SN167

Multiscopo ISTAT - Aspetti della vita quotidiana (2015)

Istat

Versione: 1.0 - 31/01/2017

UniData
Bicocca Data Archive

Website: www.unidata.unimib.it
E-mail: unidata@unimib.it
Tel.: +39 02 6448 7513
Fax: +39 02 6448 7561

Università degli Studi di Milano-Bicocca
Via Bicocca degli Arcimboldi 8
20126 - Milano (Italia)

INDICE
TABLE OF CONTENTS

Note metodologiche
Methodological Notes

ASPETTI METODOLOGICI DELL'INDAGINE

SURVEY'S METHODOLOGICAL ISSUES

Istat
Istituto Nazionale di Statistica

Aspetti della vita quotidiana

Periodo di riferimento: anno 2015

Aspetti metodologici dell'indagine

Anno 2017

INDICE

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; ${}_rf$ la frazione di campionamento;

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

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

La selezione delle famiglie da intervistare in ogni comune campione viene effettuata dalla lista anagrafica di ciascun comune senza reimmissione e con probabilità uguali.

In particolare, la tecnica di selezione è di tipo sistematico e, nell’ambito di ogni comune viene attuata attraverso le seguenti fasi:

- vengono 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 2015

REGIONI	Comuni			Famiglie			Individui	
	Campione effettivo	Campione teorico	Universo	Campione effettivo	Campione teorico	Universo (a)	Campione effettivo	Universo (a)
Piemonte	62	63	1206	1363	1876	1946	3.023	4.391
Valle d'Aosta - Vallée d'Aoste	22	22	74	495	631	61	1076	127
Liguria	26	26	235	877	1105	768	1803	1573
Lombardia	85	85	1544	1669	2.268	4.196	3.963	9.946
Trentino-Alto Adige	48	48	333	1099	1441	448	2.523	1045
Veneto	55	55	581	1113	1382	2.014	2.665	4.888
Friuli-Venezia Giulia	31	33	218	729	970	544	1596	1217
Emilia-Romagna	47	47	348	1071	1384	1957	2.414	4.425
Toscana	51	51	287	1124	1474	1609	2.575	3.736
Umbria	22	22	92	562	710	375	1329	890
Marche	36	37	239	739	964	637	1795	1544
Lazio	28	34	378	1080	1796	2.614	2.375	5.864
Abruzzo	35	36	305	764	989	552	1845	1328
Molise	23	23	136	576	665	129	1377	312
Campania	55	56	551	1361	1632	2.128	3.717	5.848
Puglia	51	51	258	1069	1276	1576	2.692	4.080
Basilicata	26	27	131	588	684	236	1421	575
Calabria	42	42	409	927	1092	797	2.299	1971
Sicilia	52	52	390	1165	1493	2.003	2.908	5.076
Sardegna	39	39	377	789	1023	712	1830	1657
Italia	**836**	**849**	**8.092**	**19.160**	**24.855**	**25.302**	**45.226**	**60.494**

(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.

L’intervista alla famiglia viene effettuata mediante tecnica Papi (Paper and Pencil interview) e 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 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 essenzialmente stime di frequenze assolute e relative, riferite alle famiglie e agli individui.

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

Il principio su cui è basato ogni metodo di stima campionaria è che le unità appartenenti al campione rappresentino anche le unità della popolazione che non sono incluse nel campione.

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

Al fine di rendere più chiara la successiva esposizione, introduciamo la seguente simbologia: *d*, indice di livello territoriale di riferimento delle stime; *i*, indice di comune; *j*, indice di famiglia; *p*, indice di componente della famiglia; *h*, indice di strato di comuni; *y*, generica variabile oggetto di indagine; Y_{hijp}, valore di *y* osservato sul componente *p* della famiglia *j* del comune *i* dello strato *h*; P_{hij} , numero di componenti della famiglia *j* del comune *i* dello strato *h*; $Y_{hij} = \sum_{p=1}^{P_{hij}} Y_{hijp}$, totale della variabile *y* osservato sulla famiglia *j* del comune *i* dello strato *h*; M_{hi}, numero di famiglie residenti nel comune *i* dello strato *h*; m_{hi}, campione di famiglie nel comune *i* dello strato *h*; N_h, totale di comuni nello strato *h*; n_h, numero di comuni campione nello strato *h* (nell’indagine in oggetto si ha n_h,=1); H_d, numero totale di strati nel generico dominio territoriale *d*.

