Nota metodologica

Premessa

L'indagine europea sulla salute (Ehis) viene condotta in tutti gli Stati dell'Unione europea con l'obiettivo di confrontare la situazione nei diversi paesi riguardo i principali aspetti delle condizioni di salute della popolazione ed il ricorso ai servizi sanitari. L'indagine è prevista dal regolamento (Ue) n. 141/2013 della Commissione, del 19 febbraio 2013 (che attua il regolamento (Ce) n. 1338/2008 del Parlamento europeo e del Consiglio relativo alle statistiche comunitarie in materia di sanità pubblica e di salute e sicurezza sul luogo di lavoro) ed è inserita nel Programma statistico nazionale 2014-2016 (cod. IST 02565).
Alla prima edizione dell'indagine (Ehis wave 1), condotta tra il 2006 ed il 2009, hanno aderito 17 Stati Membri essendo stata implementata sulla base di un *gentlemen's agreement*. L'Italia ha partecipato attivamente al lungo processo di armonizzazione per la definizione degli strumenti di rilevazione di questa prima wave, cui non ha potuto aderire per la tempistica non compatibile con la programmazione nazionale.
Per l'implementazione della seconda wave dell'indagine, uno dei principali riferimenti metodologici è stato il manuale predisposto da Eurostat, con il supporto dei paesi membri, al fine di raggiungere un elevato livello di armonizzazione (European health interview survey (Ehis wave 2) Methodological manual: http://ec.europa.eu/eurostat/documents/3859598/5926729/KS-RA-13-018-EN.PDF/26c7ea80-01d8-420e-bdc6-e9d5f6578e7c).
I risultati derivanti da questo tipo di indagine campionaria sono di grande rilevanza sociale, poiché consentono di monitorare i principali indicatori di salute utili alla programmazione sanitaria nel Paese e contribuiscono a definire le politiche europee per soddisfare i bisogni dei cittadini.

Finalità e caratteristiche dell'indagine

In l'Italia l'indagine Ehis (wave 2) è stata condotta dall'Istat nel 2015 (nei mesi da ottobre a dicembre), in ottemperanza del periodo di riferimento previsto dallo specifico regolamento.
La priorità di favorire la comparabilità a livello europeo, come principale obiettivo, ha comportato in alcuni casi la necessità di ricorrere a quesiti non sempre perfettamente sovrapponibili a quelli utilizzati nelle precedenti edizioni nazionali delle indagini sulla salute. Pertanto nella comparazione degli indicatori prodotti con la presente rilevazione si suggerisce di usare cautela e prendere visione della diversa formulazione dei quesiti contenuti nei questionari, disponibili all'indirizzo: http://www.istat.it/it/archivio/167485.
Per la gran parte dei quesiti le interviste sono state condotte secondo la tecnica Pen and paper interview (Papi) - tecnica di rilevazione che prevede l'utilizzo delle interviste faccia-a-faccia. Per un'altra parte, più esigua, è stata prevista l'autocompilazione del questionario. Nell'intervista faccia-a-faccia è stato somministrato un questionario familiare e tante schede individuali quanti sono i membri della famiglia.
La dimensione del campione è di circa 13.000 famiglie residenti in oltre 550 comuni di diversa ampiezza demografica, distribuiti su tutto il territorio nazionale. Il disegno di campionamento (descritto in dettaglio nell'Appendice B) è a due stadi con stratificazione delle unità di primo stadio (comuni). Le unità di secondo stadio sono le famiglie estratte con criterio di scelta casuale dalle liste anagrafiche comunali, secondo una strategia di campionamento volta a costituire un campione statisticamente rappresentativo della popolazione residente.
L'unità di rilevazione è costituita dalla famiglia di fatto (ff) associata alla famiglia anagrafica (fa) campionata. La famiglia di fatto è definita come l'insieme di persone che dimorano abitualmente nella stessa abitazione e sono legate da vincoli di parentela, affinità, affettività o amicizia.
Le tematiche trattate riguardano tre macro aree: lo stato di salute, i determinanti di salute e l'accesso ed utilizzo dei servizi sanitari indagati insieme al contesto socio-demografico di ciascun individuo delle famiglie intervistate. Più nello specifico i contenuti informativi, corrispondenti alle sezioni tematiche in cui sono suddivisi i questionari per l'Italia, sono i seguenti:

