Istat
Istituto Nazionale di Statistica

File ad uso pubblico
mIcro.STAT

Aspetti della vita quotidiana

Periodo di riferimento: anno 2017

Aspetti metodologici dell'indagine

Anno 2019

INDICE

1. Introduzione 3
2. Obiettivi conoscitivi 3
3. Strategia di campionamento 4
4. La rilevazione e il trattamento dei dati 8
5. La metodologia di calcolo dei pesi campionari 9
6. Valutazione del livello di precisione delle stime 13
7. La diffusione dei risultati dell’indagine 21
8. Riferimenti bibliografici 22

1. Introduzione

L’indagine campionaria “Aspetti della vita quotidiana” fa parte del sistema integrato di Indagini Multiscopo sulle famiglie avviato dal 1993 con l’obiettivo di produrre informazioni sugli individui e sulle famiglie. L’indagine viene svolta ogni anno e le informazioni raccolte consentono di conoscere le abitudini dei cittadini e i problemi che essi affrontano ogni giorno. Aree tematiche variegate si susseguono nei questionari, permettendo di capire come vivono gli individui e se sono soddisfatti del funzionamento di quei servizi di pubblica utilità che devono contribuire al miglioramento della qualità della vita. Scuola, lavoro, vita familiare e di relazione, abitazione e zona in cui si vive, tempo libero, partecipazione politica e sociale, salute, stili di vita e rapporto con i servizi sono indagati in un’ottica in cui oggettività dei comportamenti e soggettività delle aspettative, delle motivazioni, dei giudizi contribuiscono a definire l’informazione sociale.

L’indagine rientra tra quelle comprese nel Programma statistico nazionale, che raccoglie l’insieme delle rilevazioni statistiche necessarie al Paese.

2. Obiettivi conoscitivi

La popolazione di interesse dell’indagine multiscopo “Aspetti della vita quotidiana”, ossia l’insieme delle unità statistiche intorno alle quali si intende investigare, è costituita dalle famiglie residenti in Italia e dai membri che le compongono; sono pertanto esclusi i 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 al 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:
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 2.000 abitanti;
B2, comuni con 2.001-10.000 abitanti;
B3, comuni con 10.001-50.000 abitanti;
B4, comuni con oltre 50.000 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. Nell’ambito di ognuno dei domini definiti dall’incrocio della regione geografica con le sei aree A1, A2, B1, B2, B3 e B4, i comuni 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 è considerato come uno strato a sé 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.

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, 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 famiglie;
- 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 quindi calcolata 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 precedente 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 esame, 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 (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 A1, A2, B1, B2, B3 e B4 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:

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

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; ${}_r f$ 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 ordinate le famiglie dell’anagrafe del comune;
- 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 degli individui nell’universo e nel campione – Anno 2017

REGIONI	Comuni			Famiglie			Individui	
	Campione effettivo	Campione teorico	Universo	Campione effettivo	Campione teorico	Universo (a)	Campione effettivo	Universo (a)
Piemonte	63	63	1.202	1.461	1.870	2.004	3.243	4.359
Valle d'Aosta - Vallée d'Aoste	22	22	74	541	632	66	1.035	126
Liguria	26	26	235	896	1.107	756	1.864	1.555
Lombardia	84	84	1.528	1.851	2.271	4.362	4.306	9.963
Trentino-Alto Adige	48	48	294	1.140	1.442	441	2.714	1.052
Bolzano - Bozen	*23*	*23*	*116*	*524*	*698*	*212*	*1.248*	*519*
Trento	*25*	*25*	*178*	*616*	*744*	*229*	*1.466*	*533*
Veneto	55	55	579	1.206	1.377	2.039	2.885	4.868
Friuli-Venezia Giulia	33	33	216	799	972	563	1.722	1.207
Emilia-Romagna	47	47	334	1.197	1.382	1.980	2.707	4.424
Toscana	51	51	279	1.241	1.468	1.629	2.888	3.726
Umbria	22	22	92	604	711	388	1.344	884
Marche	38	38	236	830	968	631	1.997	1.531
Lazio	34	34	378	1.298	1.801	2.542	2.925	5.870
Abruzzo	37	37	305	854	989	537	2.053	1.318
Molise	23	23	136	643	662	133	1.455	309
Campania	57	57	550	1.433	1.627	2.101	3.868	5.826
Puglia	51	51	258	1.171	1.277	1.620	2.848	4.054
Basilicata	26	26	131	643	683	239	1.507	569
Calabria	43	43	409	1.032	1.093	784	2.441	1.959
Sicilia	54	54	390	1.224	1.498	1.996	3.022	5.040
Sardegna	38	38	377	890	1.021	705	2.051	1.647
Italia	**852**	**852**	**8.003**	**20.954**	**24.851**	**25.517**	**48.875**	**60.288**

(a) Stima Indagine multiscopo "Aspetti della vita quotidiana", dati in migliaia.

