UNIVERSITA' DEGLI STUDI DI PERUGIA

DIPARTIMENTO DI FILOSOFIA SCIENZE SOCIALI UMANE E DELLA FORMAZIONE

Corso di Laurea in Scienze per l'Investigazione e la Sicurezza

12. ASPETTI METODOLOGICI PER LE RILEVAZIONI STATISTICHE

Prof. Maurizio Pertichetti

Estratto dalle *Linee guida metodologiche per rilevazioni statistiche - Nozioni metodologiche di base e pratiche consigliate per rilevazioni statistiche dirette o basate su fonti amministrative - Istituto Nazionale di Statistica* il presente testo è stato, per finalità connesse alla semplificazione delle argomentazioni in esso contenute, ulteriormente riarticolato e integrato in alcune sue parti.

Indice

Premessa

L'obiettivo di questo manuale è quello di divulgare le nozioni di base riguardanti la progettazione e l'esecuzione di una rilevazione statistica, sia essa diretta che basata su fonti amministrative. I potenziali fruitori del manuale sono tutti coloro i quali, anche senza essere statistici, si trovano tuttavia nelle condizioni di voler acquisire conoscenze sui metodi di pianificazione e produzione adottati dalle indagini statistiche condotte in ambito Istat e SISTAN. Fra essi si collocano gli utenti finali dell'informazione in quanto, se si accetta l'impostazione secondo la quale la qualità del prodotto (l'informazione nel nostro caso) è guidata dalle esigenze dell'utente, diventa centrale che questi sia dotato di strumenti di tipo formativo tali da facilitare la lettura critica dei dati che vengono forniti dall'Istat o da qualsiasi altro ente del SISTAN. Questo è tanto più vero quanto più si considerino utenti non professionali, come gli operatori nel settore delle imprese o i semplici cittadini, ai quali sempre di più si cerca di facilitare un accesso più diretto all'informazione statistica, non mediato cioè dai mezzi di comunicazione di massa.

Oltre agli utenti finali il manuale risulta utile anche a coloro i quali, a fianco degli statistici, sono coinvolti nei meccanismi produttivi di una rilevazione. È infatti noto che una rilevazione necessita di un elevato grado di organizzazione e dell'apporto di numerose professionalità oltre quella dello statistico propriamente detto. Fra queste, solo per citare alcune di quelle coinvolte nelle fasi progettuali, gli esperti del fenomeno oggetto di studio, gli informatici, il personale tecnico-amministrativo di enti e delle varie articolazioni della PA. Ad un livello più esecutivo è invece utile menzionare quelli che adempiono alle fasi di contatto dei rispondenti (rilevatori), al trasporto e alla revisione del materiale raccolto, alla registrazione dei dati su supporto informatico e così via. A tutte queste figure il manuale si rivolge nel tentativo di contribuire al conseguimento di un vocabolario comune, di una visione generale del processo al quale contribuiscono e della consapevolezza di quanto il risultato del lavoro di tutti incide sul successo della rilevazione nel suo complesso e sulla qualità dell'informazione prodotta. Vale infine la pena di citare tra i potenziali fruitori gli studenti di statistica o quegli statistici i quali non siano mai stati direttamente coinvolti nei processi di reperimento, raccolta e validazione dell'informazione nell'ambito della statistica ufficiale. Ad essi infatti il manuale propone, insieme a sezioni introduttive di agile consultazione, gli spunti bibliografici necessari per gli approfondimenti desiderati.

Il manuale è organizzato in circa trenta diverse sezioni nelle quali si illustrano sia gli aspetti riguardanti la pianificazione di una rilevazione sia i temi concernenti le fasi operative. La trattazione, sia pure di base, tiene comunque conto sia dell'esperienza dell'Istat sia delle esperienze internazionali nelle materie trattate, e prevede, per le parti più operative, apposite sottosezioni rivolte ai responsabili di processo dove si forniscono raccomandazioni applicative finalizzate al conseguimento di risultati di qualità.

Progettare l’indagine

Scopo dell’indagine è quello di produrre statistiche, ovvero descrizioni riassuntive di carattere quantitativo, riguardanti il collettivo di interesse. La progettazione e l’esecuzione di un’indagine è frutto di un impegno multidisciplinare che coinvolge necessariamente un elevato numero di professionalità. L’attività di progettazione deve procedere prendendo in considerazione tutti gli aspetti coinvolti, da quelli riguardanti i fenomeni di interesse e quelli di carattere più operativi. I principali argomenti da prendere in considerazione fin dalla fase progettuale sono:

- Obiettivi, definizioni e classificazioni;
- Disegno d’indagine;

- Indagini amministrative;
- Fasi operative;
- Tempi e costi;
- Sistema di controllo della qualità;
- Elaborazioni statistiche;
- Diffusione;

Il numero degli esperti coinvolti e le relazioni esistenti fra gli argomenti da considerare sono tali da obbligare a riunire tutte le professionalità necessarie in un gruppo di progettazione il cui fine principale è quello di assicurare la collaborazione degli esperti e l'integrazione fra le soluzioni prescelte.

Il gruppo di progettazione ha come obiettivo la realizzazione di un documento di progettazione nel quale sono illustrate nel dettaglio tutte le soluzioni proposte e discusse le alternative considerate. Affinché la progettazione di un'indagine possa dirsi compiuta è inoltre necessario prevedere una o più sperimentazioni finalizzate a saggiare nella pratica le soluzioni ideate.

Obiettivi, definizioni e classificazioni

In questa sezione vengono considerati quegli aspetti definitori che più di altri sono connessi alla specifica area di interesse che si intende analizzare per mezzo dell'indagine. Questi, se non correttamente individuati, possono provocare gravi ricadute su alcune componenti della qualità come la rilevanza e l'accuratezza.

Le relazioni esistenti fra le questioni proprie del fenomeno osservato e le caratteristiche tecnico-statistiche ed operative dell'indagine sono tali e tante che risulta indispensabile la partecipazione di uno o più esperti del settore specifico all'interno del gruppo di progettazione. Di seguito una breve descrizione degli aspetti definitori che è necessario prendere in considerazione:

Fenomeno di interesse. Delimitare precisamente cosa interessa da cosa non interessa ricordando che più ampio è l'arco degli argomenti trattati, maggiori divengono le complessità da affrontare sul piano concettuale statistico ed operativo. Definire se interessa descrivere un fenomeno nella sua componente statica o in quella dinamica. Specificare se interessa confrontare i risultati con informazioni relative ad altre realtà territoriali. Specificare quali ipotesi si intende sottoporre a verifica;

Popolazione di riferimento. Individua con precisione l'insieme di unità statistiche alle quali si intende estendere i risultati dell'indagine. Specificare esattamente le caratteristiche che determinano l'inclusione (o l'esclusione) delle unità statistiche nella popolazione;

Variabili studiate. Misure di caratteristiche, solitamente elementari, riferite alle unità statistiche. Si raggruppano concettualmente in quattro grandi classi:

- Qualitative sconnesse. Assumono un insieme finito di categorie mutuamente esclusive tali che, per due differenti unità statistiche, si può definire soltanto se queste assumono la stessa o differenti categorie (sesso, stato civile);
- Qualitative ordinali. Assumono un insieme finito di categorie mutuamente esclusive tali da poter ordinare due unità statistiche secondo il possesso di caratteristiche possedute (grado di istruzione, grado di soddisfazione);

- Quantitative discrete. La caratteristica può essere descritta mediante un numero finito o infinito numerabile di valori numerici fra i quali abbia senso calcolare una differenza e/o un rapporto (numero di figli);
- Quantitative continue. La caratteristica può essere descritta mediante un'infinità non numerabile di valori fra i quali abbia senso calcolare una differenza e/o un rapporto (fatturato d'impresa).

La definizione delle variabili dovrebbe procedere attraverso una progressiva identificazione e raffinamento del fenomeno di interesse nelle sue componenti fino ad identificare gli aspetti salienti. L'obiettivo di tale procedimento dall'alto verso il basso serve a definire delle caratteristiche immediatamente utili all'obiettivo della ricerca. D'altro canto è necessario predisporre un analogo meccanismo dal basso verso l'alto considerando che le caratteristiche che si vogliono conoscere siano effettivamente misurabili sulle unità statistiche da indagare.

Classificazioni. Insieme delle categorie assunte da una variabile qualitativa sconnessa o ordinale. Definire una classificazione è un momento particolarmente critico. Ad esempio misurare il gradimento di uno spettacolo ricorrendo a quattro anziché a cinque categorie (ma anche denominando in modo appena diverso le stesse cinque categorie) può fornire risultati addirittura opposti. E' quindi opportuno, soprattutto se si desidera confrontare i risultati dell'indagine con altre fonti di informazione disponibili, ricorrere a classificazioni comunemente utilizzate. Per alcune variabili particolarmente complesse da definire (attività economiche, professioni, malattie) sono disponibili classificazioni standard riconosciute a livello internazionale.
In tutti i casi, soprattutto in quelli più complessi, nel definire una classificazione è opportuno, se possibile, procedere ad aggregazioni o raffinamenti di categorie utilizzate da classificazioni già esistenti in modo da preservare almeno in parte la confrontabilità dei risultati dell'indagine.

Disegno di indagine

La definizione del disegno di indagine mira a rispondere alle seguenti necessità:

1. Definire qual è il tipo di indagine più consono a produrre le statistiche che si desiderano;
2. Decidere tra indagine totale e campionaria e, in tal caso, disegnare ed estrarre il campione.

Consideriamo ciascuno dei due punti in maggior dettaglio.

1. Esistono una varietà di stime che può essere interessante produrre:
    - Stime di caratteristiche, attività, comportamenti, attitudini in un punto nel tempo;
    - Stime di variazione netta o lorda in due o più punti nel tempo;
    - Stime di andamenti tendenziali su più periodi temporali;
    - Stime di durata, transizioni o frequenze di accadimento per specifiche tipologie di eventi e specifici sotto-insiemi di popolazione;
    - Stime di caratteristiche basate sull'accumulo di dati nel tempo;
    - Stime di relazioni fra caratteristiche.

    Pur rimandando alla letteratura specifica per approfondimenti, è già chiaro che, a seconda delle informazioni alle quali si è interessati, è necessario fare riferimento a differenti tipi di indagine. Ricorrere all'indagine di tipo non opportuno può pregiudicare in tutto o in parte gli scopi della ricerca.

2. Raccogliere informazioni su tutte le unità statistiche appartenenti alla popolazione implica non solo un aumento insostenibile dei costi, ma anche un maggior numero di errori non campionari tali da limitare questa modalità a casi di eccezionale importanza come i Censimenti o a casi in cui le informazioni sulla totalità delle unità statistiche sono state già raccolte per motivi diversi dell'indagine, come nel caso delle indagini amministrative.

Se le considerazioni di costo/beneficio orientano la scelta verso una indagine campionaria occorre valutare i seguenti aspetti:

- identificare il metodo di selezione del campione in riferimento alla struttura degli archivi di base e alle informazioni in essi contenute, in modo da massimizzare l'efficienza delle stime prodotte, tenendo conto allo stesso tempo dei vincoli da essi imposti;
- dimensionare il campione in modo da garantire stime della precisione desiderata, dati i vincoli di costo imposti.

I due problemi elencati sono affrontati utilizzando la teoria del campionamento. La soluzione a tali problemi prende il nome di strategia di campionamento.

Indagini trasversali e longitudinali

Date le necessità conoscitive, di cui una classificazione generale è stata data riguardo al disegno di indagine, occorre predisporre le modalità di rilevazione che possano soddisfarle. Una prima grande distinzione può essere fatta tra indagini trasversali e longitudinali:

- nelle indagini trasversali si rilevano le unità statistiche raccogliendo informazioni di interesse riferite ad un particolare momento o periodo di tempo, con l'intento di stimare le caratteristiche riferite allo stato della popolazione oggetto nel momento o periodo di interesse;
- nelle indagini longitudinali invece l'obiettivo è principalmente rivolto a misurare l'evoluzione nel tempo delle caratteristiche di interesse mediante l'espediente di ricontattare le unità per analizzarne i cambiamenti.

E' importante tuttavia osservare che questa distinzione non impedisce completamente di stimare misure di cambiamento con indagini trasversali o misure di stato con indagini longitudinali, anche se ciò può essere fatto utilizzando opportune accortezze. Nel seguito, si elenca una serie di tipologie d'indagine illustrandone sia le potenzialità informative in termini di stima di caratteristiche di stato o di cambiamento.