- Dati anagrafici
- Condizioni generali di salute
- Malattie e condizioni croniche
- Infortuni e lesioni
- Limitazioni funzionali fisiche e sensoriali
- Attività di cura della persona
- Attività domestiche
- Dolore
- Benessere psicologico
- Assistenza sanitaria in regime ordinario e diurno
- Assistenza ambulatoriale e domiciliare
- Consumo di farmaci
- Prevenzione
- Difficoltà di accesso a prestazioni sanitarie
- Peso e altezza
- Attività fisica
- Consumo di frutta e verdura
- Sostegno sociale
- Cure o assistenza fornite
- Situazione lavorativa
- Assenze dal lavoro per motivi di salute
- Salute dei denti
- Consumo di tabacco
- Consumo di bevande
- Stato di salute percepito

Le tavole di dati pubblicate si riferiscono alla prevenzione, peso e altezza, attività fisica, consumo di frutta e verdura e consumo di tabacco. Entro il mese di luglio 2017, l'Istat diffonderà i dati relativi alle condizioni di salute ed all'utilizzo dei servizi sanitari.
Ulteriori indicatori sono disponibili nel database di Eurostat all'indirizzo http://ec.europa.eu/eurostat/data/database

Presentazione delle tavole
Gli indicatori che si presentano nelle tavole statistiche si riferiscono alla prevenzione dei tumori (mammografia, pap-test, ricerca del sangue occulto nelle feci), alla prevenzione della sindrome influenzale (vaccinazione) ed ai fattori di rischio per la salute (sovrappeso ed obesità, attività fisica, consumo di frutta e verdura e tabagismo).
Le tavole statistiche presentano gli indicatori riferiti alla popolazione di 15 anni e più (per l'indice di massa corporea 18 anni e più), declinati per singolo paese dell'Unione europea a 28, per classi di età e genere, territorio, titolo di studio e reddito.

Prevenzione
I quesiti relativi alla mammografia ed al pap-test chiedono all'intervistata quando ha fatto l'ultima volta il test, senza specificare se lo ha fatto in assenza di disturbi e sintomi.
La fasce di età raccomandate dalle linee guida italiane sono: per la mammografia 50-69 anni, per il pap-test 25-64 anni, per la ricerca del sangue occulto nelle feci 50-70 anni. Tuttavia, per la comparazione con gli altri paesi europei, le tavole relative al pap-test fanno riferimento alla fascia di età 20-69 anni e quelle relative alla ricerca del sangue occulto nelle feci alla fascia di età 50-74 anni.
Per quanto riguarda la frequenza con cui devono essere effettuati i test di prevenzione, poiché le raccomandazioni italiane prevedono una cadenza temporale di due anni per la mammografia, tre anni per il pap-test e due anni per la ricerca del sangue occulto nelle feci, alcune modalità di risposta sono state rese disponibili anche in modo accorpato, per rendere più fruibile l'informazione.
La vaccinazione antinfluenzale è raccomandata annualmente alle persone di 65 anni e più, oltre ad alcuni target di popolazione a rischio, che non sono presenti in tali tabelle in quanto non sono disponibili nel data base di Eurostat.

Sovrappeso ed obesità
L'eccesso ponderale viene misurato mediante l'Indice di massa corporea (vedi Glossario) e si riferisce alla popolazione adulta (18 anni e più).

