File Standard

L’indagine campionaria sulle nascite e le madri – Anno 2012

Indagine CATI (LONG FORM)

Manuale per l’utente

Istat

Istituto Nazionale di Statistica

I file standard vengono rilasciati per finalità di studio e ricerca. Per ottenere tali file è necessario registrarsi al Contact centre. Una volta effettuata la registrazione, la richiesta deve essere formulata selezionando nel Contact centre l'area "Collezioni campionarie di dati elementari (file standard) e compilando un modulo on-line.

Per informazioni sull'indagine rivolgersi a:
Istat - Servizio 'Struttura e dinamica demografica'
Viale Liegi, 13 – 00198
Roma
tel: 06.4673.7322
fax: 0646737621
e-mail: cicastag@istat.it

Il manuale, curato da Cinzia Castagnaro.

La premessa e il paragrafo 1.1 sono curati da Cinzia Castagnaro e Sabrina Prati.
I paragrafi 1.2 e 1.3 sono stati curati da Claudia Iaccarino.
Il capitolo 2 è stato curato da Claudia De Vitiis e Adriano Pareto.

I programmi per la correzione dei dati e la creazione del file standard sono stati progettati e realizzati da Claudia Iaccarino.

La rilevazione dell'indagine è stata curata da Cinzia Castagnaro e Sabrina Prati.

Indice

Premessa

Il decreto legislativo n. 322 del 6/9/1989 regola la diffusione delle informazioni statistiche prodotte nell'ambito del Sistema Statistico Nazionale al fine di garantire la riservatezza dei rispondenti. In particolare, per la diffusione di dati elementari, l'articolo 10, comma 2, dispone quanto segue: "Sono distribuite altresì, ove disponibili, su richiesta motivata e previa autorizzazione del Presidente dell'Istat, collezioni campionarie di dati elementari, resi anonimi e privi di ogni riferimento che ne permetta il collegamento con singole persone fisiche e giuridiche".

Nell'osservanza di tale disposizione e del d. lgs del 30/06/2003 n. 196 (Codice in materia di protezione dei dati personali) l'Istat ha adottato misure e tecniche che rendono impossibile, o altamente improbabile, il collegamento dei dati rilasciati con l'unità statistica a cui si riferiscono. Per tale motivo vengono apportate alcune modifiche sui file originali delle indagini, nell'intento di garantire la massima protezione ai dati, contenendo al minimo la perdita di informazioni. Le metodologie applicate si concretizzano nell'accorpamento e/o riclassificazione di modalità di variabili e nell'oscuramento di variabili. In quest'ultimo caso nei campi del tracciato record è riportata la dicitura "RISERVATO ISTAT".

Va considerato inoltre che la stessa dicitura è stata utilizzata anche per quelle informazioni che, pur essendo state oggetto di indagine, non sono risultate essere attendibili dal punto di vista campionario e quindi statisticamente non analizzabili, e per le variabili di lavorazione o controllo.

Nelle prossime pagine, dopo una breve descrizione delle fasi principali dell'Indagine Campionaria sulle Nascite, viene riportato il tracciato record che descrive le variabili contenute nel file standard. Per agevolare l'elaborazione dei dati e l'interpretazione dei risultati, negli allegati a seguire sono riportati il questionario, le classificazioni e le definizioni adottate, e vengono illustrate le caratteristiche del disegno di campionamento e la metodologia adottata per la protezione dei dati.

1. L'indagine Campionaria sulle Nascite: caratteristiche e contenuti

1.1 Le informazioni statistiche sulle nascite: nuove rilevazioni per nuove esigenze informative

Il sistema di raccolta e produzione dei dati statistici sulle nascite è stato, negli ultimi anni, fortemente modificato e rinnovato. Il processo di cambiamento, che si inquadra nella strategia dell'Istituto Nazionale di Statistica di osservare gli eventi e i comportamenti demografici in una prospettiva conoscitiva, è stato indirettamente accelerato dalla necessità di adeguare i flussi informativi alle nuove norme in materia di denuncia di nascita entrate in vigore tra il 1997 e il 1999.

Per oltre 70 anni l'Istat ha diffuso le principali informazioni statistiche sulle nascite e i parti attraverso i dati provenienti dalla rilevazione delle nascite di fonte Stato Civile, con un dettaglio informativo molto ricco ai fini della descrizione dei fenomeni. Sulla base di questa rilevazione, corrente ed esaustiva, è stato possibile fornire al paese con regolarità e accuratezza le informazioni relative alle modificazioni dei comportamenti riproduttivi avvenute nel nostro paese.

La rilevazione delle nascite ha consentito infatti per lungo tempo di monitorare con continuità e precisione la forte riduzione della fecondità, soprattutto per i figli successivi al primo, l'incremento dell'infertilità e il fortissimo innalzarsi dell'età media alla nascita del primogenito, con i conseguenti crescenti rischi non solo di infertilità, ma anche di gravidanze a maggior rischio di complicanze, particolarmente per le primipare. Essa ha inoltre garantito al paese un'informazione strutturale puntuale su alcuni fenomeni di grande rilevanza bio-demografica e socio-sanitaria, quali la natimortalità, i parti plurimi, le caratteristiche del parto rispetto alle principali caratteristiche demografiche dei genitori.

I mutamenti normativi riguardanti la dichiarazione di nascita hanno imposto la soppressione, a partire dal 1° gennaio 1999, della rilevazione individuale delle nascite di fonte Stato Civile. Ne è seguita una vera e propria azione di rigenerazione di tutta la strumentazione logica e metodologica finora utilizzata per la produzione delle statistiche sulle nascite.

