

Metodi e fonti: note metodologiche

2 Gennaio 2020

Per informazioni: statistiche@bancaditalia.it
www.bancaditalia.it/statistiche/index.html

L'indagine sui bilanci delle famiglie italiane

L'indagine sui bilanci delle famiglie italiane (IBF) è un'indagine campionaria condotta dalla Banca d'Italia sin dagli anni '60 con l'obiettivo di raccogliere informazioni sui redditi e i risparmi delle famiglie italiane.

Le motivazioni dell'indagine sono esplicitamente ricordate in uno dei primi rapporti:

“L'importanza economica che rivestono le famiglie nel nostro sistema, così come nella maggioranza di quelli ad economia di mercato, appare evidente ove si consideri che esse possiedono direttamente o indirettamente la quasi totalità della ricchezza nazionale, percepiscono quasi tutto il reddito nazionale e da esse provengono, attualmente in Italia, circa i tre quarti della domanda globale interna. Anche dal punto di vista finanziario il peso delle famiglie è notevole, dando esse origine a una parte sostanziale dei flussi finanziari e possedendo una quota notevole della ricchezza mobiliare. [...] Allo scopo ultimo di acquisire tale conoscenza e, in via preliminare, di stimare alcune grandezze concernenti il settore delle famiglie, la Banca d'Italia ha iniziato una nuova serie di ricerche campionarie sul reddito, il consumo e il risparmio delle famiglie italiane.”¹.

Nel corso degli anni l'oggetto della rilevazione si è andato estendendo per includere anche la ricchezza e altri aspetti inerenti i comportamenti economici e finanziari delle famiglie, come ad esempio l'uso dei mezzi di pagamento.

In connessione con lo svolgimento dei propri compiti istituzionali, la Banca d'Italia effettua anche attività di raccolta, produzione e pubblicazione di informazioni statistiche. I microdati delle indagini campionarie forniscono informazioni complementari a quelle derivanti dalle fonti macro (ad es., Contabilità Nazionale) in quanto consentono le analisi economiche su particolari sottopopolazioni di interesse come sulle famiglie benestanti o povere, le famiglie a basso o alto reddito, i nuclei (altamente) indebitati o le famiglie finanziariamente vulnerabili.

Questo documento² fornisce una sintesi delle principali caratteristiche metodologiche dell'indagine, con particolare enfasi su quelle riguardanti le procedure statistiche e il trattamento dei dati, allo scopo di fornire un solido *background* per i potenziali utenti dei microdati dell'indagine.

1. Il disegno campionario

L'indagine ha da sempre l'obiettivo di rappresentare la popolazione (ufficialmente) residente in Italia. Sono esclusi dalla popolazione obiettivo coloro che vivono in istituti (conventi, ospedali, prigionieri,...) o che sono illegalmente presenti sul territorio nazionale.

¹ Banca d'Italia (1966), Reddito, risparmio e alcuni consumi delle famiglie italiane nel 1965, Bollettino n. 4.

² Ultimo aggiornamento: gennaio 2020.

Il disegno di campionamento dell'indagine prevede una procedura di selezione a due stadi. Le unità di primo stadio sono i comuni³; le unità di secondo stadio sono le famiglie. Prima di procedere all'estrazione delle unità di primo stadio, queste vengono suddivise in base alla regione e alla classe di ampiezza demografica⁴ (*stratificazione* delle unità di primo stadio).

All'interno di ogni strato i comuni nei quali effettuare le interviste sono selezionati includendo tutti quelli con popolazione superiore ai 40.000 abitanti (comuni *auto-rappresentativi*) ed estraendo i comuni rimanenti con una modalità di selezione che assegna ai comuni di maggiore dimensione una probabilità più elevata di essere inclusi nel campione (*probability proportional to size* - PPS). Nella seconda fase, le famiglie da intervistare vengono estratte casualmente dalle liste anagrafiche dei comuni precedentemente selezionati, dopo aver definito un'opportuna stratificazione volta a incrementare l'efficienza del disegno (*stratificazione* delle unità di secondo stadio).

L'indagine è stata condotta fino al 1987 sulla base di rilevazioni indipendenti a cadenza pressoché annuale, mentre successivamente si è passati a rilevazioni biennali (con eccezione del 1998 e del 2019). Nel corso degli anni l'ampiezza del campione obiettivo è cresciuta fino a raggiungere le 8.000 famiglie. Nell'edizione sul 2019, sperimentalmente il campione è stato incrementato a 10.000 famiglie.

Dall'indagine sul 1989, per favorire l'analisi dell'evoluzione dei fenomeni oggetto di indagine, è stato introdotto uno schema che prevede la presenza nel campione di una quota di unità già intervistate in occasione di precedenti edizioni dell'indagine (famiglie *panel*). Nelle ultime edizioni dell'indagine tale quota rappresenta circa il 50 per cento del campione. La componente *panel* del campione, è costituita dalle famiglie presenti da almeno due rilevazioni e da una parte, estratta casualmente, di quelle rilevate solo nella precedente indagine⁵. Le famiglie non *panel* sono invece estratte casualmente dalle liste anagrafiche.

2. Il questionario e la fase di rilevazione

I questionari utilizzati nelle rilevazioni, integralmente disponibili alla pagina web "Documentazione per l'utilizzo dei microdati" della sezione dedicata all'indagine del sito internet della Banca d'Italia⁶, sono predisposti seguendo una struttura modulare. Si compongono di una parte di base, nella quale sono rilevati i fenomeni ai quali tutte le famiglie sono interessate, e di diversi allegati, in cui sono raccolte informazioni che riguardano soltanto specifici sottoinsiemi di famiglie.