[1] Madow, W.G. “On the theory of systematic sampling II”, *Annals of Mathematical Statistics*, 20, (1949): 333-354.

4. La rilevazione e il trattamento dei dati

La rilevazione, di tipo campionario, è condotta con cadenza annuale in genere nel mese di Marzo.

Per l’indagine relativa al 2017 è stata adottata per la prima volta una tecnica mista di rilevazione, CAWI[2]/PAPI[3] sequenziale. Il disegno di indagine prevede in generale che a tutte le famiglie campione sia proposta dapprima l’intervista via web e successivamente, alle famiglie non rispondenti, viene inviato il rilevatore per l’intervista PAPI. Tuttavia, per tenere sotto controllo l’effetto dell’introduzione della tecnica web, per questa prima edizione a tecnica mista è stato previsto un campione di famiglie intervistato direttamente con tecnica PAPI, come nelle precedenti edizioni. A tale scopo per ciascun comune campione le famiglie selezionate sono state casualmente suddivise in due sotto-campioni: uno, pari a circa i due terzi del totale, è stato considerato per la nuova tecnica CAWI/PAPI, l’altro, di dimensione pari a circa un terzo del totale, è stato intervistato con tecnica PAPI. In tal modo sono stati ottenuti due campioni indipendenti, selezionati nei medesimi comuni campione.

L’intervista PAPI prevede l’utilizzo di due questionari cartacei.

Il primo è il questionario base della rilevazione (modello rosa) che 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; da quattro “Schede Individuali”, una per ciascun componente della famiglia e da un “Questionario familiare” che contiene quesiti familiari ai quali risponde un solo componente adulto. Qualora i componenti siano più di quattro sono previste delle schede individuali aggiuntive di colore bianco.

Il secondo è un modello somministrato per autocompilazione (modello verde). Il modello viene consegnato dal rilevatore a ciascun componente della famiglia e contiene quesiti che possono essere agevolmente compilati in autonomia dal rispondente anche senza l’intervento diretto del rilevatore.

Le informazioni vengono fornite direttamente da tutti gli individui di 14 anni e più, mentre i bambini e i ragazzi al di sotto dei 14 anni vengono intervistati in modalità proxy, ciò significa che è un genitore o un componente maggiorenne a fornire le informazioni in loro

[2] Computer assisted web interview.
[3] Paper and pencil interview.

vece. Taluni quesiti della rilevazione, per la sensibilità dell’argomento trattato, prevedono la facoltà di non rispondere.

Trattandosi di un’indagine PAPI, i questionari 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 di frequenze assolute e relative, riferite alle famiglie e agli individui o stime di totali di variabili quantitative. Sono ottenute mediante uno stimatore di ponderazione vincolata. 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 rappresentata dall’unità medesima. Per esempio, se a un’unità campionaria viene attribuito un peso pari a 30, ciò indica che questa unità rappresenta se stessa e altre 29 unità della popolazione non incluse nel campione.

La procedura che consente di costruire i pesi finali da attribuire alle unità campionarie rispondenti, è articolata in generale 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 in opportuni sottoinsiemi di unità e 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;

3) si costruiscono i fattori correttivi che consentono di soddisfare, a livello regionale, la condizione di uguaglianza tra i totali noti di alcune variabili ausiliarie e le corrispondenti stime campionarie;

4) si calcolano, infine, i pesi finali mediante il prodotto dei pesi base per i fattori correttivi ottenuti al passo 3.

I fattori correttivi del passo 3 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). Le variabili ausiliarie considerate a livello regionale sono: tipologia comunale, classi di età, sesso, cittadinanza (italiani/stranieri), titolo di studio (totali stimati dall’indagine su Forze di Lavoro).