Indagini occasionali: si tratta di indagini pianificate allo scopo di ottenere stime riferite a caratteristiche possedute dalla popolazione in un singolo istante di tempo (es.: distribuzione per età della popolazione in un dato istante) o riferite a un periodo (es.: distribuzione del fatturato realizzato nell'arco di un anno).

Indagini ripetute (nessuna sovrapposizione fra le unità indagate nelle diverse occasioni): sono spesso chiamate indagini periodiche o ricorrenti. Secondo questa modalità un'organizzazione di indagine viene ripetuta in momenti programmati nel tempo. L'organizzazione adottata non prevede una sovrapposizione, neanche parziale, del campione di unità in differenti occasioni.

Indagini longitudinali senza rotazione: sono indagini predisposte con lo scopo di seguire un particolare gruppo di unità nel tempo, e creare un record longitudinale per ogni unità osservata. L'obiettivo è quello di studiare le modificazioni intervenute nel collettivo durante il tempo, utilizzando i cambiamenti avvenuti sui record individuali. E' importante sottolineare che mediante un'indagine longitudinale senza rotazione è possibile produrre stime riferite alla sola popolazione

di partenza dal momento che, senza disporre di ingressi di nuove unità, non si riesce a rappresentare gli eventuali mutamenti nella struttura del collettivo di riferimento.

Indagini longitudinali con rotazione: indagini disegnate per seguire un particolare gruppo di unità per un periodo di tempo, introducendo nuove unità nel campione in occasioni specificate, al fine di creare record longitudinali per ogni unità osservata e produrre analisi longitudinali. Mediante l'ingresso periodico di nuove unità nel campione è possibile mantenere il campione stesso rappresentativo della popolazione anche nelle occasioni successive alla prima. Infatti in questo modo si tiene conto che nel tempo il collettivo di interesse si modifica con l'ingresso di nuove unità (es.: nascite o immigrazioni) che, ovviamente, nella prima occasione non avevano alcuna possibilità di essere inserite in analisi. Mediante questo schema di indagine è quindi possibile produrre sia stime longitudinali, riferite alle variazioni nette intervenute e alle transizioni di stato, sia stime trasversali riferite alle popolazioni aggiornate ad ogni occasione di rilevazione.

Indagini totali e campionarie

Una delle scelte essenziali da compiere nella definizione di un disegno di indagine è quella data dall'alternativa tra un'indagine totale e un'indagine campionaria. Per indagine totale si intende una rilevazione in cui tutte le unità delle quali si possiede un indirizzo nei propri archivi di base sono interessate dalla rilevazione. La più importante fra le rilevazioni totali è senz'altro il Censimento. La particolarità del Censimento è data dal fatto che gli archivi in possesso dell'ente statistico sono costituiti da aree in cui è suddiviso l'intero territorio (sezioni di censimento). A partire dalle aree di territorio si compie una enumerazione completa delle unità statistiche di interesse (imprese, famiglie, abitazioni, ecc.) e, contestualmente, si raccolgono alcune informazioni di carattere fondamentale. Oltre al censimento, sono da citare altri due importanti casi di indagini totali: indagini in cui la popolazione di riferimento è costituita da poche unità molto importanti, nel senso che ciascuna di esse possiede una quantità rilevante della caratteristica da indagare (ad esempio il fatturato delle grandi imprese). In questo caso omettere la rilevazione anche di una sola delle unità di interesse può comportare notevoli distorsioni nelle stime. Inoltre, nel caso di popolazioni composte da pochi elementi molto importanti, è relativamente più semplice il compito di contattare e rilevare le unità. Indagini basate su dati amministrativi in cui l'informazione di interesse è già stata raccolta per finalità diverse da quella di produrre informazione statistica. Esempi di tali raccolte di dati sono: informazioni dai certificati di nascita e di assistenza al parto, archivi INPS sui lavoratori dipendenti, dati raccolti su archivi giudiziari, ecc.
Anche se dal punto di vista teorico con un'indagine totale si riescono ad ottenere misure precise dei paramenti di interesse, nella pratica i problemi connessi sono tali da limitarne l'uso all'indispensabile. Fra essi è importante citarne almeno due: l'enorme costo di rilevazione e trattamento dei dati e i problemi connessi alla qualità dei dati, primo fra tutti l'incompletezza della rilevazione dovuta all'incapacità di raggiungere tutte le unità statistiche.

Per i problemi ai quali sono soggette le rilevazioni totali si ricorre alle indagini campionarie caratterizzate dal fatto che solo una parte delle unità statistiche componenti la popolazione viene selezionata ed indagata (campione). Questo espediente, diminuendo l'onere della rilevazione, consente di destinare maggiore attenzione a tutte le attività connesse al miglioramento e al controllo della qualità dei dati raccolti. Tuttavia selezionare solo un campione implica ovviamente una minore attendibilità delle stime riferite ai parametri di interesse. E' infatti chiaro che a seconda di quali unità sono inserite nel campione prescelto, i risultati riferiti alla popolazione complessiva varieranno. Tuttavia, se la selezione del campione viene effettuata con scelta

rigorosamente casuale, è possibile misurare il livello di precisione delle stime ottenute rispetto al vero valore del parametro di interesse nella popolazione.

La definizione delle modalità di estrazione del campione, della sua dimensione e delle funzioni dei dati utilizzate per ottenere, dal campione, stime riferite alla popolazione di interesse prende il nome, come già enunciato, di strategia di campionamento. E' importante precisare che, qualora le unità da inserire nel campione siano selezionate con scelta ragionata e non con criteri di rigorosa casualità, non sarà più possibile garantire in alcun modo la rispondenza dei risultati delle analisi effettuate sui dati a requisiti statistici di affidabilità quali la correttezza e l'efficienza delle stime. Per questo motivo il significato riferito al termine "campione" sarà in questa sede sempre riferito alla selezione casuale delle unità statistiche.

Archivi di base

In questa sede per archivi di base si intendono le liste, le mappe o le altre specificazioni delle unità che costituiscono l'informazione disponibile sulle unità componenti la popolazione obiettivo riguardante una certa indagine totale o campionaria.

Gli archivi di base possono contenere o meno informazioni supplementari riguardanti le unità, come la loro dimensione o altre caratteristiche, ma devono riportare sufficienti dettagli tali che le unità possano essere localizzate e rilevate.

Nel seguito sarà fatto spesso riferimento all'influenza che gli archivi di base esercitano sulla strategia di campionamento, ma è importante osservare che le problematiche riguardanti gli archivi sono più generali e riguardano anche le indagini totali. L'accento sarà posto maggiormente sulle indagini campionarie per il solo fatto che, in questo caso, le relazioni tra archivi e campione si possono in un certo senso considerare più complesse e "nascoste". Raramente gli archivi possono essere considerati perfetti dal momento che si possono presentare problemi di incompletezza, inaccuratezza, inadeguatezza, obsolescenza, o essere soggetti a duplicazioni delle unità in esso contenute. Tali problemi saranno meglio illustrati nella sezione riguardante gli errori di copertura. In questa sezione saranno fatte alcune raccomandazioni su pratiche consigliabili al fine di prevenire, correggere e valutare gli errori di copertura.

Affinché un archivio di base possa essere considerato adeguato ad una indagine occorre considerare i seguenti elementi :

1. la popolazione obiettivo deve essere composta da un numero finito di elementi identificabili;
2. può essere condotto un campionamento su qualche insieme di unità, ma queste non necessariamente debbono essere elementi della popolazione obiettivo (campionamento a più stadi). A questo proposito un esempio è rappresentato dalle indagini Istat che rilevano le famiglie a partire dalle anagrafi anche se sono interessate a dati riferiti alla popolazione degli individui;
3. occorre definire il legame che permette di raggiungere operativamente le unità della popolazione obiettivo a partire dalle unità riportate nell'archivio di base;
4. deve essere possibile distinguere l'una dall'altra le unità componenti l'archivio in modo da poterle riconoscere al momento del contatto;
5. esistono più tipi di legame che possono collegare gli elementi costituenti l'archivio di base e le unità della Popolazione obiettivo. Tale legame contribuisce a determinare il tipo di disegno di campionamento e le procedure di stima che possono essere adottate nell'indagine (struttura degli archivi);

6. qualche strategia di campionamento o procedura di stima richiede informazioni ausiliarie sugli elementi della popolazione. In questo caso tali informazioni devono essere note per ogni elemento della popolazione obiettivo (stratificazione del campione, campione con probabilità di selezione differenti);

Struttura degli archivi

I legami che possono intercorrere fra le unità riportate negli archivi e le unità della popolazione obiettivo sono essenzialmente di quattro tipi:

- uno a uno – uno ed un solo elemento dell'archivio è associato ad una ed una sola unità appartenente alla popolazione obiettivo. Questo è il caso più semplice in cui è l'unità stessa a far parte dell'archivio.
- uno a molti – ad un elemento della lista corrispondono uno o più elementi della popolazione obiettivo ma ad ogni elemento della popolazione obiettivo corrisponde un solo elemento della lista. E' il caso delle anagrafi di popolazione, utilizzate dall'Istat per accedere alle famiglie, dalle quali si risale poi ai singoli individui che le compongono.
- molti a uno – ad un elemento della lista corrisponde un solo elemento della popolazione obiettivo ma ad un elemento della popolazione obiettivo possono corrispondere più elementi della lista. Un caso reale è fornito dall'archivio INPS sulle posizioni lavorative i cui componenti sono costituiti dai lavoratori dipendenti e le unità di interesse per l'indagine sono rappresentate dalle imprese con dipendenti. In questo caso più componenti dell'archivio possono rimandare alla stessa impresa.
- molti a molti – un elemento della lista corrisponde a uno o più elementi della popolazione obiettivo e viceversa.

Nei casi pratici si cerca di ridursi alle prime due situazioni considerate, in quanto le altre presentano numerose complicazioni sia tecniche che teoriche.
A volte gli archivi di base possono non essere centralizzati ma frazionati e collocati sul territorio. Si realizza così una gerarchia per la quale si dispone di un archivio centrale in cui sono riportate le unità presso le quali si troveranno archivi locali contenenti informazioni su altre unità e così via fino a giungere alle unità appartenenti alla popolazione obiettivo. In casi come questi è comune ricorrere alla strategia di campionamento a più stadi qualora si desideri limitare l'indagine ad un campione di unità statistiche.

Alcune Raccomandazioni

In fase di progettazione è necessario valutare possibili alternative sulla base di quanto l'archivio risulta aggiornato e rappresentativo della popolazione obiettivo e altresì valutare l'affidabilità delle informazioni in esso contenute (ad esempio gli indirizzi per il contatto delle unità). Per questo sono solitamente necessarie analisi di fattibilità basate su studi pilota. Si possono inoltre utilizzare informazioni disponibili da altre indagini che già utilizzano gli stessi archivi. Occorre valutare la possibilità di ottenere aggiornamenti dell'archivio all'epoca di riferimento dell'indagine, ad esempio abbinando archivi indipendenti. E' tuttavia da valutare attentamente il rischio che, così facendo, siano introdotte delle duplicazioni. Inoltre è bene predisporre tutte le misure possibili per identificare gli errori nell'archivio di base durante la rilevazione. Ad esempio è possibile introdurre nel questionario (o preparare appositi moduli da far compilare ai rilevatori) domande utili a contare il numero di unità non trovate o non più esistenti o a testare l'affidabilità delle informazioni contenute in archivio (n° di addetti delle imprese o loro fatturato). In particolare

per le indagini sulle imprese è bene predisporre procedure adatte a registrare le trasformazioni da esse subite nel tempo (fusioni, scorpori, cambiamenti di o di attività economica, ecc.). Per tutti i controlli il cui esito dipende dal personale sul campo (rilevatori, supervisori) inserire un argomento e delle esercitazioni pratiche nel programma di formazione del personale, motivandolo sull'importanza di individuare gli errori eventualmente presenti negli archivi di base. Per quanto concerne le indagini areali, predisporre ispezioni e confronti con altre mappe aggiornate in modo da controllare i confini delle aree identificate evitando che rimangano zone di territorio scoperte o sovrapposizioni di aree confinanti.

Altri metodi che possono risultare utili all'identificazione di errori negli archivi di base consistono nel confrontare, sul totale della popolazione o su appositi sottoinsiemi, le stime fornite dall'indagine, con quelle disponibili da altre fonti (censimento), per particolari variabili strutturali delle unità della popolazione obiettivo (età e sesso degli individui, numerosità delle famiglie, dimensione e fatturato delle imprese). Più nel dettaglio, può risultare migliore il confronto di quantità che tendono a mantenersi più stabili (es. rapporto di mascolinità) rispetto a differenze temporali, territoriali o di processo.