A partire dall'indagine sul 1998 la rilevazione dei dati è stata effettuata in prevalenza (oltre il 90 per cento nelle ultime rilevazioni) con l'aiuto del computer (CAPI, *Computer-Assisted Personal Interviewing*). Con questa metodologia i dati vengono rilevati presso le famiglie mediante un questionario elettronico, che, oltre a memorizzare le informazioni, aziona una serie di controlli consentendo di risolvere, in presenza della famiglia, le eventuali incoerenze⁷. Le restanti interviste sono state realizzate con un questionario cartaceo (PAPI, *Paper-And-pencil Personal Interviewing*) e successivamente trasferite su supporto elettronico dalla società di rilevazione, utilizzando il programma CAPI come procedura di immissione dei dati. A partire dall'indagine sul 2019, le interviste sono effettuate esclusivamente CAPI attraverso l'ausilio di un *tablet*, che automaticamente raccoglie alcuni metadati come la geo-localizzazione o la durata dell'intervista e delle sue sotto-sezioni.

³ Nelle ultime edizioni dell'indagine il numero di unità di primo stadio è variato tra i 350 e i 450 comuni.

⁴ Le classi utilizzate sono: fino a 20.000 abitanti, fra 20.000 e 40.000 abitanti, oltre 40.000 abitanti.

⁵ Come in passato, per disporre di informazioni utili per le analisi intergenerazionali sono state inoltre contattate per l'intervista le famiglie che si sono formate dalle originarie unità *panel* incluse nel campione (tipicamente quelle formate dai figli che costituiscono una nuova famiglia).

⁶ <https://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/documentazione/index.html>.

⁷ Numerose sono le cause alla base delle incoerenze che caratterizzano i dati rilevati: da parte dell'intervistato possono esserci problemi di comprensione della domanda, di memoria nel fornire una risposta adeguata o anche reticenza nel fornire informazioni percepite come riservate; da parte dell'intervistatore gli errori più frequenti riguardano la digitazione dei codici di risposta o la scrittura dei valori in un'unità di misura diversa da quella prevista nel questionario.

Nelle ultime rilevazioni la durata media dell'intervista si è attestata in media a circa 50 minuti. Tuttavia questa è abbastanza variabile all'interno del campione, risultando positivamente correlata al reddito, alla ricchezza e al numero dei componenti (e in particolare dei percettori) della famiglia.

L'esecuzione delle interviste è preceduta da un ciclo di incontri in cui funzionari della Banca d'Italia e rappresentanti della società di rilevazione impartiscono le istruzioni direttamente agli intervistatori e da una indagine pilota in cui vengono sottoposte a test le nuove domande presenti nel questionario.

3. Il trattamento dei dati personali

Le famiglie che vengono contattate per partecipare all'indagine sono selezionate dall'Istat a partire dalle liste anagrafiche. I nominativi sono estratti attraverso un criterio di casualità e sono poi comunicati alla Banca d'Italia al fine di consentire lo svolgimento dell'indagine. I dati raccolti sono utilizzati dalla Banca d'Italia per elaborare statistiche e per l'analisi della situazione economica italiana. La comunicazione di dati personali da parte dell'Istat alla Banca d'Italia è realizzata secondo le condizioni e le modalità stabilite dall'articolo 21, comma 2, del Regolamento Europeo n. 223/2009, il quale autorizza esplicitamente la "trasmissione di dati riservati" da un'autorità del Sistema statistico europeo a un membro del Sistema europeo delle Banche Centrali a condizione che tale trasmissione, come nel caso di specie, "sia necessaria ai fini dell'efficienza dello sviluppo, della produzione e della diffusione di statistiche europee o del miglioramento della loro qualità".

La partecipazione delle famiglie all'indagine è su base volontaria e pertanto il conferimento dei dati è facoltativo⁸. Le interviste raccolte sono sottoposte in primo luogo a controlli di qualità volti a selezionare quelle che possono essere considerate valide. In particolare, in presenza di un numero elevato di domande non compilate, o nel caso in cui l'intervista sia compilata da una famiglia diversa da quella che ha ricevuto l'invito a partecipare, l'intervista è annullata⁹.

Le informazioni relative alle famiglie selezionate a partecipare all'indagine sono trattate nel rispetto della vigente normativa in materia di protezione dei dati personali¹⁰ (Regolamento Europeo GDPR 679/2016 e d.lgs. n. 196/2003 e s.m.i.) e sono utilizzate dalla Banca d'Italia esclusivamente per le finalità di interesse pubblico dichiarate. Le operazioni di trattamento ai fini di conduzione dell'Indagine sono effettuate dagli addetti della Divisione Indagini campionarie autorizzati dal Capo del Servizio Analisi statistiche della Banca d'Italia e dagli addetti della società di rilevazione incaricata, che effettua le attività di raccolta dati per conto della Banca d'Italia e che, a tal fine, è stata nominata Responsabile del trattamento.