Poiché, come descritto nel paragrafo precedente, per la presente edizione le famiglie intervistate appartenevano a uno dei due campioni selezionati per le differenti tecniche di rilevazione, per il calcolo dei pesi finali si è proceduto separatamente per i due campioni per il passo 2, relativo alla correzione per mancata risposta, mentre per il passo 3 di calibrazione i due campioni sono stati messi insieme.

La correzione della mancata risposta del campione PAPI è stata effettuata utilizzando dei correttori calcolati per gruppi omogenei di rispondenti, individuati mediante un modello logistico.

Per il sotto-campione intervistato con tecnica CAWI/PAPI sequenziale, invece, sono stati definiti dei fattori correttivi dell’effetto di autoselezione dei rispondenti alla tecnica web, che permette di aggiustare le distribuzioni campionarie dei rispondenti al papi e al web. Tale correttore è stato applicato ai pesi base prima del passo di calibrazione.

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

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

La stima del totale (1) è data da

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

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[4] 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 24 totali noti, che si riferiscono alla distribuzione della popolazione regionale per sesso e otto classi di età[5], della popolazione regionale nelle sei aree A_1, A_2, B_1, B_2, B_3 e B_4 e della popolazione straniera residente in Italia per regione e sesso. Indicando, quindi, con ${}_kX$ (k=1,…,24) il totale noto della k-esima variabile ausiliaria per la generica regione geografica e con ${}_kX_{hij}$ il

[4] 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} .

[5] Le classi di età considerate sono: 0-5 anni, 6-13 anni, 14-24 anni, 25-34 anni, 35-44 anni, 45-54 anni, 55-64 anni, 65 anni e più.

valore assunto dalla k-esima variabile ausiliaria per la famiglia rispondente h_{ij}, la condizione sopra descritta è espressa dalla seguente uguaglianza

$$ {}_kX = {}_k\hat{X} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} W_{hijk}\, X_{hij} \qquad (k=1,\ldots, 24) $$

in cui H indica il numero complessivo di strati definiti nella regione. Se, ad esempio, ${}_8X$ indica il numero di maschi di età maggiore o uguale a sessantacinque anni, la variabile ausiliaria ${}_8X_{hij}$ 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[6]. 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 in seguito, tale stimatore riveste un ruolo

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

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 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:

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

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

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

Come è stato descritto in precedenza, le stime prodotte dall’indagine sono state ottenute mediante uno stimatore di ponderazione vincolata definito in base 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 $\hat{V}ar(\hat{Y}_d)$ 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 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:

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

dove Z_{hij} è la variabile linearizzata espressa come $Z_{hij} = Y_{hij} - \mathbf{X}'_{\mathbf{hij}}\beta$, essendo $\mathbf{X}_{\mathbf{hij}} = \left({}_1x_{hij},\ldots,{}_k x_{hij},\ldots,{}_K x_{hij}\right)'$ 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:

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

Dalla (6) risulta che la stima della varianza della stima $\hat{Y}_d$ viene calcolata come somma della stima delle varianze dei singoli strati, 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:

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

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:

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

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

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:

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

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, $\hat{V}ar(\hat{Y}_d)$, 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.

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 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.

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 a 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 agli individui dell’Italia Nord occidentale, l’errore relativo corrispondente si ottiene introducendo nella (12) i valori dei parametri a e b riportati nella riga del Nord-ovest del prospetto 2 alla voce Persone.

I prospetti 3 e 4, presentati in aggiunta, consentono di rendere più agevole il calcolo degli errori campionari. Essi riguardano, rispettivamente, le famiglie e gli individui e 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 – Anno 2017