Attività fisica
Secondo le raccomandazioni dell'Organizzazione mondiale della sanità (Who Hepa -Health enhancing physical activity- recommendations) la popolazione adulta e anziana dovrebbe fare almeno 150 minuti a settimana di

attività fisica aerobica nel tempo libero. Pertanto l'indicatore presentato nelle tavole si riferisce ai minuti dedicati settimanalmente alla pratica dell'attività fisica aerobica nel tempo libero (vedi Glossario).

Consumo di frutta e verdura
Gli indicatori relativi al consumo di frutta e verdura (vedi Glossario) si riferiscono alla frequenza di consumo (almeno una volta al giorno, da 1 a 3 volte a settimana, da 4 a 6 volte a settimana, raramente o mai). Secondo le linee guida internazionali si raccomanda di consumare frutta e verdura almeno una volta al giorno.

Abitudine al fumo
Gli indicatori relativi all'abitudine al fumo si riferiscono alla prevalenza di fumatori (abituali od occasionali), alla prevalenza di forti fumatori ed all'esposizione al fumo di tabacco in ambienti chiusi (vedi Glossario).

Strategia di campionamento e livello di precisione dei risultati

1. Obiettivi conoscitivi

La *popolazione di interesse* dell'indagine in oggetto, ossia l'insieme delle unità statistiche intorno alle quali si intende investigare, è costituita dalle famiglie residenti in Italia e 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 e Italia insulare);
- le regioni geografiche (a eccezione del Trentino-Alto Adige le cui stime sono prodotte separatamente per le province di Bolzano e Trento);
- la tipologia comunale ottenuta suddividendo i comuni italiani in quattro classi formate in base a caratteristiche socio-economiche e demografiche:

A) comuni appartenenti all'area metropolitana suddivisi in:

A_1 comuni centro dell'area metropolitana: Torino, Milano, Venezia, Genova, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania, Cagliari;
A_2 comuni che gravitano intorno ai comuni centro dell'area metropolitana;
B) comuni non appartenenti all'area metropolitana suddivisi in:

B_1 comuni aventi fino a 10.000 abitanti;
B_2 comuni con oltre 10.001 abitanti.

2. Strategia di campionamento

2.1 Premessa

Per definire il disegno campionario dell'Indagine Ehis-wave 2, l'Istat ha realizzato uno studio preliminare in cui si è tenuto conto delle indicazioni riportate nel manuale metodologico predisposto da Eurostat[1] e dell'esperienza consolidata, ultratrentennale, sulle indagini campionarie sulle famiglie nell'ambito delle

[1]European health interview survey (Ehis wave 2) Methodological manual http://ec.europa.eu/eurostat/documents/3859598/5926729/KS-RA-13-018-EN.PDF/26c7ea80-01d8-420e-bdc6-e9d5f6578e7c.

statistiche sociali dell'Istituto, nonché dei vincoli di budget, con una valutazione complessiva dei costi derivanti dalle interviste sulle famiglie e quelle sugli individui.
Un valore aggiunto dell'approccio delle indagini sociali per famiglia, anziché per individuo, è la possibilità di raccogliere le informazioni su tutti i componenti della famiglia, studiare le relazioni all'interno del gruppo e nel contempo avere un abbattimento dei costi, nel caso di tecniche di interviste faccia a faccia. Un elemento rilevante del costo complessivo dell'intervista è, infatti, il costo per raggiungere la famiglia campione.
Per il nostro Paese, nel regolamento europeo era stata proposta una numerosità teorica campionaria pari a 13.180 individui. Inoltre nel manuale metodologico si raccomandava che in caso di adozione di un disegno campionario di famiglie, l'effetto del disegno dovuto alla selezione di famiglie (cluster), anziché di individui, dovesse essere contenuto.
I risultati dello studio hanno mostrato che il disegno di campionamento più conveniente da un punto di vista dei costi dell'indagine è quello per famiglie dal momento che, per la principale variabile target, presenza di limitazioni gravi nelle attività quotidiane, la correlazione intra-cluster, e quindi l'effetto del disegno, è risultata molto bassa, contribuendo a incrementare l'effetto del disegno da 1,059 a 1,075[2].