Tali informazioni potrebbero essere inoltre elaborate congiuntamente a dati di fonte amministrativa (banca dati reddituale, banca dati integrata della proprietà immobiliare e archivio dei rapporti con operatori finanziari dell'Agenzia delle Entrate, Sistema Informativo Integrato di Acquirente Unico S.p.A, archivi di Infocamere, Centrale dei Rischi, COVIP e INPS). Per l'effettuazione del *record linkage*, i nominativi selezionati per partecipare all'indagine verrebbero condivisi con le Istituzioni proprietarie di questi dati amministrativi, nell'ambito di attività di interesse pubblico e sulla base di disposizioni di legge che espressamente prevedono tale collaborazione. Il trattamento dei nominativi sarebbe in ogni caso effettuato unicamente da personale di queste Istituzioni espressamente autorizzato, nel rispetto dei principi di necessità, di pertinenza e di non eccedenza. Nessuna informazione economica rilevata durante l'intervista sarà conferita insieme ai nominativi a tali Istituzioni. La base dati integrata (che unisce tutte le informazioni rilevate dall'indagine con quelle amministrative) sarebbe gestita esclusivamente dagli addetti del Servizio Analisi statistiche opportunamente autorizzati dal Capo del Servizio Analisi statistiche della Banca

⁸ Le famiglie contattate per l'intervista ricevono una lettera d'invito a firma del Governatore della Banca d'Italia e un opuscolo che illustra le finalità dell'indagine e fornisce alcuni esempi di utilizzo dei dati raccolti. Le famiglie non ricevono alcun compenso per l'intervista rilasciata; in occasione della pubblicazione dei risultati dell'indagine ad esse viene inviata una lettera di ringraziamento alla quale vengono allegati alcuni articoli di giornale riportanti i commenti della stampa.

⁹ La sezione "Il controllo dei dati e l'imputazione dei dati mancanti" offre maggiori informazioni sulle procedure di controllo dei dati.

¹⁰ I riferimenti normativi in materia di protezione dei dati personali vigenti in ciascuna edizione dell'indagine erano: dall'indagine sul 2002 all'edizione sul 2014 il d.lgs. n. 196/2003 e s.m.i.; dall'indagine sul 2016 il Regolamento Europeo GDPR 679/2016 e il d.lgs. n. 196/2003 e s.m.i.

d'Italia. Ove si rendesse necessaria una condivisione di parte di tali dati integrati con le Istituzioni ricordate in precedenza, ad esempio ai fini del miglioramento della loro qualità, i dati condivisi sarebbero preliminarmente resi anonimi e potrebbero essere usati solo per finalità di ricerca.

I microdati diffusi pubblicamente sul sito internet della Banca d'Italia contengono solo informazioni campionarie e sono resi totalmente anonimi in modo tale che non sia possibile risalire ai soggetti ai quali si riferiscono. Infine vengono messe a disposizione delle tavole statistiche, in formato Excel, che analizzano per l'anno di riferimento e in serie storica un ampio insieme di fenomeni economici studiati nell'indagine.

Ai sensi del d.lgs. n.196/2003 e del Regolamento Europeo GDPR 679/2016, a ogni famiglia intervistata è riconosciuto tra l'altro il diritto di: (i) accedere ai propri dati personali; (ii) chiederne la rettifica, l'aggiornamento o la cancellazione; (iii) opporsi al loro trattamento per motivi legittimi. Tali diritti potranno essere esercitati facendone richiesta al Titolare del trattamento - Banca d'Italia - Servizio Organizzazione - via Nazionale 91, 00184 Roma, e-mail org.privacy@bancaditalia.it. Il Responsabile della protezione dei dati, per la Banca d'Italia, è contattabile presso via Nazionale 91, 00184 Roma o all' e-mail responsabile.protezione.dati@bancaditalia.it.

4. Il controllo dei dati e l'imputazione dei dati mancanti

La rilevazione mediante la metodologia CAPI ha ridotto la necessità di verifiche di coerenza effettuate a posteriori sui dati acquisiti. Le tradizionali fasi di controllo sono però state condotte per le interviste svolte con il questionario cartaceo (meno di un decimo) per le quali, come si è detto, il programma CAPI è stato utilizzato come maschera di immissione dei dati per sfruttare la sua capacità di segnalare le incoerenze.

Conclusa la fase di controllo, viene effettuata l'imputazione delle mancate risposte, che possono dipendere sia dalla reticenza dell'intervistato, sia dalla difficoltà da questi incontrata nel fornire risposta a un particolare quesito. L'imputazione si rende necessaria per tutte le variabili elementari che costituiscono le componenti di variabili aggregate sui redditi e la ricchezza delle famiglie (si veda il paragrafo 7), in quanto anche la mancanza di una sola componente non consentirebbe il calcolo (è ad esempio necessario imputare le integrazioni non monetarie come i buoni pasto per calcolare il reddito da lavoro dipendente).

Il numero dei dati imputati è in generale modesto e riguarda principalmente le variabili per le quali è possibile per le famiglie non fornire una risposta puntuale, come ad esempio le integrazioni non monetarie dei lavoratori dipendenti, il valore delle attività finanziarie (all'interno della classe di valori fornita dalla famiglia) e alcune informazioni monetarie sulla previdenza complementare.

La metodologia utilizzata per l'imputazione fa riferimento in genere a modelli di regressione, mediante i quali si stimano i valori da attribuire alle mancate risposte sulla base di altre informazioni disponibili. Al fine di evitare una eccessiva concentrazione intorno ai valori medi, al dato così stimato viene aggiunta una componente casuale, estratta da una variabile di tipo normale, con media zero e varianza pari a quella dei residui del modello di regressione utilizzato. In tal modo si intendono preservare la media e la varianza che caratterizzano i dati effettivamente rilevati¹¹.