Strategia di campionamento

Perché un campione sia rappresentativo della popolazione di provenienza occorre che gli archivi di base usati per l'estrazione siano in buono stato di aggiornamento, che la dimensione del campione sia sufficiente e che le procedure di selezione per lo specifico disegno siano appropriate. In questa sezione descriviamo alcune delle più importanti procedure di campionamento e i loro effetti sulla precisione delle stime campionarie. Saranno inoltre fornite alcune raccomandazioni riguardanti gli aspetti da considerare nel sistema dei controlli di qualità riguardo alla strategia di campionamento.

Affinché si possa estrarre un campione occorre valutare attentamente le caratteristiche degli archivi di base (denominati nel seguito anche liste) disponibili. Una volta fatto ciò sarà possibile identificare il procedimento di selezione delle unità che meglio si adatta a tali caratteristiche. Vediamo alcune delle principali modalità di campionamento che possono essere considerate.

Campionamento casuale semplice. E' la più semplice fra le modalità di campionamento. Essa equivale ad associare ad ogni unità della popolazione una biglia numerata e ad estrarre a caso da un'urna, una per volta e senza riporla, tante biglie quante sono le unità che si vogliono campionare. Affinché si possa applicare tale metodo è necessario disporre di una lista che elenchi tutte le unità statistiche della popolazione.

Campionamento sistematico. E' una variante del campionamento casuale semplice molto efficiente da realizzare quando si disponga della lista delle unità statistiche della popolazione sotto forma di file elaborabile al computer. Tale campionamento prevede di scegliere (sistematicamente) un elemento della lista ogni k che rappresenta *l'intervallo di campionamento* ed è dato dal rapporto tra *dimensione della popolazione/dimensione del campione*. Se la lista contiene 10.000 unità e si vuole ottenere un campione di 1.000 unità, significa che si deve selezionare un'unità ogni 10. La prima unità va selezionata in maniera casuale. In questo caso estraendo un numero casuale fra 1 e 10, e a partire da questo tutte quelle contrassegnate dal numero estratto più 10. Il rapporto tra *dimensione del campione /dimensione della popolazione* è invece detto *ragione di campionamento* e rappresenta la proporzione di elementi della popolazione selezionati per il campione, nel caso esposto 1/10. Sebbene molto efficiente, questo

procedimento di stima può condurre a distorsioni se l'ordine in cui le unità sono disposte tende ad avere una ricorrenza associata alla caratteristica di interesse. Consideriamo, ad esempio, una lista di abitazioni elencate, per ogni quartiere, secondo la loro dimensione. E' possibile che, effettuando un campionamento sistematico di dimensione n pari al numero dei quartieri, si possano selezionare tutte abitazioni molto grandi o molto piccole.

Stratificazione del campione. Prima di procedere all'estrazione si suddivide la popolazione in due o più gruppi secondo una o più caratteristiche conosciute sulle unità statistiche. Si procede quindi all'estrazione delle unità per ogni gruppo (strato). Questa modalità di pianificazione del campione consente di ottenere stime più precise, a parità di dimensione del campione, rispetto al campione casuale semplice purché all'interno degli strati le unità statistiche siano fra loro omogenee riguardo alle variabili oggetto di studio. Per poter applicare tale tecnica è necessario che le caratteristiche usate nella formazione degli strati sia disponibile sulla lista per ogni unità della popolazione.

Campionamento a più stadi. Quando non sia disponibile una lista complessiva delle unità della popolazione è possibile ricorrere al campionamento a più stadi. Un esempio di tale situazione è dato dall'anagrafe che non esiste come unico archivio nazionale ma è suddivisa negli 8.103 comuni italiani. In questo caso si procede dapprima ad estrarre un campione di comuni (unità di primo stadio) e quindi, per ogni comune selezionato, un campione casuale di famiglie (unità di secondo stadio) da ciascuna lista anagrafica. A questo tipo di campionamento si ricorre in generale per necessità in quanto le stime con esso ottenibili sono di solito meno efficienti (maggior variabilità campionaria) di quelle calcolate applicando un campione casuale semplice.

Campionamento areale. Si tratta di una procedura di campionamento utilizzata quando non si dispone di una lista per la selezione delle unità, ma queste sono dislocate sul territorio. In questo caso si procede ad una suddivisione in parti (aree) dell'intero territorio e all'estrazione di un campione di aree. Quindi si esplorano le aree campionate, allo scopo di enumerare esaustivamente le unità presenti al loro interno e produrre delle liste complete. Infine, dalle liste prodotte, si estraggono le unità campione da contattare per la rilevazione vera e propria. Dal punto di vista teorico il campionamento areale deve essere considerato una forma particolare di campionamento a più stadi.

Le modalità di campionamento descritte sono di norma applicabili in maniera modulare, possono cioè essere adottate anche insieme nei casi pratici. Ad esempio nelle indagini ISTAT sulle famiglie si ricorre ad un campionamento a due stadi in cui le unità di primo stadio (i Comuni) sono stratificate secondo la zona geografica ed estratti con probabilità proporzionale alla dimensione. Una volta selezionato il campione di comuni si passa ad estrarre, per ciascun comune, il campione di famiglie applicando la tecnica del campionamento sistematico alle rispettive liste anagrafiche.
Ad ogni modalità, o insieme di modalità di campionamento prescelte sono associati degli appositi metodi di stima, cioè funzioni dei dati raccolti sul campione tali da fornire le stime relative alla popolazione ed il loro grado di precisione. Le funzioni di calcolo delle stime e della loro precisione sono basate sul calcolo delle probabilità e trattate nell'ambito della teoria dei campioni.

Alcune raccomandazioni
E' importante che la strategia di campionamento adottata sia testata, monitorata e validata al fine di valutarne la rispondenza agli obiettivi iniziali e l'adeguatezza rispetto a successive occasioni di indagine. A tal fine è bene considerare più disegni di campionamento alternativi e valutarli alla luce di informazioni disponibili quali censimenti, indagini precedenti, dati amministrativi o appositi studi pilota. Per mezzo di tali analisi è possibile raffinare la scelta delle variabili di stratificazione, la

dimensione del campione, o l'allocazione degli strati, avendo prefissato la dimensione dell'errore campionario che si è disposti a sopportare. E' opportuno che le indagini ricorrenti permettano una certa flessibilità nel disegno in maniera da far fronte a necessità quali l'aggiornamento delle probabilità di selezione o una riduzione della dimensione campionaria.
E' bene prevedere una rotazione del campione qualora si desideri fornire stime di variazioni efficienti e si voglia limitare il carico della rilevazione sulle unità statistiche.
E' inoltre opportuno considerare nella fase di disegno del campione anche problemi connessi agli errori non campionari quali l'impossibilità di contattare qualche unità, il contatto di unità non appartenenti alla popolazione (ad esempio un'impresa dove ci si aspetta una famiglia) o il rifiuto a partecipare all'indagine. In generale è meglio rinunciare ad adottare la strategia più efficiente, se si ha ragione di ritenerla difficilmente applicabile, per evitare che siano introdotti errori nella selezione del campione dei quali è difficile valutare gli effetti sulle stime.
Per le indagini ricorrenti dovrebbe essere monitorata l'efficienza del disegno di campionamento nel tempo. Infatti, per effetto di modificazioni, intervenute nella popolazione, la strategia di campionamento potrebbe divenire inadeguata e necessitare di ritocchi ad esempio nella dimensione del campione o nell'allocazione degli strati.

Tecniche di indagine

Con il termine tecnica di indagine si intende l'insieme delle modalità di contatto delle unità statistiche interessate dalla rilevazione e di reperimento delle informazioni oggetto di interesse. La scelta della tecnica di indagine più idonea a raccogliere le informazioni oggetto della ricerca è uno degli aspetti di maggiore importanza nella pianificazione e nell'esecuzione di una indagine ed è strettamente connessa ad altre caratteristiche quali il fenomeno indagato, gli archivi di base, il strategia di campionamento, l'organizzazione del personale sul campo, i costi e i tempi attesi.
Inoltre non sono da sottovalutare le implicazioni della tecnica di indagine prescelta sulla qualità dei dati, in termini di mancate risposte e di errori di misura. La complessità delle scelte e le relazioni sopra menzionate possono essere facilmente illustrate mediante qualche esempio:

- il contatto postale è difficilmente eseguibile se non si dispone di una lista di indirizzi affidabile. In questo caso è meglio ricorrere ad una indagine areale;
- se si vogliono ottenere alti tassi di risposta è meglio ricorrere ad interviste personali condotte da rilevatori esperti;
- domande su argomenti delicati (es. reddito, comportamenti sessuali, reati contro la persona) sono sottoposte a minore reticenza se condotte per telefono o mediante un questionario autocompilato;

Nelle indagini longitudinali, al fine di limitare l'onere per il rispondente, può essere opportuno far seguire ad un primo contatto effettuato mediante intervista diretta, interviste telefoniche per le successive occasioni di rilevazione.

Di seguito si elencano le principali tecniche di indagine in uso per condurre una rilevazione, considerandone i più importanti vantaggi ed aspetti critici:

- Intervista diretta (o faccia a faccia);
- Intervista telefonica;
- Questionario postale autocompilato;
- Diario;
- Dati amministrativi;

- Osservazione diretta;
- Tecniche miste;
- Nuove tecnologie.

Intervista diretta (o faccia a faccia)
L'intervista viene condotta da un rilevatore che legge le domande e le opzioni di risposta nell'esatto ordine e con lo stesso linguaggio adottati nel questionario riportandovi quindi le risposte così come sono fornite dal rispondente.

Vantaggi
- Si presta meglio ad alcuni disegni di indagine (es.: censimenti e campionamento areale)
- Maggiore possibilità di contattare e convincere il rispondente a collaborare
- Si identifica esattamente il rispondente
- Possibilità di istruire il rispondente sul significato delle domande e sul modo corretto di fornire le risposte
- Flessibilità negli strumenti utilizzabili (audiovisivi, sezioni autocompilate, tecniche di probing, ..)
- Interviste di maggiore durata

Svantaggi
- Costosa da implementare
- Necessita di una organizzazione capillare sul territorio
- Richiede tempi più lunghi di altri metodi per la raccolta dei dati
- Maggiori rischi di condizionamento

Intervista telefonica
L'intervista viene condotta al telefono da un intervistatore che legge le domande e le opzioni di risposta nell'esatto ordine e con lo stesso linguaggio adottati nel questionario riportandovi quindi le risposte così come sono fornite dal rispondente.

Vantaggi
- Costi minori rispetto all'intervista faccia a faccia
- Tempestività della raccolta dati
- Non è richiesta un'organizzazione sul territorio
- Maggiore possibilità di controllo dell'operato dei rilevatori
- Possibilità di contatto anche per le persone che non si trovano in casa in orari "canonici"
- Bassi rischi di condizionamento e maggiore possibilità di porre quesiti delicati

Svantaggi
- Impossibilità di contattare le famiglie senza telefono
- Il rispondente non è identificato con certezza
- Limitazioni nella lunghezza del questionario e nell'aiuto fornito ai rispondenti

Questionario postale autocompilato
Il rispondente riceve il questionario a mezzo posta o corriere e provvede a compilarlo nelle parti ad esso spettanti e a rispedirlo indietro o eventualmente a riconsegnarlo ad un addetto che lo ritira a domicilio.

Vantaggi
- Bassi costi di realizzazione

- E’ richiesta un’organizzazione minore
- Bassi rischi di condizionamento
- Adatta per porre quesiti delicati
- Disponibilità di tempo per reperire eventuale documentazione necessaria alla compilazione
- Possibile sottoporre più categorie di risposta

Svantaggi
- Tempi lunghi di raccolta
- Impossibilità di identificare con certezza il rispondente
- Autoselezione dei rispondenti
- Minore capacità di ottenere la partecipazione all’indagine (il tema deve essere coinvolgente)
- Più difficile aiutare i rispondenti nella comprensione delle domande e nella compilazione del questionario (importanza della grafica)

Diario
E’ un particolare tipo di questionario strutturato appositamente per registrare eventi frequenti e di scarsa importanza quali spese di bassa entità o attività quotidiane. L’organizzazione di tale strumento è tale da permettere la registrazione degli eventi nel momento della giornata in cui essi avvengono in modo tale da non dover ricorrere ad uno sforzo di memoria, con una conseguente sottonotifica degli eventi, nello svolgimento di una intervista di tipo classico.

