

***Indagine Multiscopo sulle
Famiglie
Modulo sull'uso delle
tecnologie dell'informazione
e della comunicazione
Anno 2008***

Manuale utente e tracciato record



ISTITUTO NAZIONALE DI STATISTICA

INDAGINE MULTISCOPO SULLE FAMIGLIE
MODULO SULL'UTILIZZO DELLE TECNOLOGIE DELLA COMUNICAZIONE E DELL'INFORMAZIONE
ANNO 2008
DOCUMENTAZIONE TECNICA E DESCRIZIONE DEL FILE

PREMESSA

Il Decreto Legislativo n. 322 del 6/9/1989 regola la diffusione delle informazioni statistiche prodotte nell'ambito del Sistema Statistico Nazionale al fine di garantire la riservatezza dei rispondenti. In particolare, per la diffusione di dati elementari, l'articolo 10, comma 2, dispone quanto segue: "Sono distribuite altresì ove disponibili, su richiesta motivata e previa autorizzazione del Presidente dell'Istat, collezioni campionarie di dati elementari, resi anonimi e privi di ogni riferimento che ne permetta il collegamento con singole persone fisiche e giuridiche".

Nell'osservanza di tale Decreto Legislativo e del D.lg. del 30/6/2003 n. 196 l'Istat ha adottato misure e tecniche che rendono impossibile, o altamente improbabile, il collegamento dei dati rilasciati con l'unità statistica a cui si riferiscono. Per tale motivo sono state apportate alcune modifiche sui files originali delle indagini, nell'intento di garantire la massima protezione ai dati contenendo al minimo l'eventuale perdita di informazioni.

Le metodologie applicate si concretizzano nell'accorpamento e/o riclassificazione di modalità di variabili e nell'oscuramento di variabili. In quest'ultimo caso nei campi del tracciato record è riportata la dicitura **"RISERVATO ISTAT"**.

Va considerato, inoltre, che la stessa dicitura è stata utilizzata anche per quelle variabili non attendibili dal punto di vista campionario e quindi non analizzabili statisticamente.

FINALITÀ E CARATTERISTICHE DELL'INDAGINE

Il modulo sull'uso delle tecnologie dell'informazione e della comunicazione (Ict) è definito e armonizzato a livello Eurostat. L'esigenza di raccogliere dati armonizzati sull'uso delle Ict deriva dagli obiettivi fissati nel 2000 dal Consiglio d'Europa tenutosi a Lisbona. In quel contesto si stabilì di: fare dell'Unione europea l'economia basata sulla conoscenza più competitiva e più dinamica del mondo entro il 2010, con una crescita economica sostenibile, posti di lavoro più numerosi e migliori e una maggiore coesione sociale.

A tale scopo l'Unione europea ha riconosciuto l'importanza della *e-economy* per la crescita, la produttività e l'occupazione e l'importanza di fornire ai cittadini le possibilità di accesso e le capacità necessarie per vivere e lavorare nella nuova società dell'informazione.

Il modulo, che nel corso degli anni è stato migliorato sulla base dei risultati delle precedenti sperimentazioni, permette l'approfondimento di diversi temi. I principali argomenti affrontati sono:

- possesso di beni Ict e di accesso ad Internet in casa;
- motivi per cui non si possiede un accesso Internet da casa;
- frequenza e luogo di utilizzo del personal computer;
- frequenza e luogo di utilizzo di Internet;
- attività svolte tramite Internet;
- partecipazione ad attività formative a distanza tramite Internet (*e-learning*);
- interazione, tramite Internet, con siti di istituzioni pubbliche (*e-government*);
- utilizzo di Internet per contattare medici o istituzioni sanitarie o ottenere informazioni relative a problematiche sanitarie (*e-health*);
- commercio elettronico e tipi di acquisti effettuati on-line;
- problemi di sicurezza riscontrati durante l'utilizzo di Internet;
- livello di abilità informatica (*e-skills*).

Il campione è a due stadi con stratificazione delle unità di primo stadio (Comuni). Le informazioni sono state raccolte con intervista diretta per una parte dei quesiti. Nei casi in cui l'individuo non fosse disponibile all'intervista per particolari motivi, le informazioni sono state fornite da un altro componente della famiglia. Per un'altra parte dei quesiti è stata prevista l'autocompilazione.

