File Standard

Indagine Multiscopo sulle Famiglie
Famiglia e soggetti sociali
Anno 2009

Manuale utente e tracciato record

Istat

Istituto Nazionale di Statistica

INDAGINE MULTISCOPO SULLE FAMIGLIE
FAMIGLIA E SOGGETTI SOCIALI - ANNO 2009
DOCUMENTAZIONE TECNICA E DESCRIZIONE DEL FILE

Premessa

Il Decreto Legislativo n. 322 del 6 settembre 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 Decreto Legislativo e della Legge n. 675 del 31 dicembre 1996 l'Istat ha adottato misure e tecniche che rendono impossibile, o altamente improbabile, il collegamento dei dati rilasciati con l'unità statistica cui si riferiscono. Per tale motivo sono state apportate alcune modifiche sui files originali delle indagini, nell'intento di garantire la massima protezione ai dati contenendo al minimo l'eventuale 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 variabili non attendibili dal punto di vista campionario e quindi non analizzabili statisticamente.

Finalità e caratteristiche dell'indagine

A partire dal dicembre 1993 l'Istat ha avviato il nuovo corso delle Indagini Multiscopo sulle Famiglie. Ogni anno, accanto all'indagine "Aspetti della vita quotidiana", si affiancano un'indagine a cadenza quinquennale, che approfondisce tematiche particolari, e un'indagine trimestrale su "Viaggi, vacanze e vita quotidiana".

Nel dicembre del 2009 è stata condotta, per la terza volta, l'indagine quinquennale "Famiglia e soggetti sociali" (la prima era stata realizzata nel giugno del 1998, la seconda nel novembre 2003). Tra i principali contenuti informativi si devono ricordare: le strutture familiari e i "pendolari" della famiglia; le reti di parentela, le reti di aiuto informale e gli aiuti ricevuti in occasione di eventi critici; l'affidamento dei bambini; la vita di coppia e le nozze; l'uscita dalla famiglia di origine e il ciclo di vita; la permanenza dei giovani adulti in famiglia; il percorso formativo, le carriere lavorative, la ricerca di lavoro, le eventuali interruzioni lavorative, la mobilità sociale e le intenzioni per il futuro.

Il campione è a due stadi con stratificazione delle unità di primo stadio (Comuni). L'indagine ha raggiunto approssimativamente 18.000 famiglie per un totale di circa 44.000 individui. Per una parte dei quesiti, le informazioni sono state raccolte per intervista diretta. Nei casi in cui l'individuo, per qualsiasi motivo, non sia stato disponibile all'intervista, le informazioni sono state fornite da un altro componente della famiglia. Per un'altra serie di quesiti è stata invece prevista l'autocompilazione diretta del questionario da parte del rispondente.

L'unità di rilevazione è costituita dalla famiglia di fatto (FF) associata alla famiglia anagrafica (FA) campionata. La famiglia di fatto è definita come quell'insieme di persone che:

1. hanno la loro dimora abituale nella stessa abitazione del capofamiglia anagrafico;
2. hanno con tale persona una relazione di matrimonio, parentela, affinità, adozione, tutela o affettiva. Si noti come per l'individuazione di una FF siano più importanti i concetti di "abitazione" e "dimora abituale", che non l'effettiva registrazione anagrafica degli individui conviventi.

All'interno di ciascuna FF possono essere individuati nessuno, uno o più nuclei familiari. La definizione di nucleo familiare è più restrittiva di quella di famiglia. Infatti, per nucleo familiare si intende:

1. coppia, coniugata o convivente, con o senza figli mai sposati, né conviventi coniugalmente, né aventi figli propri;
2. un solo genitore con uno o più figli mai sposati, né conviventi coniugalmente, né aventi figli propri.

I componenti la famiglia di fatto che non soddisfano i precedenti requisiti, sono considerati come "membri isolati".

Avvertenze per l'utilizzazione del file

Per gli utenti esterni all'ISTAT vengono messi a disposizione dei files con le seguenti caratteristiche:

Anno 2009

lunghezza record: 3.643
numero records individuali: 43.850
(uno per ciascuna persona intervistata)

Ogni record contiene una prima parte di informazioni sull'individuo, una seconda parte sulla famiglia di appartenenza e una terza parte contenente alcune variabili create (cioè non rilevate direttamente). A seconda della selezione che si opera sul file è possibile effettuare elaborazioni sulle seguenti unità di analisi:

a) *individui*

ogni componente è individuato dal numero progressivo della famiglia e dal suo numero d'ordine all'interno della stessa. Il numero totale di appartenenti al campione è pari al numero di records: 43.850. Per selezionare i componenti della stessa famiglia si considerano tutti i records individuali che hanno lo stesso numero generale progressivo della famiglia. Per selezionare i componenti appartenenti allo stesso nucleo si considerano tutti i records che hanno lo stesso numero progressivo della famiglia e lo stesso numero d'ordine del nucleo;

b) *famiglie*

volendo analizzare le famiglie occorre selezionare solo il primo componente di ciascuna utilizzando il numero d'ordine all'interno della famiglia. Il totale delle famiglie è pari a 17.788;

c) *nuclei familiari*

l'analisi dei nuclei familiari è possibile, invece, selezionando la persona di riferimento di ciascun nucleo. Il totale dei nuclei è di 12.636.

Costruzione delle stime ed errori di campionamento

Le informazioni riportate nei files sono di carattere campionario. Per ottenere stime relative all'intera popolazione oggetto d'indagine è necessario moltiplicare ciascuna informazione per il coefficiente di riporto all'universo.

Tali coefficienti sono stati determinati in modo da poter essere utilizzati indifferentemente per costruire sia stime relative alle persone sia stime riferite alle famiglie.

L'indagine ha la finalità di fornire stime riferite a:

1. l'intero territorio nazionale;
2. le cinque ripartizioni geografiche (Nord-ovest, Nord-est, Centro, Sud, Isole);
3. le regioni geografiche (ad eccezione del Trentino Alto Adige le cui stime sono prodotte distintamente per le province di Bolzano e Trento);
4. sei aree basate sulla tipologia socio-demografica dei comuni.

Per garantire la riservatezza e per limiti connessi alla numerosità campionaria non è possibile fornire contemporaneamente i dati su tutte e tre le suddivisioni territoriali suindicate. Sono disponibili, quindi, due files: uno contenente i codici di regione e ripartizione e l'altro con i codici di ripartizione ed area socio-demografica (dominio).

Nel diffondere i risultati di un'indagine campionaria occorre fornire agli utilizzatori le informazioni necessarie per valutare l'attendibilità delle stime ottenibili. Ad ogni stima corrisponde un errore campionario relativo; ciò significa che per consentire un uso corretto delle stime sarebbe necessario fornire per ogni stima il corrispondente errore campionario relativo. Questo, tuttavia, comporterebbe notevoli difficoltà per l'utilizzatore, dovute al fatto che la tutela della riservatezza impedisce di fornire i codici identificativi territoriali sui quali è basato il disegno dell'indagine. Per questo si ricorre ad una presentazione sintetica degli errori tramite il metodo dei modelli regressivi. Questo metodo si basa sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore relativo.

Si riportano in allegato le informazioni relative al campionamento e al calcolo degli errori di stima da cui è possibile individuare gli esempi di calcolo degli errori campionari. In seguito sono accluse le tavole per il calcolo degli errori relativi ai dati contenuti nei files standard, per stime sugli individui e sulle famiglie.