5. Le mancate interviste

Il campione di famiglie effettivamente intervistate è composto da circa 8.000 famiglie, di cui circa la metà *panel*. Per ottenere queste interviste, è stato necessario contattare nelle ultime rilevazioni circa 15.000 famiglie. Il tasso di partecipazione all'indagine è poco più del 50 per cento, storicamente più elevato per le famiglie *panel* (circa l'80 per cento) rispetto a quelle non *panel* (circa il 35 per cento).

Il motivo più frequente all'origine della mancata intervista è l'indisponibilità. Storicamente in circa un quarto dei casi le famiglie sono risultate irreperibili, ovvero non è stato possibile contattare

¹¹ Si veda il paragrafo 10.6 in *Survey Methodology*, a cura di Groves R. M., Fowler F. J., Couper M. P., Lepkowski J. M., Singer E. e Tourangeau R., New York, Wiley, 2004.

telefonicamente la famiglia e non è stato trovato nessuno in casa in tre visite effettuate dagli intervistatori in giorni e orari diversi.

Le mancate interviste costituiscono un problema nelle indagini statistiche in quanto possono condurre a campioni nei quali i segmenti di popolazione meno disposti a collaborare sono sottorappresentati, producendo stime distorte. Gli stimatori utilizzati si basano sulle informazioni di coloro che hanno partecipato alla rilevazione e possono presentare una distorsione che cresce all'aumentare della mancata risposta e del divario tra il valore atteso della variabile oggetto di studio per i rispondenti e per i non rispondenti¹².

Nel corso degli anni sono state messe a punto diverse strategie volte a limitare il fenomeno della mancata risposta¹³. Prima della rilevazione le famiglie ricevono una lettera volta a sensibilizzarle sull'importanza dell'indagine e a rassicurarle sulla riservatezza delle informazioni fornite; le famiglie hanno a disposizione un numero verde della società di rilevazione e i contatti della Banca d'Italia per richiedere eventuali chiarimenti; l'intervistatore consegna alle famiglie un opuscolo illustrativo sulle finalità dell'indagine; gli intervistatori sono selezionati con sempre maggiore accuratezza tra quelli di maggiore esperienza; il numero totale di nominativi a disposizione di ciascun intervistatore per raggiungere il proprio obiettivo di famiglie intervistate è stato progressivamente ridotto.

Per limitare a posteriori i possibili effetti negativi della mancata partecipazione sono state inoltre adottate alcune misure. In primo luogo, le famiglie che non è stato possibile intervistare sono state sostituite con altre estratte con criteri casuali negli stessi comuni. La sostituzione avviene attraverso una rigida procedura sotto il controllo della Banca d'Italia che stabilisce sia quando un nominativo può essere sostituito sia il nuovo nominativo da contattare. In secondo luogo, al termine della rilevazione, è stata effettuata una stratificazione a posteriori sulla base di alcune caratteristiche individuali dei soggetti intervistati, che consente di riequilibrare, all'interno del campione, il peso dei diversi segmenti della popolazione (si veda sezione 8).

La difficoltà a ottenere l'intervista è crescente al crescere del reddito, della ricchezza e del titolo di studio del capofamiglia.

6. La valutazione della qualità delle risposte

Sebbene per un'indagine sul reddito e la ricchezza, la mancata risposta costituisca un fenomeno rilevante nel determinare problemi di qualità delle stime, numerosi altri fattori sono indicati in letteratura come potenziali elementi di disturbo.

Un aspetto che può influire sulla qualità delle stime riguarda la reticenza delle famiglie a dichiarare le proprie fonti di reddito o le attività finanziarie o reali possedute. Sebbene la partecipazione all'indagine sia volontaria e il contenuto della rilevazione noto all'intervistato sin dall'inizio, è possibile che talvolta questi non risponda con totale sincerità alle domande più delicate, quali quelle riguardanti il reddito o la ricchezza.

Per disporre di elementi di giudizio su tale fenomeno, che per sua natura risulta difficilmente investigabile, è stato richiesto agli intervistatori di formulare, immediatamente dopo l'intervista, un giudizio sintetico sulla presunta attendibilità delle risposte su reddito e ricchezza, in base alla rispondenza tra i dati forniti e gli elementi oggettivi a loro disposizione (zona e tipologia dell'immobile di residenza, tenore di vita desumibile dagli arredi, ecc.)¹⁴. Ulteriori elementi per valutare l'attendibilità delle risposte fornite dagli intervistati possono essere ottenuti comparando le stime desunte dall'indagine con i dati della Contabilità Nazionale. Queste comparazioni vanno effettuate con

¹² Si veda G. D'Alessio e I. Faiella, *Nonresponse behaviour in the Bank of Italy's Survey of Household Income and Wealth*, Banca d'Italia, Temi di discussione, n. 462, 2002.

¹³ Il fenomeno della riduzione della partecipazione alle indagini campionarie è comune ad altri paesi. Si veda E. de Leeuw e W. de Heer, *Trends in Household Survey Nonresponse: A Longitudinal and International Comparison*, in *Survey Nonresponse*, a cura di Groves R. M., Dillman D. A., Eltinge J. L. e Little R. J. A., New York, Wiley, 2002, pp. 41-54; si veda inoltre il paragrafo 6.5 in *Survey Methodology*, a cura di Groves R. M., Fowler F. J., Couper M. P., Lepkowsky J. M., Singer E. e Tourangeau R., New York, Wiley, 2004.

