



Famiglie, soggetti sociali e ciclo di vita
Periodo di riferimento: anno 2016

Aspetti metodologici dell'indagine

INDICE

1. Introduzione	3
2. Obiettivi conoscitivi	3
3. Strategia di campionamento	4
4. La rilevazione e il trattamento dei dati	7
5. La metodologia di calcolo dei pesi campionari	8
6. Valutazione del livello di precisione delle stime	8
7. La diffusione dei risultati dell'indagine	21
8. Riferimenti bibliografici.....	21

1. Introduzione

L'indagine campionaria Famiglia e soggetti sociali fa parte del Sistema integrato di indagini sociali (Sistema di indagini Multiscopo sulle famiglie) ed è condotta ogni cinque anni a partire dal 1998.

Costituisce una delle principali fonti statistiche sulla struttura familiare e sulle caratteristiche sociali della famiglia in Italia. Le differenti tematiche che si susseguono nei questionari permettono di approfondire vari punti del vivere quotidiano: il ciclo di vita individuale e familiare, i rapporti interni alla famiglia, le reti di relazione con parenti, amici e vicinato, il sostegno ricevuto dalle famiglie e l'aiuto dato a persone non coabitanti, la cura e l'affidamento dei bambini, la vita di coppia e le prime nozze, la permanenza dei giovani in famiglia e le intenzioni di lasciare la famiglia di origine, le intenzioni riproduttive, la mobilità sociale, l'economia familiare e il lavoro domestico, i servizi assistenziali alla famiglia, il rapporto con il mondo del lavoro e la ricerca del lavoro, i percorsi lavorativi e formativi.

Nel 2016 l'indagine è stata svolta a dicembre su un campione di 32 mila 585 individui di 18 anni e più distribuiti in 852 comuni italiani di diversa ampiezza demografica. Le interviste sono state effettuate da rilevatori comunali presso l'abitazione dell'individuo campione, estratto casualmente dalle liste anagrafiche del comune.

2. Obiettivi conoscitivi

La popolazione di interesse dell'indagine in oggetto, ossia l'insieme delle unità statistiche intorn o alle quali si intende investigare, è costituita dagli individui maggiorenni residenti in Italia, al netto dei membri permanenti delle convivenze.

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 quattro classi formate in base a caratteristiche socio-economiche e demografiche:

A) comuni appartenenti all'area metropolitana suddivisi in:

A1, comuni centro dell'area metropolitana: Torino, Milano, Venezia, Genova, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania, Cagliari;

A2, comuni che gravitano intorno ai comuni centro dell'area metropolitana;

B) comuni non appartenenti all'area metropolitana suddivisi in:

B1 comuni aventi fino a 10.000 abitanti;

B2 comuni con oltre 10.000 abitanti.

3. Strategia di campionamento

3.1. Descrizione generale del disegno di campionamento

Il disegno di campionamento è di tipo complesso e si avvale di due differenti schemi di campionamento, basati sulla struttura clusterizzata della popolazione in comuni.

Nell'ambito di ognuno dei domini definiti dall'incrocio della regione geografica con le quattro aree A1, A2, B1, B2, 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 a uno stadio stratificato nel quale gli individui campione di ogni comune sono selezionati con campione casuale semplice.

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 gli individui.

I comuni vengono selezionati con probabilità proporzionali alla loro dimensione demografica e senza reimmissione, mentre gli individui vengono estratti con probabilità uguali e senza reimmissione.

La lista di campionamento per la selezione del campione è costituita dall'archivio unificato delle liste anagrafiche comunali (LAC). Tale archivio contiene le informazioni degli individui e delle famiglie residenti sul territorio italiano: per ciascun individuo sono riportate, oltre

alle variabili identificative – compreso il codice fiscale - l’indirizzo, la data di nascita, il sesso, la cittadinanza e l’anno di iscrizione in anagrafe.

3.2. 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, 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 individui, prefissata a livello nazionale essenzialmente in base a criteri di costo ed operativi, è pari a circa 32.000;
- 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 individui 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, sia quella delle stime a livello di ciascuno dei domini territoriali descritti nel primo paragrafo.