ZONE TERRITORIALI	Famiglie			Persone		
	a	b	R^2 (%)	a	b	R^2 (%)
ITALIA	**8,356772**	**-1,078490**	**97,57**	**9,957049**	**-1,170596**	**92,43**
RIPARTIZIONI GEOGRAFICHE						
Nord	9,658367	-1,159086	91,50	9,658367	-1,159086	91,50
Nord-ovest	*8,017001*	*-1,042261*	*96,69*	*9,545453*	*-1,149127*	*90,34*
Nord-est	*8,162825*	*-1,084437*	*96,88*	*9,198036*	*-1,150291*	*91,95*
Centro	8,389539	-1,093586	97,30	10,182968	-1,211823	92,69
Mezzogiorno	7,531258	-1,034497	96,11	9,061154	-1,129246	92,00
Sud	*7,449241*	*-1,035947*	*95,95*	*8,628870*	*-1,103309*	*91,27*
Isole	*8,167432*	*-1,082732*	*94,98*	*8,688184*	*-1,107906*	*91,69*
TIPI DI COMUNE						
A1	8,594602	-1,105506	98,06	8,725922	-1,076314	93,33
A2	8,180007	-1,064768	95,21	7,754656	-0,997076	89,35
B1	6,986037	-0,998113	92,11	6,581969	-0,934679	88,35
B2	7,463691	-1,017275	95,78	7,610750	-0,998378	90,86
B3	8,121965	-1,071520	96,84	7,715664	-1,010813	92,60
B4	8,377578	-1,104681	97,86	8,023438	-1,041293	92,39
REGIONI						
Piemonte	7,898482	-1,067113	96,99	8,294351	-1,089878	91,03
Valle d'Aosta - Vallée d'Aoste	5,243019	-1,074374	95,19	5,472313	-1,083691	90,36
Liguria	7,718100	-1,112476	96,77	8,427920	-1,162722	92,45
Lombardia	8,642518	-1,076524	96,14	9,450693	-1,127917	89,59
Trentino-Alto Adige	6,465551	-1,061414	96,72	7,042009	-1,086709	90,18
Bolzano - Bozen	6,369200	-1,053692	*93,16*	*6,763472*	*-1,045889*	*88,27*
Trento	6,214706	-1,046990	*97,10*	*7,121805*	*-1,137960*	*90,17*
Veneto	8,525798	-1,111292	97,52	8,533222	-1,091883	92,53
Friuli-Venezia Giulia	7,235386	-1,081847	96,52	7,241255	-1,070124	91,39
Emilia-Romagna	8,415647	-1,104559	95,52	9,022653	-1,142984	90,25
Toscana	8,053898	-1,092329	96,99	8,652680	-1,118053	92,08
Umbria	7,445785	-1,121184	96,47	7,840878	-1,148590	91,45
Marche	7,463609	-1,095089	96,19	7,609594	-1,097040	89,88
Lazio	8,344854	-1,073197	96,65	9,945922	-1,183140	91,72
Abruzzo	7,231073	-1,090050	96,80	7,681266	-1,111322	91,47
Molise	5,736376	-1,055576	93,69	6,174149	-1,090534	87,76
Campania	7,749502	-1,043625	94,21	8,125187	-1,049883	90,00
Puglia	8,139676	-1,093421	95,70	7,900669	-1,041436	90,95
Basilicata	6,847909	-1,122708	94,34	6,978747	-1,113177	90,26
Calabria	7,791513	-1,119153	96,05	7,458694	-1,056463	90,18
Sicilia	8,323389	-1,085262	95,18	8,673745	-1,097135	91,56
Sardegna	6,907287	-1,032254	89,62	7,564611	-1,071919	90,68