¹⁴ I giudizi degli intervistatori sono espressi con un punteggio compreso tra 1 (per nulla attendibile) e 10 (totalmente attendibile). Gli intervistatori forniscono informazioni anche sul livello di comprensione delle domande da parte dell'intervistato e sul clima complessivo nel quale si è svolta l'intervista.

cautela, tenendo conto che, almeno in parte, le differenze riscontrate possono essere dovute alle diverse definizioni adottate¹⁵. Una volta tenuto conto di questi aspetti, gli studi disponibili mostrano come le stime del reddito e la ricchezza immobiliare familiare risultano nel complesso allineate fra le due fonti. Maggiori discrepanze appaiono invece per le componenti della ricchezza finanziaria¹⁶.

7. L'aggregazione delle variabili

I principali aggregati economici, costruiti a partire dalle voci del questionario, sono riportati negli schemi qui presentati.

Piano di aggregazione delle variabili: conto del reddito

Nome	Descrizione ⁽¹⁾
Y	Reddito disponibile netto
YL	Reddito da lavoro dipendente
YL1	Retribuzioni nette
YL2	Integrazioni non monetarie
YT	Pensioni e trasferimenti netti
YTP	Pensioni e arretrati
YTP1	Pensioni
YTP2	Arretrati
YTA	Altri trasferimenti
YTA1	Assistenza economica (CIG, ecc.)
YTA2	Borse di studio
YTA3	Assegni alimenti e regali/contributi ricorrenti
YTA31	in entrata
YTA32	in uscita (-)
YM	Reddito netto da lavoro autonomo
YMA1	Reddito netto da lavoro autonomo
YMA2	Utili e altri redditi da partecipazione
YC	Reddito da capitale
YCA	Reddito da fabbricati
YCA1	Affitti effettivi
YCA2	Affitti imputati ⁽²⁾
YCF	Reddito da capitale finanziario ⁽³⁾
YCF1	Interessi attivi su depositi
YCF2	Interessi attivi su titoli di Stato
YCF3	Rendimenti altri titoli
YCF4	Interessi passivi (-)

$$Y = YL + YT + YM + YC$$

(1) Il simbolo (-) indica che la posta entra con il segno negativo nel calcolo dell'aggregato gerarchicamente superiore.

(2) Ad eccezione degli immobili utilizzati per l'attività di lavoro indipendente.

(3) Tasso di interesse * stock di capitale.

Piano di aggregazione delle variabili: conto dell'utilizzazione reddito

Nome	Descrizione
Y	Reddito disponibile netto
C	Consumi
CD	Consumi durevoli
CD1	Spese per mezzi di trasporto

¹⁵ Per un'analisi delle incoerenze tra le due fonti e per possibili esercizi di riconciliazione, si veda G. D'Alessio e A. Neri, *Income and wealth sample estimates consistent with macro aggregates: some experiments*, Banca d'Italia, Questioni di Economia e Finanza, n. 272, 2015.

¹⁶ Oltre che ai noti fenomeni di mancata risposta e *under-reporting*, parte di questa discrepanza è spiegabile da come le poste sono valutate nelle diverse fonti nonché da alcune voci, come le azioni, che per il settore delle famiglie sono stimate residualmente nei conti finanziari. Per un'analisi particolareggiata delle differenze di cui tener conto nel confrontare le stime micro dell'indagine con quelle macro della contabilità finanziaria si veda R. Bonci, G. Marchese e A. Neri, *La ricchezza finanziaria nei conti finanziari e nell'indagine sui bilanci delle famiglie italiane*, Banca d'Italia, Temi di discussione, n. 565, 2005. Per l'analisi dell'*under-reporting* si vedano: L. D'Aurizio, I. Faiella, S. Iezzi e A. Neri, *L'under-reporting della ricchezza finanziaria nell'indagine sui bilanci delle famiglie*, Banca d'Italia, Temi di discussione, n. 610, 2006; L. Cannari e G. D'Alessio, *Non-Reporting and Under-Reporting Behavior in the Bank of Italy's Survey of Household Income and Wealth*, in *Bulletin of the International Statistical Institute*, vol. LV, n. 3, 1993, p. 395-412. Sulla mancata risposta si veda G. D'Alessio e I. Faiella, *Nonresponse behaviour in the Bank of Italy's Survey of Household Income and Wealth*, Banca d'Italia, Temi di discussione, n. 462, 2002. Per un lavoro che considera entrambi gli aspetti si veda A. Neri e M.G. Ranalli, *To misreport or not to report? The case of the Italian Survey on Household Income and Wealth*, *Statistics in Transition*, 12, 2011, p.281-300.

CD2	Spese per mobili, ecc.
CN	Consumi non durevoli
S	Risparmio ⁽¹⁾
Y = C + S	
(1) Determinato a saldo.	

Piano di aggregazione delle variabili: conto del patrimonio

Nome	Descrizione ⁽¹⁾
W	Ricchezza netta
AR	Attività reali
AR1	Immobili
AR2	Aziende
AR3	Oggetti di valore
AF	Attività finanziarie
AF1	Depositi
AF2	Titoli di Stato
AF3	Altri titoli
AF4	Crediti commerciali e da altre fam.
PF	Passività finanziarie (-)
PF1	Debiti verso banche e soc. finanz.
PF2	Debiti commerciali
PF3	Debiti verso altre famiglie
Per memoria:	
BD	Beni di consumo durevoli
BD1	Mezzi di trasporto
BD2	Mobili, ecc.
W = AR + AF - PF	
(1) Il simbolo - indica che la posta entra con il segno negativo nel calcolo dell'aggregato gerarchicamente superiore.	