Vantaggi
- Non affetto da problemi di memoria per la rilevazione di eventi poco rilevanti e ad elevata frequenza (ad esempio: spese giornaliere, uso del tempo, visione di programmi TV)

Svantaggi
- Struttura del questionario complessa
- Sottonotifica degli eventi col passare del tempo di osservazione
- Rischi di condizionamento dei comportamenti da registrare
- Necessita di un rilevatore per la consegna, il ritiro e il supporto alla compilazione

Dati amministrativi. (Vedere anche Indagini Amministrative)
Dati, riferiti a soggetti individuali, raccolti allo scopo di intraprendere decisioni o azioni che riguardano gli individui medesimi (es. licenze, assicurazioni tributi, regolamenti, pagamenti, ...).

Vantaggi
- Relativamente economici da utilizzare a fini statistici
- Nessun disturbo ai rispondenti
- Spesso riguardano la totalità della popolazione e sono utili per costituire archivi

Svantaggi
- Possibili distorsioni dovute alla non coincidenza fra le definizioni usate per i dati amministrativi e quelli interessanti ai fini statistici
- Le leggi che regolano la raccolta possono cambiare pregiudicando la confrontabilità dei dati nel tempo
- Lo statistico non è in grado di controllare la qualità della raccolta dei dati
- Le informazioni utili ai fini statistici sono spesso raccolte in modo inaccurato perché non di primaria importanza ai fini amministrativi

Osservazione diretta
L'informazione viene raccolta dal rilevatore per mezzo dei propri sensi o mediante strumenti di misurazione fisici (applicazioni in antropologia, psicologia, geologia, telerilevamento, ...).

Vantaggi
- Preferibile qualora l'informazione fornita da un rispondente non sia considerata sufficientemente precisa (ambito sperimentale)

Svantaggi
- L'interazione fra osservatore e oggetto osservato riproduce gli stessi problemi di condizionamentoche si possono riscontrare con l'uso di rilevatori

Tecniche miste
Si utilizzano quando una sola tecnica di rilevazione non si comporta bene in tutte le situazioni pratiche.

Esempi di tecniche miste:
- Indagine postale + indagine diretta sui non rispondenti all'indagine postale
- Indagine telefonica + indagine diretta su coloro che non possiedono il telefono
- Indagine diretta + questionario individuale
- Diario + intervista finale
- Prima intervista diretta e successive con modalità telefonica
- Dati amministrativi + controllo campionario con questionario postale autocompilato.

Nuove tecnologie a supporto delle tecniche di indagine:
- CATI (Computer Assisted Telephone Interviewing)
- CAPI (Computer Assisted Personal Interviewing)

Il questionario è contenuto nel computer cosicché le domande vengono poste così come compaiono sullo schermo e le risposte sono registrate direttamente su supporto magnetico

Vantaggi
- Alcuni controlli di qualità sono eseguiti dal computer al momento dell'immissione con un conseguente risparmio nelle successive fasi di controllo di qualità
- Si gestiscono facilmente questionari molto articolati
- Possono essere predisposte formulazioni alternative delle domande
- Si accorciano i tempi di completamento dell'indagine (soprattutto nel CATI)

Svantaggi
- Occorre dotare i rilevatori di un Computer portatile (CAPI)
- E' necessario un maggiore addestramento dei rilevatori
- Problemi di hardware (CAPI - pesante, lento, batterie, ...)

Questionario (da Istat, 1989 - vol. 2)

Il questionario di indagine è lo strumento di misura con il quale si raccolgono le informazioni sulle variabili qualitative e quantitative oggetto di indagine. Il questionario deve essere visto come uno strumento di comunicazione finalizzato a facilitare l'interazione fra il ricercatore, il rilevatore e il rispondente. Affinché possa svolgere la propria funzione occorre che il questionario sia uno strumento standardizzato; ovvero domande e comunicazione devono essere identiche per tutti i rispondenti al fine che le informazioni raccolte siano confrontabili fra loro.

Le operazioni che devono essere curate per la realizzazione di un questionario possono essere schematizzate come segue:

- Definizione degli obiettivi e concettualizzazione;
- Definizione esatta di quali sono i temi che interessano l'indagine escludendo quelli che non sono di interesse primario;
- Preparazione della lista delle variabili (e non direttamente le domande) da raccogliere rispetto ai temi di interesse identificati in precedenza;
- Preparazione di un piano provvisorio delle analisi statistiche da compiere per accertarsi che i contenuti necessari allo studio siano tutti espressi.

Per la redazione del questionario è necessario:

- Stabilire la successione logica dei temi trattati (le sezioni del questionario).
  Affinché la comprensione del questionario non risulti ambigua è importante che il rispondente inquadri il contesto nel quale le domande si collocano. Per questo motivo occorre che la sequenza degli argomenti affrontati sia il più possibile coerente evitando che si verifichino salti radicali. Occorre tuttavia considerare che l'ordine stabilito nella sequenza degli argomenti può condizionare la risposta, creando distorsioni nei dati. Ad esempio se si vuole un'opinione spontanea sulla soddisfazione nel lavoro è bene non anteporre domande sulle caratteristiche specifiche del lavoro svolto che potrebbero focalizzare l'attenzione su alcuni aspetti particolarmente gradevoli o sgradevoli.
  I quesiti che implicano uno sforzo di memoria andrebbero collocati verso la metà del questionario, per evitare che all'inizio il rispondente non sia ancora disponibile a tale impegno e alla fine sia troppo stanco.
  I quesiti su temi delicati da affrontare andrebbero invece collocati verso la fine, per sfruttare la maggiore confidenza e disponibilità ormai acquisita e per non rischiare che un rifiuto a rispondere possa compromettere l'acquisizione delle informazioni collocate sull'ultima parte di questionario.

- Predisporre le domande filtro.
  Le domande filtro permettono di saltare uno o più quesiti successivi se sono verificate alcune condizioni. Tale necessità si manifesta quando:
  - occorre indirizzare gruppi particolari di rispondenti verso domande specificatamente rivolte a loro. Ad esempio per sottoporre gruppi differenti di domande per chi si dichiara occupato e per chi si dichiara non occupato;
  - si vuole evitare di scendere in domande dettagliate quando ciò è inutile. Ad esempio per non sottoporre un blocco di domande riguardanti le vacanze svolte nell'anno a coloro che dichiarano di non aver svolto vacanze nell'anno;
  - si vogliono evitare condizionamenti nella risposta. Ad esempio non si desidera chiedere opinioni sull'ultimo libro letto nei 12 mesi a chi non ha letto nessun libro nei 12 mesi, per non provocare risposte date allo scopo di non fare "brutta figura".

- Definire la sequenza di domande su uno stesso tema.
  La sequenza con la quale le domande sono poste è uno degli aspetti del questionario mediante il quale si può aiutare il rispondente nel compito di fornire le informazioni volute. Inoltre è necessario tenere presente che spesso la sequenza con la quale le domande appaiono non è "neutra" dal momento che si possono verificare condizionamenti non voluti privilegiando un ordine nei quesiti piuttosto che un altro.

Per aiutare i rispondenti nel loro compito è importante tenere presenti due stili nell'ordinamento dei quesiti:

- la successione a imbuto: si passa da domande generali a domande più particolari per dare tempo al rispondente di focalizzare l'attenzione sul tema proposto. Serve ad aiutare la memoria e a registrare opinioni non meditate;
- la successione ad imbuto rovesciato: si antepongono le domande specifiche a quelle più generali. Utili quando si desidera raccogliere opinioni meditate su un determinato argomento.

- Formulare i quesiti

  Il linguaggio utilizzato nelle domande è un aspetto critico per la riuscita di un questionario. Infatti anche piccole variazioni di linguaggio possono causare grandi effetti.

  In uno studio del 1981 un campione di famiglie è stato diviso in due sottogruppi casuali.

  Al primo sottogruppo è stata sottoposta la seguente domanda:

  Pensa che negli Stati Uniti debbano essere <u>proibiti</u> discorsi pubblici favorevoli al comunismo? (409 rispondenti);

  mentre al secondo sottogruppo è stato chiesto:

  Pensa che negli Stati Uniti debbano essere <u>permessi</u> discorsi pubblici favorevoli al comunismo? (432 rispondenti).

  Sebbene le due domande abbiano un significato esattamente opposto (la risposta "si" alla prima domanda corrisponde alla risposta "no" nella seconda) la percentuale di "si" per la prima domanda è stata del 39.3% mentre la percentuale di "no" alla seconda è stata del 56.3% con una differenza, statisticamente significativa, del 17%. Tale differenza, non attesa nel caso si considerino domande con significato esattamente opposto, può essere attribuita all'importanza del significato attribuito dai rispondenti ai termini "proibire" e "permettere".

  In molti casi anche l'ordine con il quale sono proposte le domande può influenzare la risposta. Ad esempio consideriamo le seguenti due domande:

  Domanda A: *Pensa che si dovrebbe lasciare che i giornalisti dei paesi comunisti in servizio negli Stati Uniti spediscano ai propri giornali le notizie così come le apprendono?*

  Domanda B: *Pensa che si dovrebbe lasciare che i giornalisti degli Stati Uniti in servizio nei paesi comunisti spediscano ai propri giornali le notizie così come le apprendono?*

  Quando le due domande furono proposte, con ordine invertito, a due campioni casuali di rispondenti di nazionalità statunitense (anno 1950) si ottennero i seguenti risultati:

  prima domanda A (54,7%) poi domanda B (63,7%);

  prima domanda B (81,9%) poi domanda A (74,6%).

  Una forte differenza in termini di percentuale, quella che le due domande presentano se proposte in diverso ordine, che può palesemente essere attribuita al fatto che i rispondenti si predispongono in maniera differente nelle due situazioni.

  E' inoltre importante che le domande siano formulate in modo da contenere informazioni sufficienti a non risultare ambigue. Infatti se si vuole che i gli intervistati rispondano tutti alla medesima domanda bisogna evitare che gli intervistatori siano costretti ad aggiungere parole per specificare una domanda incompleta.

  Ad esempio porre la domanda, *"La mattina consuma una colazione?"* presenta il problema di non chiarire da cosa sia costituita una colazione; non è chiaro fino a che ora del mattino un pasto possa essere considerato una colazione; non è chiaro se la domanda si riferisce ad un consumo abituale o a un giorno preciso. Meglio proporre il quesito, leggermente più lungo ma più definito, nella seguente forma: *"Per i nostri scopi consideri colazione un pasto costituito almeno da una bevanda (Te, latte, caffè, ...) e un alimento come brioches, cereali, biscotti, toast*

o frutta, consumato prima delle 10 del mattino. Secondo questa definizione negli scorsi 7 giorni quante volte ha consumato una colazione?"
Un altro tranello in cui non bisogna cadere è quello di usare un linguaggio dispregiativo o elogiativo (es.: *la scorsa domenica è stato a messa, come prescrive la Chiesa?*) oppure troppo complesso (es.: *Secondo lei negli ultimi dieci anni la propensione a sposarsi è aumentata, diminuita oppure rimasta uguale?*). Inoltre occorre evitare che i quesiti proposti contengano più domande in una volta sola (es.: *Si ritiene soddisfatto delle mansioni svolte e della posizione occupata nel suo attuale lavoro?*).

Porre attenzione alla formulazione dei quesiti retrospettivi.
I quesiti retrospettivi sottopongono il rispondente ad uno sforzo di memoria che può provocare due problemi:
- se l'evento avvenuto nel passato viene omesso per dimenticanza si sottovaluta l'entità del fenomeno da misurare;
- se un evento viene erroneamente localizzato all'interno del periodo di interesse si sopravvaluta l'entità del fenomeno (effetto telescopio).

Per questo motivo deve essere posta molta attenzione alla scelta del periodo di riferimento della domanda e alla corretta formulazione del quesito. In generale un buon quesito retrospettivo ha lo scopo di sollecitare la memoria del rispondente senza influenzarne i ricordi. Perciò è bene:
- ridurre il più possibile il periodo di riferimento;
- porre una batteria di domande per collocare temporalmente i ricordi del rispondente;
- proporre un buon numero di alternative di risposta per sollecitare la memoria;
- ricorrere ad un diario.

