La qualità del processo di produzione dei dati statistici

1 Aprile 2004

Sintesi

- Le fasi di un'indagine statistica
- Qualità dell'informazione statistica
- Esempi applicativi
  - Uso di dati amministrativi a fini statistici
  - Un'indagine di controllo del censimento della popolazione
  - Monitoraggio della rete di rilevazione
  - Controllo statistico della qualità della registrazione dati

L'indagine statistica come processo di produzione

Scopo dell'indagine è quello di produrre descrizioni riassuntive di carattere quantitativo (statistiche), riguardanti il collettivo di interesse

La progettazione di un'indagine statistica è un impegno multidisciplinare che prevede l'impiego di professionalità di tipo differente

Aspetti di progettazione

- Obiettivi, definizioni e classificazioni;
- Disegno d'indagine;
- Tempi e costi;
- Fasi operative;
- Impiego di fonti amministrative;
- Sistema di controllo della qualità;
- Elaborazioni statistiche;
- Diffusione.

Obiettivi, definizioni e classificazioni

- Fenomeno di interesse
- Collettivo di riferimento
- Tipologia delle variabili studiate
  - identificative
  - di classificazione
  - d'interesse
  - di controllo
- Classificazioni

Disegno di indagine

- Tipo di indagine in relazione agli obiettivi
  - Longitudinale o trasversale
- Campionamento
  - Semplice o complesso
- Archivi
  - Centralizzati o delocalizzati
  - Informatizzati o cartacei
- Tecnica di indagine
  - Intervista diretta, telefonica, postale, indagine estimativa
  - Uso delle nuove tecnologie

Tempi e costi

Le relazioni tra tempi e costi sono un fattore critico per la riuscita di un'indagine e sono fortemente connessi alla qualità dei dati prodotti

- poche risorse possono provocare cadute di qualità tali da tradursi in perdite economiche
- eccessive risorse possono tradursi in sprechi invece che in maggiore qualità
- maggiori risorse si possono tradurre in diminuzione di tempi a parità di qualità (es. più rilevatori)
- maggiore qualità può dilatare i tempi a parità di risorse e viceversa (es. più, o meno, controlli prima della diffusione dei dati)

Fasi operative

Ciclo produttivo dell'indagine che va dalla misurazione delle caratteristiche di interesse sulle unità selezionate fino alla disponibilità dei dati aggregati

- Rilevazione;
- Codifica dei quesiti aperti;
- Registrazione dati su supporto magnetico;
- Revisione automatica e/o interattiva;
- Elaborazioni statistiche;
- Validazione.

Rilevazione

Le unità selezionate per l'indagine vengono contattate allo scopo di raccogliere l'informazione rilevante ai fini dello studio.

Obiettivi:

- individuare l'unità di rilevazione (famiglia, impresa,…) e convincerla a partecipare all'indagine;
- raccogliere l'informazione in modo neutrale, cioè senza influenzare il rispondente;
- lasciare una buona impressione per facilitare eventuali contatti futuri (indagini longitudinali, ritorni sul campo, indagini di controllo).

Rilevazione (segue)

Gli aspetti fondamentali che devono essere considerati sono:

- predisposizione del questionario e dei modelli ausiliari;
- tempistica e interazione fra gli enti preposti alla rilevazione;
- campagne di sensibilizzazione dei rispondenti;
- formazione del personale;
- supervisione delle operazioni e recupero delle informazioni incomplete.

Codifica dei quesiti aperti

Trasposizione di informazioni sotto forma di linguaggio libero in un insieme finito di codici rispondenti ad una classificazione precostituita.

Vi si ricorre quando il rispondente non saprebbe collocare in modo corretto l'informazione secondo la classificazione richiesta, a causa della sua notevole complessità.

- ATECO [Istat, (1999)] - Cl. delle attività economiche
- ICD-9 [ONU, (1977)] - Cl. delle cause di morte

Attività demandata a personale appositamente formato che sappia usare software dedicato.

Registrazione su supporto informatico

Consiste nel convertire le informazioni disponibili su questionario cartaceo raccolte presso i rispondenti, su un supporto di formato elaborabile dalle procedure predisposte per l'indagine.

Un operatore digita su una tastiera esattamente ciò che legge sul questionario cartaceo.

Solitamente si impiega personale esterno non specializzato e per questo la fase di registrazione può essere fonte di gravi errori.

E' possibile utilizzare forme di registrazione controllata.

Revisione automatica

Consiste nell'individuare e rimuovere gli **errori di misurazione** presenti nei dati.

- Gli **errori di misurazione** riguardano direttamente i dati elementari manifestandosi come differenze tra i valori "veri" e i valori osservati di una variabile di interesse.
- Gli **errori di misurazione** possono presentarsi in ogni fase del processo di produzione dell'informazione statistica.
- Gli **errori di misurazione** possono provocare distorsioni nelle distribuzioni delle variabili investigate, nelle stime finali dei dati (totali, medie, ecc.) e in tutte le analisi statistiche effettuate sui dati non corretti.

