File Dati

Indagine Multiscopo sulle Famiglie

Sicurezza delle donne Anno 2006

Manuale utente e tracciato record

ISTAT

ISTITUTO NAZIONALE DI STATISTICA

INDAGINE MULTISCOPO SULLE FAMIGLIE
SICUREZZA DELLE DONNE ANNO 2006

DOCUMENTAZIONE TECNICA E DESCRIZIONE DEL FILE

1. 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 della Legge n. 675 del 31/12/1996 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.

L'indagine ha come obiettivo prioritario la rilevazione delle violenze sessuali e quindi indaga su fenomeni particolarmente rilevanti ma allo stesso tempo esigui quantitativamente. La rarità che li caratterizza comporta, quindi, che si ponga una particolare attenzione e cura nelle analisi da condurre sugli stessi.

Malgrado infatti la numerosità campionaria elevata e rappresentativa di tutto il territorio nazionale, le stime fatte dovranno essere validate con il calcolo degli errori campionari, diversi a seconda del dominio di stima, per poterne assicurare la significatività statistica da affiancare a quella contenutistica.

2. FINALITÀ E CARATTERISTICHE DELL'INDAGINE

Nel 2006 l'ISTAT ha condotto per la prima volta l'Indagine sulla sicurezza delle donne riguardante la rilevazione del fenomeno della violenza contro le donne in Italia in tutte le sue diverse forme, in termini di prevalenza ed incidenza, ed in particolare la violenza domestica; le caratteristiche di coloro che ne sono coinvolte e le conseguenze per le vittime; il numero oscuro delle violenze e le violenze subite prima dei 16 anni.

In particolare l'indagine fa luce sul sommerso delle violenze fisiche e sessuali e sulle modalità di accadimento delle stesse, permette di costruire il profilo delle vittime, fornisce notizie sul come, dove e quando queste sono state vittimizzate e cosa le espone di più.

Il disegno di campionamento ha previsto un campione casuale a due stadi con stratificazione delle unità primarie, nell'ambito della regione, per tipo di comune. Le unità primarie sono costituite dagli indirizzi telefonici presenti sull'Archivio informatizzato degli abbonati alla rete telefonica fissa. Le unità secondarie sono le donne (dai 16 anni ai 70 anni) che sono state estratte tra quelle facenti parte della famiglia estratta.

Le famiglie che sono state intervistate sono state 25.065.

Le informazioni sono state raccolte con intervista telefonica, mediante tecnica C.A.T.I., nel periodo gennaio – ottobre 2006. Le intervistatrici addette alla somministrazione dei questionari sono state circa 63.

Le unità di rilevazione sono le famiglie di fatto (FF) e le donne dai 16 ai 70 anni. La famiglia di fatto è definita come quell'insieme di persone che :

1. hanno la loro dimora abituale nella stessa abitazione e
2. sono legate da una relazione di parentela, affinità, affettività o amicizia.

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".

Il questionario è articolato in sezioni, che facilitano la possibilità di rintracciare i diversi contenuti al suo interno. Tre sezioni di screening hanno l'obiettivo di rilevare se la donna abbia subito violenza fisica o sessuale da un uomo non partner, da un partner attuale o da un partner precedente, nel caso affermativo alla donna viene richiesto, per ogni tipo di violenza subita, quante ne ha subite e il periodo in cui si è verificato l'ultimo episodio. Se questo si è verificato negli ultimi 12 mesi, quante volte la violenza è accaduta negli ultimi 12 mesi.

Nel caso di violenza subita prima dei 16 anni viene rilevato quante volte e con che frequenza è accaduta prima dei 16 anni di età, l'autore della violenza, la gravità e con chi ne ha parlato.

Due sezioni approfondiscono invece l'ultimo episodio della violenza subita, una per il non partner ed una per il partner. In quest'ultima vengono raccolte anche altre informazioni inerenti alcuni elementi caratteristici delle violenze ripetute.

Altre due sezioni raccolgono le informazioni sulle caratteristiche del partner attuale e sulle caratteristiche del partner precedente, soltanto se violento nei confronti della donna. All'interno di queste vi sono i quesiti sulla violenza psicologica e alcuni sulla violenza economica.

Infine la sezione H riguarda le domande inerenti la famiglia di origine, propria e dei partner, in merito alla violenza subita ed assistita.

Struttura del questionario:

O	REGOLE GENERALI
A	SCHEDA DI CONTATTO
B	ABITUDINI, CARATTERISTICHE E STATO CIVILE DELL'INTERVISTATA (a tutte le donne)
SCR_NP	SCREENING DI VIOLENZA SUBITA DA UN NON PARTNER (a tutte le donne)
REP_NP	REPORT DI VIOLENZA SUBITA DA UN NON PARTNER (se ci sono episodi di violenza da un non partner)
CRT_PR	CARATTERISTICHE DEL PARTNER ATTUALE/ ULTIMO PARTNER (se presente un partner attuale oppure se ha avuto un solo partner precedente)
SCR_PR	SCREENING DI VIOLENZA SUBITA DAL PARTNER ATTUALE/ULTIMO PARTNER (se presente un partner attuale oppure se ha avuto un solo partner precedente)
SCR_EX	SCREENING DI VIOLENZA SUBITA DAL PARTNER PRECEDENTE (se ha partner precedenti)
CRT_EX	CARATTERISTICHE DEL PARTNER PRECEDENTE (se ha partner precedenti VIOLENTI)
REP_PR	REPORT DI VIOLENZA SUBITA DA UN PARTNER (se ci sono episodi di violenza da partner attuale o precedente o ultimo partner)
H	STORIA DI VIOLENZE PREGRESSE SUBITE NELLA FAMIGLIA D'ORIGINE: DAI PARTNER DELL'INTERVISTATA E DALL'INTERVISTATA (a tutte le donne, filtro sui partner)
I	ALTRI COMPONENTI E CONCLUSIONI (se numero di componenti della famiglia maggiore di 1; conclusioni a tutte le donne)