8. Le stime campionarie

Data la generica variabile oggetto di studio y , uno stimatore corretto¹⁷ della media della popolazione è lo stimatore di Horwitz-Thompson-Narain dato da:

$$(1) \quad \bar{y} = \frac{\sum_h \sum_i \sum_j y_{hij} w_{hij}}{\sum_h \sum_i \sum_j w_{hij}} \quad j = 1, \dots, n_{hi} \quad i = 1, \dots, a_h \quad h = 1, \dots, H$$

dove y_{hij} rappresenta il valore della variabile y osservato per la j -esima famiglia intervistata nel i -esimo comune dell' h -esimo strato, mentre w_{hij} costituisce il suo peso campionario.

Il peso da attribuire alle unità campionarie è ottenuto per mezzo di una procedura articolata nelle seguenti fasi¹⁸:

¹⁷ Uno stimatore è "corretto" (*unbiased*) quando il suo valore medio coincide con il parametro della popolazione. Lo stimatore (1) è in realtà "approssimativamente" corretto, ossia presenta una distorsione di entità limitata che tende a zero al crescere della numerosità campionaria. Tale stimatore, presenta una distorsione massima pari al coefficiente di variazione campionario dei pesi (normalmente intorno all'1 per cento per le stime familiari e allo 0,7 per cento per quelle individuali). Si veda L. Kish, *Survey Sampling*, New York, Wiley, 1995.

¹⁸ Per una descrizione dettagliata della struttura di ponderazione adottata nell'indagine e degli effetti dei pesi sulla varianza campionaria degli stimatori si veda I. Faiella e R. Gambacorta, *The weighting process in the SHIW*, Banca d'Italia, Temi di discussione, n. 636, 2007.

- a) inizialmente viene calcolato il *peso da disegno* come reciproco della probabilità di inclusione delle unità ($w^{(0)}$);
- b) si ottiene poi un peso corretto per la mancata risposta totale ($w^{(1)}$) moltiplicando $w^{(0)}$ per l'inverso del tasso di risposta dello strato cui ciascuna famiglia appartiene;
- c) il peso $w^{(1)}$ viene inoltre modificato per tener conto della presenza della componente *panel* (ottenendo il peso $w^{(2)}$);
- d) infine, $w^{(2)}$ viene aggiustato sulla base di informazioni ausiliarie (*post-stratificazione*), giungendo al peso finale $w^{(3)}$.

Inizialmente, in accordo con il piano di campionamento, a ciascun individuo della famiglia viene attribuito un peso, definito come l'inverso della probabilità di inclusione nel campione (*peso da disegno*).

Il peso da disegno viene aggiustato per tener conto del rifiuto di alcune famiglie di partecipare all'indagine. La correzione per la mancata risposta consiste nell'inflazionare i pesi per il rapporto tra la dimensione del campione teorico e quella del campione effettivo all'interno di ciascun strato di secondo stadio (*Weighting Class Adjustment*)¹⁹. Come si è visto nel paragrafo 1, il disegno dell'indagine prevede che una parte del campione sia costituita da famiglie intervistate nelle rilevazioni precedenti (famiglie *panel*). Le famiglie *panel* possono presentare caratteristiche socio-demografiche in parte diverse rispetto a quelle dell'intero campione, sostanzialmente a causa del processo di deterioramento della componente *panel* legato alla non partecipazione in successive *wave* (*attrition*). Per correggere tale possibile fonte di distorsione delle stime, si procede a una stratificazione a posteriori della componente *panel* del campione sulla base di alcune caratteristiche relative alla precedente indagine. In particolare l'aggiustamento è basato sull'incrocio fra classi di reddito (a 4 modalità) e condizione professionale del capofamiglia (a 3 modalità)²⁰. Inoltre, il peso viene modificato per riprodurre la quota ottimale della componente *panel* (stimata in circa il 50 per cento del campione). Infine, i pesi vengono calibrati utilizzando informazioni esterne che, correlate con le variabili oggetto di indagine, hanno lo scopo di migliorare l'accuratezza delle stime e limitare ulteriormente le distorsioni legate alla mancata risposta. La correzione viene condotta imponendo al campione finale l'allineamento alle caratteristiche della popolazione per quanto riguarda il sesso, la classe di età (meno di 26 anni, da 26 a 45, da 46 a 65, oltre 65 anni), l'area geografica (Nord, Centro, Sud e Isole) e la dimensione del comune di residenza (fino a 20.000 abitanti, da 20.000 a 40.000, da 40.000 a 500.000, oltre 500.000 abitanti)²¹.

9. Gli errori standard

L'analisi dei dati campionari ha lo scopo di ottenere stime di uno o più parametri della popolazione (come ad es. medie, totali e coefficienti di regressione). Una valutazione della variabilità campionaria è necessaria per poter costruire degli intervalli di confidenza delle stime puntuali²² e per condurre test di ipotesi.

¹⁹ La correzione per la mancata partecipazione permette di correggere anche per i casi in cui le famiglie non sono state trovate dall'intervistatore all'indirizzo anagrafico a causa di trasferimenti, decessi o di indirizzi errati.

