File Standard

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

Indagine CATI+PAPI STRANIERE

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, 1.3 e 1.4 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.

I paragrafi 1.2, 1.3 e 1.4 e il capitolo 2 sono stati ripresi dall'ebook "Avere figli in Italia negli anni 2000. Approfondimenti dalle indagini campionarie sulle nascite e sulle madri", Istat, temi, dicembre 2014.

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

[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/).

durante il 2012; un modulo di approfondimento è dedicato proprio all'interazione maternità-lavoro; nell'edizione 2012, i contenuti sono stati rivisitati e arricchiti grazie anche alla collaborazione con l'Isfol nell'ambito di una specifica convenzione tra i due Enti.

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.

In questa edizione, oltre all'indagine CATI svolta in modo da garantire la continuità con le edizioni precedenti, è stata effettuata un'Indagine sulle madri straniere con tecnica PAPI[2].

I dati diffusi in questo file mettono insieme l'Indagine CATI e la PAPI, per rendere i risultati rappresentativi, grazie alla messa a punto di un 'peso unico' di riporto all'universo, anche della componente straniera, fondamentale per lo studio e interpretazione dei fenomeni trattati (i principali risultati sono pubblicati nell'ebook *"Avere figli in Italia negli anni 2000. Approfondimenti dalle indagini campionarie sulle nascite e sulle madri", Istat, temi, dicembre 2014,* http://www.istat.it/it/archivio/147180).

Data infatti la crescente rilevanza delle nascite da almeno un genitore straniero, la tipologia di coppia (genitori entrambi italiani, coppie miste costituite da un partner italiano e uno straniero, genitori entrambi stranieri), è divenuta una variabile strutturale di rilievo per l'analisi dei comportamenti riproduttivi.

I principali risultati dell'Indagine sono pubblicati nell'ebook "Avere figli in Italia negli anni 2000. Approfondimenti dalle indagini campionarie sulle nascite e sulle madri", Istat, temi, dicembre 2014.

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] Per quanto riguarda la tecnica di indagine e la strategia campionaria, lo studio ha preso l'avvio dalla metodologia utilizzata per le 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. I risultati degli abbinamenti, tra l'archivio contenente le unità della popolazione di interesse e le liste di telefonia fissa, per il reperimento delle utenze telefoniche, hanno evidenziato, tuttavia, che la popolazione straniera sfugge alla tecnica CATI. Si è deciso pertanto di suddividere la popolazione di riferimento dell'indagine in due sottopopolazioni sulla base della cittadinanza dei nati (italiani o stranieri) e di condurre la rilevazione con due modalità diverse.

Sul collettivo dei 'nati italiani' (ovvero nati da almeno un genitore italiano) è stata condotta come per le precedenti edizioni una indagine con interviste CATI, mentre sul collettivo dei nati stranieri (ovvero con entrambi i genitori stranieri) la rilevazione è stata condotta mediante intervista faccia a faccia con questionario cartaceo somministrato da intervistatori comunali appositamente formati. Del collettivo dei 'nati stranieri' fanno parte i bambini nati da entrambi i genitori stranieri in comuni nei quali siano nati nel 2010 almeno 50 bambini da entrambi i genitori stranieri, ossia i comuni nei quali si concentra maggiormente il fenomeno di nascite di bambini stranieri; invece, i nati stranieri nei comuni di dimensione inferiore a detta soglia sono stati inclusi nella rilevazione CATI, dal momento che per ragioni organizzative non era possibile condurre una rilevazione con intervista diretta su comuni piccoli.

Tale circostanza ha determinato la necessità di utilizzare sui due collettivi due disegni di campionamento differenti.