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

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

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

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

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

AVVERTENZE PER L'UTILIZZAZIONE DEL FILE

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

	Anno 2008
lunghezza record:	592
numero records individuali:	48.861
(uno per ciascuna persona intervistata)	

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

a) individui

ogni componente è individuato dal numero progressivo della famiglia e dal suo numero d'ordine all'interno della stessa. Il numero totale di appartenenti al campione è pari al numero di records: 48.861. Per selezionare i componenti della stessa famiglia si considerano tutti i records individuali che hanno lo stesso numero generale progressivo della famiglia. Per selezionare i componenti appartenenti allo stesso nucleo si considerano tutti i records che hanno lo stesso numero progressivo della famiglia e lo stesso numero d'ordine del nucleo;

b) famiglie

volendo analizzare le famiglie occorre selezionare solo il primo componente di ciascuna utilizzando il numero d'ordine all'interno della famiglia. Il totale delle famiglie è pari a 19.573;

c) nuclei familiari

l'analisi dei nuclei familiari è possibile, invece, selezionando la persona di riferimento di ciascun nucleo. Il totale dei nuclei è di 14.000.

COSTRUZIONE DELLE STIME ED ERRORI DI CAMPIONAMENTO

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

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

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

1. l'intero territorio nazionale;
2. le cinque ripartizioni geografiche (Italia Nord-Occidentale, Italia Nord-Orientale, Italia Centrale, Italia Meridionale, Italia Insulare);
3. le regioni geografiche (ad eccezione del Trentino Alto Adige le cui stime sono prodotte distintamente per le province di Bolzano e Trento);

4. sei aree basate sulla tipologia socio-demografica dei comuni.

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

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

Si riporta in allegato l'appendice dei volumi Istat della serie Multiscopo, contenente le informazioni relative al campionamento e al calcolo degli errori di stima da cui è possibile individuare gli esempi di calcolo degli errori campionari. In seguito sono accluse le tavole per il calcolo degli errori relativi ai dati contenuti nei files standard, per stime sugli individui e sulle famiglie.

Glossario

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

- **l'età** è espressa in anni compiuti;
- **il titolo di studio** è quello più elevato conseguito;
- **la condizione** è quella dichiarata come unica o prevalente dalle persone di 15 anni e più.
Si precisa inoltre che per:
 - occupato** si intende chi possiede un'occupazione in proprio o alle dipendenze da cui trae un profitto o una retribuzione (utile, onorario, stipendio, salario) o chi collabora con un familiare che svolge un'attività lavorativa in conto proprio senza avere un regolare contratto di lavoro (coadiuvante);
 - persona in cerca di occupazione** si intende chi ha perduto una precedente occupazione alle dipendenze, o chi non ha mai esercitato un'attività lavorativa ed è alla ricerca attiva di un'occupazione che è in grado di accettare se gli viene offerta;
 - casalinga** è chi si dedica prevalentemente alle faccende domestiche;
 - studente** è chi si dedica prevalentemente allo studio;
 - ritirato dal lavoro** è chi ha cessato un'attività lavorativa per raggiunti limiti di età, invalidità o altra causa; la figura del ritirato dal lavoro non coincide necessariamente con quella del pensionato in quanto, non sempre, il ritirato dal lavoro gode di una pensione;
 - in altra condizione** è chi si trova in condizione diversa da quelle sopra elencate (militare, inabile al lavoro, benestante, detenuto, ecc.).
- **la posizione nella professione** è quella dichiarata come unica o prevalente dagli occupati di 15 anni e più che viene aggregata nel modo seguente:
 - dirigenti, imprenditori, liberi professionisti;**
 - direttivi, quadri, impiegati;**
 - capo operai, operai, subalterni** (inclusi apprendisti, lavoratori a domicilio per conto di imprese);
 - lavoratori in proprio, coadiuvanti** (inclusi soci di cooperative di produzione di beni e/o prestazioni di servizio).
- **le ripartizioni geografiche** costituiscono una suddivisione geografica del territorio e sono così articolate:
 - Italia nord-occidentale** comprende: Piemonte, Valle d'Aosta, Lombardia, Liguria.
 - Italia nord-orientale** comprende: Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia Romagna.
 - Italia centrale** comprende: Toscana, Umbria, Marche, Lazio.
 - Italia meridionale** comprende: Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria.
 - Italia insulare** comprende: Sicilia, Sardegna.
- **il tipo di comune**
I comuni italiani sono suddivisi nelle seguenti classi:
 - comuni centro delle aree metropolitane:** si tratta dei comuni di Torino, Milano, Venezia, Genova, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania, Cagliari;
 - comuni appartenenti alla periferia delle aree metropolitane:** costituiscono i comuni delle cinture urbane;
 - altri comuni:** suddivisi per dimensione demografica (fino a 2.000 abitanti, da 2.001 a 10.000, da 10.001 a 50.000 e oltre i 50.000);

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

- **famiglia e nucleo familiare**

Si precisa che:

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

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

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

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, 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 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.