Esempio di tre modi di porre un quesito retrospettivo:
Riferire l'informazione ad un preciso momento nel passato;
- *(Censimento 20/10/91) "Indicare la condizione professionale o non professionale posseduta nell'Ottobre 1986."*;

Riferire l'informazione ad un periodo di tempo nel passato;
- *"Negli ultimi tre mesi è stato ricoverato in Ospedale, in una casa di cura convenzionata o in una casa di cura privata?";*

Registrare la data in cui è avvenuto l'ultimo evento di interesse;
- *facendo riferimento al matrimonio in corso o all'ultimo matrimonio indicare la data (mese e anno) di celebrazione del matrimonio*.

Formulare le domande delicate.
Alcuni argomenti sono psicologicamente difficili da indagare. Fra questi possiamo ad esempio annoverare: consumo di alcool, reddito, contraccezione, comportamenti sessuali, presenza di portatori di handicap in famiglia. Per questo è necessario che le domande siano formulate nel modo opportuno, come ad esempio:
- utilizzare una serie di domande di "approccio": *Alcune donne si sottopongono ad operazione per non avere più figli. Ha mai sentito parlare di tale metodo? Si è mai sottoposta a tale operazione?*;
- premettere osservazioni che informino sui comportamenti o li giustifichino: *Le è stato possibile recarsi a votare?*;
- ricorrere all'autocompilazione;
- porre le domande in forma indiretta: *Secondo lei di quanto avrebbe bisogno al mese una famiglia composta come la sua e nella stessa condizione per vivere in questa città, senza lussi, ma senza farsi mancare il necessario?* .

- Decidere l'organizzazione delle risposte
  Il modo in cui si registra la risposta alla domanda formulata deve essere considerato con la stessa attenzione posta nella predisposizione dei quesiti. Si possono identificare diversi tipi di struttura per una risposta:
  - Risposte a domande aperte: la risposta viene fornita dall'intervistato con parole proprie senza alcun suggerimento.
    Vantaggi:
    - non condizionano la risposta;
    - particolarmente utili quando occorre esplorare situazioni sconosciute;
    - utili per trattare quesiti delicati.

    Svantaggi:
    - implicano molto lavoro di registrazione e codifica;
    - riportano "luoghi comuni" in mancanza di opinioni ben definite;
    - non saranno compilate da individui che hanno difficoltà a scrivere o concettualizzare.
  - Risposte a domande strutturate (o a domande chiuse): è prevista una serie di risposte predefinite tra le quali il rispondente deve scegliere.
    Vantaggi:
    - riduce i tempi di codifica e registrazione;
    - aiuto al rispondente;
    - standardizza la domanda.

    Svantaggi:
    - troppe opzioni concentrano l'attenzione sulle ultime (Intervista diretta e telefonica);
    - poche opzioni possono trascurare fatti importanti;
    - il rispondente può rispondere a caso.
  - Domande a risposta multipla: le domande a risposta multipla sono domande strutturate che ammettono più di una risposta fra quelle predisposte;
  - Domande gerarchizzate: sono domande strutturate per le quali le opzioni di risposta devono essere ordinate secondo una scala di preferenze.
    Per ridurre gli svantaggi delle domande strutturate:
    - le diverse risposte possono essere elencare in appositi "cartellini" da sottoporre al rispondente (solo nel caso dell'intervista diretta);
    - introduzione della modalità di risposta "non so". Per gli indecisi evita una risposta data a caso, ma può indurre il rispondente alla pigrizia. Per questo, nel caso di intervista faccia a faccia, è bene associare tecniche di sollecitazione alla risposta da parte dei rilevatori;
    - accettare risposte aperte e lasciare all'intervistatore il compito di attribuire la risposta ad una delle modalità predisposta. Sussistono tuttavia rischi connessi alla interpretazione delle risposte da parte dei rilevatori.

Per la verifica del questionario, prima di rilasciarne la versione definitiva:

- occorre valutare se:
  - risponde alle esigenze conoscitive dell'indagine;
  - sono state omesse domande;
  - i riferimenti spaziali e temporali dei quesiti sono sufficienti;
  - linguaggio e struttura delle domande sono adeguati;
  - è facilmente comprensibile per gli intervistati e semplice da gestire per gli intervistatori.
- occorre mettere in atto una serie di controlli:
  - revisione estesa da parte di esperti del fenomeno;

- pre-test: rilevatori esperti intervistano un campione ragionato di individui per raccogliere elementi utili a valutare completezza, chiarezza e gestibilità del questionario;
- test di alternative: si sperimentano versioni alternative del questionario su piccoli campioni indipendenti di unità statistiche;
- indagine pilota: versione completa dell'indagine su scala ridotta per verificare il gradi di integrazione tra le fasi dell'indagine ed effettuare eventuali ultimi ritocchi anche sul questionario.

Tempi e Costi

La programmazione dei tempi e dei costi di esecuzione dell'indagine è un fattore critico per la riuscita della stessa. Tali variabili, infatti, oltre ad influenzarsi reciprocamente, sono fortemente connesse alla qualità dell'informazione prodotta.

Nella pratica l'elemento di costo viene visto come un vincolo al quale la progettazione deve sottostare senza tenere conto, in molti casi, del livello di errori che risorse carenti possono indurre nelle operazioni programmate. Se infatti una disponibilità illimitata di risorse può indurre a sprechi non sostenibili, un impegno di costo troppo limitato può altresì portare al fallimento degli obiettivi dell'indagine con perdite potenzialmente anche maggiori.

In tale contesto occorre inserire anche i tempi di esecuzione dell'indagine, tenendo conto della necessità di disporre di dati utilizzabili in un momento il più prossimo possibile a quello di riferimento dell'informazione raccolta (tempestività). La domanda di tempestività può essere indotta sia dall'urgenza dell'informazione, allo scopo ad esempio di prendere decisioni strategiche, sia da una rapidità di mutamento nel fenomeno osservato, tale da ridurre l'obsolescenza dell'informazione prodotta.

Anche la tempestività può essere messa in relazione con il costo sostenuto e la qualità dei dati prodotti. E' infatti lecito chiedersi se, al prezzo di un maggior impiego di risorse, si possa anticipare la diffusione a parità di qualità o viceversa, tenendo fisse le risorse impiegate si possa aumentare la qualità dei dati prodotti, posticipando i tempi di produzione. Ad esempio si può ritenere che, aumentando il numero di rilevatori in un'intervista diretta o telefonica, si possa comprimere il tempo di rilevazione; oppure la qualità dell'informazione prodotta potrebbe essere migliorata conducendo analisi supplementari sui dati al prezzo di un aumento dei tempi di lavorazione. Al contrario si potrebbe decidere di sopportare la diffusione di dati a qualità inferiore, per sopperire all'urgenza di informazione, diffondendo dati preliminari ad indagine non ancora conclusa.

Al fine di migliorare la pianificazione di tempi e costi d'indagine si raccomanda di considerare dapprima le singole fasi operative e quindi di valutarne attentamente l'integrazione. Inoltre occorre predisporre nel sistema dei controlli di qualità un adeguato monitoraggio delle risorse impiegate in ciascuna attività condotta, e dei loro tempi di esecuzione, mettendo tali informazioni a confronto con gli altri indicatori di qualità prodotti. Tali informazioni torneranno infatti utili sia in fase di validazione, per identificare inefficienze e colli di bottiglia, sia in successive fasi di progettazione della stessa o di altre indagini.

Sistema dei controlli di qualità

Il sistema dei controlli di qualità è costituito da un insieme di azioni predisposte nell'indagine e finalizzate al trattamento dell'errore non campionario.

Le azioni costituenti un sistema di controlli di qualità sono riunite in tre grandi classi:

- Azioni preventive, predisposte al fine di rendere meno probabile l'insorgere dell'errore attraverso l'esecuzione di pratiche che forniscano garanzie in tal senso. Ad esempio l'invio di una lettera di preavviso ai rispondenti o l'istituzione di un numero verde per le richieste di

chiarimento sono due operazioni che dovrebbero servire a facilitare le operazioni di risposte e quindi dovrebbero diminuire le mancate risposte all'indagine;

- Azioni di controllo in corso d'opera, predisposte al fine di individuare e correggere gli errori nel momento in cui questi insorgono durante il processo di produzione. L'uso dei programmi per la registrazione controllata dei dati costituisce un esempio di tali azioni. Un altro esempio è dato dall'applicazione delle tecniche di identificazione automatica degli errori, le quali servono ad individuarne la presenza di incoerenze nei dati (es.: un professionista con la sola licenza elementare) e la conseguente correzione, ad esempio, per mezzo di un ritorno sul rispondente, o almeno il ripristino dell'informazione con valori accettabili;
- Azioni di valutazione, predisposte per quantificare il livello di errore non campionario contenuto nei dati prodotti. Tali azioni implicano l'elaborazione di dati raccolti durante l'esecuzione del processo di produzione, ovvero la conduzione di prove ausiliarie o vere e proprie indagini di controllo. A seconda della natura dell'azione di valutazione si ottiene una misura dell'errore che può andare dalla semplice valutazione di quantità ad esso associate (indicatore di qualità) quali i tassi di risposta, a misure dirette di componenti dell'errore totale quali, ad esempio, la varianza semplice di risposta, ottenibile con una reintervista delle unità statistiche.

Per mezzo del sistema dei controlli di qualità si può ottenere da un lato il miglioramento dei parametri componenti le dimensioni della qualità, e dall'altro la validazione dei dati dell'indagine.

Gruppo di progettazione

L'elevato grado di complessità della fase di progettazione e la multidisciplinarietà delle conoscenze richieste rende indispensabile la formazione di un gruppo di lavoro in cui tutti gli aspetti, da quelli concettuali a quelli operativi, siano affrontati alla presenza di tutti i rappresentanti delle diverse aree di esperienza coinvolte (Statistics Canada, 1987).

Al fine di non lasciare scoperti aspetti che, se non adeguatamente affrontati nella fase di progettazione, rischiano di introdurre carenze nell'indagine è necessario comprendere nel gruppo di progettazione competenze professionali specifiche riferite:

- alla conoscenza del fenomeno oggetto di indagine;
- alla progettazione del questionario;
- al disegno di campionamento;
- ai controlli di qualità;
- alla pianificazione degli aspetti amministrativo-contabili;
- all'organizzazione del lavoro sul campo;
- alla progettazione delle applicazioni informatiche;
- alla diffusione.

Il gruppo di progettazione ha come obiettivo quello di definire gli scopi conoscitivi dell'indagine, adottare definizioni e concetti operativi e pianificare gli aspetti applicativi del processo di produzione. Particolare attenzione nell'ambito del gruppo dovrà essere data all'integrazione fra i concetti e le procedure definiti, per assicurare il funzionamento dell'intero sistema e non solo la coerenza interna delle singole parti di esso.

Documento di progettazione

L'attività di pianificazione del gruppo di progettazione deve essere approfonditamente dettagliata in un documento di progettazione che deve risultare articolato negli aspetti concettuali e in quelli relativi all'implementazione dell'indagine, considerando tuttavia le relazioni esistenti fra i due diversi piani di descrizione.

In particolare è molto importante discutere l'impatto che le definizioni e le procedure di indagine hanno sulle componenti della qualità dell'informazione prodotta. Rilevanti ai fini della stesura del documento di progettazione risultano:

- obiettivi: contestualizzazione del fenomeno oggetto di indagine e analisi delle informazioni già disponibili da altre fonti;
- definizioni e concetti: descrizione delle definizioni e dei concetti adottati con particolare riferimento alle loro relazioni con gli obiettivi ed alle problematiche riguardanti il passaggio dalle definizioni teoriche all'applicabilità pratica;
- analisi dei confronti praticabili (e non) fra i dati dell'indagine e quelli disponibili da altre fonti;
- classificazioni: standard adottati e problemi di riconducibilità ad altri standard in termini di possibilità di integrazione fra dati;
- periodicità e tempestività: pianificazione del disegno di indagine in relazione all'obiettivo di raccogliere dati trasversali e/o longitudinali; valutazione del tempo intercorrente fra il periodo di riferimento dei dati e l'istante di rilascio dei dati pubblicati;
- liste e archivi: scelta e descrizione delle liste da utilizzare per identificare la popolazione obiettivo; analisi della completezza e della ridondanza delle liste utilizzate; valutazioni concernenti la presenza di errori nelle informazioni disponibili, tali da precludere il contatto delle unità di rilevazione, il calcolo di pesi di riporto all'universo o l'assegnazione delle unità a strati;
- campionamento: definizione del disegno di campionamento in relazione alle liste di base disponibili ed agli obiettivi dell'indagine; analisi dei problemi di applicabilità del disegno teorico alle situazioni pratiche;
- strumenti di raccolta: descrizione degli strumenti utilizzati per la raccolta delle informazioni presso le unità statistiche (questionari e/o documenti amministrativi) e degli eventuali modelli ausiliari di aiuto alle operazioni di contatto o di ritorno sul campo;

Sperimentazioni (della fase progettuale)

La fase progettuale non può dirsi conclusa senza predisporre un momento di verifica delle soluzioni considerate.