Revisione automatica (segue)

Gli errori di misurazione possono essere classificati secondo la loro causa

Errori sistematici

da attribuirsi a difetti strutturali o organizzativi del processo di produzione dell'informazione statistica, alla struttura del modello, o al sistema di registrazione adottati; si manifestano nella maggior parte delle osservazioni come deviazioni "in una stessa direzione" dal valore vero di una o più variabili rilevate

Errori casuali

errori la cui origine è da attribuirsi a fattori aleatori non direttamente individuabili; nel caso di **variabili quantitative**, si ipotizza spesso una distribuzione normale a media nulla, mentre nel caso di **variabili qualitative** si suppone che i valori errati non alterino la distribuzione di frequenze relativa ai dati corretti

Revisione automatica (segue)

Un'altra classificazione li distingue secondo l'effetto

- **Fuori dominio:** il valore rilevato non appartiene ad un insieme predefinito di valori ammissibili. Errore tipico delle var. qualitative.
- **Valore anomalo (outlier)** il valore rilevato presenta catteristiche significativamente diverse da quelle assunte nella maggior parte delle altre unità
- **Valore incompatibile:** i valori di una o più variabili contraddicono predefinite regole di natura logica e/o relazioni di tipo matematico
- **Mancata risposta parziale (MRP):** per una data unità mancano i valori di un sottoinsieme di variabili richieste.

Revisione automatica (segue)

Viene effettuata mediante procedure denominate di controllo e correzione. Tali procedure possono essere schematicamente riassunte in due fasi principali

- La fase di individuazione degli errori
  - errori sistematici: analisi esplorative
  - valori anomali: test di accettazione sulle distribuzioni di interesse
  - errori casuali: controlli di coerenza
- La fase di correzione degli errori
  - Ritorni sul campo
  - Metodi deterministici $\mathbf{y}=b_0+b_1\mathbf{x}$
  - Metodi probabilistici $\mathbf{y}=b_0+b_1\mathbf{x}+\varepsilon$

Elaborazioni statistiche

Costituite dalla produzione ed interpretazione delle stime a partire dai dati elementari. Oltre a costituire i risultati dell'indagine corredati dalla una loro interpretazione sono molto utili per:

- predisporre nuove indagini sulla base dei risultati di studi pilota o di precedenti indagini;
- formulare obiettivi realistici riguardanti la qualità;
- identificare problemi e requisiti del processo di produzione.

Elaborazioni statistiche (segue)

Analisi preliminari

- studio della documentazione riguardante definizioni, concetti e procedure operative
- contatti con il personale responsabile delle varie fasi di indagine
- studio delle procedure di correzione e inclusione in analisi dei record adatti all'elaborazione

Analisi dei dati

- produzione delle stime utilizzando anche tecniche di integrazione tra micro o macro dati
- conduzione di analisi descrittive semplici
- applicazione di metodi di analisi multivariata e di tecniche diagnostiche per la valutazione dell'adattamento dei modelli ai dati
- condivisione del lavoro con esperti dei metodi statistici

Validazione

Processo attraverso il quale si valuta lo scarto tra gli obiettivi di qualità programmati per l'indagine e i risultati effettivamente conseguiti.

- Produrre documentazione per permettere all'utente una valutazione della conformità tra i dati e i propri obiettivi;
- Analizzare la conformità tra i dati prodotti e quelli disponibili da altre fonti;
- Studiare la qualità del prodotto;
- Studiare la qualità del processo di produzione.

Diffusione

Fase nella quale l'informazione statistica viene resa disponibile all'utenza

La diffusione è cruciale per un prodotto di buona qualità

L'Istat predispone varie forme di diffusione per adattarsi alle esigenze dei diversi utenti:

- Pubblicazioni generali e specifiche;
- Basi dati accessibili on line;
- File standad;
- Elaborazioni *ad hoc*;
- Laboratorio ADEle (per l'analisi dei dati elementari)

La qualità dei dati statistici

Secondo le norme ISO 8402-1984 la qualità di un bene o servizio è definita come

"Il possesso della totalità delle caratteristiche che portano al soddisfacimento delle esigenze, esplicite o implicite, dell'utente".

Prodotto e Processo

Prodotto: informazione statistica

Processo: procedimento che dall'informazione "grezza" raccolta sulle unità statistiche conduce alle stime riguardanti la popolazione oggetto

La qualità è solitamente riferita al prodotto o al processo ma tra le due esiste una relazione strettissima

Dimensioni della qualità

Dal punto di vista del prodotto possiamo classificare 7 componenti della qualità

1. Rilevanza (o pertinenza)

   Capacità dell'informazione di soddisfare le esigenze conoscitive degli utenti

2. Accuratezza

   Grado di corrispondenza fra la stima ottenuta dall'indagine e il vero (ma ignoto) valore della caratteristica oggetto di interesse nella popolazione obiettivo

Dimensioni della qualità (segue)

3. Tempestività e puntualità

   Intervallo di tempo intercorrente fra il momento della diffusione dell'informazione prodotta e l'epoca di riferimento della stessa

4. Accessibilità e chiarezza (o trasparenza)

   Corrisponde alla semplicità per l'utente di reperire, acquisire e comprendere l'informazione disponibile in relazione alle proprie finalità

Dimensioni della qualità (segue)

5. Confrontabilità

   Possibilità di paragonare nel tempo e nello spazio le statistiche riguardanti il fenomeno di interesse

6. Coerenza

   Possibilità di combinare le inferenze semplici in induzioni più complesse senza che le conclusioni risultino in contrasto fra loro

7. Completezza

   Capacità dei processi di integrarsi per fornire un quadro informativo soddisfacente del dominio di interesse

Il sistema di controllo

Dal punto di vista del processo di produzione distinguiamo tre tipi di azione

- Azioni di prevenzione

  (lettera di preavviso, istruzioni ai rilevatori, registrazione controllata,…)

- Azioni di correzione in corso d'opera

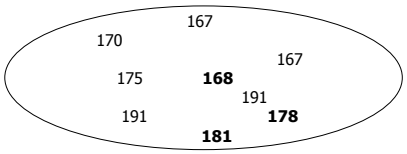
  (ritorni sul campo, revisione automatica,…)

- Azioni di valutazione a posteriori

  (reinterviste, analisi fonti alternative, studi della soddisfazione degli utenti,...)

