

APPENDICE A

Strategia di campionamento e livello di precisione dei risultati nell'indagine di ritorno del 2007

“Criticità dei percorsi lavorativi in un’ottica di genere”

1. Obiettivi dell'indagine di ritorno

La *popolazione di interesse* dell'indagine in oggetto, ossia l'insieme delle unità statistiche intorno alle quali si intende investigare, è costituita dagli individui di età compresa tra 18 e 64 anni intervistati nell'ambito dell'indagine “Famiglia e soggetti sociali” del 2003 e appartenenti alle famiglie residenti in Italia, al netto dei 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.

L'obiettivo principale dell'indagine è quello di verificare, a distanza di 3 anni, quali sono stati i percorsi lavorativi e familiari degli individui; pertanto l'indagine fa riferimento alla popolazione *compresente* tra il 2003 e il 2007 ed è stata effettuata a partire dal campione intervistato per l'indagine “Famiglia e soggetti sociali”¹ nell'anno 2003 (circa 19.000 famiglie e 50.000 individui); da questo è stato estratto un campione di circa 10.000 individui per i quali, quindi, sono disponibili le informazioni riferite sia al 2003 che al 2007.

Il *periodo di riferimento* è prevalentemente il momento dell'intervista, ossia il 2007, con alcuni confronti con quanto dichiarato nel 2003 e accaduto nel triennio.

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);
- 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.

Il disegno di campionamento è un disegno complesso, definito in due fasi:

- la prima fase è costituita dal disegno campionario dell'indagine “Famiglia e soggetti sociali” del 2003 (descritto nel seguito);
- la seconda fase ha visto l'estrazione degli individui facenti parte dei 18-64enni che costituivano il campione dell'indagine “Famiglia e soggetti sociali” del 2003.

¹ Per ulteriori approfondimenti metodologici cfr. Appendice B

2. Prima fase di selezione del campione: l'indagine "Famiglia e soggetti sociali"²

2.1 Obiettivi conoscitivi dell'indagine

La *popolazione di interesse* dell'indagine "Famiglia e soggetti sociali", ossia l'insieme delle unità statistiche intorno alle quali si intende investigare, è costituita dalle famiglie residenti in Italia e dagli individui ad esse appartenenti, al netto dei 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 il momento dell'intervista (fine 2003).

I *domini di studio*, ossia gli ambiti rispetto ai quali sono riferiti i parametri di popolazione oggetto di stima, sono:

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

2.2 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 italiani sono suddivisi in due sottoinsiemi sulla base della popolazione residente:

- l'insieme dei comuni Auto Rappresentativi (che indicheremo d'ora in avanti come comuni AR) costituito dai comuni di maggiore dimensione demografica;
- l'insieme dei comuni Non Auto Rappresentativi (o NAR) costituito dai rimanenti comuni.

Nell'ambito dell'insieme dei comuni AR, ciascun comune viene considerato come uno strato a se stante e viene adottato un disegno 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.

2.3 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

² Per approfondimenti si veda la collana Informazioni n. 18 anno 2006 "Strutture familiari e opinioni su famiglia e figli" Indagine multiscopo sulle famiglie "Famiglia e soggetti sociali", Anno 2003.

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 20.000;
- il numero di comuni campione interessati non deve essere superiore a 900 in modo da consentire un buon lavoro di controllo e supervisione.

L'allocazione del campione di famiglie e di comuni tra le varie regioni è stata poi definita adottando un criterio di compromesso tale da garantire sia l'affidabilità delle stime a livello nazionale che quella delle stime a livello di ciascuno dei domini territoriali descritti nel paragrafo 4.2.1.

2.4 Stratificazione e selezione delle unità campionarie

Nell'indagine 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; per l'indagine in oggetto tale numero è stato posto pari a 23;
- formazione di strati aventi ampiezza approssimativamente costante in termini di popolazione residente.

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 Madow³.

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

3. Seconda fase di selezione del campione: l'indagine "Criticità dei percorsi lavorativi in un'ottica di genere"

A partire dal campione di individui in età compresa tra 18-64 anni selezionato nel 2003 con l'indagine "Famiglia e soggetti sociali", è stato estratto un campione casuale semplice di 10.000 individui.

La popolazione di riferimento dell'indagine (costituita dagli individui del 2003 che ancora sono residenti in Italia nel 2007) è stata definita a partire dalla popolazione del 2003, depurandola dalle uscite (morti e migrazioni) stimate, sulla base delle fonti ufficiali, nel triennio relativo. Le fonti ufficiali di cui ci si è avvalsi sono:

- i Bilanci demografici per il triennio in esame, per conoscere la popolazione residente e i decessi nel triennio;
- l'indagine campionaria sulle Cause di morte (anno 2004), per stimare la distribuzione dei morti per le sole classi di età 18-64 nel triennio⁴;
- i Trasferimenti di residenza per l'estero nel 2002-2004, per stimare la distribuzione degli usciti in età 18-64 dalla popolazione residente.

Nel prospetto 1 viene riportata la distribuzione per ripartizione dell'universo e del campione degli individui.

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

⁴ Le classi di età considerate sono: 18-24, 25-34, 35-44, 45-54, 55-64.