2. Strategia di campionamento

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

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

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

2.2 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 che quella delle stime a livello di ciascuno dei domini territoriali descritti nel paragrafo 1.

2.3 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 N_r ;
- 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 A_1, A_2, B_1, B_2, B_3 e B_4 di ciascuna regione geografica, si articola nelle seguenti fasi:

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

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

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

- suddivisione di tutti i comuni nei due sottoinsiemi A_r e N_r : i comuni di dimensione superiore o uguale

a λ_r sono definiti come comuni A_r 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 λ_r .

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

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

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

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

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

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

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

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

REGIONI	Comuni		Famiglie		Individui	
	Campione	Universo (a)	Campione	Universo (a)	Campione	Universo (a)
Piemonte	61	1.206	1.360	1.911	3.069	4.311
Valle d'Aosta - Vallée d'Aoste	20	74	465	56	1.022	124
Lombardia	84	1.546	1.660	3.914	4.072	9.477
Trentino-Alto Adige	48	339	1.092	394	2.744	984
<i>Bolzano - Bozen</i>	23	116	574	187	1.492	483
<i>Trento</i>	25	223	518	207	1.252	501
Veneto	54	581	1.139	1.888	2.855	4.727
Friuli-Venezia Giulia	31	219	720	515	1.681	1.198
Liguria	25	235	830	752	1.767	1.594
Emilia-Romagna	47	341	1.127	1.831	2.568	4.191
Toscana	51	287	1.115	1.496	2.699	3.613
Umbria	22	92	588	345	1.499	867
Marche	37	246	804	615	1.998	1.527
Lazio	33	378	1.021	2.163	2.576	5.448
Abruzzo	35	305	719	499	1.879	1.303
Molise	24	136	572	125	1.470	319
Campania	54	551	1.321	1.968	3.838	5.771
Puglia	50	258	1.092	1.469	3.016	4.054
Basilicata	27	131	597	216	1.670	589
Calabria	42	409	914	742	2.443	1.988
Sicilia	52	390	1.253	1.899	3.336	4.994
Sardegna	39	377	781	622	2.051	1.651
Italia	836	8.101	19.170	23.421	48.253	58.730

(a) Stima Indagine multiscopo "Aspetti della vita quotidiana"

2.4 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. 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

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

La stima del totale (1) è data da

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

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 unità² ed effettuare, a livello del dominio di interesse, la somma dei prodotti così ottenuti.

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

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

Nell'indagine in oggetto vengono definiti per ciascuna regione geografica 18 totali noti, che si riferiscono alla distribuzione della popolazione regionale per sesso e sei classi di età³ e della popolazione regionale nelle sei aree A_1, A_2, B_1, B_2, B_3 e B_4 . Indicando, quindi, con ${}_kX$ ($k=1, \dots, 18$) il totale noto della k -esima variabile ausiliaria per la generica regione geografica e con ${}_kX_{hij}$ il valore assunto dalla k -esima variabile ausiliaria per la famiglia rispondente hij , la condizione sopra descritta è espressa dalla seguente uguaglianza

$${}_kX = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hijk} X_{hij} \quad (k=1, \dots, 18)$$

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

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

- 1) si calcolano i *pesi diretti* come reciproco della probabilità di inclusione delle unità;
- 2) si calcolano i fattori correttivi per mancata risposta totale, come l'inverso del tasso di risposta del comune cui ciascuna unità appartiene;
- 3) si ottengono i *pesi base*, o pesi corretti per mancata risposta totale, moltiplicando i pesi diretti per i corrispondenti fattori correttivi per mancata risposta totale;
- 4) si costruiscono i fattori correttivi che consentono di soddisfare, a livello regionale, la condizione di uguaglianza tra i totali noti delle variabili ausiliarie e le corrispondenti stime campionarie;
- 5) si calcolano, infine, i pesi finali mediante il prodotto dei pesi base per i fattori correttivi ottenuti al passo 4.

I fattori correttivi del passo 4 sono ottenuti dalla risoluzione di un problema di minimo vincolato, in cui la funzione da minimizzare è una funzione di distanza (opportunamente prescelta) tra i pesi base e i pesi finali e i vincoli sono definiti dalla condizione di uguaglianza tra stime campionarie dei totali noti di popolazione e valori noti degli stessi. La funzione di distanza prescelta è la funzione logaritmica troncata; l'adozione di tale funzione garantisce che i pesi finali siano positivi e contenuti in un predeterminato intervallo di valori possibili, eliminando in tal modo i pesi positivi estremi (troppo grandi o troppo piccoli).