(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 – Anno 2017

STIME	Italia	Nord	Nord-ovest	Nord-est	Centro	Mezzogiorno	Sud	Isole	A1	A2	B1	B2	B3	B4
20.000	36,6	32,4	32,9	29,9	32,3	27,7	26,7	28,1	32,6	31,8	25,4	27,8	30,1	31,0
30.000	29,2	26,2	26,7	23,9	25,8	22,5	21,7	22,7	26,6	26,2	21,3	23,1	24,6	25,0
40.000	24,9	22,5	23,0	20,4	22,0	19,4	18,8	19,5	23,0	22,8	18,8	20,2	21,3	21,4
50.000	22,0	20,0	20,5	18,0	19,5	17,3	16,7	17,4	20,6	20,5	17,0	18,3	19,0	19,0
60.000	19,9	18,2	18,6	16,3	17,6	15,8	15,3	15,8	18,8	18,8	15,8	16,8	17,4	17,3
70.000	18,2	16,8	17,2	15,0	16,2	14,6	14,1	14,6	17,4	17,4	14,7	15,6	16,1	15,9
80.000	16,9	15,6	16,1	13,9	15,0	13,6	13,2	13,6	16,3	16,3	13,9	14,7	15,0	14,8
90.000	15,9	14,7	15,1	13,0	14,1	12,8	12,4	12,8	15,3	15,4	13,2	13,9	14,2	13,9
100.000	15,0	13,9	14,3	12,3	13,3	12,2	11,8	12,1	14,5	14,7	12,6	13,3	13,5	13,2
200.000	10,2	9,7	10,0	8,4	9,1	8,5	8,3	8,4	10,2	10,5	9,3	9,7	9,5	9,1
300.000	8,1	7,8	8,1	6,7	7,3	6,9	6,7	6,8	8,4	8,6	7,8	8,0	7,8	7,3
400.000	6,9	6,7	7,0	5,7	6,2	6,0	5,8	5,8	7,2	7,5	6,9	7,0	6,7	6,3
500.000	6,1	6,0	6,2	5,0	5,5	5,3	5,2	5,2	6,5	6,8	6,3	6,3	6,0	5,6
750.000	4,9	4,8	5,1	4,0	4,4	4,3	4,2	4,2	5,3	5,6	5,3	5,3	4,9	4,5
1.000.000	4,2	4,1	4,4	3,4	3,7	3,7	3,7	3,6	4,6	4,8	4,7	4,6	4,3	3,9
2.000.000	2,8	2,9	3,1	2,3	2,6	2,6	2,6	2,5	3,2	3,5	3,4	3,3	3,0	2,7
3.000.000	2,3	2,3	2,5	1,9	2,0	2,1	2,1	2,0	2,6	2,9	2,9	2,8	2,5	2,2
4.000.000	1,9	2,0	2,1	1,6	1,7	1,8	1,8	1,7	2,3	2,5	2,6	2,4	2,1	1,9
5.000.000	1,7	1,8	1,9	1,4	1,5	1,6	1,6	1,5	2,0	2,2	2,3	2,2	1,9	1,6
7.500.000	1,4	1,4	1,5	1,1	1,2	1,3	1,3	1,2	1,7	1,8	1,9	1,8	1,6	1,3
10.000.000	1,2	1,2	1,3	1,0	1,1	1,1	1,1	1,1	1,4	1,6	1,7	1,6	1,3	1,1
15.000.000	0,9	1,0	1,1	0,8	0,8	0,9	0,9	0,9	1,2	1,3	1,4	1,3	1,1	0,9
20.000.000	0,8	0,9	0,9	0,7	0,7	0,8	0,8	0,7	1,0	1,1	1,3	1,2	1,0	0,8
25.000.000	0,7	0,8	0,8	0,6	0,6	0,7	0,7	0,7	0,9	1,0	1,2	1,0	0,9	0,7

STIME	Piemonte	Valle d'Aosta	Liguria	Lombardia	Trentino-Alto Adige	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Emilia-Romagna	Toscana	Umbria
20.000	27,3	7,7	21,0	38,7	14,9	16,3	12,8	30,5	18,3	29,5	27,8	16,5
30.000	22,2	6,2	16,7	31,2	12,2	13,4	10,3	24,4	14,7	23,7	22,3	13,2
40.000	19,2	5,4	14,3	26,7	10,5	11,7	8,8	20,8	12,6	20,3	19,0	11,2
50.000	17,1	4,8	12,6	23,7	9,4	10,5	7,8	18,4	11,2	18,0	16,8	9,9
60.000	15,6	4,4	11,4	21,5	8,6	9,6	7,1	16,7	10,2	16,4	15,3	9,0
70.000	14,4	4,1	10,4	19,8	7,9	8,9	6,5	15,3	9,3	15,1	14,0	8,2
80.000	13,4	3,8	9,7	18,5	7,4	8,3	6,0	14,2	8,7	14,0	13,0	7,6
90.000	12,7	3,6	9,0	17,3	7,0	7,9	5,7	13,3	8,2	13,2	12,2	7,2
100.000	12,0	3,4	8,5	16,4	6,6	7,5	5,4	12,6	7,7	12,4	11,5	6,7
200.000	8,4	2,4	5,8	11,3	4,7	5,4	3,7	8,6	5,3	8,6	7,9	4,6
300.000	6,8	1,9	4,6	9,1	3,8	4,4	3,0	6,9	4,3	6,9	6,3	3,7
400.000	5,9	1,7	3,9	7,8	3,3	3,8	2,5	5,9	3,7	5,9	5,4	3,1
500.000	5,3	1,5	3,5	6,9	2,9	3,4	2,2	5,2	3,2	5,3	4,8	2,8
750.000	4,3	1,2	2,8	5,6	2,4	2,8	1,8	4,1	2,6	4,2	3,8	2,2
1.000.000	3,7	1,0	2,3	4,8	2,1	2,5	1,5	3,5	2,2	3,6	3,3	1,9
2.000.000	2,6	0,7	1,6	3,3	1,5	1,8	1,1	2,4	1,5	2,5	2,3	1,3