Le sperimentazioni dovrebbero essere finalizzate a valutare: l'adeguatezza e la comprensibilità dei concetti e delle definizioni adottate nei casi pratici; il questionario di indagine; la migliore fra più possibili soluzioni di specifici problemi; l'adeguatezza delle previsione riguardanti tempi e costi necessari allo svolgimento delle attività predisposte.

Ovviamente limiti di bilancio e di tempo possono impedire l'esecuzione di sperimentazioni per tutti i singoli aspetti riguardanti l'indagine. In questo caso si dovrebbero tuttavia identificare tutte le fasi critiche e, almeno per quelle, predisporre esperimenti, anche limitati, per valutare la possibilità che gravi problemi sorti in questi frangenti possano pregiudicare o influenzare i risultati dell'indagine.

Per le altre operazioni si dovrebbero comunque discutere tutti i possibili aspetti critici che possano indurre problemi di tempo, costi o qualità dei dati, tenendo conto anche dei possibili confronti con studi quantitativi condotti in epoche precedenti o in contesti assimilabili al proprio.

In questo paragrafo non si intende scendere in ulteriori dettagli, per i quali si rimanda alla letteratura sull'argomento, ma si ritiene opportuno sottolineare almeno la differenza tra due importanti modalità di verifica delle progettazione di un'indagine: il test di soluzioni alternative e l'indagine pilota.

Nel test di soluzioni alternative un campione contenuto di unità statistiche viene suddiviso in un numero di sottogruppi pari al numero di diverse alternative da saggiare. Tale suddivisione in sottogruppi deve essere operata rispettando un criterio di causalità nell'assegnazione delle unità ai gruppi. Quindi, dopo aver applicato il metodo opportuno alle unità appartenenti ai gruppi, si misura una caratteristica quantitativa (variabile risposta) che possa rappresentare in modo adeguato la bontà delle alternative prescelte. La scelta dell'alternativa di maggior successo può essere valutata applicando un test statistico alle differenze riscontrate sulle misure riassuntive calcolate sulle unità appartenenti a ciascun gruppo. La caratteristica fondamentale del test di alternative è quella di prendere in considerazione un singolo aspetto da valutare, enucleandolo dal contesto, e di predisporre un esperimento piccolo e relativamente poco costoso.
L'indagine pilota è in tutto e per tutto una esecuzione dell'indagine su scala molto ridotta. L'indagine pilota viene eseguita dopo uno o più test di alternative svolti su aspetti specifici e, senza avere lo scopo di saggiare alternative, è finalizzato piuttosto a verificare che l'insieme delle soluzioni prescelte sia adeguato in una situazione reale e che l'interazione fra esse non provochi problemi. Svolta con le stesse modalità dell'indagine vera e propria, l'indagine pilota permette di identificare aspetti critici non considerati in fase di progettazione, facilitando la correzione in tempo utile degli eventuali problemi.

Fasi operative

Con il generico termine di "fasi operative" si intende individuare tutta la parte del ciclo produttivo di un indagine che va dalla misurazione delle caratteristiche di interesse sulle unità selezionate fino alla disponibilità dei dati per le analisi statistiche. In questa sede si ritiene di distinguere le seguenti fasi operative:

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

Occorre osservare che la classificazione adottata, comoda ai fini esplicativi, può nella realtà essere suddivisa ulteriormente in sotto-fasi o non prevedere una o più fasi tra quelle elencate.
Ad esempio, qualora il rilevatore si avvalga di un computer portatile per la conduzione di un intervista faccia a faccia, la fase di registrazione dei dati, viene eliminata e quella di revisione automatica risulta semplificata.

Rilevazione

Nella fase di rilevazione, le unità selezionate per l'indagine vengono contattate allo scopo di raccogliere l'informazione rilevante ai fini dello studio. Le modalità di contatto e raccolta dati presso le unità di rilevazione dipendono dalla tecnica di indagine adottata e hanno implicazioni sia sui costi che sulla qualità dei dati.
Indipendentemente dalla tecnica adottata, la rilevazione ha tre obiettivi fondamentali:

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

Affinché tali obiettivi siano raggiunti occorre che l'attività di rilevazione sia preparata con cura, predisponendo condizioni ambientali che ne facilitino la riuscita, strumenti e procedure il più possibile semplici ed efficienti e meccanismi tempestivi di individuazione dei problemi e recupero delle informazioni che altrimenti andrebbero perdute.
Gli aspetti fondamentali che devono essere considerati sono:

- predisposizione del questionario e dei modelli ausiliari, strumentali alle operazioni di contatto delle unità di rilevazione e di gestione della raccolta;
- 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.

Con questo elenco, senza la pretesa di esaurire l'argomento, si intende soltanto porre l'accento sulle maggiori problematiche delle quali tenere conto nella fase di rilevazione. Nel seguito si forniscono delle raccomandazioni su alcuni degli aspetti più delicati da considerare.

Alcune Raccomandazioni

Al fine di creare un clima favorevole alla conduzione della rilevazione è opportuno informare e sensibilizzare la popolazione obiettivo servendosi degli organi di stampa e/o di associazioni di categoria (imprese). Occorre inoltre preavvisare le unità selezionate per la rilevazione vera e propria per mezzo di lettere nelle quali siano evidenziati: lo scopo della ricerca, i benefici dell'informazione raccolta per il collettivo esaminato, il contributo individuale ad un interesse collettivo, la riservatezza della raccolta e l'inserimento casuale fra le unità contattate (solo indagini campionarie). Al fine di agevolare il compito ai rispondenti è anche auspicabile fornire sempre un recapito telefonico, meglio se gratuito, cui rivolgersi per ulteriori richieste di chiarimenti, commenti o suggerimenti.

Il personale coinvolto nelle operazioni dovrebbe essere informato adeguatamente sulle modalità dell'intero processo e non solo sul segmento di propria responsabilità. In particolare, se è previsto l'impiego di rilevatori, questi andrebbero informati sulla gravità delle mancate risposte ed andrebbe loro enfatizzata l'importanza di ottenere questionari completi. Adeguata attenzione dovrebbe inoltre essere posta sul corretto atteggiamento da tenere per aiutare i rispondenti durante l'intervista (es.: ausilio alla memoria) senza, nel contempo, influenzare le risposte.

Nel caso di indagini areali predisporre nel dettaglio le modalità di percorso delle aree e di identificazione delle unità da enumerare.

Esercitazioni pratiche e gruppi di discussione andrebbero predisposte al fine di standardizzare il comportamento dei rilevatori adeguandolo alle procedure previste.

Dovrebbero, inoltre, essere definiti nei dettagli i controlli da effettuare per giudicare il grado di completezza del questionario ed identificare palesi incongruenze eventualmente contenute in esso. Prevedere in questi casi un ritorno presso i rispondenti per la correzione dei dati. In ogni caso, mai dare ai rilevatori istruzioni per apportare correzioni ai dati raccolti in assenza dei rispondenti.

Per fronteggiare i casi in cui i rispondenti non vengono immediatamente trovati occorre predisporre un piano di contatti successivi (questa raccomandazione è valida anche nel caso di indagini telefoniche).

Assicurarsi che i contatti (telefonici e personali) avvengano in diversi orari e giorni della settimana (anche nel fine-settimana). Qualora sia impiegato un numero consistente di rilevatori, prevedere l'impiego di supervisori finalizzati a monitorare la correttezza delle procedure eseguite.

Predisporre sempre un piano di ritorni sul campo (ad esempio telefonici) per assicurarsi che le interviste abbiano avuto luogo.
Tenere periodiche riunioni insieme ai rilevatori per evidenziare e risolvere eventuali problemi non previsti in fase di progettazione.
Se i costi e l'organizzazione lo permettono, utilizzare tecniche di raccolta dati assistite dal computer (CATI, CAPI).
Nel caso di indagini postali predisporre sempre buste pre-affrancate e tentare di acquisire un recapito telefonico per eventuali ritorni sul campo. Circa dieci giorni dopo l'invio dei questionari spedire una lettera in cui si ringrazia per la partecipazione e, si ricorda, se non lo si fosse già fatto, di rispedire indietro il questionario compilato. Predisporre quindi un piano di solleciti finalizzato a diminuire le mancate risposte totali; se è disponibile il recapito telefonico, e se i costi in bilancio lo permettono, effettuare un sollecito telefonico dei non rispondenti. Prevedere, almeno su un sotto-campione di unità, un invio mediante raccomandata per poter calcolare la percentuale di mancati contatti (ricevute non tornate) distinguendola dai non rispondenti (ricevuta tornata, questionario non tornato). Registrare la data di ritorno dei questionari postali per analizzare le curve di risposta nel tempo.
Infine possono essere citate alcune attività volte alla valutazione della qualità conseguita; raccogliere, conservare ed analizzare le informazioni sul numero di contatti necessari ad ottenere la risposta o i motivi di mancata risposta; calcolare i tassi di mancata risposta sul totale del campione e su specifici sotto insiemi; esplicitare le quantità e le funzioni di calcolo utilizzate nel computo dei tassi; registrare i tassi di mancata risposta e calcolare gli andamenti nel tempo; acquisire e analizzare informazioni sui non rispondenti (almeno un campione di essi) usando dati disponibili sugli archivi di base, eventuali fonti esterne o ritorni sul campo effettuati con più efficienti (ad esempio mandando un rilevatore presso alcuni non rispondenti al questionario postale).

Registrazione su supporto informatico

La fase di registrazione su supporto informatico consiste nel convertire le informazioni raccolte presso i rispondenti, e disponibili su questionario cartaceo, su supporto di formato interpretabile dalle procedure informatiche predisposte dall'indagine. Nastri magnetici, floppy disc, CD Rom e DVD sono solo alcuni esempi di supporti disponibili per contenere i dati di indagine.
Solitamente questa operazione consiste nell'immissione dei dati al computer da parte di un operatore che digita su una tastiera esattamente ciò che legge sul questionario cartaceo. L'operazione, che non richiede un'elevata conoscenza dell'indagine e delle sue caratteristiche, è normalmente svolta da personale non specializzato. Per questo motivo la fase di registrazione dei dati deve essere considerata una notevole fonte di errore potenziale.
Le operazioni possono essere condotte secondo differenti modalità organizzative, caratterizzate dal grado di standardizzazione e controllo che si riesce ad esercitare sul personale ad esse preposto. Si va da una situazione in cui la registrazione è effettuata in proprio dall'ente gestore dell'indagine, ed una in cui questa viene appaltata ad una ditta esterna, per finire con il caso in cui la registrazione è distribuita sul territorio e affidata agli enti che curano la rilevazione in loco.
Anche le modalità tecniche possono variare dal caso più semplice in cui l'operatore digita i dati su una maschera d'acquisizione che non prevede alcun avviso di errore, fino ad una situazione di registrazione controllata in cui l'operatore viene avvisato nel caso vengano commessi evidenti errori di immissione, come ad esempio lasciare vuoto un campo obbligatorio. E' evidente che, anche nel caso siano stati previsti controlli accurati, qualche errore potrà comunque non essere rilevato come ad esempio nel caso in cui l'anno 1929 sia erroneamente digitato 1992.

In alcuni casi la fase di registrazione può essere assente, come quando la rilevazione viene effettuata in modalità assistita dal computer (CATI, CAPI). In altri casi l'operazione può essere sostituita dalla lettura ottica dei questionari dove l'operatore acquisisce un ruolo, tecnologicamente più specializzato, di supervisione delle operazioni svolte dalla macchina.
Come per tutte le fasi di un'indagine anche per la registrazione esiste il rischio che siano introdotti errori nello svolgimento del lavoro.
Tra gli errori, costituiti da ogni differenza fra quanto registrato e quanto riportato sul questionario, ne possono essere per la loro tipicità, riportati alcuni:

- quantità monetarie erroneamente divise o moltiplicate per fattori fissi (solitamente 1.000);
- scambi di tasti (es.: 27 al posto di 72) o errore di digitazione di tasti contigui es.: F invece di G);
- slineamenti: errori determinati dalla dimenticanza nell'immissione di una variabile. Da tale variabile in poi tutti i dati successivi sono registrati in posizione errata, un campo più a sinistra del dovuto, generando un'intera sequenza di errori di registrazione.