Tutti i metodi di stima che scaturiscono dalla risoluzione di un problema di minimo vincolato del tipo sopra descritto rientrano in una classe generale di stimatori nota come stimatori di ponderazione vincolata.⁴ Un importante stimatore appartenente a tale classe, che si ottiene utilizzando la funzione di distanza euclidea, è lo *stimatore di regressione generalizzata*. Come verrà chiarito meglio nel paragrafo 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.

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, 25-44 anni, 45-64 anni, 65 anni e più.

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

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

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

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

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

Come è stato descritto nel paragrafo 2.4, le stime prodotte dall'indagine sono state ottenute mediante uno stimatore di ponderazione vincolata definito in base a una funzione di distanza di tipo logaritmico troncato. Poiché, lo stimatore adottato non è funzione lineare dei dati campionari, per la stima della varianza $\hat{\text{Var}}(\hat{Y}_d)$ si è utilizzato il metodo proposto da Woodruff; in base a tale metodo, che ricorre all'espressione linearizzata in serie di Taylor, è possibile ricavare la varianza di ogni stimatore non lineare (funzione regolare di totali) calcolando la varianza dell'espressione linearizzata ottenuta. In particolare, per la definizione dell'espressione linearizzata dello stimatore ci si è riferiti allo stimatore di regressione generalizzata, sfruttando la convergenza asintotica di tutti gli stimatori di ponderazione vincolata a tale stimatore, poiché nel caso di stimatori di ponderazione vincolata che utilizzano funzioni distanza differenti dalla distanza euclidea (che conduce allo stimatore di regressione generalizzata) non è possibile derivare l'espressione linearizzata dello stimatore.

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

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

dove Z_{hij} è la variabile linearizzata espressa come $Z_{hij} = Y_{hij} - \mathbf{X}'_{hij} \beta$, essendo $\mathbf{X}_{hij} = (X_{hij,1}, \dots, X_{hij,K})'$ il vettore contenente i valori delle K (K=18) variabili ausiliarie, osservati per la generica famiglia hij e β , il vettore dei coefficienti di regressione del modello lineare che lega la variabile di interesse y alle K variabili ausiliarie x. In base alla (5), si ha, quindi, che la stima della varianza della stima \hat{Y}_d è ottenuta mediante la seguente relazione

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

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

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

in cui H_{AR} e H_{NAR} indicano rispettivamente il numero di strati 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 omissso) la varianza è stimata mediante la seguente espressione:

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

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

Negli strati 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:

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

dove le quantità sono espresse come:

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

Utilizzando le espressioni (8) e (9) è possibile, infine, calcolare la varianza di campionamento, $\hat{Var}(\hat{Y}_d)$, in base alla (7) e calcolare, quindi, in base alla (3) ed alla (4) rispettivamente l'errore di campionamento assoluto e l'errore di campionamento relativo.

Gli errori campionari espressi dalla (3) e dalla (4) consentono di valutare il grado di precisione delle stime; inoltre, l'errore assoluto permette di costruire un intervallo di confidenza, che, con livello di fiducia P contiene il parametro oggetto di stima, l'intervallo viene espresso come:

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

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

3.2 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 3.1 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.

3.3 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:

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

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:

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

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 seconda riga del prospetto 2 alla voce Persone (a = 8,886722, b = -1,121521).

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 dominio territoriale di riferimento.