3. AVVERTENZE PER L'UTILIZZAZIONE DEL FILE

Il file dati ha le seguenti caratteristiche:

Anno:	2006	
lunghezza record:	24.511	
numero records individuali:	25.065	(uno per ciascuna persona intervistata/famiglia)

4. DESCRIZIONE DEL TRACCIATO RECORD

Sul tracciato record viene descritta ogni singola variabile presente nel file dati[1]: la sua POSIZIONE NEL FILE rappresentata dal numero della colonna (o dalla colonna di inizio e di fine per la variabili più lunghe di un byte), la DESCRIZIONE DELLA VARIABILE, le MODALITÀ DI RISPOSTA e la loro CODIFICA o i corrispondenti CAMPI DI VARIAZIONE.

Oltre alle domande presenti sul questionario, sono rese disponibili nel tracciato record (nella sua parte iniziale) anche altre variabili utili ai fini della elaborazione: “ANNO DI INDAGINE”, “CODICE DI INDAGINE” e i coefficienti di riporto all’universo, le variabili che definiscono la struttura della famiglia e le caratteristiche strutturali dell’intervistata.

Altre variabili, invece, sono frutto di successive elaborazioni. In particolare, essendo presenti dei complessi filtri di accesso ad alcuni quesiti, le variabili filtro costruite vengono rilasciate nel file e sono rintracciabili nel tracciato record immediatamente prima della domanda di riferimento. Nel caso in cui lo stesso filtro riguardi PIÙ DOMANDE della stessa sezione, il filtro viene indicato una sola volta prima della prima domanda cui si riferisce.

Ad esempio:

col.	start	-	end	ETICHETTA_VARIABILE	NOMEVAR	N_COD	ETICHETTA_CODICE
col.	24471	-	24471	FILTRO QUESITI REP_PR16, 17, 25	FILTRO_rpraiuto	Codice	FILTRO_rpraiuto

Nel tracciato, inoltre, le variabili relative al “partner precedente” – caratteristiche e variabili di screening – raccolgono sia le variabili contenute nelle sezioni del questionario dedicate, appunto, al “partner precedente” che quelle contenute nelle sezioni dell’“ultimo partner”. Poiché con questa denominazione viene indicato un partner avuto in passato quando questi è anche l’unico partner (marito o convivente) avuto dalla donna, nell’intervista si è ritenuto più opportuno raccogliere i dati relativi a questo partner in maniera analoga a quelli di un “partner attuale”, ma poiché di fatto si tratta di una relazione affettiva che si è conclusa, i dati relativi vengono elaborati insieme a quelli degli ex partner. Si deve tenere presente che le caratteristiche dell’ultimo partner, così come quelle dei partner in generale precedenti, sono raccolte solo nel caso abbiano avuto comportamenti fisicamente o sessualmente violenti nei confronti della donna.

Ogni record contiene le variabili codice di indagine, anno di indagine, progressivo di famiglia, le informazioni territoriali - dominio, ripartizione geografica e/o regione - i coefficienti di riporto all’universo per le elaborazioni sugli individui, il numero dei componenti della famiglia e le caratteristiche relative a tutti i componenti della famiglia - che sono state fornite come proxy dalla donna intervistata. A queste seguono la struttura familiare, ricostruita in base alla relazione di parentela tra i componenti della famiglia (relazione di parentela nel nucleo, numero del nucleo se in presenza di famiglie con più nuclei, posizione del componente nel nucleo, tipo di nucleo, tipologia familiare). Seguono le caratteristiche della donna intervistata (sesso,

1 Si ricorda che il file dati è in formato ASCII e pertanto per la sua visualizzazione e/o elaborazione necessita dell’utilizzo di pacchetti statistici

anno di nascita, età – calcolata a partire dall'anno di nascita -, stato civile, titolo di studio, condizione professionale, lavoro svolto in passato, posizione nella professione, attività economica, dove ha vissuto l'adolescenza).

Riguardo l'ordinamento delle variabili nel tracciato record, si fa presente che nella descrizione del tracciato si è cercato di rispettare la sequenza e l'ordine delle domande nel questionario e non l'ordinamento della posizione nel file.

5. ISTRUZIONI SUI NUCLEI E LA TIPOLOGIA FAMILIARIE

Le variabili inerenti i nuclei e le tipologie familiari sono:

- **Variabile** tipo nucleo (foglio "Caratteristiche componenti" nel file excel del tracciato)

Il tipo nucleo è un codice volto a differenziare alcune particolari tipologie di nuclei.