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 L'approccio multi-canale: l'integrazione di due questionari

La terza edizione dell'Indagine Campionaria sulle Nascite e le Madri è stata condotta, come si è detto, con tecniche diverse a seconda dell'unità di rilevazione prevista: le interviste alle madri italiane sono state condotte mediante tecnica Cati[3] (il questionario elettronico è stato sviluppato in-house mediante l'utilizzo del software Blaise), mentre per le madri straniere si è utilizzata la tecnica Papi.

Per effettuare le interviste faccia a faccia si è dovuto procedere a una riduzione del questionario

[3] 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.

utilizzato con la tecnica Cati. Mentre il questionario elettronico consente di prevedere svincoli e filtri automatici, come anche controlli in fase di inserimento delle risposte e verifiche di compatibilità fra variabili, il questionario cartaceo deve necessariamente essere più snello e semplice, per evitare il più possibile errori in fase di compilazione. Una volta terminata l'acquisizione dei dati è stato quindi indispensabile procedere all'omogeneizzazione dei due questionari come di seguito sinteticamente illustrato.

- Una delle prime operazioni necessarie per l'integrazione è stata l'individuazione delle mancate risposte: mentre nel Cati non è possibile avere valori mancanti per un quesito per il quale, in base al percorso effettuato durante l'intervista, sia prevista una risposta, lo stesso non accade nell'intervista Papi, per la quale un valore mancante può essere tanto un "Non sa/Non risponde" quanto una domanda a cui effettivamente l'intervistata non deve rispondere. È stato quindi indispensabile distinguere i valori mancanti corrispondenti ad una volontà della madre di non rispondere ad un dato quesito, da quelli che, invece, erano mancanti perché il percorso non prevedeva che il quesito venisse posto.

- Si è poi proceduto all'individuazione dei membri della famiglia della donna intervistata. Mentre, infatti, nel questionario Cati si chiedeva di quante persone fosse composta la famiglia e, per alcuni di essi (figli conviventi e non e partner), si chiedevano alcune informazioni, nel questionario cartaceo si è scelto di concentrare tutti i dati anagrafici dei vari componenti in un'unica tabella riepilogativa iniziale. A partire da questa, grazie alla data di nascita e alla relazione di parentela con la donna intervistata, è stato individuato il bambino per cui la madre è stata intervistata. Allo stesso modo è stato individuato l'eventuale partner (marito o compagno) della donna e gli eventuali altri figli conviventi e non e gli altri familiari.

- Sono state quindi omogeneizzate le variabili che presentavano modalità differenti nei due questionari: è il caso dello stato civile, del motivo di acquisizione della cittadinanza italiana, del titolo di studio e del settore di attività economica.

- I due archivi, a questo punto sono stati uniti, facendo però in modo che, per ciascun record, fosse sempre possibile risalire al tipo di indagine da cui provenivano.

- A ciascun record dell'archivio unificato è stato associato un peso per il riporto all'universo (cfr. par. 3.1).

Il passaggio successivo all'unione dei dati provenienti dalle due indagini è stata la verifica dei percorsi, l'individuazione degli errori e/o delle incompatibilità e la conseguente correzione.

1.3 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.

Diverso è il caso delle interviste PAPI che hanno richiesto un'attività di validazione più massiva per identificare e correggere gli errori una volta portata a termine l'indagine.

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 possono essere di tipo probabilistico o deterministico: 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 quindi necessario stabilire quale debba essere il valore "corretto" da assegnare alla variabile per la quale si è verificata la condizione di errore.

1.4 Individuazione e correzione degli errori

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.

La maggior parte degli interventi sono stati effettuati, dunque, sulle interviste Papi, per le quali si è riscontrato almeno un errore o incompatibilità per record. In media, per ogni record Papi sono stati effettuati 8,8 interventi correttivi. Al contrario per i dati dell'indagine Cati sono stati effettuati in media 1,5 interventi per armonizzare i percorsi in accordo con le regole semplificate utilizzate per il Papi.