STIME	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
20.000	19,3	33,6	18,1	10,1	29,1	28,0	12,8	20,7	29,5	20,3
30.000	15,6	27,1	14,6	8,2	23,9	22,6	10,4	16,9	24,1	16,5
40.000	13,4	23,3	12,6	7,1	20,8	19,5	8,9	14,6	20,9	14,3
50.000	11,9	20,7	11,2	6,3	18,7	17,4	7,9	13,1	18,6	12,7
60.000	10,8	18,8	10,2	5,8	17,1	15,8	7,2	11,9	17,0	11,6
70.000	9,9	17,4	9,4	5,3	15,9	14,6	6,7	11,0	15,7	10,7
80.000	9,3	16,2	8,8	5,0	14,9	13,6	6,2	10,3	14,7	10,0
90.000	8,7	15,2	8,3	4,7	14,1	12,8	5,9	9,7	13,9	9,4
100.000	8,2	14,4	7,8	4,4	13,4	12,1	5,5	9,2	13,2	8,9
200.000	5,7	10,0	5,5	3,1	9,6	8,4	3,9	6,5	9,3	6,3
300.000	4,6	8,0	4,4	2,5	7,9	6,8	3,1	5,3	7,6	5,1
400.000	4,0	6,9	3,8	2,2	6,9	5,9	2,7	4,6	6,6	4,4
500.000	3,5	6,1	3,4	1,9	6,2	5,2	2,4	4,1	5,9	3,9
750.000	2,8	5,0	2,7	1,6	5,1	4,2	1,9	3,4	4,8	3,2
1.000.000	2,4	4,3	2,4	1,4	4,4	3,7	1,7	2,9	4,1	2,8
2.000.000	1,7	3,0	1,6	1,0	3,2	2,5	1,2	2,1	2,9	1,9

Prospetto 4 – Valori interpolati degli errori campionari relativi percentuali delle stime riferite alle persone per totale Italia, ripartizione geografica, tipo di comune e regione – Anno 2017