Questi errori sono tanto più gravi, dal punto di vista dell'informazione statistica, quanto più importanti per l'analisi sono le variabili in essi implicate. In questo ambito rientrano certamente i codici identificativi, un errore nei quali pregiudica l'identificazione univoca delle unità statistiche o la loro collocazione negli strati di appartenenza.

Alcune raccomandazioni
La fase di registrazione deve, come le altre, essere progettata prevedendo un'accurata definizione delle procedure operative, della formazione e del controllo di qualità del personale e tenendo conto che, almeno in parte, gli errori introdotti in una determinata operazione dipendono dal modo cui sono state progettate ed eseguire le fasi precedenti.
La pratica del questionario dovrebbe essere progettata in modo tale da semplificare la leggibilità all'operatore della registrazione. E' in ogni caso indispensabile pre-codificare le operazioni di risposta. Inoltre si dovrebbe evitare che la lettura del questionario risulti monotona, ad esempio sfalsando, quando possibile, gli spazi dedicati alla barratura delle risposte.
La progettazione del tracciato record deve tenere conto della variabilità del fenomeno. Ad esempio è indispensabile che il campo relativo all'età degli individui occupi tre byte in modo da non confondere i pluricentenari con i bambini. Allo stesso scopo il tracciato record deve essere corredato da un piano di registrazione dove sono riportati i codici ammissibili per ogni variabile. Occorre prevedere un codice non ambiguo per indicare le mancate risposte dal momento che altrimenti possono sorgere ambiguità qualora non si faccia esplicitamente la distinzione fra tali codici, gli zeri e i blank.
Ogni qualvolta sia possibile dovrebbe essere prevista la registrazione controllata in modo che i gravi errori possano essere immediatamente identificati e corretti. E' bene tuttavia sottolineare che un errore identificato dal programma di immissione deve essere corretto solo se a provocarlo è stato l'operatore della registrazione. In caso contrario l'errore deve essere ammesso per non costringere l'operatore ad apportare una correzione che non è in grado di eseguire. Per questo motivo i programmi di registrazione controllata devono segnalare gli errori, senza però impedirne l'immissione. La formazione degli operatori della registrazione rappresenta comunque uno degli aspetti indispensabili dei quali tenere conto. In tale occasione devono essere sviluppati gli argomenti relativi al tracciato record ed al piano di registrazione, considerando tutte le possibili condizioni di errore e le possibili soluzioni da adottare. La formazione dovrebbe essere corredata da esempi ed esercitazioni ed il grado di apprendimento degli operatori dovrebbe essere testato. Gli operatori dovrebbero inoltre, durante lo svolgimento del lavoro, essere messi a conoscenza delle quantità e della qualità del lavoro svolto.

Adeguate procedure di test della registrazione dovrebbero essere basate sulla ripetizione della registrazione su un campione di questionari e sul confronto fra la prima e la seconda registrazione per l'identificazione delle incongruenze e l'identificazione degli errori. La produzione di indicatori della qualità della registrazione può essere quindi basata sul rapporto tra byte errati e byte controllati. Indicatori più specifici possono essere calcolati in riferimento a particolari variabili o tenendo conto dei lotti di questionari elaborati dai diversi operatori.

Revisione automatica

Si definisce revisione automatica la fase di individuazione e intervento di imputazione, sui valori mancanti o incongruenti nelle variabili rilevate, per mezzo di procedure informatizzate. Tali valori, ai quali si farà nel seguito riferimento come errori, sono tutti e soli quelli che conducono a violazioni di regole logico formali, denominate regole di compatibilità, relative ai limiti imposti sul campo di variazione delle singole variabili, alle relazioni intercorrenti fra le variabili e alle relazioni formali stabilite dalle norme di compilazione dei modelli cartacei. Comprendiamo nella fase di revisione automatica anche le procedure di revisione interattiva, nelle quali viene automatizzata la sola fase di individuazione dell'errore, lasciando ad un operatore il compito di eseguire le correzioni al terminale. Come nel caso della revisione manuale, l'obiettivo di questa fase è quello di effettuare correzioni nei dati, in modo da minimizzare l'effetto degli errori riscontrati sulle successive fasi di elaborazione e sull'informazione prodotta.

Nel seguito, al posto del termine improprio di correzione, verrà usato quello, mutuato dall'inglese imputation, di imputazione. Infatti, a meno di non ritornare presso il rispondente, qualsiasi intervento di eliminazione delle condizioni di errore verificate nei dati non assicura il ripristino del vero valore presentato per l'unità statistica in questione.

Le modalità di conduzione della revisione

I programmi di revisione automatica sono costituiti da procedure di individuazione dell'errore e da procedure per la sua imputazione. Tali procedure possono essere classificate sulla tipologia di errori trattati. Gli errori possono essere infatti suddivisi, a seconda della loro natura, in errori sistematici o errori casuali:

- gli errori sistematici sono tutti quegli errori per i quali si può supporre che, per sottopopolazioni identificabili, il valore corretto con il quale effettuare l'imputazione sia unico;
- gli errori casuali, viceversa, sono tali che comunque siano identificate le sottopopolazioni di unità si deve attendere un margine di variabilità residuo rispetto alle possibili correzioni effettuabili.

Per quanto riguarda gli errori sistematici può essere fatto un esempio relativo alla rilevazione delle forze di lavoro. Per gli individui al di sotto dei quattordici anni infatti, per la legge italiana non si può far parte della popolazione attiva. Pertanto eventuali minori di tale età che si dichiarino occupati o in cerca di occupazione vengono automaticamente inclusi nella popolazione non attiva.
Questa scelta corrisponde ad ipotizzare un errore sistematico per la sottopopolazione degli individui in età inferiore ai quattordici anni quando non sia dichiarata l'appartenenza alla popolazione non attiva. Ovviamente, sottostante a tale scelta, c'è l'ipotesi che non si verifichino errori nella dichiarazione della data di nascita.
Al contrario un errore si può ritenere casuale qualora permanga, fra le possibili correzioni applicabili, una variabilità residua a prescindere dalla sottopopolazione identificata. Sempre nello stesso caso trattato, ad esempio, si può pensare che certi valori della professione siano incompatibili con il titolo di studio di un individuo in età lavorativa (14 anni o più), ma che il valore

corretto possa essere, con una certa distribuzione di probabilità, uno in una serie di modalità possibili.
In alcune situazioni inoltre non si possono fare ipotesi forti su quale variabile sia errata fra due o più variabili che concorrono nel generare l'incongruenza, se ad esempio la professione dichiarata, il titolo di studio o entrambe.
I metodi adottati per effettuare l'imputazione delle incompatibilità hanno tutti l'obiettivo di riportare i dati alla condizione di ammissibilità, apportando modificazioni tali da influire il meno possibile sulle stime di interesse. In generale per effettuare le imputazioni si può ricorrere ad un nuovo contatto dell'unità statistica per acquisire il valore vero, ad informazioni possedute rispetto a periodi precedenti, o alla sostituzione dell'informazione incongruente con altra, relativa ad unità simili a quella per la quale si è registrato l'errore. Quest'ultima modalità, denominata imputazione probabilistica, viene sovente utilizzata per correggere grandi moli di dati raccolti su unità statistiche abbastanza omogenee fra loro e si caratterizza per il basso costo, ma deve essere applicata con estrema attenzione affinché non siano violati importanti parametri di correttezza metodologica. A tale proposito l'Istat utilizza un software generalizzato per l'imputazione probabilistica dei dati, sviluppato in proprio e denominato SCIA, che consente di compiere tale tipo di revisione rispettando i requisiti metodologici.
Alla categorizzazione basata sulla natura degli errori corrispondono due classi di procedure per l'imputazione dei dati: le prime "correggono" gli errori sistematici attraverso l'applicazione di una serie di regole deterministiche del tipo "SE-ALLORA", mentre le seconde intervengono sugli errori casuali modificando il minimo insieme di informazioni, tale cioè da riportare nella regione ammissibile l'informazione raccolta, in modo da riprodurre la stessa variabilità osservata sui dati non affetti da errore e influenzare il meno possibile le stime finali.
Poiché in un file di dati possono coesistere sia errori deterministici che errori casuali l'ordine che deve essere seguito nell'applicare le procedure di revisione automatica prevede l'esecuzione preliminare delle procedure per l'individuazione e l'imputazione degli errori sistematici, seguita da quella delle procedure probabilistiche per il trattamento degli errori casuali.
Alcuni metodi di revisione rinunciano invece ad intervenire su tutti gli errori, limitandosi a trattare solo quelli più influenti sulle stime di interesse e lasciando intatti tutti gli altri. Il ricorso a tali metodi, che vanno sotto il nome di editing selettivo è particolarmente appropriato quando l'influenza sul fenomeno da parte delle unità rilevate è molto differente e si ha interesse a correggere con cura solo le unità più importanti, anche utilizzando metodi costosi come il ritorno sul campo. E' importante osservare che queste tecniche sono applicate prevalentemente sotto forma di revisione interattiva, visto che dopo l'identificazione degli errori si cerca solitamente di ristabilire proprio il valore vero ricontattando le unità di interesse. Ad esempio, in un'analisi su una popolazione di imprese, è possibile applicare dapprima le tecniche di editing selettivo sulle aziende più grandi in termini di fatturato, intervenendo successivamente con procedure probabilistiche sulle aziende più piccole e numerose.

Fonti d'errore
Occorre subito osservare che nessun programma di revisione automatica è in grado di individuare e imputare qualsiasi errore nei dati. In generale solo la classe degli errori che violano le regole di compatibilità predisposte, che denominiamo errori individuabili, potrà essere scoperta e quindi essere sottoposta alle opportune elaborazioni aventi lo scopo di risolvere le incongruenze riscontrate. Si è già detto tuttavia che tale modificazione non ripristina necessariamente l'informazione vera, ma piuttosto la modifica in modo tale che, sulla base di una serie di regole logiche che si suppongono valide per i dati raccolti, questa sia riportata ad un valore più vicino a quello reale.

Pertanto il processo di revisione automatica può essere visto come un modo per aumentare la qualità dei dati raccolti, incorporando in essi una serie di conoscenze, esprimibili sotto forma di proposizioni logiche, relative al fenomeno indagato e al processo di produzione dell'informazione. Per questo la scelta di correggere i dati dovrebbe essere presa soltanto se si giudica che gli errori individuabili siano tali da rendere troppo bassa la qualità dell'informazione rispetto ai livelli prestabiliti e se si pensa che l'insieme delle informazioni ausiliarie che si possiedono, qualora applicate sotto forma di regole di compatibilità all'insieme dei dati, permettono di correggere i dati di migliorare la qualità dell'informazione raccolta. In generale i termini del problema devono essere posti non tanto sull'esistenza di tali informazioni, quanto sulla loro corretta identificazione e strutturazione. Infatti la definizione di regole logico formali parzialmente non appropriate, o l'applicazione di procedure inadeguate può risolversi in gravi distorsioni nelle stime.

La definizione non corretta di un insieme di regole di compatibilità, invece di permettere l'individuazione degli errori, potrebbe essere fonte di ulteriori problemi. Infatti si possono introdurre distorsioni nel caso in cui i diversi errori che possono affliggere i dati siano affrontati solo in modo parziale, ad esempio trattando in modo accurato alcune condizioni di errore e trascurandone altre. Inoltre fra le molte regole di compatibilità che possono essere definite per una singola indagine, alcune possono essere in contrasto con altre, contribuendo a generare situazioni di incoerenza. D'altro canto la definizione di un insieme ridondante di regole di compatibilità, ancorché fra loro coerenti, può determinare un eccesso di correzioni contravvenendo al principio per il quale è meglio intervenire il meno possibile con correzioni nei dati.

Il trattamento di alcuni errori con metodi impropri può essere un'altra fonte di problemi. Infatti trattare gli errori deterministici con metodi di imputazione adatti agli errori casuali è un modo certo per introdurre distorsioni significative nei dati. Inoltre la trattazione di alcuni errori nella fase di revisione automatica può non costituire la scelta ottimale rappresentando una azione impropria. Ad esempio nel caso in cui si possa effettuare la registrazione controllata dei dati è bene adottare tale modalità, altrimenti la procedura di revisione automatica, pur individuando gli errori imputerà in modo non efficiente quelli provocati dalla registrazione. Per questi errori la correzione nel momento in cui sono generati consentirebbe infatti di ripristinare proprio il valore corretto.