3.3. 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 in oggetto, i comuni vengono stratificati in base alla loro dimensione demografica e nel rispetto delle seguenti condizioni:

- autoponderazione del campione a livello regionale (ossia tutti gli individui di una stessa regione devono avere la stessa probabilità di entrare nel campione);
- selezione di un comune campione nell'ambito di ciascuno strato definito sui comuni dell'insieme NAR;
- scelta di un numero minimo di individui da intervistare in ciascun comune campione; per l'indagine in oggetto tale numero è stato posto pari a 30;
- 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 A1, A2, B1, B2, 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 maggiorenne residente; determinazione di una soglia di popolazione per la definizione dei comuni AR, mediante la relazione:

$$r\lambda = \frac{r\bar{m}}{rf}$$

in cui per la generica regione geografica r si è indicato con: $r\bar{m}$ il numero minimo di individui da intervistare in ciascun comune campione; rf la frazione di campionamento; suddivisione dei 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 maggiorenne residente, approssimativamente costante e all'incirca pari alla soglia $r\lambda$.

Effettuata la stratificazione, i comuni AR sono stati inclusi con certezza nel campione; per quanto riguarda, invece, i comuni NAR, nell'ambito di ogni strato è stato estratto un

comune campione con probabilità proporzionale alla dimensione demografica, mediante la procedura di selezione sistematica proposta da Madow¹.

La selezione degli individui da intervistare in ogni comune campione è stata effettuata dalla lista anagrafica di ciascun comune, contenuta nell’archivio unificato LAC, senza reimmissione e con probabilità uguali.

Nel prospetto 1 viene riportata la distribuzione regionale dell’universo e del campione dei comuni e degli individui.

Prospetto 1 - Distribuzione regionale dei comuni e degli individui nell’universo e nel campione - Anno 2016

REGIONI	Comuni		Individui	
	Universo	Campione	Universo	Campione
Piemonte	1.206	61	3.695.163	2.273
Valle d’Aosta/Vallée d’Aoste	74	19	105.901	733
Lombardia	1.530	84	8.294.784	2.938
Bolzano	116	23	414.923	1.734
Trento	210	25	400.284	852
Veneto	579	58	4.068.409	1.907
Friuli-Venezia Giulia	216	32	1.031.070	1.204
Liguria	235	23	1.343.853	1.205
Emilia-Romagna	340	47	3.696.895	1.737
Toscana	279	52	3.157.324	1.886
Umbria	92	23	751.738	953
Marche	236	39	1.295.922	1.329
Lazio	378	35	4.914.214	1.852
Abruzzo	305	38	1.120.844	1.370
Molise	136	22	266.480	900
Campania	550	58	4.764.134	2.144
Puglia	258	53	3.381.956	1.804
Basilicata	131	27	480.334	939
Calabria	409	42	1.636.753	1.485
Sicilia	390	53	4.180.548	1.990
Sardegna	377	38	1.418.958	1.350
Italia	8.047	852	50.420.487	32.585

4. La rilevazione e il trattamento dei dati

L’intervista viene effettuata mediante tecnica Papi (Paper and Pencil interview) e prevede l’utilizzo di un questionario cartaceo.

Il questionario viene somministrato mediante intervista faccia a faccia. Questo modello è composto: da una “Scheda Generale”, in cui si rilevano le relazioni di parentela ed altre informazioni di natura socio-demografica e socio-economica relative ai componenti della famiglia.

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

Le informazioni vengono fornite direttamente da tutti gli individui di 18 anni e più. Taluni quesiti della rilevazione, per la sensibilità dell'argomento trattato, prevedono la facoltà di non rispondere.

Trattandosi di un'indagine PAPI, il questionario sono sottoposti a registrazione. A conclusione della registrazione dei dati, o meglio contestualmente ad essa poiché la registrazione procede per lotti distinti di questionari, prende avvio la fase di controllo della qualità dei dati raccolti e di validazione degli stessi, che ha il duplice obiettivo di garantire la qualità delle stime prodotte e produrre un archivio di dati elementari privo di incoerenze. Questi obiettivi vengono perseguiti attraverso un complesso e reiterato processo:

- di esplorazione dei dati, basato su una reportistica che ne evidenzia anomalie e incoerenze;
- di correzione delle incompatibilità rilevate tramite l'applicazione di opportuni interventi di correzione, sia deterministica, sia probabilistica.

Tutte le procedure di correzione sono poi valutate mediante analisi delle distribuzioni semplici e congiunte, con la determinazione dell'impatto delle procedure sulle stime finali, con le analisi di indicatori sulla frequenza di attivazione delle regole di compatibilità e di indicatori sulla frequenza di correzione per le variabili e con la valutazione delle tipologie di errore individuate (mancate risposte parziali, errori sistematici, errori casuali, valori anomali).

5. La metodologia di calcolo dei pesi campionari

Le stime prodotte dall'indagine sono essenzialmente stime di frequenze assolute e relative, riferite alle famiglie e agli individui. Di conseguenza a ciascun individuo del campione sono stati attribuiti due diversi pesi:

- un peso individuale, per la produzione di stime relative alle variabili riferite agli individui;
- un peso familiare, per la produzione di stime relative alle variabili di tipo familiare.