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

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

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

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

ZONE TERRITORIALI	Famiglie			Persone		
	a	b	R^2 (%)	a	b	R^2 (%)
ITALIA	8,67348028	-1,0989437	97,4781225	9,5966274	-1,1619918	91,7135549
RIPARTIZIONI GEOGRAFICHE						
Nord-ovest	8,73562677	-1,0958617	96,2349875	9,5608085	-1,156208	90,941289
Nord-est	8,48262537	-1,1083822	97,5203464	9,60001006	-1,1976843	92,0530573
Centro	8,36282511	-1,0890065	97,246049	9,24381825	-1,1578536	91,8977783
Sud	7,64064953	-1,0420232	96,4251534	8,55047986	-1,1150104	89,3039803
Isole	7,8702824	-1,0617232	96,0603174	8,69189678	-1,1308014	91,4105398
TIPI DI COMUNE						
A1	8,73634233	-1,1081117	98,2171402	9,87617035	-1,2031698	94,128853
A2	8,25264889	-1,0744569	95,4141872	9,26464964	-1,1587907	89,6282552
B1	6,96254298	-0,9865878	91,3101469	7,51785102	-1,0392055	84,9356917
B2	7,99757313	-1,0573593	95,2578865	8,75007524	-1,113832	89,1125748
B3	7,85989875	-1,0493024	95,8298225	9,42065379	-1,1663373	89,7932689
B4	8,14478725	-1,0967399	96,6785588	9,39355558	-1,1981049	92,9537701
REGIONI						
Piemonte	8,04594448	-1,0769022	97,3	9,01681377	-1,1626979	90,4987803
Valle d' Aosta vallée d' Aoste	5,22025764	-1,0784689	92,3	5,73381652	-1,1425724	90,177418
Lombardia	8,92130881	-1,0950872	95,7	9,7112403	-1,1558717	90,504658
<i>Bolzano</i>	6,17003283	-1,0585966	95,6	7,28967882	-1,1718884	88,9540193
<i>Trento</i>	6,92608867	-1,1070633	95,8	7,65338291	-1,200209	87,995953
Veneto	8,67072192	-1,1208147	96,0	9,42271824	-1,1827785	90,9738799
Friuli-Venezia Giulia	7,34876253	-1,085513	95,8	7,75724971	-1,1295055	90,9483589
Liguria	7,84964978	-1,1111107	96,8	8,40267228	-1,1641007	91,569252
Emilia-Romagna	8,35775328	-1,0981052	97,1	9,61528543	-1,2117316	91,5718663
Toscana	8,14711225	-1,0902589	97,3	8,84941452	-1,1539889	91,8674158
Umbria	7,1712669	-1,0965376	96,2	7,45465129	-1,125218	90,4563361
Marche	7,08892037	-1,0536949	96,7	7,87569279	-1,1365416	89,5163665
Lazio	8,47067484	-1,0800942	95,6	9,40441414	-1,1566455	91,2762902
Abruzzo	7,09011625	-1,0565537	93,4	7,87086244	-1,1342389	89,000181
Molise	6,23979278	-1,1127969	93,4	6,6915671	-1,1609754	89,8292429
Campania	8,18435515	-1,0717527	93,7	8,83295951	-1,131467	87,6479363
Puglia	7,63307661	-1,0403145	97,0	8,29184886	-1,0954893	89,1332769
Basilicata	6,51047255	-1,0736313	88,5	6,89224986	-1,1168564	89,1493024
Calabria	7,11996408	-1,0445675	94,4	7,60297582	-1,0863824	88,5770732
Sicilia	7,9606475	-1,0594464	95,5	8,73433069	-1,1265524	91,317784
Sardegna	7,67337517	-1,1065678	95,2	8,23581559	-1,1552867	90,8050971

- (a) Italia nord-occidentale: Piemonte, Valle d'Aosta, Lombardia, Liguria; Italia nord-orientale: Bolzano, Trento, Veneto, Friuli-Venezia Giulia, Emilia-Romagna; Italia centrale: Toscana, Umbria, Marche, Lazio; Italia meridionale: Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria; Italia insulare: Sicilia, Sardegna.
- (b) Comuni tipo A1: Area urbana centro; Tipo A2: Area urbana periferia; Tipo B1: comuni fino a 2.000 abitanti; Tipo B2: da 2.001 a 10.000 abitanti; Tipo B3: da 10.001 a 50.000 abitanti; Tipo B4: oltre 50.000 abitanti.

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

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

STIME	Piemonte	Valle d'Aosta	Lombardia	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Liguria	Emilia-Romagna	Toscana	Umbria
20.000	27,0	6,5	38,2	11,6	13,3	29,7	18,3	20,7	28,4	26,6	15,8
30.000	21,7	5,2	30,6	9,3	10,6	23,6	14,6	16,5	22,7	21,3	12,7
40.000	18,6	4,5	26,1	8,0	9,0	20,1	12,5	14,1	19,4	18,2	10,8
50.000	16,5	4,0	23,1	7,1	8,0	17,8	11,1	12,4	17,2	16,1	9,6
60.000	14,9	3,6	20,9	6,5	7,2	16,0	10,1	11,2	15,5	14,6	8,7
70.000	13,8	-	19,2	6,0	6,6	14,7	9,2	10,3	14,3	13,4	8,0
80.000	12,8	-	17,9	5,6	6,2	13,6	8,6	9,6	13,3	12,5	7,4
90.000	12,0	-	16,8	5,2	5,8	12,8	8,1	9,0	12,4	11,7	6,9
100.000	11,3	-	15,8	4,9	5,4	12,0	7,6	8,4	11,7	11,1	6,5
200.000	7,8	-	10,8	-	-	8,2	5,2	5,7	8,0	7,6	4,5
300.000	6,3	-	8,7	-	-	6,5	4,2	4,6	6,4	6,1	3,6
400.000	5,4	-	7,4	-	-	5,5	3,6	3,9	5,5	5,2	-
500.000	4,8	-	6,6	-	-	4,9	3,2	3,5	4,9	4,6	-
750.000	3,8	-	5,3	-	-	3,9	-	-	3,9	3,7	-
1.000.000	3,3	-	4,5	-	-	3,3	-	-	3,3	3,2	-
2.000.000	2,3	-	3,1	-	-	-	-	-	-	-	-