Accuratezza - l'errore campionario

Si misurano solo una parte, scelta con procedura di selezione casuale, delle unità della popolazione. Quindi in genere la stima ottenuta non corrisponde al vero valore posseduto dalla popolazione.

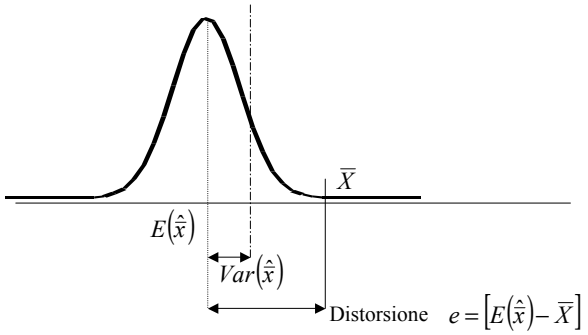


$\overline{X} = 176{,}4$

$\hat{\bar{x}} = 175{,}7$

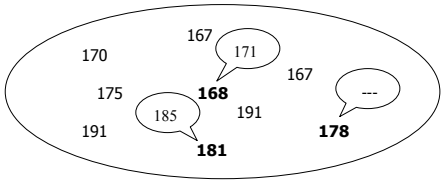
Distribuzione campionaria delle stime

Se immaginiamo di estrarre tutti i possibili campioni dalla popolazione le stime ottenute ogni volta si disporranno in prossimità del vero valore di interesse



Accuratezza- l'errore non campionario

Quando tentiamo di misurare una unità possiamo produrre un errore: o perché non riusciamo a raccogliere alcun valore o perché raccogliamo un valore diverso da quello vero



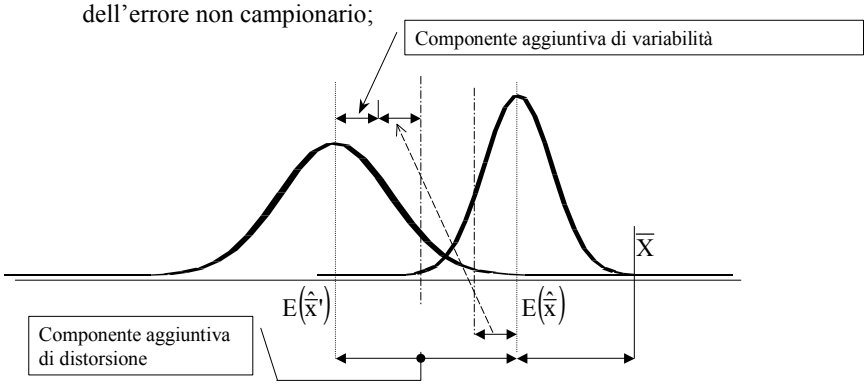
Media nella popolazione $\overline{X} = 176{,}4$

Media campionaria $\hat{\bar{x}} = 175{,}7$

Media campionaria affetta da errore non campionario $\hat{\bar{x}}' = 178{,}0$

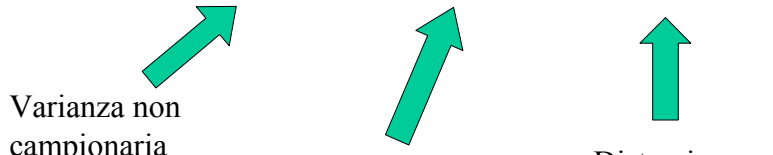
Distribuzione campionaria delle stime affetta dall'errore non campionario

La distribuzione della stima campionaria si modifica per effetto dell'errore non campionario;



Errore totale

$$E_{pm}\left(\hat{y}-Y\right)^2 = E_p\left(V_m\left(\hat{y}|c\right)\right) + V_p\left(E_m\left(\hat{y}|c\right)\right) + \left[E_{pm}\left(\hat{y}\right)-Y\right]^2$$



Varianza non campionaria

Varianza di campionamento

Distorsione

Classificazione degli errori non campionari

- Errori di lista
  - Sotto-copertura
  - Sovra-copertura
  - Errori che pregiudicano il contatto
- Mancate risposte
  - Totali
  - Parziali
- Misurazione
  - Errori di rilevazione
  - Errori di processo

Misure e indicatori dell'accuratezza

- Misure
  - ☺ Permettono una valutazione della qualità del prodotto
  - ☹ Sono costose
- Indicatore di qualità
  - ☺ Sono economici
  - ☺ Suggeriscono il punto del processo dove intervenire
  - ☹ Si riferiscono solo indirettamente alla qualità del prodotto

Misurare l'accuratezza – 1

- Errori di copertura
  - Sovracopertura
    - analisi delle condizioni di eleggibilità
  - Sottocopertura ed errori di lista
    - confronto con fonti alternative (es. dati amministrativi)

Misurare l'accuratezza – 2

- Mancate risposte
  - totali
    - campione di non rispondenti
    - caratterizzazione dei non rispondenti
    - analisi dei motivi di non risposta
  - parziali
    - analisi dell'errore di misurazione indotta dalle procedure di imputazione dei valori mancanti

Misurare l'accuratezza – 3

- Errori di misurazione
  - reinterviste
    - con riconciliazione degli errori
    - senza riconciliazione degli errori
  - randomizzazione
    - test di alternative
    - compenetrazione delle assegnazioni

Indicatori di processo

- Stato delle liste di riferimento
  - variabili identificative (contatto)
  - variabili di classificazione (eleggibilità)
- Rilevazione
  - mancate risposte per rilevatori, comuni, motivi
- Registrazione
  - byte errati
- Revisione
  - confronto dati prima e dopo la revisione
- Tempi (ritardi su date programmate)
- Costi (preventivo – consuntivo)

14° Censimento generale della popolazione

Finalità

- conteggio esaustivo delle persone **residenti** (popolazione legale) e **presenti** sul territorio italiano alla data del 21 ottobre 2001.
- revisione e aggiornamento dei registri anagrafici della popolazione residente (dati amministrativi).