Le stime sono ottenute mediante uno stimatore di ponderazione vincolata, che è il metodo di stima standard 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 ad ogni unità campionaria un peso che indica il numero di unità della popolazione rappresentate dall'unità medesima. Se, ad esempio, ad un'unità campionaria viene attribuito un peso pari a 100, vuol dire che

questa unità rappresenta se stessa ed altre 99 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;

h , indice di strato (che indica l'incrocio di regione, tipologia comunale e lo strato dei comuni all'interno del medesimo dominio);

q , indice di individuo;

y , generica variabile oggetto di indagine;

y_{hq} , valore di y osservato sull'individuo q dello strato h ;

N_h , numero di individui di 18 anni e oltre nello strato h ;

n_h , numero di individui campione nello strato h ;

H_d , numero di strati nel dominio di interesse d .

Per semplicità si omette per il momento l'indice di comune i appartenente allo strato h .

Ipotizziamo di voler stimare, con riferimento ad un generico dominio d (ad esempio una regione geografica) il totale della variabile y oggetto di indagine, espresso dalla seguente relazione:

$${}_d Y = \sum_{h=1}^{H_d} \sum_{q=1}^{N_h} Y_{hq} \quad (1)$$

Una stima del totale (1) è data dalla seguente espressione:

$${}_d \hat{Y} = \sum_{h=1}^{H_d} \hat{Y}_h = \sum_{h=1}^{H_d} \sum_{q=1}^{n_h} y_{hq} \cdot w_{hq} \quad (2)$$

in cui w_{hq} rappresenta il peso finale da attribuire all'individuo campione q dello strato h .

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

I pesi finali da attribuire alle unità campionarie sono ottenuti per mezzo di una procedura complessa che ha le seguenti finalità: correggere l'effetto distorsivo dovuto agli errori di lista e al fenomeno della mancata risposta totale; tenere conto della conoscenza di alcuni totali noti sulla popolazione oggetto di studio, nel senso che le stime campionarie di tali totali devono coincidere con i rispettivi valori noti.

La procedura per la costruzione dei pesi finali da attribuire alle unità campionarie è preceduta dal calcolo del peso diretto, ottenuto come reciproco della probabilità di inclusione di ogni unità campionaria.

Calcolo dei pesi diretti

All'interno di ciascun comune, l'estrazione degli individui campione è stata effettuata in modo da assegnare uguale probabilità di selezione a maschi e femmine maggiorenni residenti nel comune. Indicato con :

N_{hi} il numero di individui maggiorenni di sesso s nel comune i dello strato h ,

n_{his} il numero di individui campione maggiorenni di sesso s selezionati nel comune i dello strato h ,

il peso diretto della generico individuo q residente nel comune i dello strato h viene calcolato come:

$$d_{hiq} = \frac{N_h}{N_{hi}} \cdot \frac{N_{his}}{n_{his}} \quad (3)$$

Per lo stesso individuo q , il peso diretto a livello familiare d_{hiq}^f si ottiene moltiplicando i coefficienti individuali, ottenuti come in (3), per il numero di componenti eleggibili della famiglia a cui appartiene l'individuo q selezionato; posto:

Q_{hiq} = numero di componenti eleggibili della famiglia dell'individuo q nel comune i dello strato h

$$d_{hiq}^f = d_{hiq} \cdot Q_{hiq} \quad (4)$$

Calcolo dei pesi finali

Il peso diretto viene calibrato rispetto a totali noti calcolati sulla popolazione complessiva per i quali sono disponibili dati più aggiornati da fonte demografica.

Per questa indagine sono stati calcolati due sistemi di pesi differenti per la produzione delle stime riferite agli individui e alle famiglie.

Per i pesi individuali, la popolazione di riferimento è costituita dagli individui di 18 anni e oltre secondo le liste anagrafiche comunali. I totali noti imposti a livello regionale sono i seguenti:

- a. popolazione per sesso e classi di età (14-24, 25-44, 45-64, 65 e oltre) ;
- b. popolazione per tipologia comunale (aree A1, A2, B1, B2, B3 definite nel paragrafo 1);

Per i pesi a livello familiare, la popolazione di riferimento è costituita dalla popolazione residente al netto delle convivenze e i totali noti a livello regionale sono i seguenti:

- a. popolazione per sesso e classi di età (14-24, 25-44, 45-64, 65 e oltre) ;
- b. popolazione per tipologia comunale e per dimensione familiare;

In fase di calibrazione i fattori correttivi sono ottenuti dalla risoluzione di problemi 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 e i 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.