Prospetto 1 - Distribuzione per ripartizione geografica degli individui nell'universo e nel campione - Anno 2007

RIPARTIZIONI	Universo (a)	Campione
Nord-Ovest	9.563.468	2.386
Nord-Est	6.830.680	2.403
Centro	6.888.013	1.906
Sud	8.683.617	2.488
Isole	4.128.546	814
Italia	36.094.324	9.997

(a) Stima della popolazione comprese 2003-2007

4. Procedimento per il calcolo delle stime

Le stime prodotte dall'indagine sono essenzialmente stime di frequenze assolute e relative, riferite agli individui. Le stime dell'indagine sono state 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 rappresentate 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 individuo;
- h, indice di strato di comuni;
- y, generica variabile oggetto di indagine;
- Y_{hij} , valore di y osservato sull'individuo j del comune i dello strato h;
- M_{hi} , numero di individui nel comune i dello strato h;
- m_{hi} , campione di individui nel comune i dello strato h;
- N_h , totale di comuni nello strato h;
- n_h , numero di comuni campione nello strato h;
- 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

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

in cui W_{hij} è il peso finale da attribuire agli individui j del comune i dello strato h.

Dalla precedente relazione si desume, quindi, che per ottenere la stima del totale Y (1) occorre moltiplicare il valore della variabile y assunto da ciascuna unità campionaria per il peso di tale unità 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 alcuni individui selezionati 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 ripartizione geografica 10 totali noti, che si riferiscono alla distribuzione della popolazione per sesso e cinque classi di età. Indicando, quindi, con ${}_kX$ ($k=1, \dots, 10$) il totale noto della k-esima variabile ausiliaria per la generica ripartizione geografica e con ${}_kX_{hij}$ il valore assunto dalla k-esima variabile ausiliaria per l'individuo rispondente hij , la condizione sopra descritta è espressa dalla seguente uguaglianza

$${}_kX = \hat{X} = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij} {}_kX_{hij} \quad (k=1, \dots, 10)$$

in cui H indica il numero complessivo di strati definiti nella ripartizione.

Poiché la popolazione di riferimento dell'indagine è costituita dagli individui in età 18-64 anni del 2003, che ancora sono residenti in Italia nel 2007, il calcolo dei pesi finali per l'indagine ha richiesto la conoscenza dei totali noti riferiti a tale popolazione. In base alle fonti ufficiali disponibili è stato possibile ricostruire i tali totali noti a livello di ripartizione, sesso e classi di età⁵.

Il calcolo dei *pesi finali* da attribuire alle unità campionarie intervistate nel 2007 avviene per due passi successivi:

- passo 1: calcolo dei pesi finali dell'indagine 2003
- passo 2: calcolo dei pesi finali dell'indagine 2007 a partire dai pesi finali dell'indagine 2003.

Passo 1. Calcolo dei *pesi finali* da attribuire alle unità campionarie rispondenti nel 2003:

- 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 punto 4 del passo 1.

Passo 2. Calcolo dei *pesi finali* da attribuire alle unità campionarie rispondenti dell'indagine 2007:

- 6) i *pesi diretti* sono posti uguale ai *pesi finali* definiti nel passo 1 (indagine 2003);
- 7) si costruiscono i fattori correttivi che consentono di soddisfare, a livello di ripartizione, sesso e classe di età, la condizione di uguaglianza tra i totali noti (calcolati sulla popolazione compresente 2003-2007) delle variabili ausiliarie e le corrispondenti stime campionarie, attraverso la risoluzione di un problema di minimo vincolato;
- 8) si calcolano, infine, i *pesi finali* mediante il prodotto dei pesi base per i fattori correttivi ottenuti al punto 7 del passo 2.

I fattori correttivi dei passi 4 e 7 sono ottenuti dalla risoluzione di un problema di minimo vincolato, in cui la funzione da minimizzare è una funzione di distanza (opportunosamente 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

⁵ Le classi di età considerate sono: 18-24, 25-34, 35-44, 45-54, 55-64.

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 vincolata⁶. Un importante stimatore appartenente a tale classe, che si ottiene utilizzando la funzione di distanza euclidea, è lo *stimatore di regressione generalizzata*⁷. Tale stimatore riveste un ruolo centrale in quanto è possibile dimostrare che tutti gli stimatori di ponderazione vincolata convergono asintoticamente, all'aumentare della numerosità campionaria, allo stimatore di regressione generalizzata.

5. Valutazione del livello di precisione delle stime

5.1 Metodologia di calcolo degli errori campionari

Le principali statistiche di interesse per valutare la variabilità campionaria delle stime prodotte da un'indagine sono l'errore di campionamento assoluto e l'errore di campionamento relativo. Indicando con $\hat{V}ar(\hat{Y}_d)$ la stima della varianza della generica stima \hat{Y}_d , la stima dell'errore di campionamento assoluto di \hat{Y}_d si può ottenere mediante la seguente espressione

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

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 4.4., le stime prodotte dall'indagine sono state ottenute mediante uno stimatore di ponderazione vincolata definito in base ad una funzione di distanza di tipo logaritmico troncato. Poiché, lo stimatore adottato non è funzione lineare dei dati campionari, per la stima della varianza $\hat{V}ar(\hat{Y}_d)$ si è utilizzato il metodo proposto da Woodruff; in base a tale metodo, che ricorre all'espressione linearizzata in serie di Taylor, è possibile ricavare la varianza di ogni stimatore non lineare (funzione regolare di totali) calcolando la varianza dell'espressione linearizzata ottenuta. In particolare, per la definizione dell'espressione linearizzata dello stimatore ci si è riferiti allo stimatore di regressione generalizzata, sfruttando la convergenza asintotica di tutti gli stimatori di ponderazione vincolata a tale stimatore, poiché nel caso di stimatori di ponderazione vincolata che utilizzano funzioni distanza differenti dalla distanza euclidea (che conduce allo stimatore di regressione generalizzata) non è possibile derivare l'espressione linearizzata dello stimatore. L'espressione linearizzata dello stimatore (2) è data, quindi, da

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

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 variabili ausiliarie, osservati per il generico individuo hij e $\hat{\beta}$, il vettore dei coefficienti di regressione del modello lineare che lega la variabile di interesse y alle K variabili ausiliarie x. In base alla (5), si ha, quindi, che la stima della varianza della stima \hat{Y}_d è ottenuta mediante la seguente relazione

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

⁷ Deville J.C., Samdal C.E. (1992) "Calibration Estimators in Survey Sampling", Journal of the American Statistical Association, vol. 87, pp. 376-382.