Da una rilevazione sulle nascite si è passati ad un sistema di rilevazioni che consente non solo di colmare il debito informativo creatosi, ma anche di ampliare considerevolmente la produzione di informazioni rilevanti per la comprensione dei fenomeni oggetto di osservazione, venendo così incontro alle mutate esigenze della domanda informativa. Si fa sempre più pressante, infatti, l'esigenza di approfondire le determinanti e le dinamiche che influiscono sulle scelte di maternità e di paternità, così come l'esigenza di analizzare i contesti di vita familiari e sociali in cui tali determinanti svolgono la loro azione.

Il compito di soddisfare queste nuove esigenze informative è affidato all'Indagine Campionaria sulle Nascite, che rappresenta un'assoluta novità nel settore delle statistiche demografiche e la cui prima edizione è stata effettuata nel 2002 con tecnica CATI (i principali risultati sono pubblicati nel volume "*Avere un figlio in Italia Approfondimenti tematici dall'Indagine campionaria sulle nascite Anno 2002 – Settore Informazioni*", http://www3.istat.it/dati/catalogo/20061220_00/)[1]; una seconda edizione è stata realizzata nel 2005 sempre con tecnica CATI.

I dati diffusi in questa occasione riguardano la terza edizione dell'Indagine, realizzata dall'Istat durante il 2012 con tecnica CATI con un questionario *long form*; un modulo di approfondimento è

[1] Per maggiori informazioni sugli aspetti metodologici è possibile consultare il volume "*Indagine Campionaria sulle Nascite: obiettivi, metodologia e organizzazione, Anno 2002 - Settore Metodi e Norme*" http://www.istat.it/dati/catalogo/20060317_00/).

dedicato proprio all'interazione maternità-lavoro; nell'edizione 2012, i contenuti dell'Indagine sono stati arricchiti grazie all'apporto dell'Isfol che, nell'ambito di una Convenzione Istat-Isfol del 2008, progetto "Maternità e partecipazione femminile al mercato del lavoro", ha collaborato alla riprogettazione dell'indagine anche in un'ottica retrospettiva, in modo da poter analizzare sia i comportamenti riproduttivi delle donne con almeno un figlio, sia l'interazione maternità-lavoro nel medio-lungo periodo.

La popolazione di interesse dell'indagine – ossia l'insieme delle unità statistiche relativamente alle quali si intende investigare – è costituita dai nati iscritti in anagrafe nel corso del secondo semestre 2009 e primo semestre 2010; le unità di rilevazione, invece, sono le madri di tali nati, intervistate nel 2012 a una distanza media di circa due anni dal parto.

I dati diffusi in questo file relativi all'Indagine CATI consentono, grazie alla versione *long* del questionario, la confrontabilità con i risultati delle 2 edizioni precedenti, permettendo di monitorare le principali variazioni intercorse.

L'universo dei nati della popolazione residente viene individuato dalla Rilevazione degli Iscritti in Anagrafe per Nascita. Le informazioni inserite nel modello, oltre al nato e ai genitori, riguardano l'intestatario della scheda di famiglia (con l'indirizzo completo del luogo di residenza), consentendo in tal modo di reperire le famiglie al loro indirizzo anagrafico.

I principali contenuti del questionario riguardano:

1. informazioni di carattere socio-demografico sul nato, sulla madre e sul padre (in caso di riconoscimento del figlio);
2. notizie sul parto;
3. notizie sul contesto familiare;
4. approfondimenti sulla condizione professionale della madre prima e dopo la nascita del bambino;
5. aspetti connessi alla cura del bambino e alla divisione del lavoro familiare;
6. notizie sull'abitazione e sul contesto socio-economico.

Si è quindi concordato sulla necessità di disporre di stime rappresentative a livello ripartizionale per le principali caratteristiche strutturali delle nascite.

I principali temi trattati riguardano:

- i progetti riproduttivi delle madri;
- le motivazioni per non avere altri figli;
- le variazioni intercorse nella condizione professionale delle neo-madri in seguito alla nascita dei figli;
- le difficoltà nel conciliare famiglia e attività lavorativa;
- gli aiuti su cui possono contare le neo-madri per il lavoro domestico e la cura del bambino;
- le ragioni dell'accessibilità o non-accessibilità ai servizi per l'infanzia.

Particolare attenzione è stata dedicata al lavoro della madre prima e dopo la nascita del figlio con l'obiettivo di cogliere eventuali variazioni intercorse tra l'inizio della gravidanza e il momento dell'intervista.

Sulla base di queste variazioni le intervistate possono essere distinte in quattro tipologie:

- donne che attualmente hanno lo stesso lavoro che avevano durante la gravidanza;

- donne che attualmente hanno un nuovo lavoro, diverso da quello che avevano durante la gravidanza;
- donne attualmente non occupate ma che avevano un'occupazione durante la gravidanza;
- donne attualmente non occupate e che non svolgevano un'attività lavorativa durante la gravidanza.

1.2 Informazioni errate o incompatibili

Uno degli aspetti principali nell'espletamento di un'indagine campionaria è quello che riguarda la qualità dei dati dal punto di vista della correttezza e della coerenza delle informazioni raccolte.

Varie sono le possibili cause che introducono errori durante l'intervista e altrettanto varie sono le strategie che permettono di limitarne l'introduzione. Ad esempio gli errori possono derivare dalla reticenza o dalla mancanza di interesse e/o di attenzione dei soggetti intervistati; per motivare le madri in merito alla rilevanza della loro collaborazione all'indagine e rammentare l'obbligo di risposta è stata loro inviata una lettera a firma del Presidente dell'Istat che preannunciava l'intervista e illustrava i contenuti e gli scopi dell'indagine.