Di seguito sono descritti in maniera più dettagliata gli errori più comuni che si sono riscontrati in fase di validazione delle interviste:

- Gli errori di percorso: sono gli errori più frequenti in entrambe le indagini, seppure con le dovute differenze. In media nel questionario Papi gli errori di questo tipo sono 4,4 per record, mentre solo lo 0,8 per cento dei record Cati ne sono affetti. La più alta incidenza di questo tipo di errori per le interviste Papi è facilmente spiegabile, considerando la mancanza di automazione nella gestione dei salti. È questo il caso ad esempio delle variabili Q1_10_3 e Q1_10_3_1, riguardanti l'eventuale nascita del bambino all'interno di una convivenza col padre e l'anno di inizio di questa convivenza. Questi due quesiti sono fra quelli maggiormente affetti da errore (rispettivamente l'85,5 per cento e il 69,0 per cento delle madri straniere), poiché il quesito è stato erroneamente posto a donne non nubili oppure che hanno già dichiarato la nascita del bambino all'interno del matrimonio con il padre. Si tratta quindi chiaramente di un errore di percorso, visto che, analizzando i dati, oltre il 76 per cento delle donne che hanno erroneamente risposto al quesito Q1_10_3, hanno fornito risposta "Non è nato in una convivenza" oppure "Non risponde" e per oltre l'83 per cento dei casi, "Non risponde" è stata la modalità scelta per il quesito Q1_10_3_1. Dello stesso tipo sono gli errori riscontrati sul quesito Q2_10_1 sugli aspetti principali che causano difficoltà di conciliazione lavoro/famiglia, chiesto a madri che hanno dichiarato di non avere difficoltà di conciliazione: anche in questo caso gli errori riguardano esclusivamente le madri straniere (l'errore riguarda il 23,2 per cento delle straniere intervistate) e si tratta, quasi esclusivamente, di un errore di percorso, considerando il fatto che per oltre il 98 per cento delle madri la scelta è stata "Non risponde".

- I valori fuori dominio: subito dopo quelli di percorso, i fuori dominio sono gli errori maggiormente presenti. Si presentano in media con una frequenza di uno per record nei dati Papi. Nella maggior parte dei casi si tratta di variabili con risposta "Non sa/Non risponde" non prevista. La correzione è stata effettuata considerando la coerenza con le risposte fornite ad altri quesiti correlati con quello affetto da errore.

- Le incompatibilità: sono gli errori che si sono riscontrati meno frequentemente (in media 0,9 casi per record nella Papi).

2. Disegno di campionamento

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

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.

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.

La progettazione del disegno campionario di questa edizione dell'indagine ha tenuto conto per la prima volta delle differenti possibilità di contatto e di intervista con le unità di interesse a seconda del tipo di cittadinanza. Questo ha consentito non solo di limitare il rischio di mancata inclusione nel campione delle madri straniere dovuto alle difficoltà di reperimento dei numeri telefonici, ma anche di avere tassi di risposta più elevati. Infatti mentre le madri italiane o straniere con partner italiano hanno una buona propensione al contatto e alla risposta in un'indagine di tipo telefonico, tra le madri di bambini con entrambi i genitori stranieri il tasso di risposta a un'indagine CATI è molto basso, circostanza che determina problemi di rappresentatività del campione e di correttezza delle stime che da essa vengono prodotte.

Pertanto, a partire dalla lista relativa alla popolazione di interesse, sono stati individuati due collettivi che costituiscono una partizione dell'intera popolazione obiettivo e su ognuno di essi è stato definito il disegno campionario più idoneo per la tecnica di rilevazione prescelta.