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 vincolata². Un importante stimatore appartenente a tale classe, che si ottiene utilizzando la funzione di distanza euclidea o lineare, è lo stimatore di regressione generalizzata. Come verrà chiarito meglio nel paragrafo seguente, 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.

6. Valutazione del livello di precisione delle stime

6.1. 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

² Deville J.C. e Sarndäl C.E. (1992), “Calibration Estimators in Survey Sampling”, *Journal of the American Statistical Association* 87: 376-382.

campionamento relativo. Indicando con $\hat{\text{Var}}(\hat{Y}_d)$ 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:

$$\hat{\sigma}(\hat{Y}_d) = \sqrt{\hat{\text{Var}}(\hat{Y}_d)}; \quad (3)$$

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

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

Come è stato descritto in precedenza, le stime prodotte dall'indagine sono state ottenute mediante uno stimatore di ponderazione vincolata definito in base a una funzione di distanza di tipo logaritmico troncato. Poiché, lo stimatore adottato non è funzione lineare dei dati campionari, per la stima della varianza $\hat{\text{Var}}(\hat{Y}_d)$ 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 di 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:

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

dove Z_{hij} è la variabile linearizzata espressa come $Z_{hij} = Y_{hij} - \mathbf{X}'_{hij}\beta$, essendo $\mathbf{X}_{hij} = (X_{hij,1}, \dots, X_{hij,K})'$ il vettore contenente i valori delle K ($K=24$) 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

$$\hat{\text{Var}}(\hat{Y}_d) \cong \hat{\text{Var}}(\hat{Z}_d) = \sum_{h=1}^{H_d} \hat{\text{Var}}(\hat{Z}_h) . \quad (6)$$

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

$$\hat{\text{Var}}(\hat{Y}_d) \cong \hat{\text{Var}}(\hat{Z}_d) = \sum_{h=1}^{H_{AR}} \hat{\text{Var}}(\hat{Z}_h) + \sum_{h=1}^{H_{NAR}} \hat{\text{Var}}(\hat{Z}_h) , \quad (7)$$

in cui H_{AR} e H_{NAR} indicano rispettivamente il numero di strati A_r e N_{ar} appartenenti al dominio d .

Negli strati A_r (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:

$$\sum_{h=1}^{H_{AR}} \hat{\text{Var}}(\hat{Z}_h) = \sum_{h=1}^{H_{AR}} M_h^2 \frac{(M_h - m_h)}{m_h(m_h - 1)} \sum_{j=1}^{m_h} (Z_{hj} - \bar{Z}_h)^2 , \quad (8)$$

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

Negli strati N_{ar} , 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:

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

dove le quantità sono espresse come:

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

Utilizzando le espressioni (8) e (9) è possibile, infine, calcolare la varianza di campionamento, $\hat{\text{Var}}(\hat{Y}_d)$, 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:

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

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

6.2. 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 A_r , 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 N_r , 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.

6.3. 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:

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

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 R² del modello utilizzato per l'interpolazione degli errori campionari di stime di frequenze assolute e relative per aree 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:

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

che si ricava facilmente dalla (8).

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 dominio territoriale di riferimento.

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

$$\hat{\varepsilon}(\hat{Y}_d) = \hat{\varepsilon}(\hat{Y}_d^{k-1}) - \frac{\hat{\varepsilon}(\hat{Y}_d^{k-1}) - \hat{\varepsilon}(\hat{Y}_d^k)}{\hat{Y}_d^k - \hat{Y}_d^{k-1}} (\hat{Y}_d - \hat{Y}_d^{k-1}) \quad (10)$$

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 R² (%) 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 – Anno 2016

