondere della scuola i nello strato h a cui j appartiene per la probabilità che l'individuo j risponda condizionata al fatto che la scuola i ha risposto. Ossia:

$$R_{hi} = \delta_{hij}^{-1} = \delta_{hi}^{-1} \cdot \delta_{hj|i}^{-1} = \left(\frac{r\, n_h}{n_h} \right)^{-1} \left(\frac{r\, m_{hi}}{m_{hi}} \right)^{-1} ,$$

in cui: δ_{hij} è la probabilità di risposta dell'individuo j della scuola i appartenente allo strato h, δ_{hi} è la probabilità di rispondere della scuola i nello strato h e $\delta_{hj|i}$ è la probabilità che l'individuo j risponda visto che la scuola i a cui esso appartiene ha risposto;

3) si è calcolato il *fattore correttivo per la coerenza delle stime*, che ha la finalità di far coincidere le stime campionarie dei totali di strato con i corrispettivi totali noti M_h^* :

$$C_h = \frac{M_h^*}{\hat{M}_h} = \frac{M_h^*}{\sum_{i=1}^{r n_h} \sum_{j=1}^{r m_h} D_{hij} R_{hij}} ;$$

4) si è ottenuto il *peso finale* dell'individuo j appartenente alla scuola i nello strato h moltiplicando il peso diretto D_{hij} per i due fattori correttivi R_{hi} e C_h:

$$W_{hij} = D_{hij} \cdot R_{hi} \cdot C_h .$$

Una volta assegnato a ogni individuo il coefficiente di riporto all'universo, è stato possibile ottenere le stime di interesse dei parametri di popolazione del tipo (1) come indicato nella (2).

E' utile sottolineare che lo stimatore appena illustrato rientra nella classe degli *stimatori di ponderazione vincolata*, che è il metodo di stima standard per la maggior parte delle indagini ISTAT sulle imprese e sulle famiglie. Tale classe di stimatori viene utilizzata quando si dispone di informazioni espresse in forma di totali noti di variabili ausiliarie legate alle variabili di interesse.

[2] Il fattore correttivo così calcolato tiene conto della mancata risposta totale sia delle scuole sia degli alunni.

3. Valutazione del livello di precisione delle stime

3.1 Calcolo della varianza campionaria

Le principali statistiche di interesse per valutare la variabilità campionaria delle stime prodotte dall'indagine sono l'*errore di campionamento assoluto* e l'*errore di campionamento relativo*.

La stima dell'*errore di campionamento assoluto* di $\hat{X}_d$ è definita dalla seguente espressione:

$$\hat{\sigma}(\hat{X}_d) = \sqrt{\hat{V}ar(\hat{X}_d)} \ . \tag{3}$$

La stima dell'*errore di campionamento relativo* di $\hat{X}_d$ è data da:

$$\hat{\varepsilon}(\hat{X}_d) = \frac{\hat{\sigma}(\hat{X}_d)}{\hat{X}_d} \ . \tag{4}$$

La stima della varianza di $\hat{X}_d$, indicata nella (3) come $\hat{V}ar(\hat{X}_d)$, è stata calcolata utilizzando il *metodo di linearizzazione di Woodruff*, che consente di ottenere un'espressione approssimata della varianza campionaria nel caso di stimatori, come quello qui utilizzato, che non sono funzione lineare dei dati campionari.

Gli errori campionari espressi dalla (3) e dalla (4) consentono di valutare il grado di precisione delle stime; inoltre, l'errore assoluto permette di costruire un intervallo di confidenza, nel quale con una certa probabilità si trova il parametro oggetto di stima:

$$\Pr\left\{\hat{X}_d - k\hat{\sigma}(\hat{X}_d) \le X_d \le \hat{X}_d + k\hat{\sigma}(\hat{X}_d)\right\} \ge P \ . \tag{5}$$

Nella (5) il valore di k dipende dal valore fissato per la probabilità P (ad esempio, per P=0,95 si ha k=1,96).

3.2 Presentazione sintetica degli errori campionari

Ad ogni stima $\hat{X}_d$ è associato un errore campionario relativo $\varepsilon(\hat{X}_d)$; pertanto, per consentire un uso corretto delle stime fornite dall'indagine, sarebbe necessario fornire, per ogni stima pubblicata, anche il corrispondente errore di campionamento relativo.

Tuttavia, non è possibile soddisfare questa esigenza di informazione, sia per motivi di tempi e di costi di elaborazione sia perché le tavole della pubblicazione risulterebbero eccessivamente appesantite e di non agevole consultazione per l'utente finale; inoltre, non sarebbero in ogni caso disponibili gli errori delle stime non pubblicate.

Per questi motivi, generalmente, si ricorre ad una *presentazione sintetica degli errori relativi*, basata sul *metodo dei modelli regressivi*. Tale metodo consiste nella determinazione di una funzione matematica che mette in relazione ciascuna stima con la stima del proprio errore relativo.