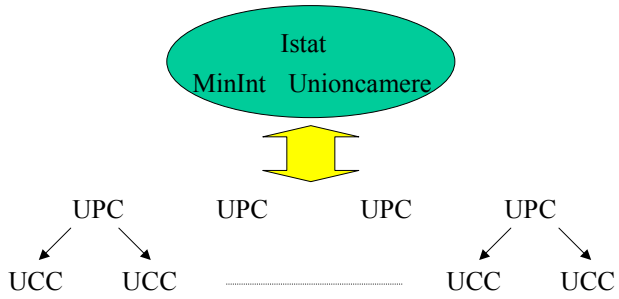
Contestualmente viene condotto il censimento degli edifici e delle abitazioni e quello delle imprese.

14° Censimento generale della popolazione

(segue)

- Unità statistiche: individui residenti e presenti in famiglie e convivenze.
- Unità di rilevazione: sezioni di censimento

Rete di rilevazione



14° Censimento generale della popolazione

(segue)

- Tecnica di indagine: il rilevatore percorre la sezione di censimento e rileva le unità statistiche sul territorio per mezzo di un questionario consegnato alle famiglie e ritirato dopo che queste lo hanno compilato;
- Modalità di registrazione su supporto magnetico: lettura ottica;
- Revisione: prima revisione sul campo con l'aiuto delle liste anagrafiche (dati amministrativi), successiva fase di revisione quantitativa condotta all'Istat confrontando i questionari registrati con i conteggi effettuati dai comuni sui modelli ausiliari e fase finale di revisione automatica per le informazioni contenute nei questionari;
- Elaborazione: costruzione di tabelle provinciali e aggregazione a livello regionale e statale;
- Diffusione: volumi a stampa e pubblicazione di un sottoinsieme delle informazioni sul sito Internet dell'Istat.

Indagine di copertura

- Obiettivo: Stimare il grado di copertura della popolazione residente conteggiata dal 14° censimento italiano della popolazione e del territorio
- Popolazione di interesse: individui e famiglie che al 21 ottobre del 2001 possiedono la dimora abituale (residenza) in alloggi situati sul territorio nazionale
- A quali popolazioni NON è interessata l'indagine di copertura: persone non abitualmente dimoranti (occasionalmente presenti o temporaneamente dimoranti presso l'alloggio), convivenze, edifici e abitazioni.

Gli errori di copertura

- Si definiscono errori di copertura tutti gli errori che provocano distorsioni nel conteggio della popolazione censita
- Gli errori di copertura si dividono in:
  - Errori di sovra-copertura
  - Errori di sotto-copertura
- Gli errori di copertura dipendono dall'area considerata dai conteggi di popolazione a causa del legame tra gli individui e il luogo della loro residenza.

Esempio : Una persona è conteggiata come residente all'indirizzo errato (localizzato in **Toscana**) mentre al vero indirizzo, collocato in **Sicilia**, non è stata enumerata.

Alcuni casi emblematici

- Residenti non censiti
- Residenti censiti nel luogo errato
- Doppie enumerazioni
- Non residenti censiti

Potrebbero verificarsi altri casi quali quelli di residenti censiti più volte in luoghi errati, ma tali casi non saranno presi in considerazione in quanto ritenuti molto rari.

Parametri di interesse

Il vero valore della popolazione (POP) può essere calcolato a partire dalla popolazione conteggiata al censimento (CEN), tenendo conto degli errori di sovra (SOV) e sotto (SOT) copertura

$$POP=CEN+SOT-SOV$$

Più che ai valori assoluti interessa conoscere il valore relativo

$$(SOV-SOT)/POP$$

E le componenti SOV/POP e SOT/POP

Domini di interesse [1]

- *Dettaglio geografico*: cinque ripartizioni territoriali, regioni
- *Dimensione demografica dei comuni*: suddivisione in quattro classi di dimensione demografica (<10, 10-50, 50-350, ≥350 in migliaia di abitanti), grandi comuni
- *Densità di popolazione all'interno dei comuni*: tipologia di località contenente la sezione (centro, nucleo, case sparse)
- *Caratteristiche riguardanti le famiglie e gli individui*: numero di componenti il nucleo familiare, età, sesso, stato civile, grado di istruzione, condizione professionale

[1] I domini pianificati dal disegno di campionamento sono sottolineati

Dual-system (II occasione campionaria)

Wolter 1986, JASA, vol. 81, pp. 338-346

CEN \ PES	+	-	tot
+	$\hat{x}_{11}$		$x_{1\cdot}$
-			
tot	$\hat{x}_{\cdot 1}$		

M; numero delle sezioni universo

m; numero delle sezioni campione

$$x_{ij11} = \begin{cases} 1 \text{ j - mo individuo dell'i - ma sezione rilevato in CEN che in PES} \\ 0 \text{ altrimenti} \end{cases}$$