2.2 Descrizione generale del disegno di campionamento

Il disegno di campionamento utilizzato 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 quattro aree A_1, A_2, B_1, B_2, i comuni sono suddivisi in due sottoinsiemi sulla base della popolazione residente:

- l'insieme dei comuni Auto rappresentativi (Ar) costituito dai comuni di maggiore dimensione demografica;
- l'insieme dei comuni Non auto rappresentativi (Nar) costituito dai rimanenti comuni.

Nell'ambito dell'insieme dei comuni Ar, ciascun comune viene 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 dal registro delle anagrafi comunali (Lac); per ogni famiglia 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 (Us) sono le famiglie anagrafiche; per ogni famiglia 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.

2.3 Definizione della dimensione campionaria

Poiché la popolazione di interesse dell'indagine europea Ehis è costituita dagli individui con più di 15 anni, al fine di realizzare il disegno di campionamento sulle famiglie, è stato utilizzato il numero medio di componenti con età superiore a 15 anni a livello regionale.
Il campione finale teorico adottato per l'indagine europea Ehis ha una dimensione in termini di famiglie pari a 12.013. Ad ogni comune campione è stato assegnato un numero minimo di interviste pari a 18. La dimensione campionaria di famiglie è stata aumentata dai responsabili dell'indagine rispetto a quella definita nel campione minimo indicato da Eurostat poiché si è tenuto conto sia di esigenze nazionali che territoriali di stima (Nuts 2 per le regioni grandi), sia della prevista caduta delle risposte dovuta ai rifiuti o alla irreperibilità delle famiglie da intervistare (errori di lista). Le valutazioni sui tassi di risposta sono state effettuate prendendo in considerazione i risultati dell'ultima indagine multiscopo "Aspetti della vita quotidiana".
La dimensione finale del campione è risultata pari a poco più di 15.900 famiglie.

[2] De Vitiis C. e Inglese F. (2014) *Sampling design italian Ehis wave II*.

Al fine di tenere sotto controllo gli errori campionari sia delle stime nazionali sia di quelle regionali, è stata utilizzata un'allocazione di compromesso tra l'allocazione uniforme e quella proporzionale alla popolazione. Il campione di individui è stato pertanto assegnato alle regioni per il 30% in modo uniforme e per il 70% in modo proporzionale. All'interno delle regioni il campione è stato distribuito in modo proporzionale tra le tipologie comunali.

2.4 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 18;
- formazione di strati aventi ampiezza approssimativamente costante in termini di popolazione residente.

Il procedimento di stratificazione, attuato all'interno di ogni dominio territoriale individuato dalle aree A_1, A_2, B_1, B_2 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[3].
La selezione delle famiglie da intervistare in ogni comune campione viene effettuata dalla Lac (Lista anagrafica comunale) senza reimmissione e con probabilità uguali. La tecnica di selezione delle unità campionarie è di tipo sistematico.
Nel prospetto 1 viene riportata la distribuzione regionale dell'universo e del campione dei comuni, delle famiglie e degli individui.

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

Prospetto 1 – Distribuzione regionale dei comuni, delle famiglie e degli individui nell'universo e nel campione