²⁰ Per tenere conto dell'*attrition* si sarebbero potuti utilizzare modelli statistici per la mancata partecipazione, come proposto da A. Giraldo, E. Rettore e U. Trivellato, *Attrition bias in the Bank of Italy's Survey of Households' Income and Wealth*, International Conference on Quality in Official Statistics, Stockholm, 14-15 Maggio 2001. Va comunque notato che, l'utilizzo di modelli o di tecniche di riponderazione nell'aggiustamento per la mancata risposta producono un impatto sostanzialmente simile, e normalmente contenuto, sulle stime campionarie. Si veda G. Kalton e I. Flores Cervantes, *Weighting Methods*, in *Journal of Official Statistics*, vol. 19, n. 2, 2003, pp. 81-97.

²¹ La tecnica utilizzata consente di allineare in modo simultaneo i pesi campionari alla distribuzione di alcune caratteristiche note da fonti esterne considerando esclusivamente le probabilità marginali. Tale metodo è noto come *Iterative Proportional Fitting* (o *Raking*). Si veda G. Kalton e I. Flores Cervantes, *Weighting Methods*, in *Journal of Official Statistics*, vol. 19, n. 2, 2003, pp. 81-97.

²² Un intervallo di confidenza è un range di valori che include con una certa probabilità il parametro che si vuole stimare. Questo intervallo è sostanzialmente una misura di affidabilità circa la collocazione più probabile della stima che si ottiene applicando lo stimatore ad un certo campione. Data la numerosità campionaria dell'indagine, un intervallo di confidenza per la media può essere costruito come: *media della variabile* $\pm 1.96 \cdot \text{errore standard della media}$.

Per determinare la variabilità campionaria degli stimatori considerati è necessario utilizzare metodi che consentano di tenere conto sia della procedura con cui gli elementi del campione sono stati selezionati sia della natura di tali stimatori²³.

La complessità del disegno campionario adottato e la non linearità della maggior parte degli stimatori utilizzati non consentono l'utilizzo delle formule analitiche, incluse quelle ottenute tramite tecniche di linearizzazione della varianza. Si è pertanto fatto ricorso al metodo di replicazione *jackknife*, che ricava informazioni sulla variabilità riproducendo le fasi di stima su *repliche* del campione originario²⁴.

Per poter procedere alla stima dell'errore standard è necessario stabilire un disegno coerente con il processo di selezione delle unità campionarie, ma che non presenti problemi computazionali²⁵. In primo luogo tutti i comuni con oltre 40 mila abitanti (comuni *auto-rappresentativi*) sono introdotti in uno strato a sé, in quanto non contribuiscono in alcun modo al processo di randomizzazione nel primo stadio di selezione. Le famiglie campionarie di ciascuno di questi comuni sono poi raggruppate in due gruppi formati casualmente. I rimanenti comuni sono accorpati nel loro strato di origine in modo da avere due comuni per strato. Alla fine di questo procedimento vengono individuati circa 350 "pseudo" strati con al loro interno due "pseudo" unità primarie.

La varianza è computata utilizzando il metodo *jackknife*:

1. il numero c di replicazioni è pari al numero di "pseudo" strati;
2. in ciascuna replicazione, all'interno di un solo "pseudo" strato è messo a zero il peso della prima "pseudo" unità primaria e il peso campionaria dell'altra è aumentato di un fattore in grado di compensare il peso delle osservazioni dell'unità cancellata;
3. con tale peso sono calcolati, per ciascuna replicazione, gli stimatori di interesse $\hat{\theta}_{(i)}$;
4. dato che il disegno per la stima della varianza presenta due unità per strato, la stima dell'errore standard è calcolata come la radice quadrata della somma degli scarti al quadrato tra stima delle replicazioni e stima sul campione complessivo $\hat{\theta}$:

$$(6) \quad \text{stderr}_j = \sqrt{\sum_{i=1}^c (\hat{\theta}_{(i)} - \hat{\theta})^2} \quad 26.$$

Una misura utile per valutare di quanto la varianza degli stimatori, calcolata tenendo in considerazione la complessità del piano di campionamento, diverga da quella ottenuta nell'ipotesi di campionamento casuale semplice è il rapporto tra queste due valutazioni della varianza campionaria: per il generico stimatore $\hat{\theta}$ il cosiddetto *design effect* (*deff*) è pari a:

$$(7) \quad \text{deff}(\hat{\theta}) = \frac{V(\hat{\theta})_{\text{compl}}}{V(\hat{\theta})_{\text{ccs}}}$$

Le stime di questi coefficienti per alcune variabili utilizzate nell'indagine sono riportate nella tavole statistiche sull'anno di riferimento.

²³ I problemi connessi all'inferenza con piani campionari complessi e il modello adottato per la stima della varianza sono descritti in dettaglio in I. Faiella, *Accounting for sampling design in the SHIW*, Banca d'Italia, Temi di discussione, n. 662, 2008.

²⁴ Nell'ambito dei metodi di replicazione le tecniche maggiormente utilizzate sono il metodo dei *random groups*, il *bootstrap*, le *balanced repeated replications* e il *jackknife*. Agli ultimi due metodi vengono riconosciute proprietà superiori quando si tenga conto della struttura imposta ai dati dalla natura complessa del disegno di campionamento (stratificazione e selezione a due stadi). Si vedano K.M. Wolter, *Introduction to Variance Estimation*, Berlin, Springer Verlag, 1985; J. Shao e J. Tu, *The Jackknife and Bootstrap*, Berlin, Springer Verlag, 1995.

²⁵ Ad esempio è necessario collassare gli strati che contengono una sola unità primaria; è inoltre desiderabile che ogni strato presenti una numerosità adeguata a fornire stime stabili.