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

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

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

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

- corregge l’effetto distorsivo della mancata risposta totale dovuta all’impossibilità di intervistare alcune delle famiglie selezionate per irreperibilità o per rifiuto all’intervista;

- tiene conto della conoscenza di totali noti di importanti variabili ausiliarie (disponibili da fonti esterne all’indagine), nel senso che le stime campionarie dei totali noti delle variabili ausiliarie devono coincidere con i valori noti degli stessi.

Nell’indagine in oggetto vengono definiti per ciascuna regione geografica 24 totali noti, che si riferiscono alla distribuzione della popolazione regionale per sesso e otto classi di età[3], 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 valore assunto dalla k-esima variabile ausiliaria per la famiglia rispondente h_{ij}, la condizione sopra descritta è espressa dalla seguente uguaglianza

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

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

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

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[4]. Un importante stimatore appartenente a tale classe, che si ottiene utilizzando la funzione di distanza euclidea, è lo stimatore di regressione generalizzata. Come verrà chiarito meglio in seguito, 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.

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

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 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:

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

dove Z_{hij} è la variabile linearizzata espressa come $Z_{hij} = Y_{hij} - \mathbf{X}_{\mathbf{hij}}^{'} \beta$, essendo $\mathbf{X}_{\mathbf{hij}} = \left({}_1 X_{hij}, \ldots, {}_k X_{hij}, \ldots, {}_K X_{hij}\right)'$ 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

$$\hat{\mathrm{Var}}\left(\hat{Y}_d\right) \cong \hat{\mathrm{Var}}\left(\hat{Z}_d\right) = \sum_{h=1}^{H_d} \hat{\mathrm{Var}}\left(\hat{Z}_h\right) . \qquad (6)$$

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, $\hat{\mathrm{Var}}\left(\hat{Z}_h\right)$, della stima $\hat{Z}_h$ è differente a seconda che lo strato sia Ar oppure Nar. Possiamo, quindi scomporre come segue

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

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:

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

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

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, $\hat{V}ar\left(\hat{Y}_d\right)$, 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:

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

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 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 2015

ZONE TERRITORIALI	Famiglie			Persone		
	a	b	R^2 (%)	a	b	R^2 (%)
ITALIA	**8,655579**	**-1,093492**	**97,36**	**10,020307**	**-1,189554**	**92,66**
RIPARTIZIONI GEOGRAFICHE						
Nord	8,495263	-1,080198	97,44	10,153908	-1,206835	91,52
Nord-ovest	8,221561	-1,053673	*96,45*	10,044979	-1,199039	*91,48*
Nord-est	8,330302	-1,087440	*97,53*	9,677603	-1,198789	*90,43*
Centro	8,672146	-1,103908	96,98	9,560905	-1,173956	91,99
Mezzogiorno	7,965907	-1,062738	96,16	9,111221	-1,148761	92,86
Sud	7,452958	-1,030222	*95,55*	8,639670	-1,121657	*91,56*
Isole	7,938987	-1,055631	*94,00*	9,314472	-1,172551	*93,57*
TIPI DI COMUNE						
A1	8,680690	-1,100128	97,58	10,266874	-1,233155	92,42
A2	8,278601	-1,058192	94,23	9,884428	-1,191813	92,20
B1	7,119472	-1,020255	93,12	8,081462	-1,103910	85,56
B2	7,630177	-1,027756	96,50	9,265044	-1,156162	91,66
B3	8,195113	-1,069978	95,90	9,357765	-1,158958	90,92
B4	8,405417	-1,102277	97,82	9,494194	-1,192805	93,39
REGIONI						
Piemonte	8,013730	-1,079845	96,94	8,934823	-1,165122	91,66
Valle d'Aosta - Vallée d'Aoste	5,361123	-1,077118	94,00	5,887885	-1,153755	88,50
Liguria	7,464424	-1,081127	97,28	8,356558	-1,171984	92,49
Lombardia	8,944066	-1,094235	96,76	10,204823	-1,201036	91,12
Trentino-Alto Adige	6,854834	-1,100407	96,57	7,730369	-1,186980	90,65
Bolzano - Bozen	6,638634	-1,094874	*96,20*	7,446460	-1,180916	*88,78*
Trento	7,199817	-1,140462	*94,41*	7,598252	-1,184392	*89,47*
Veneto	8,474442	-1,095379	95,94	9,501087	-1,187608	89,69
Friuli-Venezia Giulia	7,052955	-1,061255	95,17	8,121311	-1,175742	87,10
Emilia-Romagna	8,281996	-1,079636	96,59	9,376074	-1,179854	89,61
Toscana	8,079933	-1,084564	96,63	8,786904	-1,146888	90,71
Umbria	7,407250	-1,119196	95,63	7,832229	-1,160141	91,86
Marche	7,807456	-1,118385	95,72	8,507041	-1,180621	90,40
Lazio	8,768358	-1,095200	95,91	9,632030	-1,166341	90,57
Abruzzo	7,718308	-1,130986	96,86	8,494719	-1,203253	92,09
Molise	5,965320	-1,079395	94,91	6,781180	-1,171764	86,87
Campania	8,186600	-1,079584	93,84	8,710198	-1,118931	90,94
Puglia	8,095253	-1,082001	96,50	9,021551	-1,161539	90,95
Basilicata	6,471730	-1,070246	87,75	6,927846	-1,115451	87,60
Calabria	7,081949	-1,030211	95,60	7,791949	-1,094227	88,72
Sicilia	8,070604	-1,059337	91,67	9,663470	-1,197675	92,93
Sardegna	7,139254	-1,031923	93,10	7,379490	-1,061428	91,16