	Comuni		Famiglie		Individui	
REGIONI	Universo	Campione	Universo	Campione	Universo	Campione
Piemonte	1.206	39	2.009.950	1.259	4.373.647	2.114
Valle d'Aosta	74	10	60.592	270	126.752	430
Liguria	235	17	789.715	639	1.563.327	1.007
Lombardia	1.530	77	4.285.177	2.260	9.947.315	3.485
Trentino-Alto Adige	326	24	427.389	627	1.046.644	1.186
Bolzano	116	12	209.503	313	514.744	649
Trento	210	12	217.886	314	531.900	537
Veneto	579	42	2.025.750	1.085	4.877.700	2.230
Friuli-Venezia Giulia	216	17	554.416	471	1.211.874	800
Emilia-Romagna	340	39	1.948.638	1.102	4.421.566	1.926
Toscana	279	35	1.589.385	995	3.729.525	1.836
Umbria	92	13	378.876	395	887.121	627
Marche	236	21	628.702	505	1.537.456	1.039
Lazio	378	30	2.352.294	1.502	5.858.437	2.322
Abruzzo	305	19	543.651	398	1.323.502	747
Molise	136	10	129.411	245	311.358	496
Campania	550	41	2.100.617	1.016	5.839.527	2.474
Puglia	258	35	1.534.773	820	4.069.563	1.708
Basilicata	131	12	230.608	285	572.754	616
Calabria	409	23	782.180	495	1.966.436	1.124
Sicilia	390	38	2.013.321	1.066	5.061.290	2.094
Sardegna	377	20	691.240	499	1.652.724	949
Italia	8.047	562	25.076.685	15.934	60.378.518	29.210

2.5 Procedimento per il calcolo delle stime

Le stime prodotte dall'indagine sono essenzialmente stime di frequenze assolute e relative, riferite alle famiglie e agli individui.
Le stime sono ottenute mediante uno stimatore di ponderazione vincolata, che è il metodo di stima adottato per la maggior parte delle indagini Istat sulle imprese e sulle famiglie.
Il principio su cui è basato ogni metodo di stima campionaria è che le unità appartenenti al campione rappresentino anche le unità della popolazione che non sono incluse nel campione.
Questo principio viene realizzato attribuendo a ogni unità campionaria un peso che indica il numero di unità della popolazione rappresentata dall'unità medesima.

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

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

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

La stima del totale (1) è data da

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

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 sono stati definiti i seguenti vincoli (totali noti di popolazione):

- il primo si riferisce alla distribuzione della popolazione nelle cinque ripartizioni territoriali per sesso e nove classi decennali di età;
- il secondo si riferisce alla distribuzione della popolazione nelle regioni per sesso e cinque classi di età;
- il terzo riguarda la distribuzione del totale delle famiglie per regione, stimata dall'indagine "Aspetti della vita quotidiana".

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*, d_{hij}, 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 ripartizionale e 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.

[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}.

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 euclidea; 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).
La presenza di valori estremi dei pesi finali è stata controllata tenendo conto della regola riportata nel manuale metodologico di Ehis wave 2. A tal fine è stata calcolata la quantità Q_{hij}:

$$Q_{hij} = \frac{w_{hij}}{\overline{w}_{hij}} \frac{\overline{d}_{hij}}{d_{hij}},$$

in cui d_{hij} è il peso iniziale attribuito a tutti i componenti della famiglia *j* del comune *i* dello strato *h*, $\overline{d}_{hij}$ e $\overline{w}_{hij}$ sono i pesi medi rispettivamente dei pesi iniziali e dei pesi finali.
Tale quantità è stata utilizzata per definire l'intervallo di accettazione dei valori dei pesi finali sulla base della relazione:

$$\frac{1}{C} \leq Q_{hij} \leq C,$$

dove *C* è una costante che assume valore 3.

Per l'indagine in oggetto, il peso finale è stato determinato con la procedura di calibrazione sviluppata in Re-Genesees (Zardetto, 2015) [5].
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 nel paragrafo 3, 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.

3. Valutazione del livello di precisione delle stime

3.1 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 $\hat{Var}(\hat{Y}_d)$ la stima della varianza della generica stima $\hat{Y}_d$, la stima dell'errore di campionamento assoluto di $\hat{Y}_d$ si può ottenere mediante la seguente espressione:

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

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 nel paragrafo 2.5, 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 $\hat{Var}(\hat{Y}_d)$ si è utilizzato il metodo proposto da Woodruff; in base a tale metodo, che ricorre all'espressione linearizzata in serie

[5] Zardetto D. (2015). *ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys*. Journal of Official Statistics. Volume 31, Issue 2, Pages 177–203, ISSN (Online) 2001-7367, June 2015.