Anche la difficoltà nel ricordare può essere causa di risposte errate o mancate risposte; per questo motivo si è cercato di limitare, per quanto possibile, i quesiti relativi a eventi lontani nel tempo. Ancora, fonte di errore può essere anche l'operato delle intervistatrici, che possono registrare valori non corretti o, nel caso di questionari cartacei, possono gestire i percorsi in maniera errata.

Gli errori che si riscontrano nei dati di un'indagine possono essere sia casuali sia sistematici; quelli casuali non portano a distorsioni nelle stime finali. Gli errori sistematici, invece, tendono a concentrarsi solo in alcune variabili o modalità di risposta, hanno sempre lo stesso segno e ogni ripetizione dell'indagine ne è affetta. Questi errori, quindi, causano distorsioni nei risultati finali e devono, dunque, essere attentamente tenuti sotto controllo grazie a indicatori di monitoraggio appositamente costruiti.

Il controllo e la correzione degli errori non campionari, che possono presentarsi in ogni fase del processo produttivo, richiede una strategia complessa. In questo ambito, poiché ogni metodologia di correzione a posteriori risolve solo parzialmente e a volte in modo non del tutto soddisfacente il problema, è fondamentale la prevenzione e la correzione degli errori contestualmente all'acquisizione dei dati; l'utilizzo del Cati, dunque, è di notevole supporto perché consente di:

- evitare errori e incompatibilità fra variabili gestendo in maniera automatica la navigazione condizionata all'interno del questionario;
- stabilire il range ammesso per ciascuna variabile, evitando che gli intervistatori inseriscano valori non ammessi e, quindi, errati;
- effettuare controlli di coerenza fra le risposte inserite durante l'intervista;
- utilizzare una codifica assistita, supportando gli intervistatori nel momento in cui debbano, per esempio, inserire un codice comunale (è sufficiente scrivere qualche lettera del nome del Comune per avere già a disposizione il relativo codice comunale da inserire) o un titolo di studio.

Grazie a questi controlli le interviste Cati presentano una elevata qualità in termini di correttezza e coerenza delle risposte fornite dalle intervistate.

Nel dettaglio, le correzioni hanno riguardato sia le eventuali incoerenze delle singole variabili o tra variabili logicamente collegate, sia la gestione delle mancate risposte, rappresentate dall'assenza di risposta ad uno o più quesiti nell'ambito di un'intervista effettuata.

Gli approcci per la correzione di questo tipo di errori sono stati sia di tipo probabilistico che di tipo

deterministico: in generale, l'approccio probabilistico prevede la definizione delle condizioni di errore e la correzione avviene a seguito dell'applicazione di un algoritmo probabilistico; l'approccio deterministico prevede invece che, a priori, vengano stabilite le condizioni di errore e le azioni da intraprendere per ciascuna di esse. Le regole impiegate nell'approccio deterministico sono del tipo:

SE (condizione di errore) ALLORA (azione di correzione).

È stato quindi necessario stabilire il valore "corretto" da assegnare alla variabile per la quale si è verificata la condizione di errore.

1.3 Individuazione e correzione degli errori

Le interviste sono state condotte mediante tecnica CATI (il questionario elettronico è stato sviluppato in-house mediante l'utilizzo del software Blaise); in questo modo è stato possibile prevedere svincoli e filtri automatici, come anche controlli in fase di inserimento delle risposte e verifiche di compatibilità fra variabili. L'utilizzo di un questionario elettronico ha indubbiamente permesso di avere a disposizione dati di partenza di buona qualità; ciononostante sono stati effettuati controlli sulla correttezza e compatibilità delle informazioni raccolte.

I controlli hanno riguardato gli errori di percorso (se, per esempio, la madre è single non deve rispondere ai quesiti sul partner), gli errori dovuti a valori fuori dominio (se per un quesito sono previsti valori che vanno da 1 a 5, l'operatore che effettua l'intervista non potrà registrare il valore 6) e le incompatibilità (ad es. la madre non può indicare un anno di acquisizione del titolo di studio precedente la sua data di nascita).

L'aver sviluppato internamente all'Istat il questionario elettronico ha permesso, come si è detto, di prevedere numerosi controlli di range e di coerenza (195 regole di controllo); è stato inoltre possibile verificare immediatamente la veridicità delle risposte fornite. Ne deriva che i dati dell'indagine Cati hanno richiesto un numero molto contenuto di interventi di correzione ex-post. Questi interventi si sono limitati essenzialmente ai casi in cui si è scelto di non inserire regole su alcuni quesiti e di riservarsi il controllo e la correzione alla fase successiva alle interviste. Questa scelta è stata dettata dalla necessità di non appesantire il questionario rendendone difficoltosa la navigazione durante l'intervista.

In media sono stati riscontrati errori sullo 0,5 per cento dei record. Di questi errori la maggior parte, lo 0,4 per cento, sono costituiti da errori di percorso; seguono gli errori di incoerenza, pari allo 0,04 per cento e, per ultimi, quelli di range, prossimi allo zero.

2. Strategia di campionamento e livello di precisione delle stime

1.1 Obiettivi dell'indagine

La *popolazione di interesse* dell'indagine – ossia l'insieme delle unità statistiche relativamente alle quali si intende investigare – è costituita dai nati iscritti in anagrafe nel corso del secondo semestre 2009 e primo semestre 2010; le unità di rilevazione, invece, sono le madri di tali nati, intervistate nel 2012 a una distanza media di circa due anni dal parto.