(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 – Anno 2015

STIME	Italia	Nord	Nord-ovest	Nord-est	Centro	Mezzogiorno	Sud	Isole	A1	A2	B1	B2	B3	B4
20.000	33,7	33,2	33,1	29,5	32,3	27,8	25,3	28,4	33,0	33,3	22,5	28,0	30,1	28,5
30.000	27,0	26,7	26,7	23,7	25,8	22,4	20,5	23,0	26,4	26,8	18,3	22,7	24,2	22,8
40.000	23,1	22,9	22,9	20,3	22,0	19,2	17,7	19,7	22,6	23,1	15,8	19,6	20,8	19,4
50.000	20,4	20,3	20,4	17,9	19,5	17,1	15,8	17,5	20,0	20,5	14,1	17,5	18,4	17,2
60.000	18,5	18,4	18,5	16,3	17,6	15,5	14,4	15,9	18,1	18,6	12,8	15,9	16,7	15,6
70.000	17,0	16,9	17,1	14,9	16,2	14,3	13,3	14,7	16,6	17,1	11,9	14,7	15,4	14,3
80.000	15,8	15,7	15,9	13,9	15,0	13,3	12,4	13,7	15,4	16,0	11,1	13,7	14,3	13,3
90.000	14,8	14,8	15,0	13,0	14,1	12,5	11,7	12,9	14,4	15,0	10,4	12,9	13,5	12,4
100.000	14,0	13,9	14,2	12,3	13,3	11,8	11,0	12,2	13,6	14,2	9,9	12,2	12,7	11,7
200.000	9,6	9,6	9,8	8,4	9,1	8,2	7,7	8,4	9,3	9,8	6,9	8,6	8,8	8,0
300.000	7,7	7,7	7,9	6,8	7,2	6,6	6,3	6,8	7,5	7,9	5,6	7,0	7,1	6,4
400.000	6,6	6,6	6,8	5,8	6,2	5,7	5,4	5,8	6,4	6,8	4,9	6,0	6,1	5,5
500.000	5,8	5,8	6,1	5,1	5,5	5,0	4,8	5,2	5,6	6,1	4,4	5,3	5,4	4,8
750.000	4,6	4,7	4,9	4,1	4,4	4,1	3,9	4,2	4,5	4,9	3,5	4,3	4,3	3,9
1000.000	4,0	4,0	4,2	3,5	3,7	3,5	3,4	3,6	3,8	4,2	3,1	3,7	3,7	3,3
2.000.000	2,7	2,8	2,9	2,4	2,5	2,4	2,4	2,5	2,6	2,9	2,1	2,6	2,6	2,3
3.000.000	2,2	2,2	2,4	1,9	2,0	1,9	1,9	2,0	2,1	2,3	-	2,1	2,1	1,8
4.000.000	1,9	1,9	2,0	1,7	1,7	1,7	1,7	-	1,8	-	-	1,8	1,8	1,5
5.000.000	1,6	1,7	1,8	1,5	1,5	1,5	1,5	-	-	-	-	1,6	1,6	1,4
7.500.000	1,3	1,4	1,5	-	-	1,2	1,2	-	-	-	-	1,3	1,3	-
10.000.000	1,1	1,2	-	-	-	-	-	-	-	-	-	-	-	-
15.000.000	0,9	0,9	-	-	-	-	-	-	-	-	-	-	-	-
20.000.000	0,8	-	-	-	-	-	-	-	-	-	-	-	-	-
25.000.000	0,7	-	-	-	-	-	-	-	-	-	-	-	-	-