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

di Taylor, è possibile ricavare la varianza di ogni stimatore non lineare (funzione regolare di totali) calcolando la varianza dell'espressione linearizzata ottenuta. In particolare, per la definizione dell'espressione linearizzata dello stimatore ci si è riferiti allo stimatore di regressione generalizzata, sfruttando la convergenza asintotica di tutti gli stimatori di ponderazione vincolata a tale stimatore, poiché nel caso di stimatori di ponderazione vincolata che utilizzano funzioni distanza differenti dalla distanza euclidea (che conduce allo stimatore di regressione generalizzata) non è possibile derivare l'espressione linearizzata dello stimatore.

L'espressione linearizzata dello stimatore (2) è data, quindi, da:

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

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=18) variabili ausiliarie, osservati per la generica famiglia h_{ij} 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

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

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

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

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:

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

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

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

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

dove le quantità sono espresse come:

$$\hat{Z}_{hg} = \sum_{j=1}^{m_{hi}} z_{hij} w_{hij} \qquad \text{e} \qquad \hat{Z}_g = \sum_{h=1}^{L_g} \sum_{j=1}^{m_{hi}} z_{hij} w_{hij} \,.$$

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

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

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.

3.2. 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.

Nei prospetti 2a e 3a 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 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 prima riga del prospetto 2a.

I prospetti 2b e 3b, presentati in aggiunta con riferimento agli individui, consentono di rendere più agevole il calcolo degli errori campionari. Essi contengono gli errori di campionamento relativo, per ciascun dominio territoriale di interesse, calcolati mediante la formula (12), corrispondenti alle stime di frequenze assolute.

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 il livello di stima (riportato in colonna) che più si avvicina alla stima di interesse e nel considerare come errore relativo il valore che si trova sulla riga 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 in 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 2a – Valori dei coefficienti a, b e dell'indice di determinazione R^2 delle funzioni utilizzate per le interpolazioni degli errori campionari delle stime a livello nazionale e per ripartizione geografica

RIPARTIZIONE GEOGRAFICA	a	b	R^2
Nord-ovest	8,932	-1,114	0,932
Nord-est	8,734	-1,122	0,932
Centro	8,730	-1,115	0,936
Sud	8,701	-1,105	0,933
Isole	8,266	-1,065	0,895
ITALIA	8,940	-1,110	0,949

Prospetto 2b - Valori interpolati degli errori campionari delle stime per ripartizione geografica

RIPARTIZIONE GEOGRAFICA	Valori della stima – frequenza assoluta									
	25000	50000	75000	100000	250000	500000	750000	1000000	2500000	5000000
Nord-ovest	30,92	21,02	16,77	14,29	8,58	5,83	4,65	3,96	2,38	1,62
Nord-est	26,84	18,19	14,49	12,33	7,37	5,00	3,98	3,39	2,03	1,37
Centro	27,79	18,89	15,07	12,83	7,70	5,23	4,17	3,56	2,13	1,45
Sud	28,82	19,65	15,70	13,40	8,08	5,51	4,40	3,75	2,26	1,54
Isole	28,37	19,62	15,81	13,56	8,33	5,76	4,64	3,98	2,44	1,69
ITALIA	31,597	21,503	17,169	14,634	8,799	5,988	4,781	4,075	2,450	1,668

Prospetto 3a – Valori dei coefficienti a, b e dell'indice di determinazione R^2 delle funzioni utilizzate per le interpolazioni degli errori campionari delle stime per regione geografica