TIPO NUCLEO	DESCRIZIONE
0	Persona isolata
1	Coppia con figli
2	Coppia senza figli
3	Monogenitore maschio
4	Monogenitore femmina

- **Variabile** numero nucleo (foglio "Caratteristiche componenti" nel file excel del tracciato)

Il numero nucleo è un progressivo da 0 (persona singola) a n, dove n è un numero intero che identifica tutti i componenti del medesimo nucleo.

- **Variabile** relazione di parentela nel nucleo (foglio "Caratteristiche componenti" nel file excel del tracciato)

La relazione di parentela nel nucleo specifica, all'interno di un nucleo, il rapporto genitore-figli e privilegia, ove possibile, la figura femminile attribuendole la qualifica di capo nucleo.

RELAZIONE NEL NUCLEO	DESCRIZIONE
0	Persona singola
1	Capo nucleo
2	Coniuge o convivente del capo nucleo
3	Figlio
4	Membro aggregato non figlio

- **Variabile** tipologia familiare (foglio "Caratteristiche componenti" nel file excel del tracciato)

La tipologia familiare è un campo di due caratteri volto a definire sinteticamente il tipo di famiglia. Le caratteristiche salienti sono: la presenza o l'assenza, all'interno della famiglia, di nuclei. Nel caso esistano nuclei, la differenza tra le famiglie composte da un nucleo da quelle composte da più nuclei. All'interno delle famiglie con uno o più nuclei raggruppa le famiglie a seconda che vivano in coppia, abbiano figli e siano presenti persone isolate.

FAMIGLIE SENZA NUCLEI

01 Persona sola
02 Genitore con figli non celibi/nubili
03 Insieme di parenti
04 Insieme di parenti più altri
05 Insieme di persone non parenti

FAMIGLIE CON UN SOLO NUCLEO

Nucleo senza persone isolate

06 Coppia coniugata senza figli
07 Coppia non coniugata senza figli
08 Coppia coniugata con figli
09 Coppia non coniugata con figli
10 Genitore maschio celibe solo con figli
11 Genitore maschio coniugato non convivente solo con figli
12 Genitore maschio separato solo con figli
13 Genitore maschio divorziato solo con figli
14 Genitore maschio vedovo solo con figli
15 Genitore femmina nubile solo con figli
16 Genitore femmina coniugato non convivente solo con figli
17 Genitore femmina separato solo con figli
18 Genitore femmina divorziato solo con figli
19 Genitore femmina vedovo solo con figli

Nucleo con persone isolate

20 Coppia coniugata senza figli
21 Coppia non coniugata senza figli
22 Coppia coniugata con figli
23 Coppia non coniugata con figli
24 Genitore maschio celibe solo con figli
25 Genitore maschio coniugato non convivente solo con figli
26 Genitore maschio separato solo con figli
27 Genitore maschio divorziato solo con figli
28 Genitore maschio vedovo solo con figli
29 Genitore femmina nubile solo con figli
30 Genitore femmina coniugato non convivente solo con figli
31 Genitore femmina separato solo con figli
32 Genitore femmina divorziato solo con figli
33 Genitore femmina vedovo solo con figli

FAMIGLIE CON DUE NUCLEI

Famiglie con due nuclei senza persone isolate

34 A due generazioni
35 Di tipo fraterno
36 Binucleare di altro tipo

Famiglie con due nuclei con persone isolate

37 A due generazioni
38 Di tipo fraterno
39 Binucleare di altro tipo

FAMIGLIE CON TRE O PIÙ NUCLEI

40 Senza isolati
41 Con isolati

DIFFERENZE TRA VARIABILI DI TIPO A E VARIABLI DI TIPO B

I gruppi nucleo sono due:
- nel tipo A i figli sono sempre appartenenti al nucleo dei genitori anche se questi si sono ricongiunti in un secondo momento al nucleo genitoriale originario (ad esempio perché sono rientrati nella famiglia di origine dopo una separazione);
- nel tipo B sono definiti figli solo coloro che non hanno mai interrotto la loro permanenza in quel nucleo, gli altri diventano membri isolati.

Ad esempio nel tipo nucleo A = un membro appartiene ad un nucleo qualora sussistano rapporti di coppia o discendenza diretta (figlio).

Ad esempio nel tipo nucleo B = un membro appartiene ad un nucleo qualora sussistano rapporti di coppia o discendenza diretta mai interrotta (figlio minore o celibe).

ESEMPIO:

Relazione di parentela	Sesso	Stato civile	Numero nucleo A	Numero nucleo B	Tipo nucleo A	Tipo nucleo B
Pr	F	Coniugato	1	1	1	1
Coniuge	M	Coniugato	1	1	1	1
Suocero	M	Coniugato	2	2	1	2
Suocera	F	Coniugato	2	2	1	2
Cognato	M	Separato	2	0	1	0
Figlio	F	Nubile	1	1	1	1
Figlio	M	Celibe	1	1	1	1
Figlio	M	Celibe	1	1	1	1

6. TIPI DI ELABORAZIONI

A seconda della selezione che si opera sul file è possibile effettuare elaborazioni sulle seguenti unità di analisi: individui, vittime, tipo di violenza, caratteristiche della violenza.