STIME	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
20.000	18,8	32,9	18,5	9,2	29,7	26,3	12,7	19,9	28,2	19,3
30.000	15,2	26,4	14,9	7,3	23,9	21,3	10,2	16,1	22,8	15,5
40.000	13,0	22,6	12,8	6,2	20,5	18,4	8,8	13,9	19,5	13,2
50.000	11,6	20,0	11,4	5,5	18,2	16,3	7,8	12,4	17,4	11,7
60.000	10,5	18,2	10,4	5,0	16,5	14,9	7,1	11,2	15,8	10,5
70.000	9,7	16,7	9,6	4,6	15,2	13,7	6,5	10,4	14,5	9,7
80.000	9,0	15,5	8,9	4,2	14,1	12,8	6,0	9,7	13,5	9,0
90.000	8,5	14,6	8,4	4,0	13,3	12,0	5,7	9,1	12,7	8,4
100.000	8,0	13,8	7,9	3,7	12,5	11,4	5,4	8,6	12,0	7,9
200.000	5,6	9,5	5,5	-	8,6	7,9	3,7	6,0	8,3	5,4
300.000	4,5	7,6	4,4	-	7,0	6,4	-	4,8	6,7	4,3
400.000	3,9	6,5	3,8	-	6,0	5,5	-	4,2	5,8	3,7
500.000	3,4	5,8	-	-	5,3	4,9	-	3,7	5,1	-
750.000	-	4,6	-	-	4,3	4,0	-	-	4,1	-
1.000.000	-	4,0	-	-	3,6	3,4	-	-	3,6	-
2.000.000	-	2,7	-	-	2,5	-	-	-	-	-

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

STIME	Italia	Nord- ovest	Nord-est	Centro	Sud	Isole	A1	A2	B1	B2	B3	B4
20.000	38,5	38,9	32,3	32,9	28,8	28,6	36,1	33,1	25,0	32,0	34,5	29,1
30.000	30,4	30,8	25,3	26,0	22,9	22,7	28,3	26,2	20,2	25,5	27,2	22,8
40.000	25,7	26,0	21,3	22,0	19,5	19,3	23,8	22,2	17,4	21,7	23,0	19,2
50.000	22,6	22,9	18,6	19,4	17,3	17,0	20,8	19,5	15,5	19,2	20,2	16,8
60.000	20,3	20,6	16,7	17,4	15,6	15,3	18,6	17,5	14,1	17,3	18,2	15,0
70.000	18,6	18,8	15,2	15,9	14,3	14,1	17,0	16,0	13,0	15,9	16,6	13,7
80.000	17,2	17,4	14,1	14,7	13,3	13,0	15,7	14,8	12,2	14,8	15,4	12,7
90.000	16,1	16,3	13,1	13,8	12,4	12,2	14,6	13,8	11,4	13,8	14,3	11,8
100.000	15,1	15,3	12,3	13,0	11,7	11,5	13,7	13,0	10,8	13,0	13,5	11,1
200.000	10,1	10,3	8,1	8,7	8,0	7,8	9,0	8,7	7,6	8,9	9,0	7,3
300.000	8,0	8,1	6,4	6,9	6,4	6,2	7,1	6,9	6,1	7,1	7,1	5,7
400.000	6,7	6,9	5,4	5,8	5,4	5,2	5,9	5,8	5,3	6,0	6,0	4,8
500.000	5,9	6,0	4,7	5,1	4,8	4,6	5,2	5,1	4,7	5,3	5,3	4,2
750.000	4,7	4,8	3,7	4,0	3,8	3,7	4,1	4,1	3,8	4,2	4,2	3,3
1.000.000	4,0	4,1	3,1	3,4	3,2	3,1	3,4	3,4	3,3	3,6	3,5	2,8
2.000.000	2,6	2,7	2,0	2,3	2,2	2,1	2,3	2,3	2,3	2,5	2,4	1,8
3.000.000	2,1	2,1	1,6	1,8	1,8	1,7	1,8	1,8	1,8	2,0	1,9	1,4
4.000.000	1,8	1,8	1,4	1,5	1,5	1,4	1,5	1,5	-	1,7	1,6	1,2
5.000.000	1,6	1,6	1,2	1,3	1,3	1,3	1,3	1,4	-	1,5	1,4	1,1
7.500.000	1,2	1,3	0,9	1,1	1,1	-	1,0	1,1	-	1,2	1,1	0,8
10.000.000	1,0	1,1	-	0,9	0,9	-	-	-	-	1,0	0,9	-
15.000.000	0,8	-	-	-	-	-	-	-	-	-	-	-
20.000.000	0,7	-	-	-	-	-	-	-	-	-	-	-
25.000.000	0,6	-	-	-	-	-	-	-	-	-	-	-