$x_{i\cdot 1}$; totale individui rilevati in PES nella i - ma sezione

Dual-system: lo stimatore

$$\hat{x}_{11} = \frac{M}{m}\sum_{i}^{m}\sum_{j} x_{ij11} \qquad \hat{x}_{\cdot 1} = \frac{M}{m}\sum_{i}^{m} x_{i\cdot 1}$$

$$\hat{N} = \frac{x_{1\cdot}\hat{x}_{\cdot 1}}{\hat{x}_{11}}$$

$$E\left(\hat{N}\right) \doteq N + \frac{p_{2\cdot}p_{\cdot 2}}{p_{1\cdot}p_{\cdot 1}} + \left(\frac{1-f}{f}\right)\frac{p_{2\cdot}}{p_{1\cdot}p_{\cdot 1}}$$

$$V\left(\hat{N}\right) \doteq N\left[\frac{p_{2\cdot}p_{\cdot 2}}{p_{1\cdot}p_{\cdot 1}} + \left(\frac{1-f}{f}\right)\frac{p_{2\cdot}}{p_{1\cdot}p_{\cdot 1}}\right]$$

Aspetti operativi

- ✓ Campione areale a due stadi con stratificazione sia delle unità di primo stadio (comuni) che di secondo stadio (sezioni di censimento). Campione di 98 comuni e di circa 1200 sezioni di censimento, per un numero atteso di famiglie pari a circa 68000.
- ✓ Tecnica di indagine: ripetizione delle operazioni di censimento all'interno delle sezioni campione.
- ✓ I nomi, i cognomi e gli indirizzi sui fogli di famiglia del censimento sono fotocopiati e acquisiti per facilitare le operazioni di abbinamento
- ✓ L'indagine di copertura è svolta a breve distanza dal censimento (3-21 dicembre) per limitare le variazioni della popolazione residente.

Tecnica di indagine

Un rilevatore, diverso da quello cui è stata assegnata la sezione in occasione del censimento, percorre la sezione e:

- enumera gli edifici;
- entra in ogni edificio e identifica le abitazioni non occupate e quelle occupate da famiglie o individui;
- consegna i questionari che dovranno essere compilati dai rispettivi intestatari (membri di riferimento delle famiglie);
- raccoglie i questionari una volta compilati.

I "Lembi" dei Fogli di famiglia di censimento raccolti nella sezione sono fotocopiati e inviati in registrazione

Tutto il materiale ausiliario utilizzato nella sezione in occasione del censimento e dell'indagine di copertura viene raccolto e consegnato all'Istat

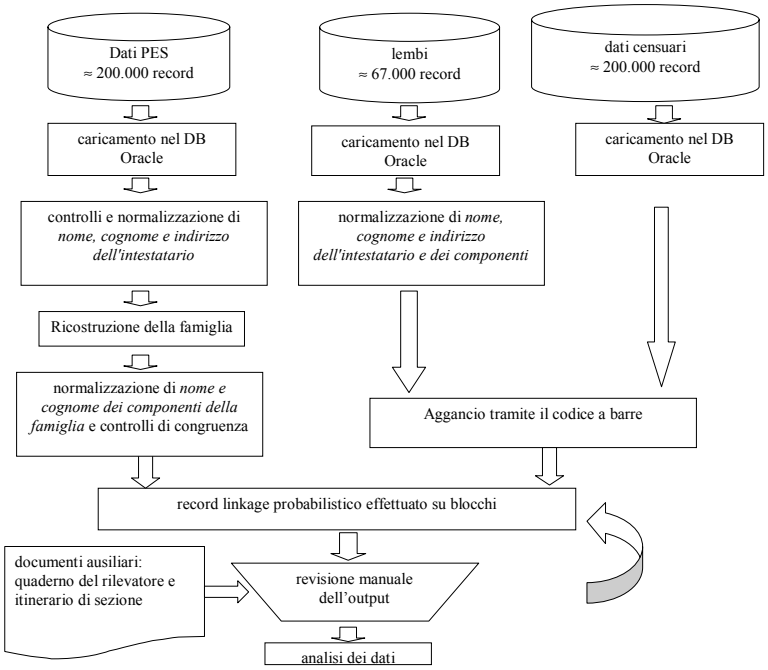
Il questionario dell'indagine di copertura

- Nomi, cognomi e indirizzo dei componenti della famiglia
- Principali caratteristiche socio-demografiche dei componenti delle famiglie

  Relazione di parentela con l'intestatario, data di nascita, sesso, luogo di nascita, stato civile, livello di istruzione, condizione professionale, posizione nella professione
- Tipo di fruizione dell'alloggio al 21 ottobre 2001 (residenza, dimora temporanea, presenza, assenza)
- Quesiti riguardanti errori che abbiano causato sovra o sotto-copertura durante il censimento

Il record linkage

- L'abbinamento fra i record del censimento e quelli dell'indagine di copertura è finalizzato all'applicazione del metodo Dual-system
- L'abbinamento viene effettuato tramite gli identificativi personali e le caratteristiche socio-demografiche comuni ai record degli individui rilevati al censimento e all'indagine di copertura
- L'abbinamento verrà effettuato con un algoritmo di nuova concezione basato su criteri probabilistici e sviluppato nell'ambito di una collaborazione fra ricercatori dell'Istat e dell'Università di Roma "La Sapienza"
- La durata prevista dell'operazione di abbinamento su due insiemi di circa 190.000 individui è stimata in 8-10 mesi



Stimare la sovra-copertura

		Area di attribuzione			
		Area 1	Area 2	...	Area k
Area di accertamento	Area 1	$\hat{s}_{11}$	$\hat{s}_{12}$	...	$\hat{s}_{1k}$
	Area 2	$\hat{s}_{21}$	$\hat{s}_{22}$	...	$\hat{s}_{2k}$
	...	...	...	$\hat{s}_{ij}$	...
	Area k	$\hat{s}_{k1}$	$\hat{s}_{k2}$	...	$\hat{s}_{kk}$

$\hat{s}_{ij}$ numero di casi di sovra-copertura accertati nell'area i e attribuibili all'area j,

$\hat{s}_{.j} = \sum_i \hat{s}_{ij}$ numero totale di casi di sovra-copertura per l'area j