Sul collettivo delle madri italiane, non esistendo la necessità di concentrare il campione sul territorio, è stato possibile definire, come fatto per le precedenti edizioni dell'indagine sulle nascite, un disegno di campionamento ad uno stadio stratificato. Invece per le madri straniere è stato necessario utilizzare un disegno a due stadi (in cui le unità di primo stadio sono i comuni) come è solitamente necessario fare quando l'intervista deve avvenire *faccia a faccia*. È utile ricordare che, in generale, utilizzare un disegno di campionamento ad uno stadio stratificato è preferibile perché determina un guadagno nell'efficienza delle stime rispetto ad un disegno a due stadi, nel quale le stime risentono dell'associazione tra le unità appartenenti stesso comune. In questo caso tuttavia, la scelta di un disegno a due stadi ha costituito una soluzione per ridurre l'impatto distorsivo derivante da una mancata risposta elevata in una parte rilevante della popolazione di interesse.

2.2 Disegno campionario per la sotto-popolazione delle madri italiane

L'universo complessivo di riferimento ammonta a 541.862 nati, 480.251 dei quali sono nati da genitori di cui almeno uno è italiano[4]. Per tale sottopopolazione, 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 che costituiscono i principali domini di interesse: la classe di età della madre e la regione di residenza, presenti entrambe nell'archivio di selezione.

La stratificazione in base all'età della madre è stata effettuata secondo una classificazione in cinque classi: fino a 24 anni, 25-29, 30-34, 35-39, 40 e oltre. L'incrocio di tale classificazione con la regione di residenza ha dato luogo alla definizione di 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 questo scopo è stata utilizzata una metodologia basata su una generalizzazione del metodo di allocazione multivariata di Bethel (Bethel, 1989) al caso di più tipologie di domini di stima (si veda Falorsi *et al*., 1998). Tale studio è stato effettuato 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

4 Si veda nota 1.

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 sovra-campionamento che consiste nel selezionare, per la rilevazione, un numero di unità campionarie superiore a quello progettato, tenendo conto del tasso di caduta osservato nell'indagine precedente. Il campione realizzato è di 17.603 unità.

2.3 Disegno campionario per la sotto-popolazione delle madri straniere

Il disegno di campionamento è a due stadi di selezione con stratificazione delle unità di primo stadio. Le unità di primo stadio sono rappresentate dai comuni, stratificati per dimensione in termini di nati.

Le unità di secondo stadio sono le madri di bambini con entrambi i genitori stranieri nati nella seconda metà del 2009 e nella prima metà del 2010, classificate in base alla macro-area geografica di provenienza. Le macro-aree definite sono: U.E. più altri paesi europei, America settentrionale e Oceania; Europa centro-orientale; Africa; Asia; America centro-meridionale.

La numerosità campionaria complessiva 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 che per questa sotto-popolazione sono la ripartizione geografica e la macro-area di cittadinanza della madre. La dimensione complessiva del campione di madri da intervistare è stata fissata pari a 2.000 unità; per ragioni di tipo organizzativo legate alla rilevazione, si è deciso di considerare come universo di primo stadio i comuni in cui sono avvenute almeno 50 nascite da madri straniere nel periodo di riferimento.

L'universo di primo stadio è pertanto costituito dai 180 comuni in cui si sono registrate 28.953 nascite da genitori stranieri, variamente distribuite tra le macro-aree di provenienza della madre.

L'allocazione delle 2.000 interviste tra le cinque macro-aree di provenienza è stata effettuata proporzionalmente al peso delle macro-aree stesse.

Infine, avendo posto pari a circa 40 il numero di madri da intervistare in ciascun comune campione, si è ottenuto che il numero di comuni campione in cui effettuare le interviste deve risultare uguale a 50.

2.3.1 Stratificazione e selezione delle unità di primo stadio

L'obiettivo della stratificazione è quello di formare gruppi (o strati) di unità caratterizzate, relativamente alle variabili oggetto d'indagine, da massima omogeneità interna agli strati e massima eterogeneità fra gli strati. Il raggiungimento di tale obiettivo si traduce in termini statistici in un guadagno nella precisione delle stime, ossia in una riduzione dell'errore campionario a parità di numerosità campionaria.