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

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

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

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 omesso) la varianza è stimata mediante la seguente espressione

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

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

Negli strati 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

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

dove le quantità sono espresse come

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

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

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

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

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

5.2 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 ad una presentazione sintetica degli errori relativi, basata sul *metodo dei modelli regressivi*. Questo metodo si basa sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore relativo.

Nella presente indagine, il modello utilizzato per le stime di frequenze assolute e relative, è del tipo seguente:

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

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 e tipologia comunale.

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 alle persone dell'Italia Nord Occidentale, l'errore relativo corrispondente si ottiene introducendo nella (12) i valori dei parametri a e b riportati nella seconda riga del prospetto 2 (a = 9,51202, b = -1,07702).

Il prospetto 3, presentato in aggiunta, consente di rendere più agevole il calcolo degli errori campionari. Esso riguarda gli individui ed ha 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 R² (%) delle funzioni utilizzate per le interpolazioni degli errori campionari delle stime riferite agli individui per totale Italia, ripartizione geografica e tipologia comunale

ZONE TERRITORIALI	PERSONE		
	a	b	R ² (%)
ITALIA	10,38642	-1,13804	96,8
RIPARTIZIONI GEOGRAFICHE (a)			
Nord-ovest	9,51202	-1,07702	95,3
Nord-est	10,14542	-1,15872	95,5
Centro	10,11610	-1,13938	95,2
Sud	10,10136	-1,13911	93,9
Isole	10,03172	-1,11533	95,1
TIPI DI COMUNE (b)			
A1	10,08305	-1,11992	95,8
A2	10,08882	-1,12098	93,1
B1	9,63746	-1,14173	93,7
B2	9,51349	-1,10200	95,2
B3	9,86373	-1,11273	95,3
B4	10,06504	-1,14151	96,0

- (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 agli individui per totale Italia, ripartizione geografica e tipo di comune

STIME	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	A1	A2	B1	B2	B3	B4
20.000	64,3	56,2	51,4	55,8	55,4	60,2	60,4	60,3	43,4	49,7	56,1	53,8
50.000	38,2	34,3	30,2	33,1	32,9	36,1	36,2	36,1	25,7	30,0	33,7	31,9
60.000	34,4	31,1	27,2	29,8	29,7	32,6	32,7	32,6	23,2	27,1	30,4	28,7
70.000	31,5	28,6	24,9	27,3	27,2	30,0	30,0	29,9	21,2	24,9	27,9	26,3
80.000	29,2	26,6	23,0	25,3	25,2	27,8	27,8	27,7	19,7	23,1	25,9	24,4
90.000	27,3	25,0	21,5	23,7	23,5	26,0	26,0	25,9	18,4	21,7	24,3	22,8
100.000	25,7	23,6	20,2	22,3	22,2	24,5	24,5	24,5	17,3	20,5	22,9	21,5
200.000	17,3	16,3	13,5	15,0	14,9	16,7	16,6	16,6	11,7	14,0	15,6	14,5
300.000	13,8	13,1	10,7	11,9	11,9	13,3	13,3	13,2	9,2	11,2	12,4	11,5
400.000	11,7	11,2	9,1	10,1	10,1	11,3	11,3	11,2	7,8	9,5	10,6	9,7
500.000	10,3	9,9	8,0	8,9	8,9	10,0	10,0	9,9	6,9	8,4	9,4	8,6
750.000	8,2	8,0	6,3	7,1	7,0	8,0	7,9	7,9	5,5	6,7	7,5	6,8
1.000.000	6,9	6,8	5,3	6,0	6,0	6,8	6,8	6,7	4,7	5,8	6,4	5,8
2.000.000	4,7	4,7	3,6	4,0	4,0	4,6	4,6	4,6	3,1	3,9	4,3	3,9
3.000.000	3,7	3,8	2,8	3,2	3,2	3,7	3,7	3,6	2,5	3,1	3,5	3,1
4.000.000	3,2	3,2	2,4	2,7	2,7	3,1	3,1	3,1	2,1	2,7	2,9	2,6
5.000.000	2,8	2,9	2,1	2,4	2,4	2,8	2,7	2,7	1,9	2,4	2,6	2,3
7.500.000	2,2	2,3	1,7	1,9	1,9	2,2	2,2	2,2	1,5	1,9	2,1	1,8
10.000.000	1,9	2,0	1,4	1,6	1,6	1,9	1,9	1,9	1,2	1,6	1,8	1,5
15.000.000	1,5	1,6	1,1	1,3	1,3	1,5	1,5	1,5	1,0	1,3	1,4	1,2

APPENDICE B

Strategia di campionamento e livello di precisione dei risultati nell'indagine multiscopo del 2003 “Famiglia e soggetti sociali”

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 dagli individui ad esse appartenenti, al netto dei 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 il momento dell'intervista.