Utilizzo integrato di fonti statistiche e amministrative: Il caso delle indagini strutturali sulle imprese

Le imprese italiane sono classificate secondo la loro dimensione in termini di addetti:

Piccole imprese 1-19 addetti (4 milioni, 9 milioni di addetti)
Medie imprese 20-99 addetti (65 mila, 2.4 milioni di addetti)
Grandi imprese 100+ addetti (9000, 3,8 milioni di addetti)

Statistiche strutturali sulle imprese

Statistiche sui risultati economici annuali delle imprese dell'industria e dei servizi disaggregate per settore di attività economica dimensione aziendale e localizzazione territoriale prodotte in ottemperanza al regolamento *Ue n.58/97 SBS – Structural Business Statistics*

- Indagine PMI sulle piccole e medie imprese (campione 116000 imprese, 3% del totale)
- Indagine SCI sulle grandi imprese (totale 9000 imprese)

Questionario postale autocompilato; dati sui conti economici, occupazione, costo del personale, investimenti, fenomeni emergenti (es. commercio elettronico)

Errore non campionario

- Errori di copertura dovuti alla rapida obsolescenza della lista delle imprese attive (ASIA) aggiornata a due anni prima dell'epoca di riferimento.
  - Nuove imprese
  - Imprese che hanno cambiato classe dimensionale
- Errori di mancata risposta totale dovuti a mancato ritorno dei questionari (40-45%)

Fonte amministrativa ausiliaria

Bilanci di impresa (BIL)

Società di capitale (circa 400000 imprese)

- 44% dell'occupazione
- 41% dell'occupazione per le imprese nella classe 10-19 addetti
- 85% dell'occupazione per le imprese nella classe 100+ addetti

Sono tenute a consegnare alle camere di commercio i propri bilanci civilistici)

- Armonizzati
- Tempestivi (ottobre dell'anno successivo a quello di riferimento)
- Su supporto informatico (lettura ottica)
- Comparabili (65% delle variabili di SCI è adeguatamente rappresentato)

Sperimentazione sui dati SCI99

Fonti utilizzate

ASIA97

ha fornito la lista di base per il contatto delle imprese da coinvolgere in SCI99

SCI99

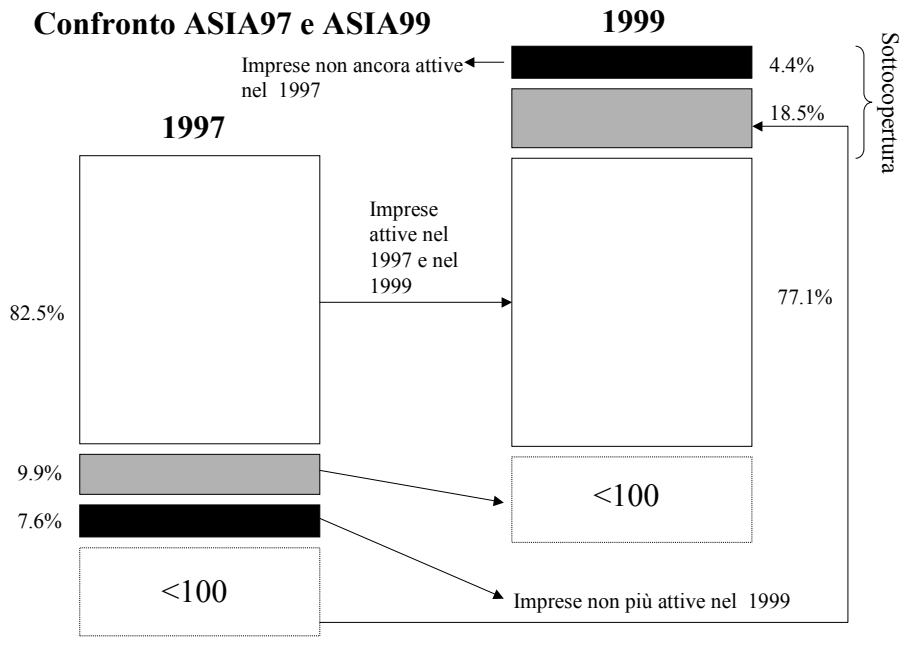
ha ottenuto le informazioni di interesse per le imprese rispondenti

BIL99

ha fornito le informazioni per integrare le mancate risposte

ASIA99

Utilizzato come termine di confronto a fine periodo



Fonte di confronto ASIA99

BIL99 \ SCI99	-	+
-	N2 7.5%	R2 4.6%
+	N1 49.6%	R1 38.3%

Procedura di imputazione - 1

- Integrazione delle mancate risposte all'indagine SCI99 (insiemi N1 ed N2) con la tecnica del donatore di minima distanza a parità di classe ATECO, classe di addetti e regione.
- Stima di modelli lineari per valutare l'associazione tra 28 variabili riassuntive di SCI99 e le corrispondenti variabili BIL99 sull'insieme R1 dei rispondenti all'indagine SCI99 contemporaneamente presenti tra i record BIL99

$$SCI99_i=\alpha_i+\beta_i \cdot BIL99_i+\varepsilon_i \text{ con } i=1,\ldots,28$$

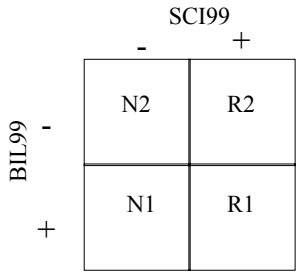
Procedura di imputazione - 2

- Uso a fini predittivi dei modelli stimati al passo precedente per imputare le variabili nell'insieme N1 dei non rispondenti a SCI99 che sono trovati in BIL99

$$\hat{SCI99}_i=\alpha_i+\beta_i \cdot BIL99_i+\varepsilon_i \text{ con } i=1,\ldots,28$$