Per la selezione delle 50 unità di primo stadio, i 180 comuni dell'universo sono stati stratificati rispetto al numero totale di nascite da madri straniere nel comune, in modo tale che venissero rispettate le seguenti condizioni:

- auto-ponderazione del campione a livello degli strati, che equivale alla condizione che tutte unità finali appartenenti a uno strato abbiano la stessa probabilità di essere selezionate nel campione;
- formazione di strati aventi ampiezza approssimativamente costante in termini di nati;
- bilanciamento rispetto ai totali di popolazione dei nati per macro-area.

Indipendentemente dalla loro regione o ripartizione geografica di appartenenza, i comuni sono stati stratificati sulla base del numero di nati in modo da ottenere strati di dimensione approssimativamente costante. I 15 comuni in cui sono avvenute più di 300 nascite, corrispondenti alle maggiori aree

metropolitane, sono stati considerati come autorappresentativi ed considerati ognuno uno strato a sé; i restanti comuni sono stati equidistribuiti tra 6 strati.

I comuni campione sono stati estratti, con probabilità uguali da ogni strato, mediante selezione bilanciata (Deville e Tillé, 2004) in modo da ottenere che le stime dirette del numero di nati a livello di macro-area di cittadinanza della madre risultassero coincidenti con i rispettivi totali di popolazione. Per la selezione è stata utilizza la Macro SAS Cube.

2.3.2 Selezione delle unità di secondo stadio

La selezione delle unità di secondo stadio (i nati) nei comuni campione è stata effettuata seguendo il duplice criterio di realizzare l'auto-ponderazione delle unità finali negli strati e rispettare l'allocazione del campione a livello di macro-area di cittadinanza.

A tale scopo le numerosità campionarie complessive per comune sono state distribuite tra le macro-aree secondo un criterio proporzionale. Tuttavia le quantità così calcolate non riproducono, a livello di macro-area, le numerosità campionarie prefissate; sono stati pertanto effettuati degli aggiustamenti in più passi in modo da ottenere le numerosità campionarie ottimali di interviste a livello di macro-area e, in seconda fase, a livello di comune. Una volta definite le numerosità campionarie per comune e macro-area, la selezione delle unità finali è stata effettuata con probabilità uguali. Il campione realizzato è stato di 1.653 unità.

2.4 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 nel disegno per le madri italiane, h=1,...,H;
- N_h , numero dei nati dello strato h;
- n_h , numerosità campionaria nello strato h;
- g, indice dello strato dei comuni nel disegno per le madri straniere, g=1,…,G;
- M_g , numero di comuni dello strato g;
- m_g , numero di comuni campione dello strato g;
- N_g , numero dei nati dello strato g;
- N_{gj} , numero dei nati stranieri del comune j dello strato g;

- ${}_aN_{gj}$, numero dei nati stranieri del comune j dello strato g di macro-area a;
- ${}_an_{gj}$, numero di nati stranieri campione del comune j dello strato g di macro-area a;
- *y* , generica variabile oggetto di indagine;
- Y_i , 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

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

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.

2.4.1 Costruzione dei coefficienti di riporto all'universo

Il peso da attribuire alle unità campionarie è stato ottenuto per mezzo di una procedura complessa che i) corregge l'effetto distorsivo della mancata risposta totale dovuta all'impossibilità di intervistare alcune delle unità selezionate per irreperibilità o per rifiuto all'intervista; ii) tiene conto della conoscenza di 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, riferite alla madre, sono l'età, lo stato civile e la cittadinanza. 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 ordine di nascita;
- totale popolazione per ripartizione e titolo di studio della madre;

- totale popolazione per ripartizione e macro-area.