STIME	Piemonte	Valle d'Aosta - Vallée d'Aoste	Liguria	Lombardia	Trentino-Alto Adige	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Emilia-Romagna	Toscana	Umbria
20.000	26,2	7,0	19,8	38,8	13,2	12,2	12,9	30,5	17,8	30,0	26,4	15,9
30.000	21,0	5,7	15,9	31,1	10,6	9,8	10,2	24,4	14,3	24,1	21,2	12,7
40.000	18,0	4,8	13,6	26,6	9,0	8,4	8,7	20,9	12,3	20,6	18,2	10,8
50.000	16,0	4,3	12,0	23,5	8,0	7,4	7,7	18,5	10,9	18,3	16,1	9,5
60.000	14,5	3,9	10,9	21,3	7,2	6,7	6,9	16,7	9,9	16,6	14,6	8,6
70.000	13,3	-	10,0	19,6	6,6	6,2	6,3	15,4	9,1	15,2	13,4	7,9
80.000	12,4	-	9,3	18,2	6,2	5,7	5,9	14,3	8,5	14,2	12,5	7,3
90.000	11,6	-	8,8	17,0	5,8	5,4	5,5	13,4	8,0	13,3	11,7	6,9
100.000	11,0	-	8,3	16,1	5,5	5,1	5,2	12,6	7,6	12,6	11,0	6,5
200.000	7,6	-	5,7	11,0	3,7	3,5	3,5	8,6	5,2	8,6	7,6	4,4
300.000	6,1	-	4,6	8,8	3,0	-	-	6,9	4,2	6,9	6,1	3,5
400.000	5,2	-	3,9	7,5	2,5	-	-	5,9	3,6	5,9	5,2	3,0
500.000	4,6	-	3,5	6,7	-	-	-	5,2	3,2	5,3	4,6	-
750.000	3,7	-	2,8	5,3	-	-	-	4,2	2,6	4,2	3,7	-
1000.000	3,2	-	2,4	4,6	-	-	-	3,6	-	3,6	3,2	-
2.000.000	2,2	-	-	3,1	-	-	-	2,5	-	2,5	2,2	-

STIME	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
20.000	19,5	35,4	17,5	9,4	28,6	27,0	12,7	21,0	29,8	21,4
30.000	15,6	28,3	13,9	7,6	23,0	21,7	10,2	17,0	24,1	17,4
40.000	13,2	24,2	11,8	6,5	19,7	18,5	8,8	14,7	20,7	15,0
50.000	11,7	21,4	10,4	5,7	17,4	16,4	7,8	13,1	18,3	13,4
60.000	10,6	19,4	9,4	5,2	15,8	14,9	7,1	11,9	16,7	12,2
70.000	9,7	17,8	8,6	4,8	14,5	13,7	6,5	11,0	15,4	11,2
80.000	9,0	16,6	8,0	4,5	13,5	12,7	6,0	10,3	14,3	10,5
90.000	8,4	15,5	7,5	4,2	12,7	12,0	5,7	9,7	13,4	9,9
100.000	7,9	14,7	7,1	4,0	12,0	11,3	5,4	9,2	12,7	9,3
200.000	5,4	10,0	4,8	2,7	8,2	7,8	3,7	6,4	8,8	6,5
300.000	4,3	8,0	3,8	-	6,6	6,2	3,0	5,2	7,1	5,3
400.000	3,7	6,9	3,2	-	5,7	5,3	-	4,5	6,1	4,6
500.000	3,2	6,1	2,8	-	5,0	4,7	-	4,0	5,4	4,1
750.000	2,6	4,9	2,3	-	4,0	3,8	-	3,2	4,4	3,3
1000.000	-	4,2	-	-	3,5	3,2	-	-	3,8	-
2.000.000	-	2,8	-	-	2,4	-	-	-	2,6	-