Ai fini dello studio del disegno campionario, le principali variabili oggetto di indagine sono l'ordine di nascita ed il tipo di filiazione. I domini di studio, ossia gli ambiti rispetto ai quali sono riferiti i parametri di popolazione oggetto di stima, sono le classi quinquennali di età della madre e, da un punto di vista territoriale, le venti regioni geografiche (con le province autonome di Bolzano e Trento considerate separatamente). Le stime dell'indagine, pertanto, sono prodotte con riferimento a tali domini o a incroci e aggregazioni ottenibili a partire da questi.

Per quanto riguarda la tecnica di indagine e la strategia campionaria, è stata utilizzata la medesima metodologia delle due edizioni precedenti dell'indagine (condotte nel 2002 e nel 2005), ossia rilevazione mediante tecnica CATI associata a un disegno campionario a uno stadio stratificato.

1.2 Disegno di campionamento

2.2.1 Lista di campionamento e informazioni disponibili per lo studio del disegno

La *lista di campionamento* per la selezione delle unità campionarie è costituita dall'archivio aggiornato di tutti i nati della popolazione residente in Italia nell'anno di riferimento, costruito a partire dalla *rilevazione degli iscritti in anagrafe per nascita*. In tale archivio, per ciascun nato sono riportate, oltre alle variabili identificative, all'indirizzo e al numero di telefono, informazioni di tipo territoriale (comune e provincia) e informazioni relative all'età della madre, che identificano i domini di stima e sono utili per la definizione di un disegno campionario stratificato.

1.2.2 Disegno campionario

L'universo complessivo di riferimento ammonta 541.862 nati. Per tale popolazione, rilevata mediante tecnica CATI, è stato definito un disegno a uno stadio stratificato. La stratificazione delle unità della popolazione è stata condotta sulla base dell'incrocio delle due variabili, presenti nell'archivio di selezione, che costituiscono i principali domini di interesse:

- classe di età della madre;
- regione di residenza.

Le classi utilizzate per la stratificazione in base all'età della madre sono le seguenti:

- fino a 24 anni,
- 25-29 anni;
- 30-34 anni;
- 35-39 anni;
- 40 anni e oltre.

L'incrocio di tale classificazione con la regione di residenza ha dato luogo alla definizione di

5x21=105 strati. Ciascun dominio di stima è così ottenibile come aggregazione di strati.

La numerosità campionaria complessiva è stata fissata in 17.716 unità. La distribuzione del campione tra gli strati è stata determinata in modo da garantire che gli errori di campionamento attesi delle principali stime riferite ai diversi domini di interesse non superassero prefissati livelli.

A tal scopo è stata utilizzata una metodologia basata su una generalizzazione del metodo di allocazione multivariata di Bethel al caso di più tipologie di domini di stima. Tale studio è stato condotto sulla base degli errori campionari di sei stime a livello di due diverse tipologie di domini di stima.

Una volta definite le numerosità campionarie teoriche negli strati, la selezione delle unità campionarie è stata effettuata, da ciascuno strato, senza reimmissione e con probabilità uguali.

Per garantire il raggiungimento del numero di interviste previste dal disegno campionario, è stato utilizzato il metodo del sovracampionamento che consiste nel selezionare un numero di unità campionarie superiore a quello progettato, tenendo conto del tasso di caduta osservato nell'indagine precedente.

1.3 Procedimento per il calcolo delle stime

Le stime prodotte dall'indagine sono essenzialmente stime di frequenze assolute e relative, riferite ai nati nel periodo di riferimento. Una stima di interesse è data, ad esempio, dal numero totale di nati da madri che lavorano al momento dell'indagine.

Le stime sono ottenute mediante uno stimatore di ponderazione vincolata, che è il metodo di stima adottato per la maggior parte delle indagini ISTAT sulle imprese e sulle famiglie.

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 a ogni unità campionaria un peso che indica il numero di unità della popolazione rappresentate dall'unità medesima. Se, per esempio, a un'unità campionaria viene attribuito un peso pari a 30, allora questa unità rappresenta se stessa e altre 29 unità della popolazione che non sono state incluse nel campione.

Al fine di rendere più chiara la successiva esposizione, introduciamo la seguente simbologia:

- d, indice del livello di riferimento delle stime (dominio di interesse), d=1,…, D;
- i, indice di unità (nato);
- h, indice dello strato, h=1,...,H;
- Nh , numero dei nati dello strato h;
- nh , numerosità campionaria nello strato h;
- y , generica variabile oggetto di indagine;
- Yi , valore osservato della variabile y sull'i-mo nato.

Se, ad esempio, *y* rappresenta la condizione lavorativa della madre (espressa dalle due modalità: lavora, non lavora), si avrà $Y_i = 1$ se la madre del nato i-mo lavora e $Y_i = 0$ altrimenti.

Si supponga di voler stimare con riferimento a un generico dominio d, il totale della variabile in esame, espresso dalla relazione:

$$Y_d = \sum_{i \in d} Y_i \qquad (1)$$

La stima del totale (1) è data da

$$\hat{Y} = \sum_{i \in d} W_i Y_i \,, \qquad (2)$$

in cui W_i è il peso finale da attribuire all'i-ma unità del campione.

Dalla precedente relazione si desume, quindi, che per ottenere la stima del totale (1) occorre moltiplicare il valore della variabile y assunto da ciascuna unità campionaria per il peso di tale unità ed effettuare, a livello del dominio di interesse, la somma dei prodotti così ottenuti.

1.3.1 Costruzione dei coefficienti di riporto all'universo

Il peso da attribuire alle unità campionarie è stato ottenuto mediante una procedura complessa che prevede le seguenti operazioni:

- correzione dell'effetto distorsivo della mancata risposta totale dovuto all'impossibilità di intervistare alcune delle unità selezionate per irreperibilità o rifiuto all'intervista;
- ricostruzione dei totali noti di importanti variabili ausiliarie correlate con le variabili d'indagine, nel senso che le stime campionarie dei totali noti delle variabili ausiliarie devono coincidere con i valori noti degli stessi.