STIME	Italia	Nord	Nord-ovest	Nord-est	Centro	Mezzogiorno	Sud	Isole	A1	A2	B1	B2	B3	B4
20.000	44,1	40,2	40,0	33,4	40,3	34,6	31,7	31,9	38,0	34,6	26,3	32,0	31,7	31,8
30.000	34,8	31,8	31,7	26,4	31,5	27,5	25,3	25,5	30,6	28,3	21,7	26,2	25,9	25,8
40.000	29,4	26,9	26,8	22,4	26,5	23,4	21,6	21,7	26,2	24,5	19,0	22,7	22,4	22,2
50.000	25,8	23,7	23,6	19,7	23,1	20,6	19,1	19,2	23,2	21,9	17,1	20,3	20,0	19,8
60.000	23,2	21,3	21,3	17,7	20,7	18,6	17,3	17,4	21,1	20,0	15,7	18,5	18,2	18,0
70.000	21,2	19,5	19,5	16,2	18,9	17,1	15,9	15,9	19,4	18,6	14,6	17,1	16,9	16,6
80.000	19,6	18,0	18,0	15,0	17,4	15,8	14,8	14,8	18,0	17,4	13,7	16,0	15,8	15,5
90.000	18,3	16,8	16,8	14,1	16,2	14,8	13,8	13,9	16,9	16,4	13,0	15,1	14,8	14,5
100.000	17,2	15,8	15,8	13,2	15,2	13,9	13,0	13,1	16,0	15,5	12,4	14,3	14,1	13,8
200.000	11,5	10,6	10,6	8,9	10,0	9,4	8,9	8,9	11,0	11,0	9,0	10,1	9,9	9,6
300.000	9,0	8,4	8,4	7,0	7,8	7,5	7,1	7,1	8,9	9,0	7,4	8,3	8,1	7,8
400.000	7,6	7,1	7,1	6,0	6,6	6,4	6,1	6,1	7,6	7,8	6,5	7,2	7,0	6,7
500.000	6,7	6,2	6,3	5,2	5,7	5,6	5,4	5,4	6,7	7,0	5,8	6,4	6,2	6,0
750.000	5,3	4,9	5,0	4,2	4,5	4,5	4,3	4,3	5,4	5,7	4,8	5,2	5,1	4,8
1.000.000	4,5	4,2	4,2	3,5	3,8	3,8	3,7	3,7	4,6	4,9	4,2	4,5	4,4	4,2
2.000.000	3,0	2,8	2,8	2,4	2,5	2,6	2,5	2,5	3,2	3,5	3,1	3,2	3,1	2,9
3.000.000	2,4	2,2	2,2	1,9	1,9	2,0	2,0	2,0	2,6	2,8	2,5	2,6	2,5	2,3
4.000.000	2,0	1,9	1,9	1,6	1,6	1,7	1,7	1,7	2,2	2,5	2,2	2,3	2,2	2,0
5.000.000	1,7	1,6	1,7	1,4	1,4	1,5	1,5	1,5	1,9	2,2	2,0	2,0	1,9	1,8
7.500.000	1,4	1,3	1,3	1,1	1,1	1,2	1,2	1,2	1,6	1,8	1,6	1,7	1,6	1,5
10.000.000	1,2	1,1	1,1	0,9	0,9	1,0	1,0	1,0	1,3	1,6	1,4	1,4	1,4	1,3
15.000.000	0,9	0,9	0,9	0,7	0,7	0,8	0,8	0,8	1,1	1,3	1,2	1,2	1,1	1,0
20.000.000	0,8	0,7	0,8	0,6	0,6	0,7	0,7	0,7	0,9	1,1	1,0	1,0	1,0	0,9
25.000.000	0,7	0,6	0,7	0,6	0,5	0,6	0,6	0,6	0,8	1,0	0,9	0,9	0,9	0,8

STIME	Piemonte	Valle d'Aosta	Liguria	Lombardia	Trentino-Alto Adige	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Emilia-Romagna	Toscana	Umbria
20.000	28,7	7,2	21,4	42,3	15,6	16,6	12,6	32,0	18,7	31,7	29,8	17,1
30.000	23,0	5,8	16,9	33,7	12,5	13,4	10,0	25,6	15,0	25,2	23,8	13,5
40.000	19,6	5,0	14,3	28,6	10,7	11,5	8,5	21,9	12,9	21,3	20,2	11,5
50.000	17,4	4,4	12,5	25,2	9,5	10,3	7,5	19,4	11,4	18,8	17,9	10,1
60.000	15,8	4,0	11,3	22,8	8,6	9,3	6,7	17,6	10,4	16,9	16,1	9,1
70.000	14,5	3,7	10,3	20,9	7,9	8,6	6,2	16,1	9,5	15,5	14,8	8,3
80.000	13,5	3,4	9,5	19,4	7,3	8,0	5,7	15,0	8,9	14,4	13,7	7,7
90.000	12,6	3,2	8,9	18,1	6,9	7,5	5,3	14,1	8,3	13,4	12,9	7,2
100.000	11,9	3,0	8,4	17,1	6,5	7,1	5,0	13,3	7,9	12,6	12,1	6,8
200.000	8,2	2,1	5,6	11,6	4,5	5,0	3,4	9,1	5,4	8,5	8,2	4,6
300.000	6,6	1,7	4,4	9,2	3,6	4,0	2,7	7,3	4,4	6,7	6,6	3,6
400.000	5,6	1,4	3,7	7,8	3,1	3,5	2,3	6,2	3,8	5,7	5,6	3,1
500.000	5,0	1,3	3,3	6,9	2,7	3,1	2,0	5,5	3,3	5,0	4,9	2,7
750.000	4,0	1,0	2,6	5,5	2,2	2,5	1,6	4,4	2,7	4,0	3,9	2,1
1.000.000	3,4	0,9	2,2	4,7	1,9	2,1	1,4	3,8	2,3	3,4	3,3	1,8
2.000.000	2,3	0,6	1,5	3,2	1,3	1,5	0,9	2,6	1,6	2,3	2,3	1,2
3.000.000	1,9	0,5	1,2	2,5	1,0	1,2	0,7	2,1	1,3	1,8	1,8	1,0
4.000.000	1,6	0,4	1,0	2,1	0,9	1,0	0,6	1,8	1,1	1,5	1,5	0,8
5.000.000	1,4	0,4	0,9	1,9	0,8	0,9	0,5	1,6	1,0	1,4	1,4	0,7