	Famiglie			Persone		
	A	B	R2 (%)	A	B	R2 (%)
Italia	62.200	-0.9088	0.9673	143.823	-14.851	0.9431
RIPARTIZIONI						
GEOGRAFICHE						
(a)						
Nord-ovest	634.619	-0.89122	0.95802	1.263.533	-140.029	0.93933
Nord-est	565.456	-0.86728	0.95158	1.131.676	-134.802	0.93642
Centro	557.893	-0.83064	0.92624	1.326.082	-145.244	0.93143
Sud	535.061	-0.85132	0.94958	1.462.580	-156.397	0.93211
Isole	560.693	-0.85638	0.94692	1.506.902	-161.139	0.94760
TIPI DI COMUNE (b)						
A1	651.614	-0.89400	0.93627	1.131.618	-128.036	0.95997
A2	600.295	-0.85555	0.92487	1.188.808	-132.894	0.93668
B1	413.089	-0.70949	0.86811	973.589	-122.256	0.95296
B2	570.105	-0.86759	0.95054	1.102.391	-127.077	0.94058
B3	596.968	-0.88739	0.96649	1.155.037	-130.548	0.96068
B4	543.167	-0.83192	0.94618	1.133.080	-130.163	0.95957
REGIONI						
Piemonte	577.501	-0.85734	0.94135	1.210.527	-141.490	0.93268
Valle d'Aosta/Vallée d'Aoste	432.004	-0.93712	0.93631	862.051	-145.448	0.94622
Lombardia	654.390	-0.88855	0.95448	1.218.610	-137.090	0.94019
Bolzano	414.576	-0.88840	0.93251	941.480	-144.170	0.96523
Trento	553.416	-0.95814	0.95288	971.898	-139.924	0.95301
Veneto	607.098	-0.88397	0.94069	1.114.716	-135.091	0.91952
Friuli-Venezia Giulia	507.272	-0.86163	0.94322	1.067.442	-141.737	0.94243
Liguria	627.449	-0.93156	0.94770	1.126.469	-141.973	0.90677
Emilia-Romagna	590.085	-0.87027	0.94733	1.148.142	-137.486	0.93288
Toscana	621.242	-0.90244	0.94972	1.170.125	-139.429	0.93881
Umbria	516.111	-0.88437	0.95361	1.063.640	-144.852	0.93876
Marche	510.004	-0.84899	0.94664	1.229.033	-152.062	0.94176
Lazio	664.064	-0.88909	0.94060	1.287.572	-142.285	0.93249
Abruzzo	535.995	-0.88969	0.93766	1.177.139	-148.345	0.93060
Molise	458.893	-0.91322	0.91941	957.014	-147.117	0.94863
Campania	613.556	-0.89412	0.95462	1.373.082	-151.528	0.91645
Puglia	619.894	-0.92207	0.95233	1.402.066	-157.927	0.92605
Basilicata	481.767	-0.88767	0.93595	1.097.996	-153.336	0.93528
Calabria	570.432	-0.91202	0.95320	1.162.763	-144.819	0.88218
Sicilia	620.082	-0.89985	0.93799	1.491.499	-160.382	0.94870
Sardegna	566.174	-0.88339	0.92883	1.195.426	-147.492	0.90713

(a) Italia nord-occidentale: Piemonte, Valle d'Aosta, Liguria, Lombardia; 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	24.9	28.9	23.1	26.6	21.4	23.8	31.1	29.1	23.5	23.6	24.4	24.6
30.000	20.7	24.2	19.3	22.5	18.0	20.0	25.9	24.5	20.4	19.8	20.4	20.8
40.000	18.2	21.2	17.1	20.0	16.0	17.7	22.8	21.6	18.4	17.4	18.0	18.4
50.000	16.4	19.2	15.5	18.2	14.5	16.1	20.6	19.7	17.0	15.8	16.3	16.8
60.000	15.1	17.7	14.3	16.9	13.4	14.8	19.0	18.2	15.9	14.6	15.0	15.6
70.000	14.1	16.6	13.4	15.8	12.6	13.9	17.7	17.0	15.1	13.7	14.0	14.6
80.000	13.3	15.6	12.6	15.0	11.9	13.1	16.7	16.1	14.4	12.9	13.2	13.8
90.000	12.6	14.8	12.0	14.3	11.3	12.5	15.9	15.3	13.8	12.3	12.5	13.1
100.000	12.0	14.1	11.5	13.6	10.8	11.9	15.1	14.6	13.3	11.7	12.0	12.6
200.000	8.7	10.4	8.5	10.2	8.0	8.9	11.1	10.9	10.4	8.7	8.8	9.4
300.000	7.3	8.7	7.1	8.6	6.8	7.5	9.3	9.1	9.0	7.3	7.3	8.0
400.000	6.4	7.6	6.3	7.7	6.0	6.6	8.1	8.1	8.1	6.4	6.5	7.1
500.000	5.8	6.9	5.7	7.0	5.4	6.0	7.4	7.3	7.5	5.8	5.9	6.4
750.000	4.8	5.8	4.8	5.9	4.6	5.0	6.1	6.2	6.5	4.9	4.9	5.4
1.000.000	4.2	5.1	4.2	5.2	4.1	4.5	5.4	5.5	5.9	4.3	4.3	4.8
2.000.000	3.1	3.7	3.1	3.9	3.0	3.3	4.0	4.1	4.6	3.2	3.2	3.6
3.000.000	2.6	3.1	2.6	3.3	2.5	-	3.3	3.4	4.0	2.7	2.6	3.1
4.000.000	2.2	2.7	2.3	2.9	2.2	-	-	-	-	-	-	-
5.000.000	2.0	2.5	2.1	2.7	2.0	-	-	-	-	-	-	-
7.500.000	1.7	-	-	-	-	-	-	-	-	-	-	-
10.000.000	1.5	-	-	-	-	-	-	-	-	-	-	-
15.000.000	1.2	-	-	-	-	-	-	-	-	-	-	-
20.000.000	1.1	-	-	-	-	-	-	-	-	-	-	-