Nell'indagine sulle nascite sono stati definiti i totali noti sulla base delle informazioni contenute nell'archivio di selezione; tali informazioni, utilizzate come variabili ausiliarie, sono note sia per le unità rispondenti, sia per le unità non rispondenti all'indagine e costituiscono la base per la costruzione di fattori correttivi per mancata risposta totale.

Le variabili ausiliarie considerate sono l'età, lo stato civile, la cittadinanza e il titolo di studio della madre e l'ordine di nascita. I totali noti utilizzati sono i seguenti:

- totale popolazione per ripartizione geografica e singolo anno di età (fino a 18, 19, ..., 44, 45 e oltre);
- totale popolazione per ripartizione, stato civile (coniugata, non coniugata) e 5 classi di età;
- totale popolazione per regione e 5 classi di età;
- totale popolazione per ripartizione e cittadinanza (italiana, straniera);
- totale popolazione per ripartizione e titolo di studio (basso, medio, alto);
- totale popolazione per ripartizione e ordine di nascita (primo, secondo, terzo e oltre).

Indicando con ${}_kX$ il totale noto della k-esima variabile ausiliaria e con ${}_kX_{hi}$ il valore assunto dalla variabile k nell'unità rispondente i dello strato h, la condizione di uguaglianza tra il valore del totale noto e la stima campionaria del totale stesso ${}_k\hat{X}$ è espressa dalla seguente relazione:

$${}_k\hat{X} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} {}_kX_{hi} W_{hi} = {}_kX \qquad (k=1,\ldots., K)$$

dove:

- Whi è il peso finale dell'unità i dello strato h;
- H è il numero complessivo di strati;
- nh è il numero unità rispondenti nello strato h.

Le variabili X_{hi} sono variabili dicotomiche, quindi se, per esempio, ${}_1X$ indica il numero di nati da madri di età inferiore o uguale a 18 anni nella prima ripartizione geografica, la variabile ausiliaria ${}_1X_{hi}$ assume il valore 1 se l'unità i dello strato h corrisponde a un nato da madre di età inferiore o uguale a 18 anni e appartenente alla prima ripartizione geografica e il valore 0 altrimenti.

La procedura che consente di costruire i pesi finali, da attribuire alle unità campionarie rispondenti, è articolata nelle seguenti fasi:

1. calcolo dei pesi diretti come reciproco della probabilità di inclusione delle unità, uguale per tutte le unità di uno stesso strato e fornita dall'espressione:

$$\pi_{hi}^{*} = N_h / n_h^{*}$$

dove N_h è il numero di unità dello strato h e n_h^{*} è il numero di unità estratte nello strato h;

2. calcolo dei fattori correttivi per mancata risposta totale, come inverso del tasso di risposta all'interno dello strato a cui ciascuna unità appartiene:

$$c_{hi} = n_h^{*} / n_h$$

3. calcolo dei pesi base, o pesi corretti per mancata risposta totale, moltiplicando i pesi diretti per i corrispondenti fattori correttivi per mancata risposta totale:

$$\pi_{hi} = \left(N_h / n_h^{*}\right)\left(n_h^{*} / n_h\right) = N_h / n_h$$

4. costruzione dei fattori correttivi γhi che consentono di soddisfare la condizione di uguaglianza tra i totali noti delle variabili ausiliarie e le corrispondenti stime campionarie;

5. determinazione dei pesi finali mediante il prodotto dei pesi base per i fattori correttivi ottenuti al passo 4:

$$W_{hi} = \pi_{hi} \times \gamma_{hi}$$

I fattori correttivi del passo 4 sono ottenuti mediante la risoluzione di un problema di minimo vincolato, in cui la funzione da minimizzare è una funzione di distanza (opportunamente prescelta) tra i pesi base e i pesi finali e i vincoli sono definiti dalla condizione di uguaglianza tra le stime campionarie dei totali noti della popolazione e i valori degli stessi. La funzione di distanza prescelta è la funzione logaritmica troncata; l'adozione di tale funzione garantisce che i pesi finali siano positivi e contenuti in un predeterminato intervallo di valori possibili, riducendo in tal modo i valori estremi (troppo grandi o troppo piccoli).

I metodi di stima che scaturiscono dalla risoluzione del problema di minimo vincolato sopra descritto rientrano nella classe generale di stimatori noti come *stimatori di ponderazione vincolata*[2].

1.4 Valutazione del livello di precisione delle stime

2.4.1 Metodologia di calcolo degli errori campionari

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

[2] Nella letteratura in lingua anglosassone sull'argomento tali stimatori sono noti come *calibration estimators.*

con $\hat{V}ar(\hat{Y}_d)$ la stima della varianza della generica stima $\hat{Y}_d$, la stima dell'errore di campionamento assoluto di $\hat{Y}_d$ si può ottenere mediante la relazione

$$\hat{\sigma}(\hat{Y}_d) = \sqrt{\hat{V}ar(\hat{Y}_d)}; \qquad (3)$$

la stima dell'errore di campionamento relativo di $\hat{Y}_d$ è invece definita dall'espressione

$$\hat{\varepsilon}(\hat{Y}_d) = \frac{\hat{\sigma}(\hat{Y}_d)}{\hat{Y}_d}. \qquad (4)$$