Alcune note:

a) Il numero totale di appartenenti al campione è pari al numero di records; per calcolare le stime si dovrà utilizzare il coefficiente di riporto all'universo; il coefficiente deve essere diviso per 10^6 per avere dati in unità e 10^9 per ottenere dati in migliaia;

b) per ogni sezione riguardante la violenza, il suo approfondimento e le caratteristiche del partner, sono stati costruiti dei filtri che identificano il target di popolazione che deve entrare in quella determinata sezione. I filtri di accesso alla sezione sono rintracciabili nel tracciato record all'inizio di ogni sezione di riferimento.

c) alcune variabili già oggetto di pubblicazione (nella statistica in breve e nelle tavole on line) sono state diffuse in modo rielaborato; è questo il caso delle variabili "denuncia la violenza da partner" e "ferite da partner", che aggregano i dati relativi all'ultimo episodio con i dati attinenti le violenze ripetute da partner (rintracciabili nell'ultima parte della sezione REP_PR). Ad esempio per le denunce sono elaborati insieme i quesiti REP_PR18 (lei o qualcuno altro ha denunciato il fatto alla polizia o ad altre autorità giudiziarie) e REP_PR33 (ha mai riferito alle forze dell'ordine i fatti che lei ha subito).

7. 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.

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

1. l'intero territorio nazionale;
2. le 5 ripartizioni (nord-ovest, nord-est, centro, sud, isole);
3. le sei aree basate sulla tipologia socio-demografica dei comuni (comuni centro delle aree metropolitane, comuni della periferia delle aree metropolitane, comuni con meno di 2.000 abitanti, comuni con 2.001-10.000 abitanti, comuni con 10.001-50.000 abitanti, comuni con più di 50.000 abitanti);
4. le diverse regioni.

Ogni stima è condizionata ad una attenta valutazione dell’errore campionario.

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, in quanto 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.

8. STRATEGIA DI CAMPIONAMENTO E VALUTAZIONE DEGLI ERRORI CAMPIONARI

8.1 - Introduzione

La *popolazione di interesse* dell'indagine è costituita dalle donne di età compresa tra 16 e 70 anni residenti in Italia. L'indagine è stata svolta mediante intervista telefonica e ha utilizzato come lista di selezione l'archivio degli abbonati Telecom al telefono; le *unità di campionamento* sono, pertanto, i numeri telefonici appartenenti a detto archivio.

L'indagine ha la finalità di fornire stime con diversi riferimenti territoriali:

- l'intero territorio nazionale;
- le cinque ripartizioni geografiche (Nord-ovest, Nord-est, Centro, Sud e Isole);
- le regioni geografiche;
- sei aree basate sulla tipologia socio-demografica dei comuni, così definite:
  - A, *area metropolitana* suddivisa in :
    - A_1 , comuni centro dell'area metropolitana: Torino, Milano, Venezia, Genova, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania e Cagliari;
    - A_2 , comuni che gravitano intorno al centro dell'area metropolitana;
  - B, area *non metropolitana* suddivisa in :
    - B1 , comuni aventi fino a 2 mila abitanti;
    - B2 , comuni con 2.001-10 mila abitanti;
    - B3 , comuni con 10.001-50 mila abitanti;
    - B4 , comuni con oltre 50 mila abitanti.

La *base di campionamento* adottata, ovvero la lista di selezione delle unità campionarie, è l'*archivio informatizzato ufficiale delle famiglie abbonate alla rete della telefonia fissa.* Tale scelta è motivata dal fatto che le informazioni dell'archivio in oggetto sono contenute in un *file* che viene costantemente aggiornato sulle variazioni degli intestatari e degli indirizzi telefonici; esso è, inoltre, di agevole utilizzo per la selezione delle unità campionarie in quanto si presta facilmente alla scelta di diversi criteri di ordinamento.

Le informazioni relative a ciascun indirizzo, utilizzabili per la stratificazione delle unità della popolazione di riferimento, sono essenzialmente di tipo territoriale; esse sono la provincia, il comune, la sezione di censimento, la via, il numero civico, l'ampiezza del comune di appartenenza, in termini demografici e in termini di numero di indirizzi.

Poiché non tutte le famiglie presenti nella lista contengono unità eleggibili, è stato necessario selezionare dalla lista un numero di indirizzi più elevato rispetto alla numerosità campionaria progettata, determinato sulla base di una stima della percentuale di famiglie con donne eleggibili.

Per una discussione più approfondita sulle caratteristiche della lista di selezione e sui problemi che dall'uso di tale lista derivano si può far riferimento al volume "*Indagini Sociali Telefoniche: Metodologia ed Esperienze della Statistica Ufficiale*", *anno 2000, Metodi e Norme,* ISTAT.

8.2 - Descrizione del disegno di campionamento

Il disegno di campionamento è a *due stadi* con stratificazione delle unità di primo stadio. Le unità di primo stadio sono gli indirizzi telefonici dell'archivio di selezione e, quindi, le famiglie ad essi corrispondenti. Le unità di secondo stadio sono le donne eleggibili: per ciascuna famiglia selezionata al primo stadio si seleziona un'unità campionaria tra i componenti eleggibili della famiglia (donne tra i 16 e i 70 anni).

Gli indirizzi telefonici sono stati stratificati per regione geografica e per tipologia di comune.