I *domini di studio*, ossia gli ambiti rispetto ai quali sono riferiti i parametri di popolazione oggetto di stima, sono:

- l'intero territorio nazionale;
- le cinque ripartizioni geografiche (Italia Nord-Occidentale, Italia Nord-Orientale, Italia Centrale, Italia Meridionale, Italia Insulare);
- le regioni geografiche (ad eccezione del Trentino Alto Adige le cui stime sono prodotte separatamente per le province di Bolzano e Trento);
- la tipologia comunale ottenuta suddividendo i comuni italiani in 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

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 italiani sono suddivisi in due sottoinsiemi sulla base della popolazione residente:

- l'insieme dei comuni Auto Rappresentativi (che indicheremo d'ora in avanti come comuni Ar) costituito dai comuni di maggiore dimensione demografica;
- l'insieme dei comuni Non Auto Rappresentativi (o Nar) costituito dai rimanenti comuni.

Nell'ambito dell'insieme dei comuni Ar, ciascun comune viene considerato come uno strato a se stante e viene adottato un disegno 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.

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;
- 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 poi definita 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 primo paragrafo.

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 Multiscopo, 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; per l'indagine in oggetto 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:

$${}_r\lambda = \frac{{}_r\bar{m} \cdot {}_r\delta}{{}_r f}$$

in cui per la generica regione geografica r si è indicato con: ${}_r\bar{m}$ 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 A_r e Nar : i comuni di dimensione superiore o uguale a ${}_r\lambda$ 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 ${}_r\lambda$.

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 Madow⁸.

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 :

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

⁸ Madow, W.G. (1949) "On the theory of systematic sampling II", Ann. Math. Stat., 20, 333-354

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

Prospetto 1 - Distribuzione regionale dei comuni, delle famiglie e delle persone nell'universo e nel campione- Anno 2003

REGIONI	Comuni		Famiglie		Individui	
	Universo	Campione	Universo (a)	Campione	Universo	Campione
Piemonte	1.206	59	1.820.856	1.338	4.212.593	3.154
Valle d'Aosta	74	20	53.240	480	120.618	1.121
Lombardia	1.546	78	3.705.353	1.658	9.120.847	4.006
Bolzano	116	22	175.523	585	465.055	1.432
Trento	223	24	200.345	554	482.804	1.444
Veneto	581	51	1.723.530	1.128	4.575.134	2.974
Friuli-Venezia Giulia	219	30	509.761	673	1.182.458	1.650
Liguria	235	25	711.480	818	1.560.949	1.776
Emilia-Romagna	341	45	1.704.107	1.121	4.030.978	2.675
Toscana	287	49	1.417.327	1.168	3.514.253	2.929
Umbria	92	22	319.724	610	837.778	1.601
Marche	246	35	570.384	832	1.489.232	2.138
Lazio	378	25	2.161.262	1.016	5.130.141	2.347
Abruzzo	305	34	474.534	763	1.275.059	2.069
Molise	136	23	123.424	579	319.925	1.476
Campania	551	47	1.936.331	1.223	5.730.469	3.640
Puglia	258	48	1.422.253	1.152	4.020.911	3.370
Basilicata	131	24	210.288	578	594.441	1.590
Calabria	409	41	718.352	965	2.001.799	2.683
Sicilia	390	49	1.820.087	1.202	4.974.247	3.313
Sardegna	377	36	582.379	784	1.633.240	2.153
Italia	8.101	787	22.360.539	19.227	57.272.931	49.541

(a) Stima Indagine Multiscopo

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 rappresentate 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

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

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 unità⁹ 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 età¹⁰ e della popolazione regionale nelle sei aree A_1, A_2, B_1, B_2, B_3 e B_4 . Indicando, quindi, con ${}_k X$ ($k=1, \dots, 18$) il totale noto della k -esima variabile ausiliaria per la generica regione geografica e con ${}_k X_{hij}$ il valore assunto dalla k -esima variabile ausiliaria per la famiglia rispondente hij , la condizione sopra descritta è espressa dalla seguente uguaglianza

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

in cui H indica il numero complessivo di strati definiti nella regione. Se, ad esempio, ${}_6 X$ indica il numero di maschi di età maggiore o uguale a sessantacinque anni, la variabile ausiliaria ${}_6 X_{hij}$ 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).

⁹ 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} .

¹⁰ Le classi di età considerate sono: 0-5, 6-13, 14-24, 25-44, 45-64, più di 65 anni.

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 vincolata¹¹. 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 successivo, tale stimatore riveste un ruolo centrale in quanto è possibile dimostrare¹² 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

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 precedente, le stime prodotte dall'indagine sono state ottenute mediante uno stimatore di ponderazione vincolata definito in base ad 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

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

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 $\hat{\beta}$, 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

¹¹ Nella letteratura in lingua anglosassone sull'argomento tali stimatori sono noti come *calibration estimators*.

¹² Deville J.C., Sarndal C.E. (1992) "Calibration Estimators in Survey Sampling", Journal of the American Statistical Association, vol. 87, pp. 376-382.

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

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 omesso) la varianza è stimata mediante la seguente espressione

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

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

Negli strati 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

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

dove le quantità sono espresse come

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

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

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

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

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

Fondamenti statistici della procedura per il calcolo degli errori campionari

Per il calcolo degli errori di campionamento delle indagini condotte dall'Istat sulle famiglie e sulle imprese viene correntemente utilizzata una procedura informatica sviluppata nell'ambito dell'Istituto. Nel paragrafo precedente è stata descritta la metodologia, implementata dalla procedura, per il calcolo degli errori di campionamento delle stime prodotte dall'indagine mentre, nel presente paragrafo, vengono discussi i fondamenti statistici e i limiti della metodologia medesima.