Appendice A

Glossario

I dati generali individuali fanno riferimento alle caratteristiche delle persone all'epoca dell'intervista. In particolare:

- **l'età** è espressa in anni compiuti;

- **il titolo di studio** è quello più elevato conseguito;

- **la condizione** è quella dichiarata come unica o prevalente dalle persone di 15 anni e più.

  Si precisa inoltre che:
  per **occupato** si intende chi possiede un'occupazione, in proprio o alle dipendenze, da cui trae un profitto o una retribuzione (utile, onorario, stipendio, salario) o chi collabora con un familiare che svolge un'attività lavorativa in conto proprio senza avere un regolare contratto di lavoro (coadiuvante);

  per **persona in cerca di occupazione** (disoccupato) si intende chi ha perduto una precedente occupazione o chi non ha mai esercitato un'attività lavorativa ed è alla ricerca attiva di un'occupazione che è in grado di accettare se gli viene offerta;

  casalinga è chi si dedica prevalentemente alle faccende domestiche;

  studente è chi si dedica prevalentemente allo studio;

  ritirato dal lavoro è chi ha cessato un'attività lavorativa per raggiunti limiti di età, invalidità o altra causa; la figura del ritirato dal lavoro non coincide necessariamente con quella del pensionato in quanto, non sempre, il ritirato dal lavoro gode di una pensione;

  in altra condizione è colui che si trova in condizione diversa da quelle sopra elencate (militare, inabile al lavoro, benestante, detenuto, ecc.);

- **la posizione nella professione** è quella dichiarata come unica o prevalente dagli occupati di 15 anni e più che viene aggregata nel modo seguente:

  dirigenti, imprenditori, liberi professionisti;

  direttivi, quadri, impiegati, intermedi (appartenenti alle categorie speciali);

  capo operai, operai, subalterni (inclusi apprendisti, lavoratori a domicilio per conto di imprese);

  lavoratori in proprio, coadiuvanti (inclusi soci di cooperative di produzione di beni e/o prestazioni di servizio, collaboratori coordinati e continuativi e prestatori d'opera occasionali);

- **le ripartizioni geografiche** costituiscono una suddivisione geografica del territorio e sono così articolate:

  Nord-ovest comprende: Piemonte, Valle d'Aosta, Lombardia, Liguria

  Nord-est comprende: Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna

  Centro comprende: Toscana, Umbria, Marche, Lazio

  Sud comprende: Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria

  Isole comprende: Sicilia, Sardegna

- **il tipo di comune**

  I Comuni italiani sono suddivisi nelle seguenti classi:

  comuni centro delle aree metropolitane: si tratta dei Comuni di Torino, Milano, Venezia, Genova, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania, Cagliari;

  comuni appartenenti alla periferia delle aree metropolitane: costituiscono i comuni delle cinture urbane;

  altri comuni: suddivisi per dimensione demografica (fino a 2.000 abitanti, da 2.001 a 10.000, da 10.001 a 50.000 e oltre i 50.000);

  Si precisa che la soglia dei 2.000 abitanti costituisce la dimensione demografica suggerita dagli organismi internazionali per identificare uno stile di vita tipico dei piccoli centri.

- **famiglia e nucleo familiare**

  la **famiglia** è costituita dall'insieme delle persone coabitanti legate da vincoli di matrimonio o parentela, affinità, adozione, tutela o affettivi;

  il **nucleo** è l'insieme delle persone che formano una coppia con figli celibi o nubili, una coppia senza figli, un genitore solo con figli celibi o nubili.

  Una famiglia può coincidere con un nucleo, può essere formata da un nucleo più altri membri aggregati, da più nuclei (con o senza membri aggregati), o da nessun nucleo (persone sole, famiglie composte ad esempio da due sorelle, da un genitore con figlio separato, divorziato o vedovo, ecc.);

- **nidi vuoti**

  nuclei in cui i figli non sono più presenti perché hanno ormai lasciato la famiglia di origine;

- **pendolari della famiglia**

  si definiscono tali gli individui che hanno risposto affermativamente al seguente quesito: *"Nell'ultimo anno, le è capitato di vivere in un'abitazione diversa da questa con una certa regolarità, per esempio: due giorni a settimana, oppure tutta la settimana tranne il week-end, oppure tutto il periodo delle lezioni a scuola o all'Università? (escludendo quindi i giorni di vacanza e i viaggi di lavoro occasionali)"*.

Appendice B

Strategia di campionamento e livello di precisione dei risultati

Obiettivi conoscitivi

La *popolazione di interesse* dell'indagine in oggetto, ossia l'insieme delle unità statistiche intorno alle quali si intende investigare, è costituita dalle famiglie residenti in Italia e dagli individui ad esse appartenenti, al netto dei membri permanenti delle convivenze. La famiglia è intesa come *famiglia di fatto*, ossia un insieme di persone coabitanti e legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o affettivi.

Il *periodo di riferimento* è prevalentemente costituito dai dodici mesi che precedono l'intervista, anche se per alcuni quesiti il riferimento è il momento dell'intervista.

I *domini di studio*, ossia gli ambiti rispetto ai quali sono riferiti i parametri di popolazione oggetto di stima, sono:

- l'intero territorio nazionale;
- le cinque ripartizioni geografiche (Italia Nord-Occidentale, Italia Nord-Orientale, Italia Centrale, Italia Meridionale, Italia Insulare);
- le regioni geografiche (ad eccezione del Trentino Alto Adige le cui stime sono prodotte separatamente per le province di Bolzano e Trento);
- la tipologia comunale ottenuta suddividendo i comuni italiani in sei classi formate in base a caratteristiche socio-economiche e demografiche:
  - A) *comuni appartenenti all'area metropolitana* suddivisi in:
    - A_1, *comuni centro dell'area metropolitana*: Torino, Milano, Venezia, Genova, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania, Cagliari;
    - A_2, *comuni che gravitano intorno ai comuni centro dell'area metropolitana*;
  - B) *comuni non appartenenti all'area metropolitana* suddivisi in:
    - B_1 comuni aventi fino a 2.000 abitanti;
    - B_2 comuni con 2.001-10.000 abitanti;
    - B_3 comuni con 10.001-50.000 abitanti;
    - B_4 comuni con oltre 50.000 abitanti.

Strategia di campionamento

Descrizione generale del disegno di campionamento

Il disegno di campionamento è di tipo complesso e si avvale di due differenti schemi di campionamento. Nell'ambito di ognuno dei domini definiti dall'incrocio della regione geografica con le sei aree A_1, A_2, B_1, B_2, B_3 e B_4, i comuni italiani sono suddivisi in due sottoinsiemi sulla base della popolazione residente:

- l'insieme dei comuni Auto Rappresentativi (che indicheremo d'ora in avanti come comuni Ar) costituito dai comuni di maggiore dimensione demografica;
- l'insieme dei comuni Non Auto Rappresentativi (o Nar) costituito dai rimanenti comuni.

Nell'ambito dell'insieme dei comuni Ar, ciascun comune viene considerato come uno strato a se stante e viene adottato un disegno noto con il nome di *campionamento a grappoli*. Le unità primarie di campionamento sono rappresentate dalle famiglie anagrafiche, estratte in modo sistematico dall'anagrafe del comune stesso; per ogni famiglia anagrafica inclusa nel campione vengono rilevate le caratteristiche oggetto di indagine di tutti i componenti di fatto appartenenti alla famiglia medesima.