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 2015

STIME	Italia	Nord	Nord-ovest	Nord-est	Centro	Mezzogiorno	Sud	Isole	A1	A2	B1	B2	B3	B4
20.000	41,5	40,7	40,1	33,4	35,6	32,2	29,1	31,7	37,8	38,3	24,0	33,5	34,6	31,4
30.000	32,6	31,9	31,4	26,2	28,1	25,5	23,2	25,0	29,4	30,1	19,2	26,5	27,4	24,6
40.000	27,5	26,8	26,4	22,0	23,7	21,6	19,7	21,1	24,7	25,4	16,4	22,5	23,2	20,7
50.000	24,0	23,4	23,1	19,3	20,8	19,0	17,4	18,5	21,5	22,2	14,5	19,7	20,4	18,2
60.000	21,6	21,0	20,7	17,3	18,7	17,1	15,7	16,6	19,2	19,9	13,1	17,8	18,3	16,3
70.000	19,7	19,1	18,9	15,8	17,1	15,7	14,4	15,2	17,5	18,2	12,0	16,3	16,8	14,9
80.000	18,2	17,6	17,4	14,5	15,8	14,5	13,4	14,1	16,1	16,8	11,2	15,0	15,5	13,7
90.000	17,0	16,4	16,3	13,5	14,7	13,6	12,5	13,1	15,0	15,6	10,5	14,1	14,5	12,8
100.000	15,9	15,4	15,3	12,7	13,8	12,8	11,8	12,3	14,0	14,7	9,9	13,2	13,6	12,0
200.000	10,5	10,1	10,1	8,4	9,2	8,6	8,0	8,2	9,1	9,7	6,7	8,9	9,1	7,9
300.000	8,3	7,9	7,9	6,6	7,3	6,8	6,4	6,5	7,1	7,6	5,4	7,0	7,2	6,2
400.000	7,0	6,7	6,6	5,5	6,1	5,8	5,4	5,5	6,0	6,4	4,6	5,9	6,1	5,3
500.000	6,1	5,8	5,8	4,8	5,4	5,1	4,8	4,8	5,2	5,6	4,1	5,2	5,4	4,6
750.000	4,8	4,6	4,6	3,8	4,2	4,0	3,8	3,8	4,0	4,4	3,3	4,1	4,2	3,6
1.000.000	4,0	3,8	3,8	3,2	3,6	3,4	3,2	3,2	3,4	3,7	2,8	3,5	3,6	3,0
2.000.000	2,7	2,5	2,5	2,1	2,4	2,3	2,2	2,1	2,2	2,5	1,9	2,3	2,4	2,0
3.000.000	2,1	2,0	2,0	1,7	1,9	1,8	1,8	1,7	1,7	1,9	1,5	1,9	1,9	1,6
4.000.000	1,8	1,7	1,7	1,4	1,6	1,5	1,5	1,4	1,4	1,6	1,3	1,6	1,6	1,3
5.000.000	1,6	1,5	1,5	1,2	1,4	1,4	1,3	1,2	1,3	1,4	-	1,4	1,4	1,2
7.500.000	1,2	1,1	1,1	1,0	1,1	1,1	1,0	1,0	1,0	1,1	-	1,1	1,1	0,9
10.000.000	1,0	1,0	1,0	0,8	0,9	0,9	0,9	-	0,8	-	-	0,9	0,9	0,8
15.000.000	0,8	0,7	0,8	0,6	0,7	0,7	0,7	-	-	-	-	0,7	0,7	0,6
20.000.000	0,7	0,6	0,6	-	-	0,6	-	-	-	-	-	-	-	-
25.000.000	0,6	0,6	-	-	-	0,5	-	-	-	-	-	-	-	-

STIME	Piemonte	Valle d'Aosta - Vallée d'Aoste	Liguria	Lombardia	Trentino - Alto Adige	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Emilia-Romagna	Toscana	Umbria
20.000	27,2	6,3	19,7	43,0	13,4	12,0	12,7	32,3	17,2	31,5	27,6	16,1
30.000	21,5	5,0	15,5	33,7	10,5	9,4	10,0	25,4	13,5	24,8	21,9	12,7
40.000	18,2	4,2	13,1	28,3	8,9	7,9	8,4	21,4	11,4	20,9	18,6	10,7
50.000	15,9	3,7	11,5	24,8	7,8	7,0	7,4	18,7	10,0	18,4	16,3	9,4
60.000	14,3	3,3	10,3	22,2	7,0	6,2	6,6	16,8	9,0	16,5	14,7	8,5
70.000	13,1	3,0	9,4	20,2	6,4	5,7	6,0	15,3	8,2	15,1	13,5	7,8
80.000	12,1	2,8	8,7	18,7	5,9	5,3	5,6	14,2	7,6	13,9	12,5	7,2
90.000	11,3	2,6	8,2	17,4	5,5	4,9	5,2	13,2	7,1	13,0	11,7	6,7
100.000	10,7	2,5	7,7	16,3	5,1	4,6	4,9	12,4	6,7	12,2	11,0	6,3
200.000	7,1	1,7	5,1	10,8	3,4	3,1	3,2	8,2	4,4	8,1	7,4	4,2
300.000	5,6	-	4,0	8,4	2,7	2,4	2,5	6,5	3,5	6,4	5,9	3,3
400.000	4,7	-	3,4	7,1	2,3	2,0	2,1	5,5	3,0	5,4	5,0	2,8
500.000	4,2	-	3,0	6,2	-	1,8	1,9	4,8	2,6	4,7	4,4	2,5
750.000	3,3	-	2,4	4,9	-	-	1,5	3,8	2,0	3,7	3,5	2,0
1.000.000	2,8	-	-	4,1	-	-	-	3,2	1,7	3,1	2,9	1,7
2.000.000	1,9	-	-	2,7	-	-	-	2,1	1,1	2,1	2,0	-
3.000.000	1,5	-	-	2,1	-	-	-	1,6	-	1,6	1,6	-
4.000.000	1,2	-	-	1,8	-	-	-	1,4	-	1,4	1,3	-
5.000.000	1,1	-	-	1,6	-	-	-	1,2	-	1,2	-	-