Come è stato descritto in precedenza, le stime prodotte dall'indagine sono state ottenute mediante uno stimatore di ponderazione vincolata definito in base ad una funzione di distanza di tipo logaritmico troncato. Poiché lo stimatore adottato non è funzione lineare dei dati campionari, per la stima della varianza $\hat{V}ar(\hat{Y}_d)$ si è utilizzato il metodo proposto da Woodruff[3]; in base a tale metodo, che ricorre all'espressione linearizzata in serie di Taylor, è possibile ricavare la varianza di ogni stimatore non lineare (funzione regolare di totali) calcolando la varianza dell'espressione linearizzata ottenuta. In particolare, per la definizione dell'espressione linearizzata dello stimatore ci si è riferiti allo stimatore di regressione generalizzata, sfruttando la convergenza asintotica di tutti gli stimatori di ponderazione vincolata a tale stimatore, poiché nel caso di stimatori di ponderazione vincolata che utilizzano funzioni distanza differenti dalla distanza euclidea (che conduce allo stimatore di regressione generalizzata) non è possibile derivare l'espressione linearizzata dello stimatore. L'espressione linearizzata dello stimatore (2) è data, quindi, da:

$$\hat{Y}_d \cong \hat{Z}_d = \sum_{i=1}^{n} \hat{Z}_i \qquad (5)$$

dove Z_i è la variabile linearizzata per la generica unità rispondente i, espressa come $Z_i = Y_i - X_i'\beta$, essendo $X_i = ({}_1X_i, ..., {}_kX_i, ... {}_KX_i)'$ il vettore contenente i valori delle K variabili ausiliarie, osservati per la generica unità campionaria i e $\hat{\beta}$, il vettore dei coefficienti di regressione del modello lineare che lega la variabile di interesse y alle K variabili ausiliarie *x*. In base alla (5), la stima della varianza della stima $\hat{Y}_d$ è ottenibile in generale mediante la seguente relazione:

$$\hat{V}ar(\hat{Y}_d) \cong \hat{V}ar(\hat{Z}_d) = \sum_{h=1}^{H_d} \hat{V}ar(\hat{Z}_h) + \sum_{g=1}^{G_d} \hat{V}ar(\hat{Z}_g), \qquad (6)$$

ossia la stima della varianza della stima $\hat{Y}_d$ viene calcolata come somma della stima delle varianze

[3] Woodruff R.S. (1971), A Simple method for approximating the variance of a complicate estimate, *Journal of the American Statistical Association*, 66, pp 411-414.

della variabile linearizzata nei singoli strati appartenenti al dominio d, se il dominio d è ottenibile come aggregazione di strati. Nell'indagine in oggetto, poiché la popolazione è stata partizionata in due sotto-popolazioni che sono state stratificate in modo differente, le stime della varianza riferita ai domini di stima (tranne quelle per il dominio nazionale) non possono essere ottenute utilizzando la formula (6). E' infatti necessario definire per ogni variabile di interesse *y* e ogni dominio di stima d, una corrispondente variabile ${}_dY'_i$ definita come

$$ {}_dY'_i = \begin{cases} Y_i & \text{se l'unità } i \in d \\ 0 & \text{se l'unità } i \notin d \end{cases} $$

e sostituire questa variabile al posto di Y_i nelle espressioni della variabile linearizzata dando luogo alla variabile $Z'_i = Y'_i - X_i'\beta$ per la stima della varianza campionaria utilizzando le usuali espressioni del tipo (6).

In particolare, per la parte di campione appartenente agli strati derivanti dal disegno a uno stadio stratificato, le espressioni utilizzate per la stima della varianza sono del tipo:

$$ \sum_{h=1}^{H_d} \hat{V}ar\left(\hat{Z}'_h\right) = \sum_{h=1}^{H_d} N_h^2 \frac{(N_h - n_h)}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \left(Z'_{hi} - \overline{Z}'_h\right)^2 , \qquad (7) $$

dove si è posto $\overline{Z}'_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Z'_{hi}$.

Una volta calcolata la stima della varianza campionaria, utilizzando l'espressione (7), è possibile ottenere rispettivamente l'errore di campionamento assoluto e l'errore di campionamento relativo delle stime di interesse.

Tali errori consentono di valutare il grado di precisione delle stime; inoltre, l'errore assoluto permette di costruire un intervallo di confidenza che, al livello di fiducia P, contiene il parametro oggetto di stima, tale intervallo viene espresso come:

$$ \left\{ \hat{Y}_d - k_p \hat{\sigma}(\hat{Y}_d) \le Y_d \le \hat{Y}_d + k_p \hat{\sigma}(\hat{Y}_d) \right\} \qquad (8) $$

Nella (8) il valore di k_P dipende dal valore fissato per la probabilità P; ad esempio, per P=0.95 si ha k=1.96.

1.4.2 Presentazione sintetica degli errori campionari

Poiché a ogni stima $\hat{Y}_d$ è associato un errore campionario relativo $\hat{\varepsilon}(\hat{Y}_d)$, per consentire un uso corretto delle stime fornite dall'indagine, sarebbe necessario presentare, per ciascuna stima pubblicata, anche il corrispondente errore relativo. Ciò non è possibile, 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 di stime non pubblicate, che l'utente può ricavare in modo autonomo.

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

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

$$\log\hat{\varepsilon}^2(\hat{Y}_d) = a + b\log(\hat{Y}_d) \qquad (8)$$

dove i parametri a e b sono stimati mediante il *metodo dei minimi quadrati*.

Per calcolare gli errori di campionamento è stato utilizzato il software generalizzato Genesees[4], messo a punto presso l'Istat, che consente di ottenere gli errori campionari e gli intervalli di confidenza e di costruire modelli regressivi del tipo (8) per la presentazione sintetica degli errori di campionamento.