Nell'ambito dei comuni Nar viene adottato un disegno a due stadi con stratificazione delle unità primarie. Le Unità Primarie (UP) sono i comuni, le Unità Secondarie sono le famiglie anagrafiche; per ogni famiglia anagrafica inclusa nel campione vengono rilevate le caratteristiche oggetto di indagine di tutti i componenti di fatto appartenenti alla famiglia medesima.

I comuni vengono selezionati con probabilità proporzionali alla loro dimensione demografica e senza reimmissione, mentre le famiglie vengono estratte con probabilità uguali e senza reimmissione.

Definizione della dimensione campionaria

Per un'indagine ad obiettivi plurimi, come quella in esame, è poco realistico pensare di poter disegnare una strategia campionaria che assicuri prefissati livelli di precisione di tutte le stime prodotte. La questione è complicata dal fatto che l'indagine ha la finalità di determinare stime per livelli territoriali differenti, il che comporta l'adozione di soluzioni di tipo ottimale diverse e contrastanti. Ad esempio, se l'unico ambito territoriale di pubblicazione delle stime fosse quello nazionale, una soluzione approssimativamente ottimale sarebbe quella di determinare la numerosità nazionale e ripartirla tra le regioni in modo proporzionale alla loro dimensione demografica; viceversa, avendo la finalità di produrre stime con uguale attendibilità a livello regionale, una soluzione approssimativamente ottimale sarebbe quella di selezionare un campione uguale in tutte le regioni. Quest'ultima soluzione, però, è poco efficiente per le stime a livello nazionale. Per affrontare questo problema, conformemente a quanto fatto in altri paesi, si è fatto ricorso ad una strategia che perviene alla definizione della numerosità campionaria attraverso approssimazioni successive.
In base alle considerazioni precedenti si è deciso di adottare un'ottica mista basata sia su criteri di costo ed organizzativi, sia su una valutazione degli errori campionari delle principali stime a livello nazionale e con riferimento a ciascuno dei domini territoriali di interesse.
I criteri seguiti possono essere sintetizzati nei seguenti punti:

- la dimensione del campione teorico in termini di famiglie, prefissata a livello nazionale essenzialmente in base a criteri di costo ed operativi, è pari a circa 24.000;
- il numero di comuni campione interessati non deve essere superiore a 900 in modo da consentire un buon lavoro di controllo e supervisione.

L'allocazione del campione di famiglie e di comuni tra le varie regioni è stata poi definita adottando un criterio di compromesso tale da garantire sia l'affidabilità delle stime a livello nazionale che quella delle stime a livello di ciascuno dei domini territoriali descritti nel primo paragrafo.

Stratificazione e selezione delle unità campionarie

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.
Nell'indagine Multiscopo, i comuni vengono stratificati in base alla loro dimensione demografica e nel rispetto delle seguenti condizioni:

- autoponderazione del campione a livello regionale;
- selezione di un comune campione nell'ambito di ciascuno strato definito sui comuni dell'insieme Nar;
- scelta di un numero minimo di famiglie da intervistare in ciascun comune campione; per l'indagine in oggetto tale numero è stato posto pari a 23;
- formazione di strati aventi ampiezza approssimativamente costante in termini di popolazione residente.

Il procedimento di stratificazione, attuato all'interno di ogni dominio territoriale individuato dalle aree A_1, A_2, B_1, B_2, B_3 e B_4 di ciascuna regione geografica, si articola nelle seguenti fasi:

- ordinamento dei comuni del dominio in ordine decrescente secondo la loro dimensione demografica in termini di popolazione residente;
- determinazione di una soglia di popolazione per la definizione dei comuni Ar, mediante la relazione:

$$_r\lambda = \frac{_r\overline{m}\ _r\delta}{_rf}$$

in cui per la generica regione geografica r si è indicato con: $_r\overline{m}$ il numero minimo di famiglie da intervistare in ciascun comune campione; $_r\delta$ il numero medio di componenti per famiglia; $_rf$ la frazione di campionamento;

- suddivisione di tutti i comuni nei due sottoinsiemi Ar e Nar: i comuni di dimensione superiore o uguale a $_r\lambda$ sono definiti come comuni Ar e i rimanenti come Nar;
- suddivisione dei comuni dell'insieme Nar in strati aventi dimensione, in termini di popolazione residente, approssimativamente costante e all'incirca pari alla soglia $_r\lambda$.

Effettuata la stratificazione, i comuni Ar sono inclusi con certezza nel campione; per quanto riguarda, invece, i comuni Nar, nell'ambito di ogni strato viene estratto un comune campione con probabilità proporzionale alla dimensione demografica, mediante la procedura di selezione sistematica proposta da Madow[1].

La selezione delle famiglie da intervistare in ogni comune campione viene effettuata dalla lista anagrafica di ciascun comune senza reimmissione e con probabilità uguali.
In particolare, la tecnica di selezione è di tipo sistematico e, nell'ambito di ogni comune viene attuata attraverso le seguenti fasi:

- vengono messi in sequenza i fogli delle famiglie dell'anagrafe del comune;

[1] Madow, W.G. (1949) "On the theory of systematic sampling II", Ann. Math. Stat., 20, 333-354

- si calcola il passo di campionamento e_{hi} , come rapporto tra il numero delle famiglie residenti nel comune i dello strato h e il corrispondente numero di famiglie campione, $e_{hi}=M_{hi}/m_{hi}$;
- si selezionano le m_{hi} famiglie che nella sequenza costruita al punto 1) occupano le seguenti posizioni :

$$1,\ 1+e_{hi},\ 1+2e_{hi},\ .....,\ 1+(m_{hi}-1)e_{hi}.$$

Nel prospetto 1 viene riportata la distribuzione regionale dell'universo e del campione dei comuni, delle famiglie e degli individui.

Prospetto 1 - Distribuzione regionale dei comuni, delle famiglie e delle persone nell'universo e nel campione- Anno 2009

REGIONI	Comuni		Famiglie		Individui	
	Universo	Campione	Universo (a)	Campione	Universo	Campione
Piemonte	1,206	61	1,972,573	1,261	4,404,242	2,857
Valle d'Aosta	74	21	57453	449	126793	945
Lombardia	1,546	82	4,073,270	1,561	9,757,743	3,768
Bolzano	*116*	*23*	*201884*	*535*	*498421*	*1284*
Trento	*223*	*25*	*219138*	*485*	*519008*	*1206*
Veneto	581	53	1,962,578	1,047	4,865,647	2,715
Friuli-Venezia Giulia	219	32	539,909	643	1,219,788	1,518
Liguria	235	25	753,666	756	1,602,413	1,498
Emilia-Romagna	341	46	1,887,449	1,002	4,344,813	2,314
Toscana	287	49	1,557,477	1,007	3,704,550	2,441
Umbria	92	22	361,258	546	894,539	1,410
Marche	246	37	626,194	711	1,568,343	1,803
Lazio	378	31	2,386,974	874	5,636,730	2,058
Abruzzo	305	33	533,998	673	1,332,230	1,620
Molise	136	23	126,837	567	318,744	1,393
Campania	551	51	2,068,788	1,226	5,805,510	3,446
Puglia	258	48	1,526,956	1,035	4,068,577	2,832
Basilicata	131	27	232,906	608	586,530	1,630
Calabria	409	42	766,598	867	1,999,714	2,181
Sicilia	390	52	1,950,650	1,167	5,020,548	2,990
Sardegna	377	38	658,278	768	1,663,722	1,941
Italia	**8,101**	**821**	**24,464,834**	**17,788**	**59,938,605**	**43,850**