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	25.7	8.4	32.4	9.8	13.8	26.1	17.7	22.9	25.7	25.6	16.6
30.000	21.6	6.9	27.0	8.2	11.4	21.9	14.9	18.9	21.5	21.3	13.8
40.000	19.1	6.0	23.8	7.2	9.9	19.2	13.1	16.6	19.0	18.7	12.2
50.000	17.4	5.4	21.5	6.5	8.9	17.4	11.9	14.9	17.2	16.9	11.0
60.000	16.1	5.0	19.9	6.0	8.2	16.1	11.0	13.7	15.9	15.6	10.2
70.000	15.0	-	18.6	5.6	7.6	15.0	10.3	12.8	14.9	14.5	9.5
80.000	14.2	-	17.5	5.3	7.1	14.2	9.8	12.0	14.1	13.7	9.0
90.000	13.5	-	16.6	5.0	6.7	13.4	9.3	11.3	13.4	13.0	8.5
100.000	12.9	-	15.8	4.8	6.4	12.8	8.9	10.8	12.8	12.4	8.1
200.000	9.6	-	11.6	3.5	4.6	9.4	6.6	7.8	9.4	9.1	6.0
300.000	8.1	-	9.7	-	-	7.9	5.5	6.5	7.9	7.5	5.0
400.000	7.1	-	8.6	-	-	7.0	4.9	5.7	7.0	6.6	-
500.000	6.5	-	7.7	-	-	6.3	4.4	5.1	6.3	6.0	-
750.000	5.4	-	6.5	-	-	5.3	-	4.2	5.3	5.0	-
1.000.000	4.8	-	5.7	-	-	4.6	-	-	4.7	4.4	-
2.000.000	3.6	-	4.2	-	-	3.4	-	-	3.5	3.2	-

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	19.1	33.9	17.8	10.8	25.7	23.1	13.7	18.9	25.8	21.4
30.000	16.1	28.3	14.9	9.0	21.4	19.1	11.5	15.7	21.5	17.9
40.000	14.3	24.9	13.1	7.9	18.8	16.8	10.1	13.8	18.9	15.7
50.000	13.0	22.5	11.8	7.1	17.0	15.1	9.1	12.5	17.1	14.3
60.000	12.0	20.8	10.9	6.5	15.7	13.9	8.4	11.5	15.7	13.2
70.000	11.2	19.4	10.2	6.1	14.7	13.0	7.9	10.7	14.7	12.3
80.000	10.6	18.3	9.6	5.7	13.8	12.2	7.4	10.1	13.8	11.6
90.000	10.1	17.4	9.1	5.4	13.1	11.5	7.0	9.5	13.1	11.0
100.000	9.7	16.6	8.7	5.2	12.5	11.0	6.7	9.1	12.5	10.5
200.000	7.2	12.2	6.4	-	9.2	8.0	4.9	6.6	9.2	7.7
300.000	6.1	10.2	5.3	-	7.7	6.6	-	5.5	7.6	6.5
400.000	5.4	8.9	4.7	-	6.7	5.8	-	4.8	6.7	5.7
500.000	4.9	8.1	4.3	-	6.1	5.2	-	4.4	6.1	5.2
750.000	-	6.8	-	-	5.1	4.3	-	3.6	5.0	4.3
1.000.000	-	6.0	-	-	4.5	3.8	-	-	4.4	-
2.000.000	-	4.4	-	-	3.3	-	-	-	3.2	-