- Stima delle 65 variabili componenti le 28 stimate da modello nell'insieme N1 riproporzionando le 28 stimate secondo le proporzioni ereditate dal metodo del donatore applicato al primo passaggio

Procedura di imputazione –3



R1 Dati indagine SCI99; stima dei modelli di associazione

R2 Dati indagine SCI

N1 Integrazione dati BIL e SCI

N2 Donatore dati SCI

L'archivio dei rilevatori

Indagini ISTAT sulle famiglie

- Indagine sui consumi delle famiglie
- Indagine sulle forze di lavoro
- Indagine multiscopo sugli aspetti della vita quotidiana

Coordinate centralmente dall'ISTAT

Eseguite dagli uffici comunali di statistica dei comuni selezionati nel campione

Rilevatori arruolati, istruiti e coordinati a livello comunale sotto la supervisione dell'ISTAT

L'archivio dei rilevatori

- Dal 1988 l'ISTAT mantiene un archivio con notizie sui rilevatori coinvolti nelle tre indagini sulle famiglie
- Informazioni standard per ogni indagine
- Informazioni riguardanti:
  - Dati strutturali dei comuni campione
  - Dati anagrafici e di produzione dei rilevatori
  - Informazioni sull'esito delle interviste assegnate ai rilevatori
- Obiettivi
  - Migliorare le procedure e gli strumenti di rilevazione
  - Individuare possibili fonti di errore e rimuoverle

Alcune varabili riferite ai rilevatori

- Famiglie intervistate
- Individui intervistati
- Famiglie e individui sostituiti o non intervistati, secondo il motivo
- Risposte proxy accettate
- Interviste annullate per contenuto informativo insufficiente o scadente

Famiglie assegnate, intervistate, cadute, nulle, proxy per 13 grandi Comuni
Indagine "Aspetti della vita quotidiana – 2001" (valori percentuali)

Comuni	Famiglie intervistate	Famiglie cadute	Interviste Proxy	Autocompilate proxy
Torino	61.0	39.0	29.3	27.1
Genova	68,1	31.9	29.8	22.1
Milano	65.7	34.0	27.3	18.1
Verona	79.1	20.9	36.0	15.4
Venezia	74.5	25.5	28.6	27.5
Bologna	63.9	36.1	30.1	21.1
Firenze	60.8	39.3	39.5	32.2
Roma	56.7	43.3	24.7	19.3
Napoli	69.0	28.3	33.1	27.0
Bari	81.3	18.8	32.4	23.8
Palermo	76.1	23.9	29.3	37.8
Catania	71.3	28.7	42.6	38.8
Cagliari	100.0	-----	22.9	27.4
Totale	**65.2**	**34.4**	**29.7**	**24.6**

Interviste cadute, nulle e proxy secondo il titolo di studio dei rilevatori
Indagine "Aspetti della vita quotidiana – 2001" (valori percentuali)

Titolo di studio	Numero rilevatori	Interviste cadute	Interviste nulle	Interviste proxy	Autocompilate proxy
Laurea o sup.	164	17.1	0.4	31.1	27.0
Maturità	1087	13.9	0.4	31.4	25.7
Lic. Media	219	15.3	0.4	29.5	24.6
Altro titolo	21	30.8	0.0	20.8	16.5
Totale	**1491**	**14.7**	**0.4**	**31.0**	**25.6**

Interviste falsificate: un modello statistico per il monitoraggio

- Interviste falsificate, fenomeno noto in letteratura
- Attività di monitoraggio basata su telefonate di controllo a campione alle famiglie intervistate

- Protocollo:
  - Prima telefonata di controllo a tutti i rilevatori
  - Telefonata di controllo a tutte le famiglie di un rilevatore se e solo se la prima ha dato esito positivo

Supponiamo esistano due sottopopolazioni non identificabili

F_i = 1 i-mo intervistatore falsifica
0 i-mo intervistatore non falsifica

N_i = n. interviste assegnate al i-mo rilevatore

T_i = 1 prima intervista falsificata dall'i-mo rilevatore
0 altrimenti

K_i= n. interviste falsificate dall'i-mo rilevatore dato che T_i=1

P(T=1) = probabilità che per un generico intervistatore risulti la prima telefonata falsificata

P(F=1) = probabilità che un generico intervistatore sia un falsificatore

$\pi_{F=1}$ = probabilità che un'intervista risulti falsificata dato che l'intervistatore è un falsificatore ($\pi_{F=0} = 0$)

Problema: stimare il numero atteso dei falsificatori dato il numero di interviste falsificate da quelli positivi al test T

$$\hat{\pi}_{F=1} = \frac{\sum_{i \in T} K_i}{\sum_{i \in T} N_i} \qquad T=\{i|\ T_i=1\}$$

$$P(T=1) = \pi_{F=1} P(F=1) + \pi_{F=0} P(F=0)$$

$$\hat{P}(F=1) = \frac{P(T=1)}{\pi_{F=1}} \qquad \text{dato che } \pi_{F=0} = 0$$

Intervistatore	I tel.	N. int	N. False
1	0		
2	1	20	9
3	0		
4	1	19	12
5	0		
6	0		
7	0		
8	0		
9	0		
10	0		
11	0		
12	1	25	8
13	0		
15	0		
16	0		

$$\hat{\pi}_{F=1} = \frac{\sum_{i \in T} K_i}{\sum_{i \in T} N_i} = \frac{9+12+8}{20+19+25} \cong 0.45$$

$$\hat{P}(F=1) = \frac{P(T=1)}{\pi_{F=1}} = \frac{3 \div 16}{0.45} \cong 0.41$$

$16 \cdot 0.41 \cong 6.6$ Falsificatori

I controlli sull'errore di registrazione

L'ISTAT invia in *outsourcing* i questionari cartacei da inserire su supporto magnetico