(a) Stima Indagine Multiscopo

Procedimento per il calcolo delle stime

Le stime prodotte dall'indagine sono essenzialmente stime di frequenze assolute e relative, riferite alle famiglie e agli individui. 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 di livello territoriale di riferimento delle stime; i, indice di comune; j, indice di famiglia; p, indice di componente della famiglia; h, indice di strato di comuni; y, generica variabile oggetto di indagine; Y_{hijp}, valore di y osservato sul componente p della famiglia j del comune i dello strato h; P_{hij}, numero di componenti della famiglia j del comune i dello strato h; $Y_{hij}=\sum_{p=1}^{P_{hij}} Y_{hijp}$, totale della variabile y osservato sulla famiglia j del comune i dello strato h; M_{hi}, numero di famiglie residenti nel comune i dello strato h; m_{hi}, campione di famiglie nel comune i dello strato h; N_h, totale di comuni nello strato h; n_h, numero di comuni campione nello strato h (nell'indagine in oggetto si ha $n_h=1$); H_d, numero totale di strati nel generico dominio territoriale d.

Ipotizziamo di voler stimare, con riferimento ad un generico dominio d, il totale della generica variabile y oggetto di indagine, espresso dalla seguente relazione

$$Y_d = \sum_{h=1}^{H_d} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij} \,. \qquad (1)$$

La stima del totale (1) è data da

$$\hat{Y}_d = \sum_{h=1}^{H_d} \hat{Y}_h \;, \quad \text{essendo} \quad \hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij} Y_{hij} \,, \qquad (2)$$

in cui W_{hij} è il peso finale da attribuire a tutti i componenti della famiglia j del comune i dello strato h.

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[2] ed effettuare, a livello del dominio di interesse, la somma dei prodotti così ottenuti.

Il peso da attribuire alle unità campionarie è ottenuto per mezzo di una procedura complessa che:

- corregge l'effetto distorsivo della mancata risposta totale dovuta all'impossibilità di intervistare alcune delle famiglie selezionate per irreperibilità o per rifiuto all'intervista;
- tiene conto della conoscenza di totali noti di importanti variabili ausiliarie (disponibili da fonti esterne all'indagine), nel senso che le stime campionarie dei totali noti delle variabili ausiliarie devono coincidere con i valori noti degli stessi.

Nell'indagine in oggetto vengono definiti per ciascuna regione geografica 18 totali noti, che si riferiscono alla distribuzione della popolazione regionale per sesso e sei classi di età[3] e della popolazione regionale nelle sei aree A_1, A_2, B_1, B_2, B_3 e B_4. Indicando, quindi, con ${}_kX$ (k=1,…,18) il totale noto della k-esima variabile ausiliaria per la generica regione geografica e con ${}_kX_{hij}$ il valore assunto dalla k-esima variabile ausiliaria per la famiglia rispondente hij, la condizione sopra descritta è espressa dalla seguente uguaglianza

$${}_k X = {}_k \hat{X} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij\,k} X_{hij} \quad (k=1,\ldots., 18)$$

in cui H indica il numero complessivo di strati definiti nella regione. Se, ad esempio, ${}_6X$ indica il numero di maschi di età maggiore o uguale a sessantacinque cinque anni, la variabile ausiliaria ${}_6X_{hij}$ rappresenta il numero di maschi di età maggiore o uguale a sessantacinque anni della famiglia (hij).

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

1) si calcolano i *pesi diretti* come reciproco della probabilità di inclusione delle unità;
2) si calcolano i fattori correttivi per mancata risposta totale, come l'inverso del tasso di risposta del comune cui ciascuna unità appartiene;
3) si ottengono i *pesi base*, o pesi corretti per mancata risposta totale, moltiplicando i pesi diretti per i corrispondenti fattori correttivi per mancata risposta totale;
4) si costruiscono i fattori correttivi che consentono di soddisfare, a livello regionale, 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.

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[4]. 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 successivo, tale stimatore riveste un ruolo centrale in quanto è possibile dimostrare[5] che tutti gli stimatori di ponderazione vincolata convergono asintoticamente, all'aumentare della numerosità campionaria, allo stimatore di regressione generalizzata.

---

[2] Al fine di ottenere stime coerenti per individui e famiglie i pesi finali sono definiti in modo tale che a ciascuna famiglia (hij) e a tutti i componenti della stessa sia assegnato un medesimo peso finale W_{hij} .

[3] Le classi di età considerate sono: 0-5, 6-13, 14-24, 25-44, 45-64, più di 65 anni.

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

[5] Deville J.C., Sarndal C.E. (1992) "Calibration Estimators in Survey Sampling", Journal of the American Statistical Association, vol. 87, pp. 376-382.

Valutazione del livello di precisione delle stime

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 seguente espressione

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

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

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

Come è stato descritto nel paragrafo precedente, 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; 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

$$\hat{Y}_d \cong \hat{Z}_d = \sum_{h=1}^{H_d} \hat{Z}_h \ , \quad \text{essendo} \quad \hat{Z}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} Z_{hij} W_{hij} \tag{5}$$

dove Z_{hij} è la variabile linearizzata espressa come $Z_{hij} = Y_{hij} - \mathbf{X}'_{\mathbf{hij}}\beta$, essendo $\mathbf{X}_{\mathbf{hij}} = \left({}_1X_{hij},\ldots,{}_kX_{hij},\ldots{}_KX_{hij}\right)'$ il vettore contenente i valori delle K (K=18) variabili ausiliarie, osservati per la generica famiglia hij 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), si ha, quindi, che la stima della varianza della stima $\hat{Y}_d$ è ottenuta mediante la seguente relazione

$$\hat{V}ar\left(\hat{Y}_d\right) \cong \hat{V}ar\left(\hat{Z}_d\right) = \sum_{h=1}^{H_d} \hat{V}ar\left(\hat{Z}_h\right). \tag{6}$$

Dalla (6) risulta che la stima della varianza della stima $\hat{Y}_d$ viene calcolata come somma della stima delle varianze dei singoli strati, Ar e Nar, appartenenti al dominio d. La formula di calcolo della varianza, $\hat{V}ar\left(\hat{Z}_h\right)$, della stima $\hat{Z}_h$ è differente a seconda che lo strato sia Ar oppure Nar. Possiamo, quindi scomporre come segue

$$\hat{V}ar\left(\hat{Y}_d\right) \cong \hat{V}ar\left(\hat{Z}_d\right) = \sum_{h=1}^{H_{AR}} \hat{V}ar\left(\hat{Z}_h\right) \ + \sum_{h=1}^{H_{NAR}} \hat{V}ar\left(\hat{Z}_h\right) \ , \tag{7}$$

in cui H_{AR} e H_{NAR} indicano rispettivamente il numero di strati Ar e Nar appartenenti al dominio d.

Negli strati Ar (in cui ciascun comune fa strato a sé e $N_h = n_h = 1$, l'indice i di comune diviene superfluo e viene omesso) la varianza è stimata mediante la seguente espressione

$$\sum_{h=1}^{H_{AR}} \hat{V}ar\left(\hat{Z}_h\right) = \sum_{h=1}^{H_{AR}} M_h^2 \frac{(M_h - m_h)}{m_h(m_h - 1)} \sum_{j=1}^{m_h} \left(Z_{hj} - \overline{Z}_h\right)^2 , \tag{8}$$

dove si è posto $M_h = M_{hi}$, $m_h = m_{hi}$, $Z_{hj} = Z_{hij}$ e $\overline{Z}_h = \frac{1}{m_h} \sum_{j=1}^{m_h} Z_{hj}$.