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	85.0	54.0	36.2	57.0	65.0	64.1	50.6	52.9	30.5	45.8	50.2	45.8
30.000	62.9	40.7	27.5	42.5	47.3	46.2	39.0	40.4	23.8	35.4	38.5	35.2
40.000	50.8	33.2	22.7	34.5	37.8	36.7	32.4	33.4	20.0	29.5	31.9	29.2
50.000	43.0	28.4	19.5	29.3	31.7	30.6	28.1	28.8	17.4	25.6	27.6	25.3
60.000	37.6	25.0	17.3	25.7	27.5	26.5	25.0	25.5	15.6	22.8	24.5	22.4
70.000	33.5	22.5	15.6	23.0	24.4	23.4	22.7	23.0	14.2	20.7	22.2	20.3
80.000	30.4	20.5	14.2	20.8	22.0	21.0	20.8	21.1	13.1	19.0	20.3	18.6
90.000	27.8	18.8	13.1	19.1	20.0	19.1	19.3	19.5	12.2	17.6	18.8	17.2
100.000	25.7	17.5	12.2	17.7	18.5	17.5	18.0	18.2	11.4	16.5	17.6	16.1
200.000	15.4	10.8	7.7	10.7	10.7	10.0	11.6	11.5	7.5	10.6	11.2	10.2
300.000	11.4	8.1	5.8	8.0	7.8	7.2	8.9	8.8	5.8	8.2	8.6	7.9
400.000	9.2	6.6	4.8	6.5	6.2	5.7	7.4	7.2	4.9	6.8	7.1	6.5
500.000	7.8	5.7	4.1	5.5	5.2	4.8	6.4	6.2	4.3	5.9	6.1	5.6
750.000	5.8	4.3	3.1	4.1	3.8	3.5	5.0	4.8	3.3	4.6	4.7	4.3
1.000.000	4.7	3.5	2.6	3.3	3.0	2.7	4.1	3.9	2.8	3.8	3.9	3.6
2.000.000	2.8	2.1	1.6	2.0	1.8	1.6	2.7	2.5	1.8	2.5	2.5	2.3
3.000.000	2.1	1.6	1.2	1.5	1.3	1.1	2.0	1.9	1.4	1.9	1.9	1.8
4.000.000	1.7	1.3	1.0	1.2	1.0	0.9	1.7	1.6	1.2	1.6	1.6	1.5
5.000.000	1.4	1.1	0.9	1.0	0.9	0.7	1.5	1.3	1.0	1.4	1.4	1.3
7.500.000	1.0	0.9	0.7	0.8	0.6	0.5	-	-	-	-	-	-
10.000.000	0.8	0.7	0.5	0.6	0.5	-	-	-	-	-	-	-
15.000.000	0.6	0.5	0.4	-	-	-	-	-	-	-	-	-
20.000.000	0.5	-	-	-	-	-	-	-	-	-	-	-

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	Piemonte	Valle d'Aosta	Lombardia	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Liguria	Emilia Romagna	Toscana	Umbria
20.000	38.5	5.5	49.9	8.8	12.6	32.8	18.6	24.7	34.4	34.9	15.7
30.000	28.9	4.1	37.8	6.6	9.5	24.9	14.0	18.5	26.0	26.3	11.7
40.000	23.6	3.4	31.0	5.3	7.8	20.5	11.4	15.1	21.4	21.5	9.5
50.000	20.2	2.8	26.6	4.5	6.7	17.6	9.7	12.9	18.3	18.4	8.1
60.000	17.7	2.5	23.5	4.0	5.9	15.6	8.5	11.3	16.2	16.2	7.1
70.000	15.9	2.2	21.1	3.6	5.3	14.1	7.7	10.2	14.5	14.6	6.3
80.000	14.5	2.0	19.3	3.2	4.8	12.8	7.0	9.2	13.3	13.3	5.7
90.000	13.3	1.9	17.8	3.0	4.4	11.9	6.4	8.5	12.2	12.2	5.3
100.000	12.3	1.7	16.6	2.8	4.1	11.0	5.9	7.9	11.4	11.4	4.9
200.000	7.6	-	10.3	1.7	2.5	6.9	3.6	4.8	7.1	7.0	3.0
300.000	5.7	-	7.8	1.2	1.9	5.3	2.7	3.6	5.3	5.3	2.2
400.000	4.6	-	6.4	1.0	1.6	4.3	2.2	2.9	4.4	4.3	1.8
500.000	4.0	-	5.5	0.9	-	3.7	1.9	2.5	3.8	3.7	1.5
750.000	3.0	-	4.2	-	-	2.8	1.4	1.9	2.8	2.8	1.1
1.000.000	2.4	-	3.4	-	-	2.3	-	-	2.3	2.3	-
2.000.000	1.5	-	2.1	-	-	1.5	-	-	1.5	1.4	-
3.000.000	1.1	-	1.6	-	-	1.1	-	-	1.1	1.1	-
4.000.000	0.9	-	1.3	-	-	0.9	-	-	0.9	0.9	-
5.000.000	-	-	1.1	-	-	-	-	-	0.8	-	-