La determinazione del numero totale di unità campionarie e la sua allocazione tra gli strati è in genere, per un'indagine ad obiettivi plurimi come quella in esame, un'operazione complessa. È poco realistico, infatti, pensare di poter definire un campione che assicuri prefissati livelli di precisione a tutte le stime d'interesse, considerando anche il fatto che le stime vengono prodotte con diversi riferimenti territoriali. L'allocazione ottimale delle unità del campione con riferimento ad un dato tipo di dominio può risultare in contrasto con l'allocazione ottimale con riferimento ad un altro tipo di dominio. In particolare, per quanto riguarda le stime riferite all'intero territorio nazionale l'allocazione ottimale risulta vicina a quella

proporzionale tra le diverse regioni; per quanto riguarda, invece, le stime riferite alle regioni, l'allocazione ottimale risulta prossima a quella che assegna a tutte le regioni un campione di uguale numerosità. È necessario quindi un procedimento complesso articolato in più fasi.

Dapprima, mediando tra esigenze operative e di costo ed esigenze relative all'attendibilità delle principali stime di interesse, viene definita la numerosità *n* complessiva del campione. Nella presente indagine si è fissata una numerosità campionaria complessiva di 25.000 interviste. Successivamente, sulla base di valutazioni dell'errore di campionamento atteso delle principali stime a livello regionale e nazionale, è stata determinata l'allocazione del campione tra le regioni; si è ottenuta in tal modo un'allocazione di compromesso tra l'allocazione *uguale* e quella *proporzionale*. Infine, le numerosità campionarie regionali sono state ripartite tra le diverse tipologie di comune in modo proporzionale alla popolazione residente.

L'estrazione degli indirizzi campione da ciascuno strato è stata effettuata con probabilità uguali e senza reimmissione, mediante tecnica di selezione sistematica. Per ogni famiglia rispondente risultata eleggibile, è stata selezionata l'unità a cui somministrare l'intervista mediante estrazione casuale dalla lista delle donne eleggibili della famiglia.

Nel prospetto 1 sono riportate le numerosità campionarie per regione.

Prospetto 1 - Distribuzione regionale del campione

Regioni	Indirizzi campione
Piemonte	1.357
Valle d'Aosta	906
Lombardia	1.906
Bolzano	1.387
Trento	1.021
Veneto	1.066
Friuli -Venezia Giulia	1.327
Liguria	1.279
Emilia Romagna	982
Toscana	1.049
Umbria	1.483
Marche	1.027
Lazio	927
Abruzzo	1.512
Molise	1.332
Campania	957
Puglia	1.104
Basilicata	1.423
Calabria	1.072
Sicilia	936
Sardegna	947
ITALIA	**25.000**

8.3 - Procedimento per il calcolo delle stime

Le stime sono ottenute mediante uno stimatore di ponderazione vincolata. Il principio su cui è basato ogni metodo di stima campionaria è che le unità appartenenti al campione rappresentino anche le unità della popolazione che non sono incluse nel campione. Questo principio viene realizzato attribuendo ad ogni unità campionaria un peso che indica il numero di unità della popolazione rappresentate dall'unità medesima. Se, ad esempio, ad un'unità campionaria viene attribuito un peso pari a 100, vuol dire che questa unità rappresenta se stessa ed altre 99 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; h, indice di strato; j, indice di famiglia; q indice di individuo all'interno della famiglia j; y, generica variabile oggetto di indagine; Y_{hjp} valore di y osservato sull'individuo p della famiglia j dello strato h (per stime di frequenze, y è una variabile dicotomica che assume valore 1 se l'individuo presenta la caratteristica di interesse e zero altrimenti); Q_{hj}, numero di individui eleggibili appartenenti alla famiglia j dello strato h; M_h, numero di famiglie residenti nello strato h; m_h, campione di famiglie nello strato h; p_h , numero di individui campione nello strato h (dal momento che si intervista un unico individuo in ciascuna famiglia campione si ha $p_h = m_h$); H_d, numero di strati nel dominio d.

Ipotizziamo di voler stimare, con riferimento ad un generico dominio d (ad esempio una regione geografica) il totale della variabile y oggetto di indagine, espresso dalla seguente relazione:

$$ {}_{d}Y = \sum_{h=1}^{H_d} \sum_{j=1}^{M_h} \sum_{q=1}^{Q_{hj}} Y_{hjq} \qquad (1) $$

Una stima del totale (1) è data dalla seguente espressione:

$$ {}_{d}\hat{Y} = \sum_{h=1}^{H_d} \hat{Y}_h = \sum_{h=1}^{H_d} \sum_{j=1}^{m_h} Y_{hj.} W_{hj.} , \qquad (2) $$

in cui Y_{hj} e W_{hj} rappresentano rispettivamente il valore assunto dalla variabile y e il peso finale da attribuire all'individuo campione della famiglia j dello strato h.

Dalla precedente relazione si desume, quindi, che per ottenere la stima del totale (1) occorre moltiplicare il peso finale associato a ciascuna unità campionaria per il valore della variabile y assunto da 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 ha le seguenti finalità:

- correggere l'effetto distorsivo dovuto agli errori di lista e al fenomeno della mancata risposta totale;
- tenere conto della conoscenza di alcuni totali noti sulla popolazione oggetto di studio, nel senso che le stime campionarie di tali totali devono coincidere con i rispettivi valori noti.