Negli strati Ar, nei quali si adotta un disegno di campionamento a grappoli e in cui le unità primarie (le famiglie) vengono selezionate senza reimmissione e probabilità uguali, la procedura consente di ottenere stime

della varianza campionaria che risultano corrette.

Negli strati Nar, per i quali si adotta un disegno di campionamento a due stadi con selezione delle unità primarie (comuni) senza reimmissione e probabilità variabili, la procedura consente di ottenere stime corrette della varianza campionaria qualora:

- in ciascuno strato sono selezionate due o più unità primarie;
- le unità primarie sono scelte mediante estrazioni indipendenti.

La prima condizione non viene soddisfatta in quanto, nell'indagine in oggetto, da ciascuno strato viene selezionato un solo comune campione e per stimare la varianza di campionamento si ricorre alla tecnica di *collassamento degli strati*. Questa tecnica, che consiste nel formare superstrati contenenti ciascuno un numero di strati maggiore di uno, conduce in generale ad una sovrastima della varianza di campionamento effettiva.

La seconda ipotesi implica che la selezione delle unità primarie venga effettuata con reimmissione. Anche questa assunzione non è soddisfatta per i comuni Nar e ciò comporta una sovrastima della varianza. Si osservi, tuttavia, che tale sovrastima dipende dalla frazione di campionamento di ciascuno strato Nar: è di entità trascurabile negli strati nei quali la frazione di campionamento è piccola, mentre viceversa può risultare di entità più cospicua per quegli strati in cui la frazione di campionamento è maggiore.

Presentazione sintetica degli errori campionari

Ad ogni stima \hat{Y}_d corrisponde un errore di campionamento relativo $\hat{\varepsilon}(\hat{Y}_d)$; ciò significa che per consentire una lettura corretta delle tabelle pubblicate sarebbe necessario presentare per ogni stima pubblicata il corrispondente errore di campionamento relativo. Ciò, tuttavia, non è possibile sia per limiti di tempo e di costi di elaborazione, sia perché le tavole della pubblicazione risulterebbero appesantite e di non facile consultazione per l'utente finale. Inoltre, non sarebbero comunque disponibili gli errori delle stime non pubblicate, che l'utente può ricavare in modo autonomo.

Per le ragioni sopra esposte, si ricorre frequentemente ad una presentazione sintetica degli errori relativi, basata sul *metodo dei modelli regressivi*. Questo metodo si basa sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore relativo.

Nella presente indagine, il modello utilizzato per le stime di frequenze assolute e relative, è del tipo seguente:

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

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 alle persone dell'Italia Nord Occidentale, l'errore relativo corrispondente si ottiene introducendo nella (12) i valori dei parametri a e b riportati nella seconda riga del prospetto 2 alla voce PERSONE (a = 9,352384, b = -1,138993).

I prospetti 3 e 4, presentati in aggiunta, consentono di rendere più agevole il calcolo degli errori campionari. Essi riguardano, rispettivamente, le famiglie e le persone ed hanno la seguente struttura: a) in fiancata sono elencati i valori crescenti di stima (20.000, 30.000, ..., 25.000.000); b) le colonne successive contengono gli errori di campionamento relativo, per ciascun dominio territoriale di interesse, calcolati mediante la formula (12), corrispondenti alle stime di frequenze assolute della prima colonna.

Le informazioni contenute in tali prospetti permettono di calcolare l'errore relativo di una generica stima di frequenza assoluta (o relativa) mediante due procedimenti che risultano di facile applicazione, anche se conducono a risultati meno precisi di quelli ottenibili mediante l'espressione (12). Il primo metodo consiste nell'individuare, nella prima colonna del prospetto, il livello di stima che più si avvicina alla stima di interesse e nel considerare come errore relativo il valore che si trova sulla stessa riga, nella colonna corrispondente al dominio territoriale di riferimento.

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

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

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 R² (%) delle funzioni utilizzate per le interpolazioni degli errori campionari delle stime riferite alle FAMIGLIE e alle PERSONE per totale Italia, ripartizione geografica, tipo di comune e regione