STIME	Piemonte	Valle d' Aosta	Lombardia	Bolzano	Trento	Veneto	Friuli- Venezia Giulia	Liguria	Emilia Romagna	Toscana	Umbria
20.000	28,7	6,1	42,0	11,6	12,0	31,8	18,0	21,0	30,3	27,5	15,8
30.000	22,7	4,9	33,2	9,1	9,4	25,0	14,3	16,5	23,7	21,8	12,6
40.000	19,2	4,1	28,1	7,7	7,9	21,1	12,2	14,0	19,9	18,5	10,7
50.000	16,8	3,6	24,7	6,8	7,0	18,5	10,7	12,3	17,4	16,2	9,4
60.000	15,1	3,3	22,2	6,1	6,2	16,6	9,7	11,1	15,6	14,6	8,5
70.000	13,8	3,0	20,4	5,5	5,7	15,2	8,9	10,1	14,2	13,4	7,8
80.000	12,8	2,8	18,8	5,1	5,2	14,0	8,2	9,3	13,1	12,4	7,2
90.000	12,0	2,6	17,6	4,8	4,9	13,1	7,7	8,7	12,2	11,6	6,8
100.000	11,3	2,4	16,6	4,5	4,6	12,3	7,3	8,2	11,4	10,9	6,4
200.000	7,5	-	11,1	3,0	3,0	8,1	4,9	5,5	7,5	7,3	4,3
300.000	5,9	-	8,8	2,4	2,4	6,4	3,9	4,3	5,9	5,8	3,4
400.000	5,0	-	7,4	2,0	2,0	5,4	3,3	3,7	4,9	4,9	2,9
500.000	4,4	-	6,5	-	-	4,7	2,9	3,2	4,3	4,3	2,6
750.000	3,5	-	5,2	-	-	3,7	2,3	2,5	3,4	3,4	2,1
1.000.000	3,0	-	4,4	-	-	3,1	2,0	2,1	2,8	2,9	-
2.000.000	2,0	-	2,9	-	-	2,1	-	-	1,9	1,9	-
3.000.000	1,6	-	2,3	-	-	1,6	-	-	1,5	1,5	-
4.000.000	1,3	-	2,0	-	-	1,4	-	-	-	-	-
5.000.000	-	-	1,7	-	-	-	-	-	-	-	-

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

STIME	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
20.000	18,5	35,9	18,6	9,0	30,5	27,8	12,4	20,6	29,8	20,1
30.000	14,7	28,4	14,8	7,1	24,3	22,3	9,9	16,6	23,7	15,9
40.000	12,4	24,0	12,6	6,0	20,6	19,0	8,4	14,2	20,2	13,5
50.000	11,0	21,1	11,1	5,3	18,2	16,9	7,5	12,5	17,8	11,9
60.000	9,9	19,0	10,0	4,8	16,4	15,3	6,7	11,4	16,0	10,7
70.000	9,1	17,4	9,1	4,4	15,0	14,0	6,2	10,5	14,7	9,8
80.000	8,4	16,1	8,5	4,0	13,9	13,0	5,7	9,7	13,6	9,0
90.000	7,8	15,0	7,9	3,8	13,0	12,2	5,4	9,1	12,8	8,4
100.000	7,4	14,1	7,5	3,6	12,3	11,5	5,1	8,6	12,0	7,9
200.000	5,0	9,5	5,0	2,4	8,3	7,9	3,4	5,9	8,1	5,3
300.000	4,0	7,5	4,0	1,9	6,6	6,3	2,7	4,7	6,5	4,2
400.000	3,4	6,3	3,4	-	5,6	5,4	2,3	4,1	5,5	3,6
500.000	3,0	5,6	3,0	-	4,9	4,8	2,1	3,6	4,9	3,1
750.000	2,4	4,4	2,4	-	3,9	3,8	-	2,9	3,9	2,5
1.000.000	2,0	3,7	2,0	-	3,3	3,3	-	2,5	3,3	2,1
2.000.000	-	2,5	-	-	2,3	2,2	-	1,7	2,2	-
3.000.000	-	2,0	-	-	1,8	1,8	-	-	1,8	-
4.000.000	-	1,7	-	-	1,5	1,5	-	-	1,5	-
5.000.000	-	1,5	-	-	1,3	-	-	-	1,3	-

3.4 Esempi di calcolo degli errori campionari

3.4.1 Esempi relativi alle stime delle famiglie

Esempio 1

Nel 2007, il numero delle famiglie in Lombardia che possiedono una connessione a Internet di tipo tradizionale è 527.000.