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
20.000	25.0	54.5	23.2	8.2	52.8	44.5	12.2	25.7	61.6	26.5
30.000	18.4	40.8	17.2	6.1	38.9	32.3	8.9	19.2	44.5	19.7
40.000	14.8	33.3	13.9	4.9	31.3	25.7	7.2	15.6	35.3	15.9
50.000	12.5	28.4	11.8	4.2	26.4	21.6	6.0	13.3	29.6	13.5
60.000	10.9	24.9	10.3	3.7	23.0	18.7	5.3	11.6	25.5	11.8
70.000	9.7	22.3	9.2	3.3	20.5	16.5	4.7	10.4	22.6	10.5
80.000	8.7	20.3	8.3	3.0	18.5	14.9	4.2	9.4	20.3	9.5
90.000	8.0	18.7	7.6	2.7	16.9	13.6	3.9	8.7	18.4	8.8
100.000	7.4	17.3	7.0	2.5	15.6	12.5	3.6	8.0	17.0	8.1
200.000	4.3	10.6	4.2	1.5	9.2	7.2	2.1	4.9	9.7	4.9
300.000	3.2	7.9	3.1	1.1	6.8	5.2	1.5	3.6	7.0	3.6
400.000	2.6	6.5	2.5	-	5.5	4.2	1.2	2.9	5.6	2.9
500.000	2.2	5.5	2.1	-	4.6	3.5	1.0	2.5	4.7	2.5
750.000	1.6	4.1	1.6	-	3.4	2.5	-	1.9	3.4	1.8
1.000.000	1.3	3.4	1.3	-	2.7	2.0	-	1.5	2.7	1.5
2.000.000	-	2.1	-	-	1.6	1.2	-	-	1.5	-
3.000.000	-	1.5	-	-	1.2	0.9	-	-	1.1	-
4.000.000	-	1.3	-	-	1.0	0.7	-	-	0.9	-
5.000.000	-	1.1	-	-	0.8	-	-	-	0.7	-

6.4. Esempi di calcolo per la costruzione dell'intervallo di confidenza

Nelle tabelle seguenti sono illustrate le modalità di calcolo per la costruzione dell'intervallo di confidenza. Nel primo esempio (tabella 1), l'intervallo è calcolato per una stima sulle famiglie, l'errore campionario è da ricercare nel prospetto 3. Nel secondo esempio (tabella 2), il calcolo è fatto per una stima di individui, l'errore di riferimento è nel prospetto 4.

Tabella 1- Esempio per il calcolo degli errori campionari nel caso di stime riferite alle famiglie

	Famiglie nel Nord-est che dichiarano "adeguate" le proprie risorse economiche
Stima puntuale:	3.353.050
Errore relativo (CV)	2,6/100=0,026
Stima intervallare	
Semi ampiezza dell'intervallo	$(3.353.050 \cdot 0,026) \cdot 1,96 = 170.871$
Limite inferiore dell'intervallo di confidenza	$3.353.050 - 170.871 = 3.182.179$
Limite superiore dell'intervallo di confidenza	$3.353.050 + 170.871 = 3.523.921$

Tabella 2- Esempio per il calcolo degli errori campionari nel caso di stime riferite alle persone

	Persone che hanno avuto fratelli o sorelle in Abruzzo
Stima puntuale:	984.720
Errore relativo (CV)	1,3/100=0,013
Stima intervallare	
Semi ampiezza dell'intervallo	$(984.720 \cdot 0,013) \cdot 1,96 = 25.091$
Limite inferiore dell'intervallo di confidenza	$984.720 - 25.091 = 959.629$
Limite superiore dell'intervallo di confidenza	$984.720 + 25.091 = 1.009.811$

Per avere un intervallo più preciso, l'errore campionario può essere calcolato direttamente con la funzione interpolante:

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{\exp(a + b \log(\hat{Y}))}$$

utilizzando i parametri riportati nel prospetto 2.

7. La diffusione dei risultati dell'indagine

I dati d'indagine vengono resi disponibili mediante il rilascio di file di microdati (MFR e micro.stat). Ricercatori e studiosi possono, inoltre, accedere al Laboratorio di Analisi dei Dati Elementari (“ADELE”) per effettuare di persona le proprie analisi statistiche sui microdati dell'indagine, nel rispetto delle norme sulla riservatezza dei dati personali.

I dati diffusi sono privi degli elementi identificativi del soggetto al quale si riferiscono, nonché di ogni altro elemento che consenta, anche indirettamente, il collegamento con le famiglie o gli individui intervistati.

8. Riferimenti bibliografici

Il sistema di indagini sociali multiscopo, Metodi e Norme, n. 31, Anno 2006.

Contatti

**Servizio Registro della popolazione,
statistiche demografiche e condizioni di vita**

Eleonora Meli

Tel. +39 06 4673.7595

Email elmeli@istat.it