Negli strati Nar, in cui viene estratto un solo comune campione da ogni strato, per stimare la varianza di campionamento si ricorre alla *tecnica di collassamento degli strati*. Questa tecnica consiste nel formare G gruppi contenenti ciascuno L_g $(L_g \geq 2)$ strati; la varianza viene stimata mediante la formula seguente

$$\sum_{h=1}^{H_{NAR}} \hat{V}ar(\hat{Z}_h) = \sum_{g=1}^{G} \hat{V}ar(\hat{Z}_g) = \sum_{g=1}^{G} \frac{L_g}{L_g - 1} \sum_{h=1}^{L_g} \left(\hat{Z}_{hg} - \frac{\hat{Z}_g}{L_g} \right)^2 \qquad (9)$$

dove le quantità sono espresse come

$$\hat{Z}_{hg} = \sum_{j=1}^{m_{hi}} Z_{hij} W_{hij} \quad \text{e} \quad \hat{Z}_g = \sum_{h=1}^{L_g} \sum_{j=1}^{m_{hi}} Z_{hij} W_{hij} .$$

Utilizzando le espressioni (8) e (9) è possibile, infine, calcolare la varianza di campionamento, $\hat{V}ar(\hat{Y}_d)$ in base alla (7) e calcolare, quindi, in base alla (3) ed alla (4) rispettivamente l'errore di campionamento assoluto e l'errore di campionamento relativo.
Gli errori campionari espressi dalla (3) e dalla (4) consentono di valutare il grado di precisione delle stime; inoltre, l'errore assoluto permette di costruire un intervallo di confidenza, che, con livello di fiducia P contiene il parametro oggetto di stima, l'intervallo viene espresso come

$$\left\{ \hat{Y}_d - k_p \hat{\sigma}(\hat{Y}_d) \leq Y_d \leq \hat{Y}_d + k_p \hat{\sigma}(\hat{Y}_d) \right\} \qquad (10)$$

Nella (10) il valore di k_P dipende dal valore fissato per la probabilità P; ad esempio, per P=0.95 si ha k=1.96.

Fondamenti statistici della procedura per il calcolo degli errori campionari

Per il calcolo degli errori di campionamento delle indagini condotte dall'Istat sulle famiglie e sulle imprese viene correntemente utilizzata una procedura informatica sviluppata nell'ambito dell'Istituto. Nel paragrafo precedente è stata descritta la metodologia, implementata dalla procedura, per il calcolo degli errori di campionamento delle stime prodotte dall'indagine mentre, nel presente paragrafo, vengono discussi i fondamenti statistici e i limiti della metodologia medesima.

Negli strati Ar, nei quali si adotta un disegno di campionamento a grappoli e in cui le unità primarie (le famiglie) vengono selezionate senza reimmissione e probabilità uguali, la procedura consente di ottenere stime della varianza campionaria che risultano corrette.

Negli strati Nar, per i quali si adotta un disegno di campionamento a due stadi con selezione delle unità primarie (comuni) senza reimmissione e probabilità variabili, la procedura consente di ottenere stime corrette della varianza campionaria qualora:

- in ciascuno strato sono selezionate due o più unità primarie;
- le unità primarie sono scelte mediante estrazioni indipendenti.

La prima condizione non viene soddisfatta in quanto, nell'indagine in oggetto, da ciascuno strato viene selezionato un solo comune campione e per stimare la varianza di campionamento si ricorre alla tecnica di *collassamento degli strati*. Questa tecnica, che consiste nel formare superstrati contenenti ciascuno un numero di strati maggiore di uno, conduce in generale ad una sovrastima della varianza di campionamento effettiva.

La seconda ipotesi implica che la selezione delle unità primarie venga effettuata con reimmissione. Anche questa assunzione non è soddisfatta per i comuni Nar e ciò comporta una sovrastima della varianza. Si osservi, tuttavia, che tale sovrastima dipende dalla frazione di campionamento di ciascuno strato Nar: è di entità trascurabile negli strati nei quali la frazione di campionamento è piccola, mentre viceversa può risultare di entità più cospicua per quegli strati in cui la frazione di campionamento è maggiore.

Presentazione sintetica degli errori campionari

Ad ogni stima $\hat{Y}_d$ corrisponde un errore di campionamento relativo $\hat{\varepsilon}(\hat{Y}_d)$; ciò significa che per consentire una lettura corretta delle tabelle pubblicate sarebbe necessario presentare per ogni stima pubblicata il corrispondente errore di campionamento relativo. Ciò, tuttavia, non è possibile sia per limiti di tempo e di costi di elaborazione, sia perché le tavole della pubblicazione risulterebbero appesantite e di non facile consultazione per l'utente finale. Inoltre, non sarebbero comunque disponibili gli errori delle stime non pubblicate, che l'utente può ricavare in modo autonomo.

Per le ragioni sopra esposte, si ricorre frequentemente ad una presentazione sintetica degli errori relativi, basata sul *metodo dei modelli regressivi*. Questo metodo si basa sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore relativo.

Nella presente indagine, il modello utilizzato per le stime di frequenze assolute e relative, è del tipo seguente:

$$\log\left(\hat{\varepsilon}^2(\hat{Y}_d)\right) = a + b \log(\hat{Y}_d) \qquad (11)$$

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

Nel prospetto 2 sono riportati i valori dei coefficienti a e b e dell'indice di determinazione R^2 del modello utilizzato per l'interpolazione degli errori campionari di stime di frequenze assolute e relative, per totale Italia, ripartizione geografica, tipologia comunale e regione.

Sulla base delle informazioni contenute in tale prospetto, è possibile calcolare la stima dell'errore di campionamento relativo di una determinata stima di frequenza assoluta $\hat{Y}_d$ mediante la formula:

$$\hat{\varepsilon}(\hat{Y}_d) = \sqrt{\exp\left(a + b\log(\hat{Y}_d)\right)} \qquad (12)$$

che si ricava facilmente dalla (11).

Se, per esempio, la stima $\hat{Y}_d$ si riferisce alle persone dell'Italia Nord Occidentale, l'errore relativo corrispondente si ottiene introducendo nella (12) i valori dei parametri a e b riportati nella seconda riga del prospetto 2 alla voce PERSONE (a = 9,352384, b = -1,138993).

I prospetti 3 e 4, presentati in aggiunta, consentono di rendere più agevole il calcolo degli errori campionari. Essi riguardano, rispettivamente, le famiglie e le persone ed hanno la seguente struttura: a) in fiancata sono elencati i valori crescenti di stima (20.000, 30.000, …, 25.000.000); b) le colonne successive contengono gli errori di campionamento relativo, per ciascun dominio territoriale di interesse, calcolati mediante la formula (12), corrispondenti alle stime di frequenze assolute della prima colonna.

Le informazioni contenute in tali prospetti permettono di calcolare l'errore relativo di una generica stima di frequenza assoluta (o relativa) mediante due procedimenti che risultano di facile applicazione, anche se conducono a risultati meno precisi di quelli ottenibili mediante l'espressione (12). Il primo metodo consiste nell'individuare, nella prima colonna del prospetto, il livello di stima che più si avvicina alla stima di interesse e nel considerare come errore relativo il valore che si trova sulla stessa riga, nella colonna corrispondente al domino territoriale di riferimento.