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 – Anno 2015**

STIME	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
20.000	20,3	38,3	18,1	9,0	30,6	28,9	12,8	21,8	33,3	20,9
30.000	16,0	30,2	14,2	7,1	24,4	22,8	10,2	17,5	26,1	16,8
40.000	13,5	25,6	11,9	6,0	20,7	19,3	8,7	14,9	22,0	14,5
50.000	11,8	22,5	10,4	5,2	18,3	17,0	7,6	13,2	19,3	12,8
60.000	10,6	20,2	9,3	4,7	16,5	15,3	6,9	12,0	17,3	11,7
70.000	9,7	18,5	8,5	4,3	15,2	14,0	6,3	11,0	15,7	10,7
80.000	9,0	17,1	7,8	4,0	14,1	12,9	5,9	10,2	14,5	10,0
90.000	8,4	15,9	7,3	3,7	13,2	12,1	5,5	9,6	13,5	9,4
100.000	7,9	15,0	6,9	3,5	12,4	11,4	5,2	9,0	12,7	8,9
200.000	5,2	10,0	4,5	2,3	8,4	7,6	3,5	6,2	8,4	6,2
300.000	4,1	7,9	3,5	1,8	6,7	6,0	2,8	5,0	6,6	5,0
400.000	3,5	6,7	3,0	-	5,7	5,1	2,4	4,2	5,5	4,3
500.000	3,0	5,9	2,6	-	5,0	4,5	2,1	3,7	4,8	3,8
750.000	2,4	4,6	2,0	-	4,0	3,5	1,7	3,0	3,8	3,1
1.000.000	2,0	3,9	1,7	-	3,4	3,0	-	2,6	3,2	2,6
2.000.000	1,3	2,6	1,1	-	2,3	2,0	-	1,8	2,1	1,8
3.000.000	-	2,1	-	-	1,9	1,6	-	-	1,7	-
4.000.000	-	1,7	-	-	1,6	1,3	-	-	1,4	-
5.000.000	-	1,5	-	-	1,4	1,2	-	-	1,2	-

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:	853.000
Errore relativo (CV)	4,4/100=0,044
Stima intervallare	
Semi ampiezza dell'intervallo	(853.000*0,044)*1,96 =73.563
Limite inferiore dell'intervallo di confidenza	853.000-73.563=779.437
Limite superiore dell'intervallo di confidenza	853.000+73.563=926.563

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

	Persone di 14 anni e più che fumano nel Lazio
Stima puntuale:	1.075.000
Errore relativo (CV)	3,9/100=0,039
Stima intervallare	
Semi ampiezza dell'intervallo	(1.075.000*0,039)*1,96 = 82.173
Limite inferiore dell'intervallo di confidenza	1.075.000 - 82.173 = 992.827
Limite superiore dell'intervallo di confidenza	1.075.000 + 82.173=1.157.173

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.

Ad esempio, per calcolare l’intervallo di confidenza per la stima delle persone che fumano nel Lazio, i parametri sono i seguenti:

a = 9,632030 b = -1,166341

Per $\hat{Y}$ = 1.075.000 si ha:

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{\exp\left(9{,}632030 - 1{,}166341 \times \log(1.075.000)\right)} = 0{,}0358.$$

L’errore relativo percentuale è quindi pari al 3,58% e il calcolo dell’errore assoluto e dell’intervallo di confidenza è del tutto analogo a quello riportato nelle tabelle 1 e 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

Silvia Montecolle
Tel. +39 06 4673.7361
Email montecol@istat.it

DESCRIZIONE DEL FILE

FILE DESCRIPTION

Istat
Istituto Nazionale di Statistica

Aspetti della vita quotidiana

Periodo di riferimento: anno 2015

Descrizione del file

Anno 2017

INDICE

Introduzione

I file di microdati ad uso pubblico (mIcro.STAT) sono collezioni di dati elementari, liberamente scaricabile via web (http://www.istat.it/it/archivio/microdati+ad+uso+pubblico), per le quali, a partire dal 2013, sono stati sviluppati anche i corrispondenti file per la ricerca (MFR[1]). Il contenuto informativo di questi ultimi è superiore rispetto a quello del file ad uso pubblico[2].