La tavola 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 di frequenze riferite ai nati, relative alle variabili rilevate sulle unità del campione complessivo, per ripartizione geografica, regione e classe di età della madre.

Sulla base delle informazioni contenute nella suddetta tavola, è possibile calcolare l'errore relativo di una determinata stima di frequenza assoluta $\hat{Y}_d^*$, riferita ai diversi domini, mediante la formula:

$$\hat{\varepsilon}(\hat{Y}_d^*) = \sqrt{\exp(a + b\log(\hat{Y}_d^*))} \qquad (9)$$

e costruire l'intervallo di confidenza al 95% come:

$$\left\{\hat{Y}_d^* - 1.96\cdot\hat{\varepsilon}(\hat{Y}_d^*)\cdot\hat{Y}_d^*;\ \hat{Y}_d^* + 1.96\cdot\hat{\varepsilon}(\hat{Y}_d^*)\cdot\hat{Y}_d^*\right\}.$$

Allo scopo di facilitare il calcolo degli errori campionari, nelle tavole 2, 3 e 4 sono riportati, gli errori relativi percentuali corrispondenti a valori crescenti di stime di frequenze assolute riferite ai nati calcolati introducendo nella (9) i valori di a e b riportati nella tavola 1.

Tali informazioni consentono di calcolare l'errore relativo di una generica stima di frequenza assoluta mediante due procedimenti di facile applicazione che conducono a risultati meno precisi di quelli ottenibili applicando direttamente la formula (9).

Il primo metodo consiste nell'approssimare l'errore relativo della stima di interesse $\hat{Y}_d^*$ con quello, riportato nei prospetti, corrispondente al livello di stima che più si avvicina a $\hat{Y}_d^*$.

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 $\hat{Y}_d^*$, si ricava mediante l'espressione:

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

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

---

[4] Pagliuca (a cura di), 2002, Funzioni di Genesees, Manuali Utente e Aspetti Metodologici, link: http://www.istat.it/Metodologi/index.htm (selezionare "Metodi e Software per indagini statistiche").

Tavola 1 - Valori dei coefficienti a e b e dell'indice di determinazione R^2 (%) del modello per l'interpolazione degli errori campionari delle stime riferite ai nati per ripartizione geografica, regione e classe di età della madre

DOMINIO DI STIMA	a	b	R^2
ITALIA	**4,08088**	**-1,09370**	**90,23**
RIPARTIZIONI GEOGRAFICHE			
Nord	4,26274	-1,13537	88,77
Centro	4,91740	-1,21892	86,54
Sud e Isole	3,71434	-1,07595	87,68
REGIONI			
Piemonte	4,28965	-1,14730	88,40
Valle d'Aosta	2,22713	-1,20514	70,36
Lombardia	4,19585	-1,10814	87,36
Bolzano-Bozen	3,39323	-1,27551	88,87
Trento	3,19814	-1,29410	91,68
Veneto	4,08315	-1,11486	88,27
Friuli-Venezia Giulia	3,44797	-1,23263	90,46
Liguria	3,87501	-1,24662	90,61
Emilia Romagna	4,10870	-1,11970	85,69
Toscana	4,40185	-1,14721	88,03
Umbria	3,24172	-1,21430	88,22
Marche	4,19070	-1,23133	87,23
Lazio	4,20512	-1,09667	87,04
Abruzzo	3,19125	-1,19016	92,97
Molise	2,08070	-1,21604	88,82
Campania	4,06533	-1,07343	85,47
Puglia	3,99054	-1,10821	89,41
Basilicata	2,44067	-1,18101	90,46
Calabria	3,10407	-1,09662	90,83
Sicilia	4,06863	-1,10208	89,26
Sardegna	3,19299	-1,15002	91,56
CLASSI DI ETA' DELLA MADRE			
Fino a 24	4,63769	-1,27918	87,05
25 - 29	4,82372	-1,19607	85,77
30 - 34	4,06509	-1,11916	85,49
35 - 39	4,90706	-1,19200	91,44
40 e oltre	4,63965	-1,27243	89,93

Tavola 2 - Valori interpolati degli errori campionari relativi percentuali delle stime riferite ai nati per ripartizione geografica

STIMA	RIPARTIZIONE GEOGRAFICA			Italia
	Nord	Centro	Sud e Isole	
250	36,7	40,4	32,8	37,6
500	24,7	26,5	22,6	25,7
750	19,7	20,7	18,2	20,6
1.000	16,7	17,4	15,6	17,6
1.250	14,7	15,1	13,8	15,6
1.500	13,3	13,6	12,5	14,1
1.750	12,2	12,3	11,5	13,0
2.000	11,3	11,4	10,7	12,1
2.500	9,9	9,9	9,5	10,7
5.000	6,7	6,5	6,6	7,3
10.000	4,5	4,3	4,5	5,0
20.000	3,0	2,8	3,1	3,4
30.000	2,4	2,2	2,5	2,7
40.000	2,1	1,8	2,1	2,3
50.000	1,8	1,6	1,9	2,1
60.000	1,6	1,4	1,7	1,9
70.000	1,5	1,3	1,6	1,7
80.000	1,4	1,2	1,5	1,6
90.000	1,3	1,1	1,4	1,5
100.000	1,2	1,0	1,3	1,4
150.000	1,0		1,1	1,1
200.000	0,8			1,0
250.000	0,7			0,9

Tavola 3 - Valori interpolati degli errori campionari relativi percentuali delle stime riferite ai nati per regione