Con il secondo metodo, l'errore campionario della stima $\hat{Y}_d$ si ricava mediante la seguente espressione:

$$\hat{\varepsilon}(\hat{Y}_d) = \hat{\varepsilon}(\hat{Y}_d^{k-1}) - \frac{\hat{\varepsilon}(\hat{Y}_d^{k-1}) - \hat{\varepsilon}(\hat{Y}_d^{k})}{\hat{Y}_d^{k} - \hat{Y}_d^{k-1}}(\hat{Y}_d - \hat{Y}_d^{k-1}) \qquad (13)$$

dove $\hat{Y}_d^{k-1}$ e $\hat{Y}_d^{k}$ sono i valori delle stime, riportati nella prima colonna, entro i quali è compresa la stima di interesse $\hat{Y}_d$, ed $\hat{\varepsilon}(\hat{Y}_d^{k-1})$ e $\hat{\varepsilon}(\hat{Y}_d^{k})$ i corrispondenti errori relativi.

Prospetto 2 - Valori dei coefficienti a, b e dell'indice di determinazione R^2 (%) delle funzioni utilizzate per le interpolazioni degli errori campionari delle stime riferite alle FAMIGLIE e alle PERSONE per totale Italia, ripartizione geografica, tipo di comune e regione

ZONE TERRITORIALI	Famiglie			Persone		
	a	b	R^2 (%)	a	b	R^2 (%)
Italia	**9.409733**	**-1.149873**	**97.5**	**10.313747**	**-1.208937**	**94.2**
RIPARTIZIONI GEOGRAFICHE (a)						
Nord-ovest	8.996030	-1.119907	97.4	10.286523	-1.219604	94.2
Nord-est	8.780904	-1.129590	97.7	9.700607	-1.201380	94.2
Centro	9.883370	-1.199286	97.3	10.298458	-1.221226	93.1
Sud	8.084236	-1.078241	95.5	8.974649	-1.146204	92.9
Isole	8.160411	-1.088379	95.2	8.909809	-1.142909	91.9
TIPI DI COMUNE (b)						
A1	9.638377	-1.172019	98.1	10.306184	-1.220965	94.3
A2	8.607818	-1.096894	95.6	9.783769	-1.187261	91.6
B1	6.949447	-0.996816	92.1	7.898661	-1.081174	89.5
B2	8.761392	-1.122568	95.8	9.434101	-1.174362	93.5
B3	8.619179	-1.106753	96.5	9.398508	-1.163792	93.0
B4	9.305768	-1.179879	97.4	10.017366	-1.224499	94.5
REGIONI						
Piemonte	8.548783	-1.123562	93.2	9.245709	-1.184000	94.3
Valle d'Aosta	6.026323	-1.191009	95.6	6.170943	-1.191794	94.7
Lombardia	9.194684	-1.122379	97.2	10.381688	-1.220204	93.9
- *Bolzano*	6.742691	-1.126116	94.7	7.432121	-1.196514	89.9
-Trento	7.263187	-1.147975	94.5	8.033554	-1.234355	92.7
Veneto	8.675884	-1.114872	96.5	9.403876	-1.179651	92.9
Friuli-Venezia Giulia	8.466274	-1.192996	97.4	8.344447	-1.181036	94.8
Liguria	8.083133	-1.131281	96.7	8.471890	-1.164710	93.4
Emilia-Romagna	9.140306	-1.165767	97.7	9.658934	-1.206622	94.2
Toscana	8.903462	-1.166468	97.1	9.177685	-1.178637	93.3
Umbria	7.933146	-1.178620	96.9	8.856490	-1.262574	94.2
Marche	8.483376	-1.191736	95.8	8.371847	-1.169519	92.1
Lazio	9.773085	-1.172988	96.7	10.318815	-1.207507	92.1
Abruzzo	8.157097	-1.164734	96.5	8.405947	-1.176134	93.0
Molise	6.661473	-1.184216	95.6	6.911919	-1.183496	90.5
Campania	8.311312	-1.079331	95.5	8.654089	-1.110687	92.1
Puglia	8.509217	-1.119101	94.1	9.058575	-1.170980	92.4
Basilicata	7.016818	-1.141475	94.8	7.239578	-1.160087	91.4
Calabria	8.369700	-1.169428	94.7	8.941273	-1.195706	92.0
Sicilia	8.473059	-1.107176	94.9	8.981321	-1.141191	91.6
Sardegna	6.988637	-1.041384	86.4	8.101993	-1.146037	92.9

(a) Italia nord-occidentale: Piemonte, Valle d'Aosta, Lombardia, Liguria; Italia nord-orientale: Bolzano, Trento, Veneto, Friuli-Venezia Giulia, Emilia Romagna; Italia centrale: Toscana, Umbria, Marche, Lazio; Italia meridionale: Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria; Italia insulare: Sicilia, Sardegna.

(b) Comuni tipo A1: Area urbana centro; Tipo A2: Area urbana periferia; Tipo B1: comuni fino a 2.000 abitanti; Tipo B2: da 2.001 a 10.000 abitanti; Tipo B3: da 10.001 a 50.000 abitanti; Tipo B4: oltre 50.000 abitanti.

Prospetto 3 - Valori interpolati degli errori campionari relativi percentuali delle stime riferite alle FAMIGLIE per totale Italia, ripartizione geografica, tipo di comune e regione

STIME	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	A1	A2	B1	B2	B3	B4
20.000	37.2	35.1	30.0	36.9	27.3	27.0	37.4	32.4	23.2	30.8	31.0	30.4
30.000	29.5	28.0	23.9	28.9	22.0	21.7	29.5	25.9	19.0	24.5	24.8	24.0
40.000	25.0	23.8	20.3	24.4	18.8	18.5	24.9	22.1	16.4	20.9	21.1	20.2
50.000	22.0	21.0	17.9	21.3	16.7	16.4	21.8	19.6	14.7	18.4	18.7	17.7
60.000	19.8	19.0	16.1	19.1	15.1	14.9	19.6	17.7	13.4	16.6	16.9	15.9
70.000	18.1	17.4	14.8	17.4	13.9	13.7	17.9	16.3	12.4	15.2	15.5	14.5
80.000	16.8	16.1	13.7	16.1	12.9	12.7	16.6	15.1	11.6	14.1	14.4	13.4
90.000	15.7	15.1	12.8	15.0	12.1	11.9	15.5	14.2	11.0	13.2	13.5	12.5
100.000	14.7	14.2	12.1	14.1	11.5	11.2	14.6	13.4	10.4	12.5	12.7	11.8
200.000	9.9	9.7	8.2	9.3	7.9	7.7	9.7	9.2	7.4	8.5	8.7	7.8
300.000	7.8	7.7	6.5	7.3	6.3	6.2	7.6	7.3	6.0	6.7	6.9	6.2
400.000	6.6	6.6	5.5	6.1	5.4	5.3	6.5	6.3	5.2	5.7	5.9	5.2
500.000	5.8	5.8	4.9	5.4	4.8	4.7	5.7	5.5	4.7	5.1	5.2	4.6
750.000	4.6	4.6	3.9	4.2	3.9	3.8	4.5	4.4	3.8	4.0	4.2	3.6
1.000.000	3.9	3.9	3.3	3.5	3.3	3.2	3.8	3.8	3.3	3.4	3.6	3.0
2.000.000	2.6	2.7	2.2	2.3	2.3	2.2	2.5	2.6	2.3	2.3	2.4	2.0
3.000.000	2.1	2.1	1.8	1.8	1.8	1.8	2.0	2.1	1.9	1.8	1.9	1.6
4.000.000	1.8	1.8	1.5	1.5	1.6	-	1.7	1.8	1.7	1.6	1.7	1.3
5.000.000	1.6	1.6	1.3	1.3	1.4	-	1.5	1.6	1.5	1.4	1.5	1.2
7.500.000	1.2	-	-	-	-	-	1.2	1.3	1.2	1.1	1.2	0.9
10.000.000	1.0	-	-	-	-	-	-	-	-	-	-	-
15.000.000	0.8	-	-	-	-	-	-	-	-	-	-	-
20.000.000	0.7	-	-	-	-	-	-	-	-	-	-	-