Indicando, quindi, con ${}_kX$ il k-mo totale noto e con ${}_kX_i$ il valore assunto dalla k-ma variabile ausiliaria per l'unità rispondente i, la condizione di uguaglianza tra il valore del totale noto e la stima campionaria del totale stesso è espressa dalla seguente relazione:

$${}_kX={}_k\hat{X}=\sum_{i=1}^{n}{}_kX_iW_i \quad (k=1,\ldots,K).$$

Le variabili *X* sono variabili dicotomiche, quindi se, ad esempio, ${}_kX$ indica il numero di nati da madri di età pari a 23 anni nella prima ripartizione geografica, la variabile ausiliaria ${}_kX_i$ assume il valore uno se l'unità i è un nato da madre di 23 anni e appartiene alla ripartizione 1 e valore zero altrimenti.

Poiché la popolazione complessiva è stata suddivisa in due sotto-popolazioni nelle quali sono stati definiti due differenti disegni di campionamento, nel calcolo dei pesi si è tenuto conto di questi due disegni campionari attraverso la definizione di pesi diretti (calcolati come l'inverso della probabilità di inclusione delle unità nel campione) che discendono dai disegni utilizzati.

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

1. si calcolano i pesi diretti come reciproco della probabilità di inclusione delle unità: per le unità selezionate con disegno a uno stadio stratificato, tale peso diretto è uguale per tutte le unità di uno stesso strato h ed è fornito dall'espressione:

$$d^*_{hi}=N_h/n^*_h,$$

essendo n^*_h la dimensione teorica del campione per lo strato h; invece per unità selezionate con il disegno a due stadi (le madri straniere), il peso diretto assume la seguente espressione:

$${}_ad^*_{gji}=\left(\frac{m_g}{M_g}\frac{{}_an^*_{gj}}{{}_aN_{gj}}\right)^{-1},$$

essendo ${}_an^*_{gj}$ il numero teorico di madri campione della macro-area a, per il comune j dello strato g;

2. si calcolano i fattori correttivi per mancata risposta totale, definiti, per il disegno sulle madri italiane, come l'inverso del tasso di risposta all'interno dello strato cui ciascuna unità appartiene:

$$c_{hi}=n^*_h/n_h,$$

mentre, per il campione sulle madri straniere, come inverso del tasso di risposta all'interno di celle individuate dall'incrocio di ripartizione geografica a tre modalità e quattro classi di età:

$$c_q=n^*_q/n_q$$

dove si è indicato con q la generica cella (q=1,…,12) di aggiustamento;

3. si ottengono i pesi base, o pesi corretti per mancata risposta totale, rispettivamente d_{hi} e ${}_ad_{gji}$, moltiplicando i pesi diretti per i corrispondenti fattori correttivi per mancata risposta totale;
4. si costruiscono i fattori correttivi $\square_i$ che consentono di soddisfare la condizione di

uguaglianza tra i totali noti delle variabili ausiliarie e le corrispondenti stime campionarie;

5. si calcolano, infine, i pesi finali mediante il prodotto dei pesi base per i fattori correttivi ottenuti al passo 4:

$W_i = d_{hi} \times \gamma_i$ e $W_i =_a d_{gji} \times \gamma_i$

I fattori correttivi del passo 4 sono ottenuti dalla 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 stime campionarie dei totali noti di popolazione e valori noti 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, eliminando in tal modo i pesi positivi estremi (troppo grandi o troppo piccoli).

Tutti i metodi di stima che scaturiscono dalla risoluzione di un problema di minimo vincolato del tipo sopra descritto rientrano in una classe generale di stimatori nota come stimatori di ponderazione vincolata[5]. Un importante stimatore appartenente a tale classe, che si ottiene utilizzando la funzione di distanza euclidea, è lo stimatore di regressione generalizzata. Come verrà chiarito meglio nel paragrafo 4, tale stimatore riveste un ruolo centrale perché è possibile dimostrare che tutti gli stimatori di ponderazione vincolata convergono asintoticamente, all'aumentare della numerosità campionaria, allo stimatore di regressione generalizzata.