STIMA	Piemonte	Valle d'Aosta	Lombardia	Bolzano-Bozen	Trento	Veneto	Friuli-Venezia Giulia	Liguria	Emilia Romagna	Toscana	Umbria
250	36,0	10,9	38,2	16,1	13,9	35,5	18,7	22,2	35,5	38,1	17,7
500	24,2	7,2	26,0	10,4	8,9	24,1	12,2	14,4	24,1	25,6	11,6
750	19,2	5,6	20,8	8,0	6,8	19,2	9,5	11,2	19,2	20,3	9,1
1.000	16,2	4,7	17,7	6,7	5,7	16,4	7,9	9,4	16,3	17,2	7,6
1.250	14,3		15,7	5,8	4,9	14,5	6,9	8,1	14,4	15,1	6,7
1.500	12,9		14,2	5,1	4,4	13,1	6,2	7,3	13,0	13,6	6,0
1.750	11,8		13,0	4,7	3,9	12,0	5,6	6,6	11,9	12,5	5,4
2.000	10,9		12,1	4,3	3,6	11,1	5,2	6,1	11,1	11,5	5,0
2.250	10,2		11,3	4,0	3,4	10,4	4,8	5,6	10,4	10,8	4,7
2.500	9,6		10,7	3,7	3,1	9,8	4,5	5,3	9,8	10,2	4,4
2.750	9,1		10,1	3,5	2,9	9,3	4,3	5,0	9,3	9,6	4,1
3.000	8,6		9,7	3,3	2,8	8,9	4,0	4,7	8,8	9,1	3,9
3.500	7,9		8,9	3,0	2,5	8,1	3,7	4,3	8,1	8,4	3,6
4.000	7,3		8,2	2,8	2,3	7,6	3,4	3,9	7,5	7,8	3,3
4.500	6,9		7,7	2,6	2,1	7,1	3,1	3,7	7,0	7,3	3,1
5.000	6,5		7,3	2,4	2,0	6,7	2,9	3,4	6,6	6,8	2,9
7.500	5,1		5,8			5,3	2,3	2,7	5,3	5,4	2,2
10.000	4,3		5,0			4,5		2,2	4,5	4,6	
15.000	3,4		4,0			3,6			3,6	3,6	
20.000	2,9		3,4			3,1			3,0	3,1	
40.000			2,3			2,1					

Tavola 3 *(segue)* - Valori interpolati degli errori campionari relativi percentuali delle stime riferite ai nati per regione

STIMA	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
250	27,1	39,7	18,5	9,9	39,4	34,5	13,0	22,9	36,5	20,6
500	17,7	27,1	12,2	6,5	27,2	23,5	8,6	15,6	24,9	13,8
750	13,8	21,7	9,6	5,1	21,9	18,8	6,8	12,5	19,9	11,0
1.000	11,6	18,5	8,1	4,2	18,7	16,0	5,7	10,7	17,0	9,3
1.250	10,1	16,4	7,1	3,7	16,6	14,1	5,0	9,5	15,0	8,2
1.500	9,0	14,8	6,4	3,3	15,1	12,8	4,5	8,6	13,6	7,4
1.750	8,2	13,6	5,8	3,0	13,9	11,7	4,1	7,9	12,5	6,7
2.000	7,5	12,7	5,4	2,8	12,9	10,9	3,8	7,3	11,6	6,2
2.250	7,0	11,9	5,0	2,6	12,1	10,2	3,6	6,9	10,9	5,8
2.500	6,6	11,2	4,7		11,5	9,6	3,3	6,5	10,3	5,5
2.750	6,2	10,6	4,4		10,9	9,1	3,2	6,1	9,7	5,2
3.000	5,9	10,2	4,2		10,4	8,7	3,0	5,9	9,3	4,9
3.500	5,3	9,3	3,8		9,6	8,0	2,7	5,4	8,5	4,5
4.000	4,9	8,7	3,5		8,9	7,4	2,5	5,0	7,9	4,2
4.500	4,6	8,1	3,3		8,4	7,0	2,4	4,7	7,4	3,9
5.000	4,3	7,7	3,1		7,9	6,6		4,4	7,0	3,7
7.500	3,3	6,1	2,4		6,4	5,2		3,5	5,6	2,9
10.000	2,8	5,2	2,1		5,4	4,5		3,0	4,8	2,5
15.000		4,2			4,4	3,6		2,4	3,8	
20.000		3,6			3,8	3,0			3,3	
40.000		2,5			2,6				2,2	

Tavola 4 - Valori interpolati degli errori campionari relativi percentuali delle stime riferite ai nati per classe di età della madre

STIMA	Fino a 24	25 - 29	30 - 34	35 - 39	40 e oltre
250	29,7	41,1	34,7	43,3	30,3
500	19,1	27,1	23,6	28,6	19,5
750	14,7	21,3	18,8	22,5	15,1
1.000	12,3	17,9	16,0	18,9	12,6
1.250	10,6	15,7	14,1	16,6	10,9
1.500	9,5	14,1	12,7	14,9	9,7
1.750	8,6	12,8	11,7	13,6	8,8
2.000	7,9	11,8	10,9	12,5	8,1
2.500	6,8	10,4	9,6	11,0	7,0
5.000	4,4	6,8	6,5	7,3	4,5
10.000	2,8	4,5	4,4	4,8	2,9
20.000	1,8	3,0	3,0	3,2	1,9
30.000	1,4	2,3	2,4	2,5	1,4
40.000	1,2	2,0	2,0	2,1	
50.000	1,0	1,7	1,8	1,8	
60.000		1,5	1,6	1,7	
70.000		1,4	1,5	1,5	
80.000		1,3	1,4	1,4	
90.000		1,2	1,3	1,3	
100.000		1,1	1,2	1,2	
150.000		0,9	1,0		