ZONE TERRITORIALI	Famiglie			Persone		
	a	b	R ² (%)	a	b	R ² (%)
Italia	8,659646	-1,096289	95,2	9,371995	-1,136573	84,2
RIPARTIZIONI GEOGRAFICHE (a)						
Nord-ovest	8,581638	-1,091408	95,9	9,352384	-1,138993	87,4
Nord-est	8,178906	-1,083047	95,0	8,661908	-1,113448	83,6
Centro	7,878008	-1,034465	90,8	8,956406	-1,116919	80,3
Sud	7,774754	-1,057153	93,9	8,551253	-1,104758	83,9
Isole	7,808933	-1,046630	92,9	8,374141	-1,079396	80,4
TIPI DI COMUNE (b)						
A1	8,937190	-1,132329	97,2	9,326560	-1,147439	87,1
A2	7,969334	-1,035100	90,6	8,759367	-1,088582	80,7
B1	6,721240	-0,996985	90,5	8,774188	-1,189391	87,8
B2	8,797890	-1,129127	93,9	10,261337	-1,237353	88,6
B3	8,377845	-1,072053	91,5	9,967681	-1,192492	84,5
B4	8,713524	-1,144521	97,1	8,737984	-1,133256	88,0
REGIONI						
Piemonte	8,658422	-1,138033	95,3	8,655898	-1,127006	86,9
Valle d'Aosta	5,246402	-1,093407	95,0	5,594850	-1,127750	91,2
Lombardia	8,573054	-1,075883	95,4	9,143921	-1,105513	87,3
- Bolzano	6,213041	-1,074103	95,1	7,081878	-1,151113	90,0
- Trento	7,031645	-1,138987	90,7	6,506914	-1,078298	80,9
Veneto	8,135617	-1,069791	94,2	8,467400	-1,088644	82,5
Friuli-Venezia Giulia	7,640448	-1,105811	92,8	7,478353	-1,084928	88,5
Liguria	7,758562	-1,110095	95,0	7,859412	-1,100659	87,5
Emilia-Romagna	8,263197	-1,093671	94,6	8,557233	-1,105402	84,8
Toscana	8,198323	-1,113092	95,4	8,453074	-1,120608	87,1
Umbria	7,118840	-1,114647	96,1	7,287622	-1,109017	86,5
Marche	7,294788	-1,091944	95,6	7,850890	-1,127638	86,9
Lazio	8,092067	-1,026263	87,7	8,635640	-1,065452	78,1
Abruzzo	7,148910	-1,076441	92,8	7,500997	-1,096177	87,3
Molise	5,652458	-1,034606	92,2	6,037476	-1,066349	87,3
Campania	7,865277	-1,045245	91,8	7,823541	-1,022658	82,8
Puglia	8,082287	-1,097802	93,9	8,190990	-1,084808	82,4
Basilicata	7,259191	-1,170136	95,1	7,745373	-1,198747	91,1
Calabria	7,735716	-1,127663	94,6	8,071050	-1,143025	92,6
Sicilia	8,425014	-1,092515	93,0	8,176027	-1,050251	80,6
Sardegna	6,672450	-1,003203	93,2	7,198364	-1,035618	86,1

- (c) 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.
- (d) 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	33,3	32,8	28,0	30,6	26,0	27,9	32,0	32,0	20,7	30,4	32,6	27,0
30.000	26,7	26,3	22,5	24,8	21,0	22,5	25,5	25,9	16,9	24,1	26,3	21,4
40.000	22,8	22,5	19,2	21,4	18,0	19,4	21,6	22,3	14,6	20,5	22,5	18,1
50.000	20,2	19,9	17,0	19,1	16,0	17,2	19,1	19,9	13,1	18,1	20,0	16,0
60.000	18,3	18,0	15,4	17,3	14,5	15,7	17,2	18,1	12,0	16,3	18,1	14,4
70.000	16,8	16,6	14,2	16,0	13,4	14,5	15,8	16,7	11,1	15,0	16,7	13,2
80.000	15,6	15,4	13,2	15,0	12,5	13,5	14,6	15,6	10,4	13,9	15,5	12,2
90.000	14,6	14,5	12,4	14,1	11,7	12,7	13,7	14,7	9,8	13,0	14,6	11,4
100.000	13,8	13,6	11,7	13,3	11,1	12,0	12,9	13,9	9,3	12,2	13,8	10,7
200.000	9,4	9,3	8,0	9,3	7,7	8,3	8,7	9,7	6,6	8,3	9,5	7,2
300.000	7,6	7,5	6,5	7,5	6,2	6,8	6,9	7,9	5,4	6,6	7,6	5,7
400.000	6,5	6,4	5,5	6,5	5,3	5,8	5,9	6,8	4,6	5,6	6,6	4,9
500.000	5,7	5,7	4,9	5,8	4,7	5,2	5,2	6,0	4,2	4,9	5,8	4,3
750.000	4,6	4,5	3,9	4,7	3,8	4,2	4,1	4,9	3,4	3,9	4,7	3,4
1.000.000	3,9	3,9	3,4	4,0	3,3	3,6	3,5	4,2	2,9	3,3	4,0	2,9
2.000.000	2,7	2,7	2,3	2,8	2,3	2,5	2,4	2,9	2,1	2,3	2,8	1,9
3.000.000	2,1	2,1	1,9	2,3	1,8	2,0	1,9	2,4	1,7	1,8	2,2	1,5
4.000.000	1,8	1,8	1,6	2,0	1,6	-	1,6	2,1	1,5	1,5	1,9	1,3
5.000.000	1,6	1,6	-	-	1,4	-	1,4	1,8	1,3	1,3	1,7	1,1
7.500.000	1,3	-	-	-	-	-	1,1	1,5	1,1	1,1	1,4	0,9
10.000.000	1,1	-	-	-	-	-	-	-	-	-	-	-
15.000.000	0,9	-	-	-	-	-	-	-	-	-	-	-
20.000.000	0,8	-	-	-	-	-	-	-	-	-	-	-