2.5 Valutazione del livello di precisione delle stime

2.5.1 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 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

[5] Nella letteratura in lingua anglosassone sull'argomento tali stimatori sono noti come *calibration estimators*, si veda come riferimento Deville e Sarndäl (1992).

stima della varianza $\hat{V}ar(\hat{Y}_d)$ si è utilizzato il metodo proposto da Woodruff (1971); 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:

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

ossia la stima della varianza della stima $\hat{Y}_d$ viene calcolata come somma della stima delle varianze 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). È 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 $\overline{Z}'_h = \frac{1}{n_h}\sum_{i=1}^{n_h} Z'_{hi}$.

Per l'insieme degli strati relativi al disegno a due stadi, invece, la varianza viene stimata mediante la formula seguente

$$\sum_{h=1}^{G_d} \hat{V}ar\left({}_d\hat{Z}'_g\right) = \sum_{g=1}^{G_d} \frac{m_g}{m_g - 1} \sum_{i=1}^{m_g} \left({}_d\hat{Z}'_{gi} - \frac{{}_d\hat{Z}'_g}{m_g} \right)^2 \qquad (8)$$

dove le quantità sono espresse come

$${}_d\hat{Z}'_{gi} = \sum_{j=1}^{m_{gi}} Z_{gji} W_{gji} \quad \text{e} \quad {}_d\hat{Z}'_g = \sum_{i=1}^{m_g} \sum_{j=1}^{n_{gj}} Z_{gji} W_{gji} \quad .$$

Una volta calcolata la stima della varianza campionaria, utilizzando le espressioni (7) e (8), 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:

$$\left\{ \hat{Y}_d - k_p \hat{\sigma}(\hat{Y}_d) \le Y_d \le \hat{Y}_d + k_p \hat{\sigma}(\hat{Y}_d) \right\} \qquad (9)$$

Nella (9) il valore di k_P dipende dal valore fissato per la probabilità P; ad esempio, per P=0.95 si ha k=1.96.

2.5.2 Presentazione sintetica degli errori campionari

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

Ciò non è possibile, sia per limiti di tempo 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 di stime non pubblicate, che l'utente può ricavare in modo autonomo.

Per questi 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:

$$\log\hat{\varepsilon}^2(\hat{Y}_d) = a + b \log(\hat{Y}_d) \qquad (10)$$

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

Per calcolare gli errori di campionamento è stato utilizzato il software generalizzato Genesees (Pagliuca, 2002), messo a punto presso l'Istat, che consente di calcolare gli errori campionari e gli intervalli di confidenza e, inoltre, permette di costruire modelli regressivi del tipo (10) 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:

$$\hat{\varepsilon}(\hat{Y}_d^*) = \sqrt{\exp(a + b \log(\hat{Y}_d^*))} \qquad (11)$$

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 (11) 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 (11).

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.

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
		RIPARTIZIONE GEOGRAFICA	
Nord	5,15597	-1,21683	91,91
Centro	4,38914	-1,14279	88,39
Mezzogiorno	4,42749	-1,15000	86,99
Italia	**5,08709**	**-1,21434**	**90,59**
		REGIONI	
Piemonte	4,24297	-1,08202	92,71
Valle d'Aosta	2,53797	-1,11997	89,69
Lombardia	5,27958	-1,19575	97,21
Bolzano-Bozen	3,66258	-1,21334	93,66
Trento	3,03211	-1,16750	93,70
Veneto	3,60634	-1,01169	89,98
Friuli-Venezia Giulia	3,76765	-1,18245	92,07
Liguria	4,36747	-1,23464	96,09
Emilia Romagna	4,07574	-1,05933	93,93
Toscana	4,84090	-1,16722	93,99
Umbria	2,80933	-1,03624	81,00
Marche	3,80952	-1,09562	89,79
Lazio	3,87032	-1,05089	91,53
Abruzzo	3,95359	-1,21714	93,19
Molise	2,46612	-1,16409	87,81
Campania	6,15434	-1,27633	93,35
Puglia	4,01837	-1,06946	93,62
Basilicata	3,04247	-1,21016	93,55
Calabria	4,07816	-1,17481	92,69
Sicilia	4,58634	-1,12696	92,78
Sardegna	3,68506	-1,16283	93,69
		CLASSI DI ETÀ DELLA MADRE	
Fino a 24	4,12716	-1,17515	92,45
25 - 29	4,49923	-1,10918	91,64
30 - 34	4,61504	-1,13919	92,49
35 - 39	4,39974	-1,12523	91,72
40 e oltre	4,70877	-1,23941	93,62