Per il calcolo dei pesi la popolazione di riferimento è costituita dalle donne di in età 16-70 anni, al netto delle convivenze; i totali noti imposti a livello regionale sono i seguenti:

a. popolazione per classi di età (16-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-70);
b. popolazione per tipologia comunale (aree A_1, A_2, B_1, B_2, B_3, B_4 definite nel paragrafo 1);
c. popolazione per titolo di studio[2] (nessuno o licenza elementare, licenza media o avviamento professionale, diploma superiore, laurea o titolo superiore);
d. popolazione per stato civile (nubili, coniugate, separate o divorziate, vedove);
e. popolazione per dimensione familiare (famiglie mono-componenti per età (16-49,50-70), 2 componenti, 3, 4 , 5 o più componenti)[3].

La procedura per la costruzione dei pesi finali da attribuire alle unità campionarie, è articolata nelle seguenti fasi :

1. viene dapprima calcolato il peso base (o peso diretto), ottenuto come reciproco della probabilità di inclusione di ogni unità campionaria;
2. si calcola quindi il fattore correttivo che consente di soddisfare la condizione di uguaglianza tra i totali noti della popolazione e le corrispondenti stime campionarie;
3. il peso finale è dato dal prodotto del peso base per i fattori correttivi sopra indicati.

Il fattore correttivo del punto 3. è ottenuto mediante la risoluzione di un problema di minimo vincolato, in cui la funzione da minimizzare è la distanza tra i pesi base ed i pesi finali; i vincoli sono definiti dalla condizione che le stime campionarie dei totali di popolazione sopra definiti coincidano con i valori noti degli stessi.

E' utile osservare che i vincoli c, d ed e sono stati utilizzati nonostante il fatto che non si basino su totali noti da fonte censuaria o anagrafica, ma solo su stime prodotte da un'altra indagine campionaria. Si è comunque ritenuto opportuno utilizzarli per correggere, almeno in parte, la distorsione dovuta alla sottocopertura della lista di selezione.

8.4 - Valutazione del livello di precisione delle stime

Le principali statistiche di interesse per valutare la variabilità campionaria delle stime prodotte dall'indagine sono l'errore di campionamento assoluto e l'errore di campionamento relativo.

[2] I totali noti relativi allo stato civile e al titolo di studio derivano da stime dell'indagine sulle Forze di lavoro

[3] I totali noti relativi alla dimensione familiare derivano da stime dell'indagine Multiscopo '*Aspetti della vita quotidiana*'.

Indicando con $\hat{V}ar(_d\hat{Y})$ la varianza della stima $_d\hat{Y}$, riferita al dominio d, la stima dell'errore di campionamento assoluto di $_d\hat{Y}$ si può ottenere mediante la seguente espressione:

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

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

$$\hat{\varepsilon}\,(_d\hat{Y}) = \frac{\sqrt{\hat{V}ar(_d\hat{Y})}}{_d\hat{Y}} \tag{4}$$

La stima della varianza, $\hat{V}ar(_d\hat{Y})$, viene calcolata come somma della stima della varianza dei singoli strati appartenenti al dominio d; in simboli:

$$\hat{V}ar(_d\hat{Y}\ ) = \sum_{h=1}^{H_d}\hat{V}ar\left(\hat{Y}_h\right) = \sum_{h=1}^{H_d}\frac{m_h}{m_h-1}\sum_{j=1}^{m_h}\frac{(\hat{Y}_{hj}-\hat{\bar{Y}}_h)^2}{m_h-1} \tag{5}$$

dove

$$\hat{Y}_{hj} = Y_{hj}W_{hj} \qquad \text{e} \qquad \hat{\bar{Y}}_h = \frac{1}{m_h}\sum_{j=1}^{m_h}\hat{Y}_{hj}\,.$$

Gli errori campionari delle espressioni (3) e (4), consentono di valutare il grado di precisione delle stime; inoltre, l'errore assoluto permette di costruire l'intervallo di confidenza, che, con una certa probabilità, contiene il parametro d'interesse. Con riferimento alla generica stima $\hat{Y}$ tale intervallo assume la seguente forma:

$$\Pr\left\{\hat{Y} - k\,\hat{}\,(\hat{Y}) \le Y \le \hat{Y} + k\,\hat{}\,(\hat{Y})\right\} = P \tag{6}$$

Nella (6) il valore di k dipende dal valore fissato per la probabilità P; ad esempio, per P=0,95 si ha k=1,96.

8.5. Presentazione sintetica degli errori campionari

Ad ogni stima $_d\hat{Y}$ è associato un errore campionario relativo $\hat{\varepsilon}(_d\hat{Y})$; quindi, per consentire un uso corretto delle stime fornite dall'indagine, sarebbe necessario fornire, per ogni stima pubblicata, anche 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 eccessivamente appesantite e di non agevole consultazione per l'utente finale. Inoltre, non sarebbero in ogni caso disponibili gli errori delle stime non pubblicate, che l'utente può ricavare in modo autonomo.

Per questi motivi, generalmente, si ricorre ad una presentazione sintetica degli errori relativi, basata sul *metodo dei modelli regressivi*. Tale metodo si fonda sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore relativo.

L'approccio utilizzato per la costruzione di questi modelli è diverso a seconda che si tratti di variabili qualitative o quantitative. Infatti, solo nel caso delle stime di frequenze assolute (o relative) riferite alle modalità di variabili qualitative, è possibile utilizzare dei modelli che hanno un fondamento teorico e secondo cui gli errori relativi delle stime di frequenze assolute sono funzione decrescente dei valori delle stime stesse.