Prospetto 3 segue **- Valori interpolati degli errori campionari relativi percentuali delle stime riferite alle FAMIGLIE per totale Italia, ripartizione geografica, tipo di comune e regione**

STIME	Piemonte	Valle d'Aosta	Lombardia	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Liguria	Emilia Romagna	Toscana	Umbria
20.000	27.5	5.6	38.3	11.0	12.8	29,2	18.7	21.0	30.0	26.6	15.4
30.000	21.9	4.4	30.5	8.8	10.2	23,5	14.7	16.7	23.7	21.0	12.1
40.000	18.7	3.7	25.9	7.5	8.6	20,2	12.4	14.2	20.1	17.8	10.2
50.000	16.5	3.2	22.9	6.6	7.6	17,9	10.9	12.5	17.6	15.6	9.0
60.000	14.9	-	20.7	5.9	6.8	16,2	9.7	11.3	15.8	14.0	8.1
70.000	13.6	-	18.9	5.4	6.3	15,0	8.9	10.3	14.5	12.8	7.4
80.000	12.6	-	17.6	5.1	5.8	13,9	8.2	9.6	13.4	11.8	6.8
90.000	11.8	-	16.5	4.7	5.4	13,1	7.6	9.0	12.5	11.1	6.4
100.000	11.2	-	15.5	4.5	5.1	12,4	7.2	8.5	11.8	10.4	6.0
200.000	7.6	-	10.5	-	-	8,5	4.7	5.7	7.9	6.9	4.0
300.000	6.0	-	8.4	-	-	6,9	3.7	4.5	6.2	5.5	3.1
400.000	5.1	-	7.1	-	-	5,9	3.1	3.9	5.2	4.6	-
500.000	4.5	-	6.3	-	-	5,2	2.7	3.4	4.6	4.1	-
750.000	3.6	-	5.0	-	-	4,2	-	-	3.6	3.2	-
1.000.000	3.1	-	4.3	-	-	3,6	-	-	3.1	2.7	-
2.000.000	2.1	-	2.9	-	-	-	-	-	-	-	-

Prospetto 3 segue **- Valori interpolati degli errori campionari relativi percentuali delle stime riferite alle FAMIGLIE per totale Italia, ripartizione geografica, tipo di comune e regione**

STIME	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
20.000	19.0	39.8	18.5	7.9	30.5	27.6	11.7	20.1	28.8	19.0
30.000	14.9	31.4	14.6	6.2	24.5	22.0	9.3	15.8	23.0	15.4
40.000	12.6	26.5	12.3	5.3	21.0	18.7	7.9	13.4	19.6	13.2
50.000	11.0	23.2	10.8	4.6	18.6	16.5	6.9	11.7	17.3	11.8
60.000	9.9	20.9	9.7	4.1	16.8	14.9	6.3	10.6	15.7	10.7
70.000	9.0	19.1	8.9	3.8	15.5	13.7	5.7	9.6	14.4	9.9
80.000	8.3	17.6	8.2	-	14.4	12.7	5.3	8.9	13.4	9.2
90.000	7.8	16.5	7.7	-	13.5	11.9	5.0	8.3	12.5	8.7
100.000	7.3	15.5	7.2	-	12.8	11.2	4.7	7.8	11.8	8.2
200.000	4.8	10.3	4.8	-	8.8	7.6	-	5.2	8.0	5.7
300.000	3.8	8.1	3.8	-	7.1	6.1	-	4.1	6.4	4.6
400.000	3.2	6.9	3.2	-	6.0	5.2	-	3.5	5.5	4.0
500.000	2.8	6.0	-	-	5.4	4.6	-	3.1	4.8	-
750.000	-	4.7	-	-	4.3	3.6	-	-	3.9	-
1.000.000	-	4.0	-	-	3.7	3.1	-	-	3.3	-
2.000.000	-	2.7	-	-	2.5	-	-	-	-	-

Prospetto 4 - Valori interpolati degli errori campionari relativi percentuali delle stime riferite alle PERSONE per totale Italia, ripartizione geografica, tipo di comune e regione

STIME	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	A1	A2	B1	B2	B3	B4
20.000	43.6	40.8	33.3	40.7	30.5	30.0	41.0	37.3	24.6	33.4	34.5	34.8
30.000	34.1	31.9	26.1	31.8	24.2	23.8	32.0	29.3	19.7	26.3	27.3	27.2
40.000	28.7	26.8	22.0	26.7	20.5	20.2	26.8	24.7	16.9	22.2	23.1	22.8
50.000	25.1	23.3	19.2	23.3	18.0	17.8	23.4	21.6	15.0	19.5	20.3	19.9
60.000	22.5	20.9	17.2	20.8	16.2	16.0	20.9	19.4	13.6	17.5	18.2	17.8
70.000	20.5	19.0	15.7	19.0	14.9	14.7	19.1	17.7	12.5	16.0	16.7	16.2
80.000	18.9	17.5	14.5	17.5	13.8	13.6	17.6	16.4	11.6	14.8	15.4	14.9
90.000	17.6	16.3	13.5	16.3	12.9	12.7	16.3	15.3	10.9	13.8	14.4	13.9
100.000	16.5	15.3	12.7	15.2	12.1	12.0	15.3	14.3	10.3	13.0	13.5	13.0
200.000	10.8	10.0	8.4	10.0	8.1	8.0	10.0	9.5	7.1	8.6	9.0	8.5
300.000	8.5	7.8	6.6	7.8	6.5	6.4	7.8	7.5	5.7	6.8	7.1	6.6
400.000	7.1	6.6	5.5	6.5	5.5	5.4	6.6	6.3	4.9	5.7	6.0	5.6
500.000	6.2	5.7	4.8	5.7	4.8	4.8	5.7	5.5	4.3	5.0	5.3	4.9
750.000	4.9	4.5	3.8	4.5	3.8	3.8	4.5	4.3	3.5	4.0	4.2	3.8
1.000.000	4.1	3.8	3.2	3.7	3.2	3.2	3.8	3.7	3.0	3.4	3.5	3.2
2.000.000	2.7	2.5	2.1	2.4	2.2	2.2	2.5	2.4	2.0	2.2	2.4	2.1
3.000.000	2.1	1.9	1.6	1.9	1.7	1.7	1.9	1.9	1.6	1.8	1.9	1.6
4.000.000	1.8	1.6	1.4	1.6	1.5	1.5	1.6	1.6	1.4	1.5	1.6	1.4
5.000.000	1.5	1.4	1.2	1.4	1.3	1.3	1.4	1.4	1.2	1.3	1.4	1.2
7.500.000	1.2	1.1	0.9	1.1	1.0	-	1.1	-	-	1.0	1.1	0.9
10.000.000	1.0	0.9	-	-	0.9	-	-	-	-	0.9	0.9	-
15.000.000	0.8	-	-	-	-	-	-	-	-	-	-	-
20.000.000	0.7	-	-	-	-	-	-	-	-	-	-	-
25.000.000	0.6	-	-	-	-	-	-	-	-	-	-	-