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

STIMA	RIPARTIZIONE GEOGRAFICA			Italia
	Nord	Centro	Mezzogiorno	
250	45,8	38,3	38,2	44,5
500	30,0	25,8	25,7	29,2
750	23,5	20,4	20,3	22,9
1.000	19,7	17,3	17,2	19,2
1.250	17,2	15,3	15,2	16,8
1.500	15,4	13,7	13,7	15,0
1.750	14,0	12,6	12,5	13,7

2.000	12,9	11,7	11,6	12,6
2.500	11,3	10,3	10,2	11,0
5.000	7,4	6,9	6,8	7,2
10.000	4,9	4,7	4,6	4,7
20.000	3,2	3,1	3,1	3,1
30.000	2,5	2,5	2,4	2,4
40.000	2,1	2,1	2,1	2,0
50.000	1,8	1,9	1,8	1,8
60.000	1,6	1,7	1,6	1,6
70.000	1,5	1,5	1,5	1,5
80.000	1,4	1,4	1,4	1,3
90.000	1,3	1,3	1,3	1,2
100.00	1,2	1,2	1,2	1,2
150.00	0,9		1,0	0,9
200.00	0,8			0,8
250.00	0,7			0,7

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

TIMA S	Regioni										
	Piemonte	Valle d'Aosta	Lombardia	Bolzano-Bozen	Trento	Veneto	Friuli-Venezia Giulia	Liguria	Emilia Romagna	Toscana	Umbria
25											
50											
75											
1.000											
1.250											
1.500											
1.750											
2.000											
2.250											
2.500											
2.750											
3.000											
3.500											
4.000											
4.500											
5.000											
7.500											
10.000											
15.000											
20.000											
40.000											

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

TIMA ST	Regioni

	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
25										
50										
75										
1.000										
1.250										
1.500										
1.750										
2.000										
2.500										
3.000										
4.000										
5.000										
7.500										
10.000										
15.000										
20.000										
40.000										

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

STIMA	CLASSE DI ETÀ				
	Fino a 24	25 - 29	30 - 34	35 - 39	40 e oltre
250	30,7	44,4	43,3	40,4	34,4
500	20,4	30,2	29,2	27,3	22,4
750	16,1	24,1	23,1	21,8	17,4
1.000	13,6	20,6	19,6	18,5	14,6
1.250	11,9	18,2	17,3	16,3	12,7
1.500	10,7	16,4	15,6	14,7	11,3
1.750	9,8	15,1	14,3	13,5	10,3
2.000	9,0	14,0	13,2	12,5	9,5
2.500	7,9	12,4	11,7	11,1	8,3
5.000	5,3	8,4	7,9	7,5	5,4
10.000	3,5	5,7	5,3	5,1	3,5
20.000	2,3	3,9	3,6	3,4	2,3
30.000	1,8	3,1	2,8	2,7	1,8
40.000	1,6	2,7	2,4	2,3	
50.000	1,4	2,3	2,1	2,0	
60.000		2,1	1,9	1,8	
70.000		1,9	1,7	1,7	
80.000		1,8	1,6	1,6	
90.000		1,7	1,5	1,5	
100.00		1,6	1,4	1,4	
150.00		1,3	1,1		