Si cerca il livello di stima che più si avvicina a 527.000 nella prima colonna del prospetto 3.

L'errore relativo percentuale della stima considerata, per il totale Italia, è pari al 6,6%.

L'errore assoluto sarà: $\sigma(527.000) = 0,066 \times 527.000 = 34.782$

L'intervallo di confidenza avrà come estremi : $527.000 - (1,96 \times 34.782) = 492.218$
 $527.000 + (1,96 \times 34.782) = 561.782$

Esempio 2

Considerando la stima precedente si possono ottenere valori più precisi dell'errore di campionamento operando mediante interpolazione lineare dei due livelli di stima consecutivi tra i quali è compreso il valore della stessa.

Tali livelli sono 500.000 e 750.000 ai quali corrispondono i valori 6,6% e 3,8%.

L'errore relativo corrispondente a 527.000 è pari a :

$$\sigma(527.000) = 6,6 - (6,6 - 3,8) / (750.000 - 500.000) \times (527.000 - 500.000) = 6,3\%$$

Il corrispondente errore assoluto è $0,063 \times 527.000 = 33.188$ e l'intervallo di confidenza avrà come estremi:

$$527.000 - (1,96 \times 33.188) = 493.812$$

$$527.000 + (1,96 \times 33.188) = 560.188$$

Esempio 3

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

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

i cui parametri, riportati nel prospetto 2 alla riga Lombardia, sono i seguenti:

$$a = 8,92130881 \quad b = -1,09894370282613.$$

Per $\hat{Y} = 3.453.000$ si ha:

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{\exp(8,92130881 - 1,0950872 \times \log(527.000))} = 0,0637.$$

L'errore relativo percentuale è quindi pari al 6,37% e il calcolo dell'errore assoluto e dell'intervallo di confidenza è del tutto analogo a quello degli esempi 1 e 2.

3.4.2 Esempi relativi alle stime delle persone

Esempio 1

Nel 2007, le persone in Italia nella classe di età compresa tra i 20 e 24 anni che usano Internet sono 1.061.000.

Si cerca il livello di stima che più si avvicina a 1.061.000 nella prima colonna del prospetto 4 per il totale Italia.

L'errore relativo percentuale della stima considerata è pari a 4,0 %.

L'errore assoluto sarà: $\sigma(1.061.000) = 0,040 \times 1.061.000 = 42.440$

L'intervallo di confidenza avrà come estremi : $1.061.000 - (1,96 \times 42.440) = 1.018.560$

$$1.061.000 + (1,96 \times 42.440) = 1.103.440$$

Esempio 2

Considerando la stima precedente si possono ottenere valori più precisi dell'errore di campionamento operando mediante interpolazione lineare dei due livelli di stima consecutivi tra i quali è compreso il valore della stessa.

Tali livelli sono 1.000.000 e 2.000.000 ai quali corrispondono i valori percentuali 4,0 e 2,6.

L'errore relativo corrispondente a 2.085.000 è pari a :

$$\sigma(1.061.000) = 4,0 - (4,0 - 2,6) / (2.000.000 - 1.000.000) \times (1.061.000 - 1.000.000) = 3,91\%$$

Il corrispondente errore assoluto è

$$0,0391 \times 1.061.000 = 41.534$$

e l'intervallo di confidenza avrà come estremi:

$$1.061.000 - (1,96 \times 41.534) = 1.019.466$$

$$1.061.000 + (1,96 \times 41.534) = 1.102.534$$

Esempio 3

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

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

i cui parametri, riportati nel prospetto 2 alla riga Italia, sono i seguenti:

$$a = 9,59662740147983 \quad b = -1,1619918037424.$$

Per $\hat{Y} = 1.061.000$ si ha:

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{\exp(9,59662740147983 - 1,1619918037424 \times \log(1.061.000))} = 0,0383.$$

L'errore relativo percentuale è quindi pari al 3,83% e il calcolo dell'errore assoluto e dell'intervallo di confidenza è del tutto analogo a quello degli esempi 1 e 2.