Per calcolare gli errori di campionamento è stato utilizzato un software generalizzato, messo a punto presso l'Istat, che consente di calcolare gli errori campionari e gli intervalli di confidenza e permette di costruire dei modelli regressivi per la presentazione sintetica degli errori di campionamento.

8.6 Presentazione sintetica degli errori campionari per stime di frequenze

Il modello utilizzato per le stime di frequenze assolute, con riferimento al generico dominio d, è il seguente:

$$\log \hat{\varepsilon}^2({}_d\hat{Y}) = a + b \log({}_d\hat{Y}) \qquad (7)$$

dove i parametri a e b vengono stimati mediante il metodo dei minimi quadrati.

Il prospetto 2 riporta i valori dei coefficienti a e b e dell'indice di determinazione R^2 del modello utilizzato per l'interpolazione degli errori campionari delle stime di frequenze riferite alle famiglie e alle persone, per aree territoriali.

Sulla base delle informazioni contenute nel suddetto prospetto è possibile calcolare l'errore relativo di una determinata stima di frequenza assoluta ${}_d\hat{Y}^*$, riferita ai diversi domini, mediante la formula:

$$\hat{\varepsilon}({}_d\hat{Y}^*) = \sqrt{\exp(a + b \log({}_d\hat{Y}^*))} \qquad (8)$$

e costruire l'intervallo di confidenza al 95% come:

$$\left\{ {}_d\hat{Y}^* - 1{,}96 \cdot \hat{\varepsilon}({}_d\hat{Y}^*) \cdot {}_d\hat{Y}^* ;\ {}_d\hat{Y}^* + 1{,}96 \cdot \hat{\varepsilon}({}_d\hat{Y}^*) \cdot {}_d\hat{Y}^* \right\}.$$

Allo scopo di facilitare il calcolo degli errori campionari, nel prospetto 3 sono riportati gli errori relativi percentuali corrispondenti a valori crescenti di stime di frequenze assolute calcolati introducendo nella (8) i valori di a e b riportati nel prospetto 2.

Le informazioni contenute in tale prospetto consentono di calcolare l'errore relativo di una generica stima di frequenza assoluta mediante due procedimenti di facile applicazione che, tuttavia, conducono a risultati meno precisi di quelli ottenibili applicando direttamente la formula (8).

Il primo metodo consiste nell'approssimare l'errore relativo della stima di interesse ${}_d\hat{Y}^*$ con quello, riportato nei prospetti, corrispondente al livello di stima che più si avvicina a ${}_d\hat{Y}^*$.

Il secondo metodo, più preciso del primo, si basa sull'uso di una formula di interpolazione lineare per il calcolo degli errori di stime non comprese tra i valori forniti nei prospetti. In tal caso, l'errore campionario della stima ${}_d\hat{Y}^*$, si ricava mediante l'espressione:

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

dove ${}_d\hat{Y}^{k-1}$ e ${}_d\hat{Y}^{k}$ sono i valori delle stime entro i quali è compresa la stima ${}_d\hat{Y}^*$, mentre $\hat{\varepsilon}({}_d\hat{Y}^{k-1})$ e $\hat{\varepsilon}({}_d\hat{Y}^{k})$ sono 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 di frequenze assolute per aree territoriali

	PERSONE

	a	b	R^2 (%)
ITALIA	9,239869	-1,17711	93,8
RIPARTIZIONI GEOGRAFICHE (a)			
Nord-ovest	8,279632	-1,09880	89,7
Nord-est	7,696539	-1,08854	89,5
Centro	8,588522	-1,15688	92,4
Sud	8,543144	-1,14395	91,9
Isole	8,204668	-1,11884	89,5
TIPI DI COMUNE (b)			
A1	8,765766	-1,150312	91,7
A2	9,213202	-1,199260	92,6
B1	8,023610	-1,130024	90,3
B2	8,544833	-1,146776	91,9
B3	8,732646	-1,155952	92,0
B4	8,488616	-1,151277	92,1
REGIONI			
Piemonte	8,869456	-1,190629	90,8
Valle d'Aosta	4,767639	-1,158700	88,4
Lombardia	9,520444	-1,194766	92,5
Bolzano	5,975135	-1,106340	88,7
Trento	6,584325	-1,191006	91,3
Veneto	8,950070	-1,192976	90,2
Friuli-Venezia Giulia	7,949205	-1,225341	89,9
Liguria	7,834485	-1,169833	91,9
Emilia-Romagna	8,493159	-1,162099	91,8
Toscana	8,535521	-1,173684	92,7
Umbria	7,052279	-1,170616	91,8
Marche	8,013984	-1,209651	91,5
Lazio	9,081276	-1,189965	92,1
Abruzzo	7,988770	-1,216480	92,4
Molise	6,285044	-1,216237	93,2
Campania	9,065464	-1,178238	92,1
Puglia	8,594078	-1,160083	90,7
Basilicata	6,959656	-1,208514	93,7
Calabria	7,891011	-1,153949	88,9
Sicilia	8,566980	-1,141567	89,3
Sardegna	7,469845	-1,124523	89,2

(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 mila abitanti; Tipo B2: da 2.001 a 10 mila abitanti; Tipo B3: da 10.001 a 50 mila abitanti; Tipo B4: oltre 50 mila abitanti.