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

STIME	Piemonte	Valle d'Aosta	Lombardia	Bolzano	Trento	Veneto	Friuli- Venezia Giulia	Liguria	Emilia Romagna	Toscana	Umbria
20.000	27,1	6,1	35,3	10,9	12,0	29,2	19,1	19,8	27,7	24,4	14,1
30.000	21,5	4,9	28,4	8,8	9,5	23,5	15,3	15,8	22,2	19,4	11,2
40.000	18,3	4,2	24,3	7,5	8,1	20,2	13,0	13,5	19,0	16,6	9,6
50.000	16,1	3,7	21,6	6,7	7,1	17,9	11,5	11,9	16,8	14,6	8,5
60.000	14,5	-	19,6	6,1	6,4	16,2	10,4	10,8	15,2	13,2	7,6
70.000	13,3	-	18,0	5,6	5,9	15,0	9,6	9,9	14,0	12,1	7,0
80.000	12,3	-	16,8	5,2	5,4	13,9	8,9	9,2	13,0	11,3	6,5
90.000	11,5	-	15,7	4,9	5,1	13,1	8,3	8,6	12,2	10,5	6,1
100.000	10,8	-	14,9	4,6	4,8	12,4	7,8	8,1	11,5	9,9	5,7
200.000	7,3	-	10,2	-	-	8,5	5,3	5,5	7,9	6,8	3,9
300.000	5,8	-	8,2	-	-	6,9	4,3	4,4	6,3	5,4	3,1
400.000	4,9	-	7,0	-	-	5,9	3,6	3,8	5,4	4,6	-
500.000	4,3	-	6,3	-	-	5,2	3,2	3,3	4,8	4,1	-
750.000	3,4	-	5,0	-	-	4,2	-	-	3,8	3,2	-
1.000.000	2,9	-	4,3	-	-	3,6	-	-	3,3	2,8	-
2.000.000	2,0	-	3,0	-	-	-	-	-	-	-	-

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

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

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

STIME	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	A1	A2	B1	B2	B3	B4
20.000	39,0	38,1	30,6	34,9	30,3	31,4	36,1	36,4	22,3	36,9	39,8	28,9
30.000	31,0	30,3	24,5	27,8	24,2	25,2	28,6	29,2	17,5	28,7	31,3	22,9
40.000	26,3	25,7	20,8	23,7	20,6	21,6	24,3	25,0	14,7	24,0	26,3	19,5
50.000	23,2	22,6	18,4	20,9	18,3	19,2	21,3	22,1	12,9	20,9	23,1	17,2
60.000	20,9	20,4	16,6	18,9	16,5	17,4	19,2	20,0	11,6	18,7	20,7	15,5
70.000	19,1	18,7	15,3	17,3	15,2	16,0	17,6	18,4	10,6	17,0	18,9	14,2
80.000	17,7	17,3	14,2	16,1	14,1	14,9	16,3	17,1	9,8	15,7	17,4	13,2
90.000	16,6	16,2	13,3	15,1	13,2	14,0	15,2	16,1	9,1	14,6	16,2	12,3
100.000	15,6	15,3	12,5	14,2	12,4	13,2	14,3	15,2	8,5	13,6	15,2	11,6
200.000	10,5	10,3	8,5	9,6	8,5	9,1	9,6	10,4	5,7	8,9	10,1	7,8
300.000	8,4	8,2	6,8	7,7	6,8	7,3	7,6	8,3	4,4	6,9	7,9	6,2
400.000	7,1	6,9	5,8	6,6	5,8	6,2	6,5	7,1	3,7	5,8	6,7	5,3
500.000	6,3	6,1	5,1	5,8	5,1	5,5	5,7	6,3	3,3	5,0	5,8	4,7
750.000	5,0	4,8	4,1	4,6	4,1	4,4	4,5	5,1	2,6	3,9	4,6	3,7
1.000.000	4,2	4,1	3,5	3,9	3,5	3,8	3,8	4,3	2,2	3,3	3,9	3,1
2.000.000	2,8	2,8	2,4	2,7	2,4	2,6	2,6	3,0	1,4	2,1	2,6	2,1
3.000.000	2,3	2,2	1,9	2,1	1,9	2,1	2,0	2,4	1,1	1,7	2,0	1,7
4.000.000	1,9	1,9	1,6	1,8	1,6	1,8	1,7	2,0	1,0	1,4	1,7	1,4
5.000.000	1,7	1,6	1,4	1,6	1,4	1,6	1,5	1,8	0,8	1,2	1,5	1,3
7.500.000	1,3	1,3	1,1	1,3	1,1	-	1,2	-	-	0,9	1,2	1,0
10.000.000	1,1	1,1	-	-	1,0	-	-	-	-	0,8	1,0	-
15.000.000	0,9	-	-	-	-	-	-	-	-	-	-	-
20.000.000	0,8	-	-	-	-	-	-	-	-	-	-	-
25.000.000	0,7	-	-	-	-	-	-	-	-	-	-	-

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	Piemonte	Valle d'Aosta	Lombardia	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Liguria	Emilia Romagna	Toscana	Umbria
27.000	24,1	5,2	34,4	9,7	10,6	26,7	16,6	18,5	25,6	22,5	13,3
30.000	22,7	4,9	32,4	9,1	10,0	25,2	15,7	17,5	24,2	21,2	12,6
40.000	19,3	4,2	27,7	7,7	8,5	21,6	13,4	14,9	20,6	18,1	10,7
50.000	17,0	3,7	24,4	6,8	7,6	19,1	11,9	13,2	18,2	15,9	9,5
60.000	15,4	3,3	22,1	6,1	6,9	17,3	10,8	11,9	16,5	14,4	8,6
70.000	14,1	3,0	20,3	5,6	6,3	15,9	9,9	11,0	15,1	13,2	7,9
80.000	13,1	2,8	18,9	5,2	5,9	14,8	9,2	10,2	14,1	12,3	7,3
90.000	12,2	2,6	17,7	4,9	5,5	13,9	8,6	9,6	13,2	11,5	6,8
100.000	11,5	2,5	16,7	4,6	5,2	13,1	8,2	9,0	12,4	10,8	6,5
200.000	7,8	-	11,4	3,1	3,6	9,0	5,6	6,2	8,5	7,3	4,4
300.000	6,2	-	9,1	2,4	2,9	7,2	4,5	4,9	6,8	5,8	3,5
400.000	5,3	-	7,7	2,1	2,5	6,2	3,8	4,2	5,8	5,0	3,0
500.000	4,7	-	6,8	-	-	5,5	3,4	3,7	5,1	4,4	2,6
750.000	3,7	-	5,5	-	-	4,4	2,7	3,0	4,1	3,5	2,1
1.000.000	3,2	-	4,7	-	-	3,7	2,3	2,5	3,5	3,0	-
2.000.000	2,1	-	3,2	-	-	2,6	-	-	2,4	2,0	-
3.000.000	1,7	-	2,5	-	-	2,1	-	-	1,9	1,6	-
4.000.000	1,4	-	2,2	-	-	1,8	-	-	-	-	-
5.000.000	-	-	1,9	-	-	-	-	-	-	-	-