Il modello utilizzato per le stime di frequenze assolute è il seguente:

$$\log\hat{\varepsilon}^2(\hat{X}_d) = a + b\log(\hat{X}_d) \ , \tag{6}$$

in cui i parametri a e b sono stimati mediante il metodo dei minimi quadrati.

Nella indagine in oggetto è stato stimato un modello di tipo (6) per ciascuno dei seguenti domini di interesse:

- D1. totale Italia;
- D2. ripartizioni geografiche;
- D3. ordine scolastico;
- D4. ripartizione geografica per ordine scolastico.

Per calcolare il livello di precisione delle stime prodotte dall'indagine è stato utilizzato un software generalizzato, messo a punto dall'Istat, che consente di calcolare gli errori campionari e gli intervalli di confidenza e permette, inoltre, di costruire modelli regressivi del tipo (6) per la presentazione sintetica degli errori di campionamento.

Il prospetto 1 riporta i valori dei coefficienti a e b e dell'indice di determinazione R^2 del modello utilizzato per l'interpolazione degli errori campionari delle stime riferite ai domini D1-D4.

Sulla base delle informazioni contenute in tale prospetto, è possibile calcolare l'errore relativo di una determinata stima di frequenza assoluta $\hat{X}_d^*$ nel modo di seguito descritto.

Dalla (6), mediante semplici passaggi, si ricava:

$$\hat{\varepsilon}(\hat{X}_d^*) = \sqrt{\exp(a + b \log(\hat{X}_d^*))} \ , \qquad (7)$$

se, per esempio, la generica stima $\hat{X}_d^*$ si riferisce alla ripartizione Nord, è possibile introdurre nella (7) i valori dei parametri a e b (a=3,72113, b=-1,07359) riportati nella corrispondente riga del prospetto 2 e ricavare il corrispondente errore relativo.

Una volta calcolato l'errore relativo, è possibile costruire l'intervallo di confidenza al 95% come:

$$\hat{X}_d^* - 1{,}96 \cdot \hat{\varepsilon}(\hat{X}_d^*) \cdot \hat{X}_d^* ; \hat{X}_d^* + 1{,}96 \cdot \hat{\varepsilon}(\hat{X}_d^*) \cdot \hat{X}_d^* \ .$$

Allo scopo di facilitare il calcolo degli errori campionari, nel prospetto 2 sono riportati i valori interpolati degli errori di campionamento relativi di alcune stime di frequenze relative percentuali nei vari domini di stima.

Le informazioni contenute in tale prospetto consentono di calcolare l'errore relativo di una generica stima di frequenza assoluta mediante due procedimenti che risultano di facile applicazione, anche se conducono a risultati meno precisi di quelli ottenibili applicando direttamente la formula (7). Il primo metodo consiste nell'approssimare l'errore relativo della stima di interesse con quello, riportato nei prospetti, corrispondente al livello di stima che più vi si avvicina. Il secondo metodo, più preciso del primo, si basa sull'uso di una formula di interpolazione lineare per il calcolo degli errori di stime non comprese tra i valori forniti nei prospetti. In tal caso, l'errore campionario della stima $\hat{X}_d^*$ si ricava mediante l'espressione:

$$\hat{\varepsilon}(\hat{X}_d^*) = \hat{\varepsilon}(\hat{X}_d^{k-1}) + \frac{\hat{\varepsilon}(\hat{X}_d^{k-1}) - \hat{\varepsilon}(\hat{X}_d^{k})}{\hat{X}_d^{k} - \hat{X}_d^{k-1}} (\hat{X}_d^* - \hat{X}_d^{k-1}) \ ,$$

dove $\hat{X}_d^{k-1}$ e $\hat{X}_d^{k}$ sono i valori delle stime entro i quali è compresa la stima $\hat{X}_d^*$, mentre $\hat{\varepsilon}(\hat{X}_d^{k-1})$ e $\hat{\varepsilon}(\hat{X}_d^{k})$ sono i corrispondenti errori relativi.

Prospetto 1 – Valori dei coefficienti a e b e dell'indice di determinazione R^2 (%) del modello per l'interpolazione degli errori campionari delle stime di frequenze di variabili qualitative per totale Italia, ripartizione geografica, ordine scolastico e incrocio di ripartizione geografica e ordine scolastico (Dati provvisori)

DOMINIO DI STIMA		a	b	R^2
NORD	PRIMARIE	3,96787	-1,11241	91,9
	SECONDARIE DI I GRADO	3,91270	-1,13290	93,0
	TOTALE NORD	3,72113	-1,07359	94,1
CENTRO	PRIMARIE	3,54353	-1,15846	91,9
	SECONDARIE DI I GRADO	3,32480	-1,15784	92,6
	TOTALE CENTRO	3,27627	-1,11635	93,7
SUD E ISOLE	PRIMARIE	4,17122	-1,16025	91,4
	SECONDARIE DI I GRADO	4,04967	-1,16380	92,8
	TOTALE SUD E ISOLE	3,55915	-1,08047	94,6
ITALIA	PRIMARIE	3,43638	-1,06195	94,2
	SECONDARIE DI I GRADO	3,69813	-1,10800	94,7
	TOTALE ITALIA	**3,28949**	**-1,04231**	**95,9**