STIME	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
20.000	19,6	41,2	19,0	9,9	32,1	29,9	13,2	22,3	33,4	21,7
30.000	15,7	32,4	15,1	7,9	25,9	24,2	10,6	18,0	26,8	17,5
40.000	13,4	27,4	12,9	6,8	22,3	20,9	9,0	15,4	22,9	15,0
50.000	11,9	24,0	11,4	6,0	19,8	18,6	7,9	13,7	20,2	13,3
60.000	10,8	21,5	10,3	5,4	18,0	16,9	7,2	12,5	18,3	12,1
70.000	9,9	19,7	9,5	5,0	16,6	15,6	6,6	11,5	16,8	11,1
80.000	9,2	18,2	8,8	4,6	15,5	14,5	6,1	10,7	15,6	10,3
90.000	8,6	16,9	8,2	4,4	14,6	13,7	5,7	10,1	14,6	9,7
100.000	8,1	15,9	7,8	4,1	13,8	12,9	5,4	9,5	13,8	9,2
200.000	5,6	10,6	5,3	2,8	9,6	9,0	3,7	6,6	9,5	6,3
300.000	4,4	8,3	4,2	2,3	7,7	7,3	2,9	5,3	7,6	5,1
400.000	3,8	7,0	3,6	1,9	6,7	6,3	2,5	4,6	6,5	4,4
500.000	3,4	6,1	3,2	1,7	5,9	5,6	2,2	4,1	5,7	3,9
750.000	2,7	4,8	2,5	1,4	4,8	4,5	1,8	3,3	4,6	3,1
1.000.000	2,3	4,1	2,2	1,2	4,1	3,9	1,5	2,8	3,9	2,7
2.000.000	1,6	2,7	1,5	0,8	2,9	2,7	1,0	2,0	2,7	1,8
3.000.000	1,3	2,1	1,2	0,6	2,3	2,2	0,8	1,6	2,1	1,5
4.000.000	1,1	1,8	1,0	0,6	2,0	1,9	0,7	1,4	1,8	1,3
5.000.000	1,0	1,6	0,9	0,5	1,8	1,7	0,6	1,2	1,6	1,1

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 in Sicilia che dichiarano “adeguate” le proprie risorse economiche
Stima puntuale:	874.000
Errore relativo (CV)	4,1/100=0,041
Stima intervallare	
Semi ampiezza dell’intervallo	(874.000*0, 041)*1,96 =70.235
Limite inferiore dell’intervallo di confidenza	874.000-70.235=803.765
Limite superiore dell’intervallo di confidenza	874.000+70.235=860.739

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

	Persone di 14 anni e più che fumano in Lombardia
Stima puntuale:	1.788.000
Errore relativo (CV)	3,2/100=0,032
Stima intervallare	
Semi ampiezza dell’intervallo	(1.788.000*0,032)*1,96 = 112.143
Limite inferiore dell’intervallo di confidenza	1.788.000- 112.143= 1.675.857
Limite superiore dell’intervallo di confidenza	1.788.000+ 112.143= 1.900.143

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

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

utilizzando i parametri riportati nel prospetto 2.

7. La diffusione dei risultati dell’indagine

I principali risultati dell’indagine vengono resi disponibili sul sito dell’Istat attraverso statistiche report pubblicate nei settori con argomento: “Opinioni dei cittadini”, “Salute e sanità”, “Cultura, comunicazione, tempo libero”, “Partecipazione sociale”. I risultati sono diffusi sempre sul sito anche attraverso il datawarehouse I.Stat.

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.

Ogni anno, inoltre, i dati raccolti vengono analizzati e pubblicati anche su volumi a carattere generale (Rapporto annuale, Annuario statistico italiano, Noi Italia, Italia in cifre). I volumi curati dall’Istat sono consultabili nel Catalogo editoriale.

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

Sante Orsini
Tel. +39 06 4673.7256
Email orsini@istat.it