Prospetto 4 segue **- Valori interpolati degli errori campionari relativi percentuali delle stime riferite alle PERSONE per totale Italia, ripartizione geografica, tipo di comune e regione**

STIME	Piemonte	Valle d'Aosta	Lombardia	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Liguria	Emilia Romagna	Toscana	Umbria
27.000	24.2	5.0	35.5	9.2	10.2	26.8	15.7	18.2	26.5	24.1	13.4
30.000	22.8	4.7	33.3	8.6	9.6	25.2	14.7	17.1	24.9	22.6	12.5
40.000	19.2	4.0	28.0	7.3	8.0	21.3	12.4	14.4	20.9	19.1	10.4
50.000	16.8	3.5	24.4	6.3	7.0	18.6	10.9	12.7	18.3	16.7	9.1
60.000	15.1	3.1	21.8	5.7	6.2	16.7	9.8	11.4	16.4	15.0	8.1
70.000	13.8	2.8	19.9	5.2	5.7	15.3	8.9	10.4	14.9	13.7	7.3
80.000	12.7	2.6	18.3	4.8	5.2	14.1	8.3	9.6	13.8	12.7	6.7
90.000	11.9	2.4	17.1	4.5	4.9	13.2	7.7	9.0	12.8	11.8	6.2
100.000	11.2	2.3	16.0	4.2	4.6	12.4	7.2	8.5	12.0	11.1	5.8
200.000	7.4	-	10.5	2.8	3.0	8.2	4.8	5.7	7.9	7.4	3.8
300.000	5.8	-	8.2	2.2	2.3	6.5	3.8	4.5	6.2	5.8	2.9
400.000	4.9	-	6.9	1.8	1.9	5.5	3.2	3.8	5.2	4.9	2.4
500.000	4.3	-	6.0	-	-	4.8	2.8	3.3	4.6	4.3	2.1
750.000	3.4	-	4.7	-	-	3.8	2.2	2.6	3.6	3.4	1.6
1.000.000	2.9	-	3.9	-	-	3.2	1.9	2.2	3.0	2.9	-
2.000.000	1.9	-	2.6	-	-	2.1	-	-	2.0	1.9	-
3.000.000	1.5	-	2.0	-	-	1.7	-	-	1.5	1.5	-
4.000.000	1.3	-	1.7	-	-	1.4	-	-	-	-	-
5.000.000	-	-	1.5	-	-	-	-	-	-	-	-

Prospetto 4 segue **- Valori interpolati degli errori campionari relativi percentuali delle stime riferite alle PERSONE per totale Italia, ripartizione geografica, tipo di comune e regione**

STIME	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
27.000	16.9	36.7	16.6	7.6	26.2	23.6	10.0	19.6	26.4	16.6
30.000	15.8	34.5	15.6	7.1	24.7	22.2	9.4	18.4	24.9	15.6
40.000	13.4	29.0	13.2	6.0	21.1	18.7	8.0	15.5	21.1	13.3
50.000	11.8	25.3	11.5	5.3	18.6	16.4	7.0	13.6	18.6	11.7
60.000	10.6	22.7	10.4	4.7	16.8	14.8	6.3	12.2	16.7	10.5
70.000	9.7	20.7	9.5	4.3	15.4	13.5	5.8	11.1	15.3	9.6
80.000	8.9	19.1	8.7	4.0	14.3	12.5	5.3	10.2	14.2	8.9
90.000	8.3	17.8	8.2	3.7	13.4	11.7	5.0	9.5	13.3	8.3
100.000	7.8	16.7	7.7	3.5	12.7	11.0	4.7	9.0	12.5	7.8
200.000	5.2	11.0	5.1	2.3	8.6	7.3	3.1	5.9	8.4	5.3
300.000	4.1	8.6	4.0	1.8	6.9	5.8	2.5	4.6	6.7	4.2
400.000	3.5	7.2	3.4	-	5.9	4.9	2.1	3.9	5.7	3.5
500.000	3.1	6.3	3.0	-	5.2	4.3	1.8	3.4	5.0	3.1
750.000	2.4	4.9	2.3	-	4.1	3.4	-	2.7	4.0	2.5
1.000.000	2.0	4.2	2.0	-	3.5	2.8	-	2.3	3.4	2.1
2.000.000	-	2.7	-	-	2.4	1.9	-	1.5	2.3	-
3.000.000	-	2.1	-	-	1.9	1.5	-	-	1.8	-
4.000.000	-	1.8	-	-	1.6	-	-	-	1.5	-
5.000.000	-	1.6	-	-	1.4	-	-	-	1.3	-

Esempi di calcolo degli errori campionari

Esempio 1

Si consideri la stima del numero di individui pendolari, che per la ripartizione Sud è pari a 663.570 unità.

Nella prima colonna del prospetto 4 della presente appendice metodologica, si cerca il valore più vicino a questa stima, che è pari a 750.000. In corrispondenza di tale valore, per la ripartizione Sud, è riportato un errore relativo percentuale del 3,8 per cento.

Pertanto, l'errore assoluto della stima sarà uguale a:

$\sigma(663.570) = 0{,}038 \times 663.570 = 25.216$

e l'intervallo di confidenza al 95% avrà come estremi:

663.570 – (1,96 x 25.216) = 614.147
663.570 + (1,96 x 25.216) = 712.993.

Esempio 2

Considerando la medesima stima dell'esempio 1, si possono ottenere valori più precisi dell'errore di campionamento operando mediante interpolazione lineare dei due livelli di stima consecutivi tra i quali è compreso il valore della stessa, nel prospetto 3. Tali livelli sono 500.000 e 750.000 ai quali corrispondono, rispettivamente, i valori percentuali 4,8 e 3,8. L'errore relativo corrispondente a 663.570 è pari a:

$\hat{\varepsilon}\,(663.570) = 4{,}8 - (4{,}8 - 3{,}8) \times (663.570 - 500.000) / (750.000 - 500.000) = 4{,}15$ per cento.

L'errore assoluto sarà il seguente:

$\sigma(663.570) = 0{,}0454 \times 663.570 = 27.510$

e l'intervallo di confidenza avrà come estremi:

663.570- (1,96 x 27.510) = 609.651
663.570+ (1,96 x 27.510) = 717.489.

Esempio 3

Il calcolo dell'errore dell'esempio 1 può essere effettuato, direttamente, tramite la funzione interpolante:

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{\exp\left(a + b \ln(\hat{Y})\right)}$$

i cui parametri, riportati nel prospetto 2 alla riga Toscana alla voce Famiglie, sono i seguenti:

a = 8,974649 b = -1,146204.

Per $\hat{Y}$ = 663.570 si ha:

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{exp\left(8{,}974649 - 1{,}146204 \times ln(663.570)\right)} = 0{,}040953.$$

L'errore relativo percentuale è quindi pari al 4,01 per cento e il calcolo dell'errore assoluto e dell'intervallo di confidenza è del tutto analogo a quello degli esempi 1 e 2.