Le società appaltatrici delle commesse si impegnano per contratto a garantire un determinato livello di qualità dei dati

I dati sono consegnati in "lotti" sottoposti a test prima di essere accettati

Fasi del controllo

- Acquisizione e smistamento
- Controlli formali e quantitativi (lettura file, conteggio record, identificazione record vuoti e doppioni)
- Controlli qualitativi
  - Determinazione della frazione sondata e della soglia di rifiuto
  - Estrazione del campione dal master
  - Selezione dei questionari cartacei
  - Registrazione del campione di questionari cartacei
  - Abbinamento e riconciliazione errore
  - Calcolo della statistica test e accettazione/rifiuto del lotto
- Conteggio battute utili ai fini della fatturazione

Il Test

Unità statistica; byte (ipotesi di costanza della probabilità d'errore)

N = totale dei byte nel lotto

n = numero di byte campione

e = numero di errori nel campione

$\hat{E}$ = e/f stima del numero di byte errati nel lotto

Occorre determinare

f=n/N frazione sondata

Soglia di rifiuto E_s

Sotto l'ipotesi nulla $H_0 \equiv p_0$

$$Var(\hat{E}) = \frac{N^2(1-f)}{(N-1)f} p_0(1-p_0) \qquad E(\hat{E}) = Np_0$$

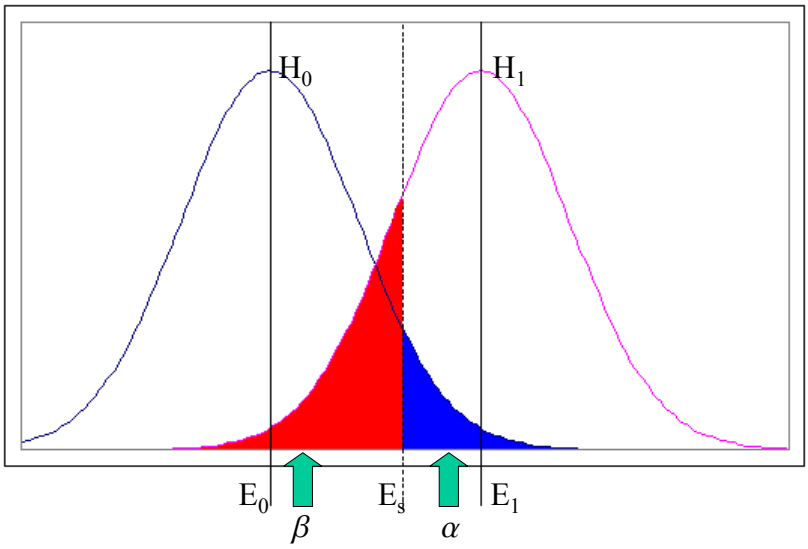
Sotto l'ipotesi alternativa $H_1 \equiv p_1$

$$Var(\hat{E}) = \frac{N^2(1-f)}{(N-1)f} p_1(1-p_1) \qquad E(\hat{E}) = Np1$$

Determiniamo i livelli di rischio accettabile per gli errori di I e II specie

$$z_0 : \Pr\{Z \le z_0 | Z \approx N(0,1)\} = 1-\alpha$$

$$z1 : \Pr\{Z \le z1 | Z \approx N(0,1)\} = \beta$$



Si risolve il sistema

$$\begin{cases} z_0 = \dfrac{\dfrac{e}{f} - Np_0}{\sqrt{\dfrac{N^2(1-f)}{(N-1)f}\,p_0(1-p_0)}} \\ z_1 = \dfrac{\dfrac{e}{f} - Np_1}{\sqrt{\dfrac{N^2(1-f)}{(N-1)f}\,p_1(1-p_1)}} \end{cases}$$

e, assumendo N≅N-1,

si ottiene

$$f = \frac{k^2}{N(p_1-p_0)^2 + k^2} \qquad e = \frac{k^2 p_0 + z_0 k(p_1-p_0)\sqrt{p_0(1-p_0)}}{(p_1-p_0) + \dfrac{1}{N}k^2}$$

con $k = z_0\sqrt{p_0(1-p_0)} - z_1\sqrt{p_1(1-p_1)}$

Un test alternativo

Obiettivo eliminare la fase di riconciliazione dell'errore

n = numero di byte campione

Sia B_{1i} l'errore nella prima digitazione del byte i-mo

Sia B_{2i} l'errore nella seconda digitazione del byte i-mo

B_{1i} e B_{2i} non sono direttamente osservabili

Sia D_i la differenza tra la prima e la seconda registrazione del byte i-mo (osservabile); allora possiamo ammettere che:

B_{1i}	B_{2i}	D_i
0	0	0
0	1	1
1	0	1
1	1	1

Sotto le ipotesi:

$$P(B_{1i}=1)=P(B_{2i}=1)=p;\;\; P(B_{1i},B_{2i})=P(B_{1i})P(B_{2i})$$

Valgono le seguenti relazioni

$$P(D_i=0)=(1-p)^2;\qquad P(D_i=1)=2p(1-p)+p^2$$

Che permettono di scrivere la verosimiglianza

$$L(D|p)=\prod_i (1-p)^{2D_i}\left(2p(1-p)+p^2\right)^{(1-D_i)}$$

e la conseguente stima di massima verosimiglianza

$$\hat{p}=\frac{n-\sqrt{n\sum_1 D_i}}{n}$$