Prospetto 3 - Valori interpolati degli errori relativi percentuali delle stime di frequenze assolute per aree territoriali

STIME	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	A1	A2	B1	B2	B3	B4
10.000	44,9	39,8	31,2	35,6	36,9	35,0	40,1	40,0	30,4	36,5	38,4	34,7
20.000	29,9	27,2	21,4	23,8	24,8	23,7	26,9	26,4	20,5	24,5	25,7	23,3
30.000	23,5	21,8	17,2	18,8	19,7	18,9	21,3	20,7	16,3	19,4	20,4	18,5
40.000	19,9	18,6	14,7	16,0	16,7	16,1	18,1	17,4	13,9	16,5	17,2	15,6
50.000	17,4	16,5	13,0	14,0	14,7	14,2	15,9	15,2	12,2	14,5	15,1	13,8
75.000	13,7	13,2	10,4	11,1	11,7	11,3	12,6	12,0	9,7	11,5	12,0	10,9
100.000	11,6	11,2	8,9	9,4	9,9	9,6	10,7	10,1	8,3	9,7	10,1	9,2
250.000	6,8	6,8	5,4	5,5	5,9	5,8	6,3	5,8	4,9	5,8	6,0	5,4
500.000	4,5	4,6	3,7	3,7	3,9	3,9	4,2	3,8	3,3	3,9	4,0	3,7
750.000	3,5	3,7	3,0	2,9	3,1	3,1	3,3	3,0	2,6	3,1	3,2	2,9
1.000.000	3,0	3,2	2,5	2,5	2,7	2,7	2,8	2,5	2,3	2,6	2,7	2,5
5.000.000	1,2	1,3	1,1	1,0	1,1	1,1	1,1	1,0	0,9	1,0	1,1	1,0

Prospetto 3 *(segue)* **- Valori interpolati degli errori relativi percentuali delle stime di frequenze assolute per aree territoriali**

STIME	Piemonte	Valle d'Aosta	Lombardia	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Liguria	Emilia Romagna	Toscana	Umbria
1.000	138,0	19,8	188,4	43,4	44,0	142,6	77,3	88,4	126,2	123,9	59,6
5.000	53,0	7,8	72,1	17,8	16,9	54,6	28,8	34,5	49,5	48,2	23,2
10.000	35,1	5,2	47,6	12,2	11,2	36,1	18,9	23,0	33,1	32,1	15,5
20.000	23,2	3,5	31,5	8,3	7,4	23,9	12,3	15,3	22,1	21,4	10,3
30.000	18,2	2,8	24,7	6,6	5,8	18,7	9,6	12,1	17,5	16,8	8,1
40.000	15,4	2,3	20,8	5,6	4,9	15,8	8,1	10,2	14,8	14,2	6,9
50.000	13,4	2,1	18,2	5,0	4,3	13,8	7,0	9,0	13,0	12,5	6,0
75.000	10,6	1,6	14,3	4,0	3,4	10,9	5,5	7,1	10,3	9,8	4,8
100.000	8,9	1,4	12,0	3,4	2,8	9,1	4,6	6,0	8,7	8,3	4,0
250.000	5,2	0,8	7,0	2,0	1,6	5,3	2,6	3,5	5,1	4,8	2,4
500.000	3,4	0,5	4,6	1,4	1,1	3,5	1,7	2,3	3,4	3,2	1,6
750.000	2,7	0,4	3,6	1,1	0,9	2,7	1,3	1,8	2,7	2,5	1,2
1.000.000	2,3	0,4	3,0	1,0	0,7	2,3	1,1	1,6	2,3	2,1	1,0

Prospetto 3 *(segue)* **- Valori interpolati degli errori relativi percentuali delle stime di frequenze assolute per aree territoriali**

STIME	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
1.000	84,3	153,8	81,3	34,7	158,9	133,7	49,9	96,1	140,6	86,2
5.000	31,8	59,0	30,5	13,0	61,6	52,6	18,9	38,0	56,1	34,9
10.000	20,9	39,1	20,0	8,6	40,9	35,2	12,4	25,4	37,8	23,6
20.000	13,8	25,9	13,1	5,6	27,2	23,5	8,2	17,1	25,4	16,0
30.000	10,8	20,3	10,3	4,4	21,4	18,6	6,4	13,5	20,2	12,7
40.000	9,1	17,1	8,6	3,7	18,1	15,7	5,4	11,4	17,1	10,8
50.000	7,9	15,0	7,5	3,2	15,9	13,8	4,7	10,1	15,1	9,6
75.000	6,2	11,8	5,9	2,5	12,5	10,9	3,7	8,0	12,0	7,6
100.000	5,2	9,9	4,9	2,1	10,5	9,2	3,1	6,7	10,1	6,5
250.000	3,0	5,8	2,8	1,2	6,1	5,4	1,8	4,0	6,0	3,9
500.000	2,0	3,8	1,9	0,8	4,1	3,6	1,2	2,7	4,0	2,6
750.000	1,5	3,0	1,4	0,6	3,2	2,9	0,9	2,1	3,2	2,1
1.000.000	1,3	2,5	1,2	0,5	2,7	2,4	0,8	1,8	2,7	1,8