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
27.000	16,1	32,7	15,9	8,9	27,1	23,7	10,6	16,6	28,1	18,6
30.000	15,2	30,9	15,0	8,4	25,7	22,4	10,0	15,6	26,6	17,6
40.000	12,9	26,5	12,8	7,2	22,2	19,2	8,4	13,3	22,8	15,1
50.000	11,4	23,5	11,3	6,4	19,8	17,0	7,3	11,7	20,3	13,5
60.000	10,3	21,4	10,2	5,8	18,0	15,4	6,6	10,5	18,5	12,3
70.000	9,4	19,7	9,4	5,3	16,7	14,1	6,0	9,6	17,0	11,3
80.000	8,7	18,3	8,7	5,0	15,6	13,2	5,5	8,9	15,9	10,6
90.000	8,2	17,2	8,2	4,7	14,6	12,3	5,2	8,3	14,9	9,9
100.000	7,7	16,3	7,7	4,4	13,9	11,7	4,8	7,9	14,1	9,4
200.000	5,2	11,3	5,3	3,1	9,7	8,0	3,2	5,3	9,8	6,6
300.000	4,1	9,1	4,2	2,5	7,9	6,4	2,5	4,2	7,9	5,3
400.000	3,5	7,8	3,6	-	6,8	5,5	2,1	3,6	6,8	4,6
500.000	3,1	6,9	3,2	-	6,1	4,9	1,8	3,1	6,1	4,1
750.000	2,5	5,6	2,6	-	5,0	3,9	-	2,5	4,9	3,3
1.000.000	2,1	4,8	2,2	-	4,3	3,3	-	2,1	4,2	2,9
2.000.000	-	3,3	-	-	3,0	2,3	-	1,4	2,9	-
3.000.000	-	2,7	-	-	2,4	1,8	-	-	2,4	-
4.000.000	-	2,3	-	-	2,1	-	-	-	2,0	-
5.000.000	-	2,0	-	-	1,9	-	-	-	1,8	-

Esempi di calcolo degli errori campionari

Esempio 1

Dai dati risulta che in Toscana la stima del numero delle famiglie i cui membri hanno l'abitudine di farsi regali (non monetari) è pari a 843.000 unità.

Nella prima colonna del prospetto 3 della presente appendice metodologica, si cerca il valore più vicino a questa stima, che è pari a 750.000. In corrispondenza di tale valore, per la Toscana, è riportato un errore relativo percentuale del 3,2 per cento.

Pertanto, l'errore assoluto della stima sarà uguale a:

$$\sigma(843.000) = 0,032 \times 843.000 = 26.976$$

e l'intervallo di confidenza avrà come estremi:

$$843.000 - (1,96 \times 26.976) = 790.127$$

$$843.000 + (1,96 \times 26.976) = 895.873.$$

Esempio 2

Dai dati risulta che nel Lazio la stima del numero di persone che hanno fratelli e/o sorelle viventi è pari a 4.257.000 unità.

Nella prima colonna del prospetto 4, si cerca il valore più vicino a questa stima, che è pari a 4.000.000. In corrispondenza di tale valore, per il Lazio, è riportato un errore relativo percentuale del 2,3 per cento.

Pertanto, l'errore assoluto della stima sarà uguale a:

$$\sigma(4.257.000) = 0,023 \times 4.257.000 = 97.911$$

e l'intervallo di confidenza avrà come estremi:

$$4.257.000 - (1,96 \times 97.911) = 4.065.094$$

$$4.257.000 + (1,96 \times 97.911) = 4.448.906.$$

Esempio 3

Considerando la stima del numero delle famiglie in Toscana dell'esempio 1, 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, nel prospetto 3. Tali livelli sono 750.000 e 1.000.000 ai quali corrispondono, rispettivamente, i valori percentuali 3,2 e 2,8. L'errore relativo corrispondente a 843.000 è pari a:

$$\hat{\varepsilon}(843.000) = 3,2 - (3,2 - 2,8) \times (843.000 - 750.000) / (1.000.000 - 750.000) = 3,05 \text{ per cento.}$$

L'errore assoluto sarà il seguente:

$$\sigma(843.000) = 0,0305 \times 843.000 = 25.722$$

e l'intervallo di confidenza avrà come estremi:

$$843.000 - (1,96 \times 25.722) = 792.586$$

$$843.000 + (1,96 \times 25.722) = 893.414.$$

Esempio 4

Il calcolo dell'errore dell'esempio 1 può essere effettuato, direttamente, tramite la funzione interpolante:

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

i cui parametri, riportati nel prospetto 2 alla riga Toscana alla voce Famiglie, sono i seguenti:

$$a = 8,198323 \quad b = -1,113092.$$

Per $\hat{Y} = 843.000$ si ha:

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{\exp(8,198323 - 1,113092 \times \ln(843.000))} = 0,030356.$$

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