Prospetto 2 - Valori interpolati degli errori campionari relativi percentuali delle stime di frequenze (percentuali e assolute) di variabili qualitative per totale Italia, ripartizione geografica, ordine scolastico e incrocio di ripartizione geografica e ordine scolastico (Dati provvisori)

Stima (%)	Italia		Nord		Centro		Sud e Isole		Primarie		Secondarie di I grado	
	Valore assol. stima	Errore relativo %	Valore assol. stima	Errore relativo %	Valore assol. stima	Errore relativo %	Valore assol. stima	Errore relativo %	Valore assol. stima	Errore relativo %	Valore assol. stima	Errore relativo %
0,5	724	16,8	330	28,6	140	32,6	253	29,8	405	23,0	318	26,1
1	1.447	11,7	660	19,7	281	22,1	507	20,5	810	15,9	637	17,8
2,5	3.618	7,2	1.649	12,0	702	13,3	1.267	12,5	2.025	9,8	1.592	10,7
5	7.235	5,0	3.299	8,3	1.403	9,0	2.533	8,6	4.051	6,8	3.184	7,3
10	14.470	3,5	6.598	5,7	2.806	6,1	5.067	5,9	8.102	4,7	6.369	5,0
15	21.705	2,8	9.897	4,6	4.209	4,9	7.600	4,7	12.153	3,8	9.553	4,0
20	28.941	2,5	13.196	3,9	5.612	4,2	10.133	4,1	16.204	3,2	12.737	3,4
25	36.176	2,2	16.495	3,5	7.015	3,7	12.666	3,6	20.255	2,9	15.921	3,0
30	43.411	2,0	19.793	3,2	8.418	3,3	15.200	3,3	24.305	2,6	19.106	2,7
35	50.646	1,8	23.092	2,9	9.821	3,0	17.733	3,0	28.356	2,4	22.290	2,5
40	57.881	1,7	26.391	2,7	11.224	2,8	20.266	2,8	32.407	2,2	25.474	2,3
45	65.116	1,6	29.690	2,6	12.627	2,6	22.799	2,6	36.458	2,1	28.658	2,2
50	72.352	1,5	32.989	2,4	14.030	2,5	25.333	2,5	40.509	2,0	31.843	2,0

Prospetto 2 (segue) **- Valori interpolati degli errori campionari relativi percentuali delle stime di frequenze (percentuali e assolute) di variabili qualitative per totale Italia, ripartizione geografica, ordine scolastico e incrocio di ripartizione geografica e ordine scolastico.** (Dati provvisori)

Stima (%)	Nord Primarie		Nord Secondarie di I grado		Centro Primarie		Centro Secondarie di I grado		Sud e Isole Primarie		Sud e Isole Secondarie di I grado	
	Valore assol. stima	Errore relativo %	Valore assol. stima	Errore relativo %	Valore assol. stima	Errore relativo %	Valore assol. stima	Errore relativo %	Valore assol. stima	Errore relativo %	Valore assol. stima	Errore relativo %
0,5	183	40,1	147	41,8	81	46,2	59	49,6	141	45,5	112	48,7
1	365	27,3	294	28,3	162	30,9	119	33,2	283	30,4	224	32,5
2,5	913	16,4	736	16,8	405	18,2	297	19,5	707	17,9	559	19,1
5	1.827	11,2	1.472	11,4	809	12,2	594	13,1	1.415	12,0	1.119	12,7
10	3.654	7,6	2.944	7,7	1.619	8,1	1.187	8,8	2.829	8,0	2.237	8,5
15	5.481	6,1	4.416	6,1	2.428	6,4	1.781	6,9	4.244	6,3	3.356	6,7
20	7.308	5,2	5.888	5,2	3.237	5,4	2.375	5,9	5.658	5,4	4.475	5,7
25	9.135	4,6	7.360	4,6	4.047	4,8	2.968	5,1	7.073	4,7	5.593	5,0
30	10.962	4,1	8.832	4,1	4.856	4,3	3.562	4,6	8.488	4,2	6.712	4,5
35	12.789	3,8	10.304	3,8	5.665	3,9	4.156	4,2	9.902	3,9	7.831	4,1
40	14.616	3,5	11.776	3,5	6.475	3,6	4.749	3,9	11.317	3,6	8.949	3,8
45	16.443	3,3	13.248	3,3	7.284	3,4	5.343	3,7	12.731	3,3	10.068	3,5
50	18.270	3,1	14.720	3,1	8.094	3,2	5.937	3,4	14.146	3,1	11.187	3,3