Per la predisposizione del file mIcro.STAT relativo all'indagine *Aspetti della vita quotidiana*, periodo di riferimento 2015, è stata adottata una opportuna metodologia, allo scopo di limitare il rischio di violazione della riservatezza. Gli scenari di intrusione considerati sono stati:

- identificazione attraverso archivi esterni, ovvero attraverso il collegamento con i dati rilasciati da altre fonti pubbliche;
- identificazione spontanea, ovvero conseguente a conoscenze *a priori* dell'utente che potrebbero permettere di attribuire correttamente i dati rilasciati alle unità della popolazione rilevata.

Le variabili coinvolte nel processo di protezione sono quelle che possono consentire l'associazione tra le informazioni e i rispondenti, ossia:

- gli *identificativi diretti*, che individuano in maniera univoca le unità statistiche di *rilevazione/analisi* (come ad esempio indirizzo e codice fiscale);
- gli *identificativi indiretti,* o *variabili chiave* (come ad esempio il numero di componenti della famiglia), che permettono di circoscrivere la popolazione cui appartengono i rispondenti e che, da sole o in combinazione con altre, possono portare alla re-identificazione di uno o più record.

Mentre i primi vengono cancellati dal file, i secondi vengono trattati allo scopo di limitare il rischio di violazione della riservatezza. Per il significato delle variabili e delle rispettive modalità si rinvia al questionario di indagine contenuto nella cartella metadati.

[1] MFR è l'acronimo per il file per la ricerca (Microdata File for Research). La documentazione metodologica di tali file è consultabile sul sito Istat (http://www.istat.it/it/archivio/file+per+la+ricerca).
[2] I ricercatori, che necessitano di informazioni maggiormente dettagliate, possono, su richiesta motivata e previa autorizzazione del Presidente dell'Istituto, richiedere il file di microdati per la ricerca.

Le misure di protezione adottate

Per la predisposizione del file ad uso pubblico *Aspetti della vita quotidiana*, periodo di riferimento 2015*,* sono state adottate le misure di protezione appresso specificate.
Il progressivo famiglia univoco a livello indagine [*profam*] è stato sostituito con un numero fittizio.
Le variabili peso [*peso*] e statura [*stat*] sono state sostituite con la variabile indice di massa corporea [*bmi*], con le modalità: sottopeso, normopeso, sovrappeso, obeso.

Soppressione di variabili:

1. Provincia di residenza [*prov*]
2. Comune di residenza [*com*]
3. Ampiezza demografica del comune di residenza [*dom*]
4. Settore di attività economica [*ateco*]
5. Frequenza scolastica [*frsc*]

Ricodifiche di variabili categoriali o quantitative:

1. Regione di residenza [*reg* nel file originario → *regmif* nel file mIcro.STAT]
   Aggregazione delle modalità *Provincia autonoma di Trento* e *Provincia autonoma di Bolzano* in un'unica modalità *Trentino-Alto Adige*

2. Età [*eta* → *etami*]
   Aggregazione dell'età in anni compiuti nelle seguenti classi di età:

Etichetta	Descrizione
001	0 – 2 anni
002	3 – 5 anni
003	6 – 10 anni
004	11 – 13 anni
005	14 – 15 anni
006	16 – 17 anni
007	18 – 19 anni
008	20 – 24 anni
009	25 – 34 anni
010	35 – 44 anni
011	45 – 54 anni
012	55 - 59 anni
013	60 – 64 anni
014	65 – 74 anni

015	75 anni e più

3. Stato civile [*stciv* → *stcivm*]

   Aggregazione delle modalità *separato/a di fatto, separato/a legalmente e divorziato/a*

4. Titolo di studio [*istr* → *istrmi*]

   Aggregazione delle modalità originarie nelle seguenti:

Etichetta	Descrizione
01	laurea e post-laurea
07	diploma
09	licenza di scuola media
10	licenza di scuola elementare, nessun titolo di studio

5. Condizione occupazionale [*cond* → *condmi*]

   Aggregazione delle modalità originarie nelle seguenti:

Etichetta	Descrizione
1	occupato
2	in cerca di occupazione
3	inattivo

6. Posizione nella professione [*posiz* → *posizm*]

   Aggregazione delle modalità originarie nelle seguenti:

Etichetta	Descrizione
1	dirigente, autonomo come imprenditore o libero professionista
2	direttivo, quadro o impiegato
3	capo operaio, operaio subalterno e assimilati, apprendista o lavorante a domicilio per conto d'impresa
4	lavoratore in proprio, socio cooperativa Produzione Beni e/o prestazioni di servizio, coadiuvante, collaborazione coordinata e continuativa (con o senza progetto) o prestazione d'opera occasionale

7. Tipologia familiare [*tipfa2* → *tipfa2mi*]

   Aggregazione delle modalità:

Etichetta	Descrizione
01	senza nuclei
02	un nucleo senza isolati coppie
03	un nucleo senza isolati monogenitore
04	un nucleo con isolati coppie
05	un nucleo con isolati monogenitore
06	due nuclei senza isolati

07	due nuclei con isolati
08	tre o più nuclei

Variabili relative al ricovero in ospedale o in istituto di cura

Per la variabile:

- *ggrica*, numero di notti in cui è stato ricoverato negli ultimi tre mesi

i valori sono stati aggregati secondo la seguente classificazione*:*

Etichetta	**Descrizione**
puntuali	Per i valori minore e uguali a 15
16	16 – 20
17	21 – 30
18	31 – 40
19	41 – 50
20	51 e oltre

Per la variabile:

- *nnrica*, numero di volte in cui è stato ricoverato negli ultimi tre mesi

è stato applicato il metodo di top-coding, in particolare i valori uguali e superiori a 5 sono stati ricodificati in 5, con significato “5 e oltre”.

Variabili relative ai mezzi di trasporto personali

Per le seguenti variabili:

- *nmotor,* numero motorini, scooter;
- *nmoto,* numero di motociclette, moto;
- *nauto,* numero automobili;

i valori uguali e superiori a 4 sono stati ricodificati in 4, con significato “4 e oltre”.

Variabili relative al consumo di bevande alcoliche

Per le seguenti variabili:

- *bicbirra,* consumo abituale al giorno di birra, in bicchieri;
- *bicvino,* consumo abituale al giorno di vino, in bicchieri;

- *bicaltro,* numero di bicchieri al giorno di aperitivi alcolici, amari o superalcolici;

i valori superiori a 5 sono stati ricodificati in 6, con significato “6 e oltre”.

Per le seguenti variabili:

- *nbicalc,* consumo di bevande alcoliche negli ultimi 12 mesi, numero bicchieri;
- *bicfuori,* complessivamente in una settimana quanti bicchieri di vino o alcolici consuma abitualmente fuori dai pasti?

i valori superiori a 20 sono stati ricodificati in 21, con significato “21 e oltre”.

Altre variabili quantitative

Per la seguente variabile:

- *nsigar,* quante sigarette fuma in media al giorno;

i valori superiori a 20 sono stati ricodificati in 21, con significato “21 e oltre”.

Per la seguente variabile:

- *nccred,* numero di carte di credito;

i valori superiori a 3 sono stati ricodificati in 4, con significato “4 e oltre”.

Per la seguente variabile:

- *stanze,* numero di stanze nell’abitazione;

i valori superiori a 9 sono stati ricodificati in 10, con significato “10 e oltre”.

Ulteriori interventi di protezione dei dati

In alcuni record sono stati modificati i valori puntuali in corrispondenza della variabile *regione*: essi sono stati sostituiti in alcuni casi con il corrispondente livello ripartizionale, in altri casi con la modalità “non disponibile” (codice 999).

Inoltre, sono stati inseriti valori mancanti in corrispondenza di una o più variabili.

Ai fini della tutela della riservatezza, le variabili individuate dal D.Lgs 30/06/03 sono state trattate mediante permutazione casuale di una parte delle osservazioni.

A causa delle misure di protezione adottate, si possono verificare scostamenti rispetto ai dati pubblicati dall’Istat.

Riferimenti bibliografici

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. e de Wolf, P.-P. (2012). Statistical Disclosure Control. Wiley.

Willenborg, L. e de Waal, T. (1996). Statistical Disclosure Control in Practice. Lecture Notes in Statistics, 111, New York: Springer-Verlag.

Willenborg, L. e de Waal, T. (2000). Elements of statistical disclosure control. Lecture Notes in Statistics, 115, New York: Springer-Verlag.

Curatori

Il documento è stato redatto da Ludovica Ioppolo.