REGIONE GEOGRAFICA	a	b	R^2
Piemonte	8,335	-1,073	0,970
Valle D'Aosta	5,886	-1,054	0,811
Lombardia	9,000	-1,095	0,970
Trentino-Alto Adige			
Bolzano	7,730	-1,134	0,835
Trento	7,941	-1,144	0,852
Veneto	8,522	-1,089	0,971
Friuli-Venezia Giulia	8,030	-1,086	0,937
Liguria	8,298	-1,107	0,955
Emilia-Romagna-	9,227	-1,149	0,971
Toscana	8,187	-1,071	0,906
Umbria	8,934	-1,195	0,902
Marche	7,403	-1,021	0,918
Lazio	8,710	-1,092	0,974
Abruzzo	8,440	-1,107	0,861
Molise	5,936	-0,944	0,853
Campania	8,233	-1,042	0,966
Puglia	8,336	-1,059	0,976
Basilicata	7,664	-1,108	0,838
Calabria	7,846	-1,038	0,927
Sicilia	7,937	-1,011	0,962
Sardegna	8,225	-1,089	0,871

Prospetto 3b - Valori interpolati degli errori campionari delle stime per regione geografica

Regione geografica	Valori della stima – frequenza assoluta									
	10000	25000	50000	75000	100000	250000	500000	750000	1000000	2500000
Piemonte	46,03	28,15	19,41	15,61	13,38	8,18	5,64	4,54	3,89	2,38
Valle D'Aosta	14,80	9,13	6,34	5,12	4,40	2,71	1,88	1,52	1,31	0,81
Lombardia	58,18	35,23	24,11	19,31	16,50	9,99	6,84	5,48	4,68	2,83
Trentino-Alto Adige										
Bolzano	25,75	15,32	10,34	8,22	6,98	4,15	2,80	2,23	1,89	1,13
Trento	27,32	16,17	10,88	8,63	7,32	4,33	2,92	2,31	1,96	1,16
Veneto	47,04	28,56	19,58	15,70	13,43	8,15	5,59	4,48	3,83	2,33
Friuli-Venezia Giulia	37,32	22,69	15,57	12,50	10,69	6,50	4,46	3,58	3,06	1,86
Liguria	38,81	23,38	15,93	12,73	10,86	6,54	4,46	3,56	3,04	1,83
Emilia-Romagna-	50,88	30,06	20,19	16,00	13,56	8,01	5,38	4,26	3,61	2,14
Toscana	43,32	26,53	18,30	14,73	12,63	7,73	5,34	4,30	3,68	2,25
Umbria	35,43	20,49	13,54	10,63	8,95	5,18	3,42	2,68	2,26	1,31
Marche	36,74	23,01	16,15	13,13	11,34	7,10	4,98	4,05	3,50	2,19
Lazio	50,97	30,90	21,17	16,96	14,50	8,79	6,02	4,83	4,12	2,50
Abruzzo	41,57	25,03	17,06	13,63	11,62	7,00	4,77	3,81	3,25	1,96
Molise	25,22	16,37	11,80	9,75	8,51	5,52	3,98	3,29	2,87	1,86
Campania	50,55	31,36	21,85	17,69	15,23	9,45	6,58	5,33	4,59	2,85
Puglia	49,17	30,26	20,96	16,91	14,52	8,94	6,19	5,00	4,29	2,64
Basilicata	28,07	16,89	11,51	9,19	7,84	4,72	3,21	2,57	2,19	1,32
Calabria	42,48	26,41	18,43	14,93	12,86	8,00	5,58	4,52	3,89	2,42
Sicilia	50,41	31,73	22,35	18,21	15,75	9,91	6,98	5,69	4,92	3,10
Sardegna	40,62	24,67	16,92	13,57	11,60	7,05	4,83	3,87	3,31	2,01

AVVERTENZE

Le ripartizioni geografiche costituiscono una suddivisione geografica del territorio e sono così articolate:

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

Nord-est: comprende Trentino-Alto Adige (Bolzano-Bozen, Trento), Veneto, Friuli-Venezia Giulia, Emilia-Romagna

*Centro***:** comprende Toscana, Umbria, Marche, Lazio

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

Isole: comprendono Sicilia, Sardegna