²⁶ Si veda L. Kish e M. Frankel, *Inference from complex samples*, in *The Journal of the Royal Statistical Society, Series B*, vol. 36 (1), 1974, pp. 1-37.

10. La distribuzione dei dati

I microdati anonimi delle indagini sui bilanci delle famiglie condotte dal 1977 al 2014 sono disponibili sul sito Internet della Banca d'Italia all'indirizzo <http://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-impres/bilanci-famiglie>²⁷. Le informazioni necessarie per lo sfruttamento dei dati (questionari delle indagini più recenti, nomi delle variabili, note per l'utilizzo dei dati, ecc.) sono disponibili nella sezione documentazione in formato PDF. I dati sono disponibili in formato SAS, STATA e in formato ASCII separato da virgola (CSV, *comma separated value*).

È prevista la distribuzione di due tipologie di archivi. Gli **archivi annuali** contengono i dati delle indagini dal 1989 in versione pressoché integrale (incluse le sezioni monografiche delle singole indagini). L'**archivio storico** riporta le informazioni a partire dal 1977, in formato omogeneo, per il sottoinsieme delle variabili ritenute utili per le analisi longitudinali.

I pesi degli archivi annuali possono essere diversi dai corrispondenti dell'archivio storico. I pesi degli archivi annuali infatti non subiscono alcun processo di revisione, mentre i pesi dell'archivio storico vengono allineati alle statistiche demografiche sulla popolazione italiana rilasciate dall'ISTAT (ad es. ricostruzioni intercensuarie), quando queste sono oggetto di revisione. La revisione dei pesi campionari può causare delle discrepanze tra le statistiche desumibili dall'attuale versione dell'archivio storico e quelle di versioni precedentemente rilasciate.

11. L'uso dei dati

L'indagine rappresenta una importante fonte di informazione sulle condizioni economiche delle famiglie e contribuisce in modo rilevante al dibattito socio-politico su tale tema. I principali risultati vengono diffusi nella collana statistiche della Banca d'Italia e hanno un ampio impatto mediatico: vengono infatti ripresi dai principali quotidiani nazionali e citati nei dibattiti politici televisivi.

I micro dati dell'indagine vengono inoltre impiegati per la realizzazione di progetti di ricerca nell'ambito delle attività istituzionali della Banca d'Italia. Esempi di studi riguardano gli effetti delle manovre di politica monetaria sui consumi delle famiglie, la distribuzione del reddito e della ricchezza, il funzionamento dei mercati finanziari e aspetti inerenti la tassazione o le riforme del sistema pensionistico. Un impiego di particolare rilevanza è la valutazione dell'impatto di particolari misure di politica economica attraverso un modello di micro-simulazione.

I dati dell'indagine, dopo essere resi totalmente anonimi, sono messi a disposizione anche di utenti esterni appartenenti ad istituzioni di ricerca pubbliche e private. Le principali pubblicazioni scientifiche basate sull'indagine si possono quantificare in circa 800 lavori²⁸.

12. L'indagine nel contesto internazionale

Negli anni più recenti l'indagine è entrata a far parte di progetti internazionali che consentono il confronto della condizione economica delle famiglie tra i principali paesi occidentali.

I primi due progetti sono il *Luxembourg Income Study* ed il *Luxembourg Wealth Study*, il cui scopo è quello di produrre archivi di dati armonizzati sul reddito e ricchezza delle famiglie (www.lisdatacenter.org/). L'armonizzazione avviene "a posteriori" individuando le grandezze rilevate in modo comparabile e creando una documentazione unica per le varie indagini in modo da rendere più facile il loro uso. I microdati sono messi a disposizione tramite un sistema di elaborazione a distanza e un *web tabulator*.

²⁷ La documentazione e le pubblicazioni sono disponibili sia in italiano sia in inglese.

²⁸ Si veda la lista dei riferimenti bibliografici dei principali lavori di ricerca che sfruttano la base dati dell'indagine all'indirizzo <http://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-impres/bilanci-famiglie>.

A partire dal 2008 l'indagine è anche parte integrante di un progetto della Banca Centrale Europea che ha lo scopo di realizzare un'indagine armonizzata sui bilanci familiari nell'area dell'euro (*Household Finance and Consumption Survey - HFCS*). L'indagine rileva le variabili necessarie a ricostruire i bilanci delle famiglie sia in termini di consistenze sia di flussi, con particolare attenzione alle componenti della ricchezza, e fornisce informazioni sul comportamento delle famiglie utili per una migliore comprensione dei meccanismi di trasmissione della politica monetaria e per la valutazione di profili di stabilità finanziaria. I micro dati anonimizzati sono a disposizione degli studiosi per finalità di ricerca (www.ecb.europa.eu/pub/economic-research/research-networks/html/researcher_hfcn.en.html). Il principale vantaggio di questo progetto è l'armonizzazione delle metodologie e delle definizioni usate dai vari paesi per condurre le indagini che ne fanno parte che consente il confronto dei risultati concernenti l'Italia con quelli degli altri paesi dell'area dell'euro²⁹.

²⁹ Sulle differenze esistenti tra i dati dell'indagine sui bilanci delle famiglie italiane e quelli dell'indagine HFCS si veda R. Gambacorta, G. Ilardi, A. Locatelli, R. Pico e C. Rampazzi, *Principali risultati dell'Household Finance and Consumption Survey: l'Italia nel confronto internazionale*, Banca d'Italia, Questioni di Economia e Finanza, n. 161, 2013.