Nel caso della revisione interattiva la non ottemperanza delle procedure prestabilite da parte di uno o più operatori dedicati alla fase costituisce un problema particolarmente grave. Infatti le distorsioni introdotte potrebbero essere anche maggiori che in altri casi, in quanto si ricorre di solito a correzioni interattive nei frangenti più delicati, quando cioè sia vitale che l'informazione venga ripristinata in modo il più aderente possibile alla realtà. Effettivamente il ricorso a tale modalità di intervento avviene di norma quando si intende correggere i dati di unità molto influenti sui risultati dell'indagine, come accade per le grandi imprese. In questo caso la procedura potrebbe prevedere dapprima un ritorno sul questionario, quindi la consultazione di archivi storici o derivanti da altre fonti ed infine, qualora l'informazione raccolta fino ad allora non fosse ritenuta affidabile, il ritorno presso l'impresa stessa. L'omissione, da parte degli operatori, di qualche operazione prevista da questa procedura potrebbe vanificare almeno in parte gli sforzi di progettazione fatti per mantenere alta la qualità dei dati raccolti.

Alcune Raccomandazioni

In sede di progettazione occorre innanzitutto valutare attentamente l'effettiva necessità di introdurre un processo di imputazione anziché limitarsi alla semplice individuazione e conteggio delle incompatibilità riscontrate nei dati. Se ad esempio devono essere calcolati dati di sommario o tabulazioni complesse su dati per i quali è richiesta la coerenza con ammontari desunti da altre

fonti, è solitamente opportuno procedere ad una fase di revisione automatica. Nel caso in cui si debbano invece applicare modelli per l'associazione potrebbe essere sufficiente eliminare dall'analisi le unità statistiche per le quali i dati risultano incompleti o incongruenti, o, in altri casi, applicare modelli che ammettono la presenza di dati incompleti.
Se si decide di procedere alla revisione automatica, occorre scegliere fra i possibili metodi alternativi, preferibilmente con l'ausilio di valutazioni quantitative basate, ad esempio, su dati di censimento, di precedenti occasioni di indagine o dati simulati, e su una analisi costi-benefici. Come principio generale è bene dare la priorità a metodi dei quali siano ben noti i principi teorici e le proprietà statistiche, siano conosciute e sperimentate le strategie applicative e possibilmente siano disponibili programmi generalizzati ben collaudati.
La procedura complessiva di revisione automatica deve essere progettata in modo tale che le diverse fasi operative siano fra loro coerenti. In termini esemplificativi si può riportare un processo nel quale siano presenti le seguenti fasi:

- applicazione di procedure per l'individuazione e l'imputazione di errori sistematici;
- applicazione di procedure di editing selettivo per l'identificazione e l'imputazione di unità influenti;
- applicazione di procedure per l'individuazione e l'imputazione di errori casuali su un set di variabili di maggiore importanza;
- applicazione di procedure per l'individuazione e l'imputazione di errori casuali su un set di variabili di minore importanza, condizionatamente alle imputazioni effettuate in precedenza.

Ciascuna delle fasi citate deve prevedere un momento di analisi e validazione seguente all'operazione vera e propria, mediante il quale individuare possibili distorsioni sistematiche, introdotte da una definizione imperfetta dell'insieme di regole di compatibilità, e risolvere di conseguenza i problemi. In fase di progettazione deve essere quindi previsto il monitoraggio di ciascuna operazione della revisione, utilizzando la documentazione prodotta sotto forma di indicatori di prestazione e fornita in output al processo.
Le procedure di revisione automatica devono essere corredate da una analisi dei valori anomali (outlier) e da strategie per il loro trattamento. Il problema dell'identificazione degli outlier è particolarmente delicato in quanto i singoli casi identificati possono essere esatti, anche se anomali in quanto lontani dalla media del sottoinsieme cui appartengono, ed è solo dalla loro eccessiva frequenza che si individua un problema sistematico. Tale problema potrebbe essere introdotto proprio dai precedenti passaggi delle procedure di revisione e deve pertanto essere attentamente considerato.

Per quanto detto le procedure di identificazione ed imputazione degli errori devono produrre indicatori utili al monitoraggio del processo di produzione. Esempi di tali indicatori sono costituiti da tabelle che riportano l'incidenza degli errori riscontrati nel complesso e secondo le singole variabili controllate. Inoltre può essere analizzata la variabilità degli indicatori tra sottogruppi di unità statistiche aggregate secondo i domini territoriali di appartenenza o secondo gli enti (es. Comuni o rilevatori) che hanno compiuto la rilevazione. La variabilità insita in tali tabelle può aiutare nell'individuazione di problemi e distorsioni introdotte dall'organizzazione dell'indagine. Fra gli indicatori che dovrebbero essere forniti, sia a livello aggregato che per sottoinsiemi di dati, sono da citare:

- tassi di mancata risposta parziale per variabile;
- tassi di attivazione delle regole di compatibilità;
- tassi di imputazione per ciascuna variabile e per modalità di imputazione adottata;
- matrici di transizione delle variabili nel processo dai dati grezzi ai dati puliti;

- dissomiglianze fra le distribuzioni semplici e doppie sulle principali variabili prima e dopo il processo di identificazione ed imputazione degli errori;
- differenze fra le stime prodotte dall'indagine calcolate sui dati grezzi e su quelli puliti.

Per facilitare il calcolo degli indicatori citati è opportuno mantenere in archivio per un periodo congruo, oltre ai dati puliti, anche il file dei dati grezzi. Se esistono problemi di spazio è possibile mantenere in archivio il solo insieme dei record per i quali è stata apportata almeno un'imputazione.

Per quanto riguarda i dati di fonte amministrativa occorre segnalarne la peculiarità, in quanto la raccolta è effettuata per scopi differenti da quelli dell'indagine statistica e tutta una serie di controlli, manuali e automatici, possono essere stati eseguiti in precedenza dagli enti che li utilizzano a scopi amministrativi. In tal caso è opportuno approfondire la conoscenza delle procedure utilizzate per identificare ed imputare gli errori, dato che queste potrebbero non essere in accordo con gli scopi della ricerca.

Per tutte le procedure manuali richieste durante l'applicazione della fase di revisione automatica è necessario comportarsi, in fase di progettazione, formazione e controllo, come per altre operazioni quali la codifica dei quesiti aperti e la registrazione dei dati su supporto magnetico. Al fine di garantire una corretta applicazione delle procedure occorre inoltre predisporre verifiche periodiche sull'esecuzione delle operazioni e sulla completezza della documentazione richiesta.

L'applicazione di procedure automatizzate per la revisione, più che per altre fasi, richiede il ricorso intensivo agli elaboratori elettronici e l'impiego di professionalità di tipo tecnico e statistico di livello elevato. Pertanto, si deve tenere conto di alcune pratiche che è opportuno seguire per ottenere un impiego ottimale delle risorse informatiche. Diamo conto nel seguito, basandoci su Statistics Canada (1987), di alcune pratiche consigliate in tal caso. Dato che tali regole si possono applicare anche a tutti gli altri casi in cui si fa ricorso a procedure informatizzate, se ne dovrebbe tenere conto anche in tali situazioni.

I programmi utilizzati per effettuare le elaborazioni devono essere documentati con riferimento alla progettazione e alla validazione. Occorre inoltre predisporre i manuali operativi di ciascuna procedura in modo che siano descritti gli *Obiettivi della procedura* :

- istruzioni per l'esecuzione dei programmi. Occorre che in questa sede siano fornite le istruzioni per la definizione dei dati in input, soprattutto, qualora i programmi siano eseguiti da enti distribuiti sul territorio e i dati siano organizzati in maniera non standard;
- descrizione dei file utilizzati dai programmi e dei controlli sui dati, tali da accertare in ogni momento che si stanno utilizzando i dati appropriati;
- istruzioni sui file di output forniti dalle procedure. In questa sede occorre fornire notizie, oltre che sul formato degli output, sulla loro importanza nel contesto del singolo programma e della procedura complessiva e sulle figure professionali responsabili della loro produzione e archiviazione;
- istruzioni su come gestire i problemi operativi o di sistema facendo riferimento anche alla documentazione relativa alla progettazione e validazione dei sistemi.

Tutto il personale operativo dovrebbe essere formato sull'uso degli elaboratori in modo che siano in grado di eseguire le mansioni loro assegnate. La formazione deve riguardare inoltre tutti gli aspetti relativi alle operazioni manuali che sono direttamente associate all'uso dei programmi e degli elaboratori.

Codifica di quesiti aperti

Con il nome di codifica viene indicata l'attività di trasposizione di informazioni pervenute sotto forma di linguaggio libero in un insieme finito di codici rispondenti ad una classificazione

precostituita. Un esempio di codifica è dato dalla trasposizione delle informazioni riguardanti il settore di attività economica delle imprese, descritto in forma colloquiale dal rispondente, nei rispettivi codici secondo la classificazione ATECO [Istat, (1999)]. Un altro esempio è dato dalla codifica delle cause di morte secondo la classificazione ICD-9 [ONU, (1977)] riportate sui certificati di morte, ed utilizzate nell'ambito della corrispondente indagine amministrativa.
Il ricorso ai quesiti aperti è motivato da quelle situazioni in cui il rispondente non saprebbe collocare in modo corretto l'informazione secondo la classificazione richiesta, a causa della sua notevole complessità. Infatti le classificazioni ATECO e ICD-9 prevedono centinaia di codici differenti e soltanto esperti codificatori sono in grado di risalire dall'informazione al codice corrispondente. In questo caso, contrariamente a quanto si fa normalmente, si rinuncia a precostituire una griglia di opzioni fisse che il rispondente deve contrassegnare, preferendo al suo posto proporre una domanda aperta.
A causa della sua complessità l'operazione di codifica è da considerarsi critica e deve essere demandata a personale formato alla bisogna. E' appena il caso di osservare come un esperto di una singola classificazione non necessariamente lo è per una differente, magari riguardante soggetti diversi.
L'operazione di codifica avviene solitamente apponendo in un apposito spazio il codice corrispondente alla descrizione riportata per esteso. Dal punto di vista tecnico le modalità mediante le quali l'operazione di classificazione viene eseguita riguardano gli strumenti di ausilio alla ricerca dei codici. Infatti anche se la codifica avviene solitamente utilizzando liste su base cartacea, la tecnologia informatica rende possibile consultare tali liste di codici su un elaboratore elettronico sfruttando tutte le possibilità di ricerca offerte da tale strumento. In alcuni casi inoltre sono disponibili programmi per la classificazione automatica che permettono di ridurre o facilitare l'intervento degli operatori umani, sfruttando la potenza di calcolo degli elaboratori elettronici.
Occorre notare che, soprattutto per motivi di disponibilità di risorse, l'apposizione dei codici può essere demandata al personale degli enti periferici. La cosa dovrebbe comunque essere limitata il più possibile e comunque a codifiche di bassa complessità dal momento che, in particolare per i dati di fonte amministrativa l'attività, sicuramente in competizione con altre considerate prioritarie, sarebbe probabilmente affetta da numerosi errori. Tuttavia, in tutti quei casi in cui il materiale viene registrato a cura degli enti periferici, la codifica sul territorio può diventare ineludibile, rendendo di conseguenza necessaria una costante attività di controllo della qualità di tale operazione.
Per quanto riguarda la possibilità che errori siano introdotti in questa fase, oltre alla possibilità che il codificatore compia un errore di interpretazione, si possono citare gli errori di trascrizione e quelli provocati dall'inadeguatezza della classificazione stessa, ad esempio per sopravvenuta obsolescenza. Inoltre errori possono essere indotti da difficoltà insite nelle modalità di ricerca dei codici, come ad esempio per una classificazione su lista cartacea, solitamente meno gestibile di una informatizzata. In tutti questi casi, oltre ad una responsabilità dell'operatore negli errori, occorre tenere conto anche delle condizioni esterne, come l'ambiente di lavoro.

Alcune Raccomandazioni

Anche per la fase di codifica l'indirizzo fondamentale verte sulla progettazione di procedure congruenti con tutte le altre fasi dell'indagine, finalizzate alla produzione di dati di qualità sufficiente per conseguire gli scopi della ricerca. A tale scopo grande importanza riveste il sistema dei controlli di qualità nel cui ambito devono essere predisposti gli strumenti di ausilio alle operazioni, la formazione, la documentazione dell'attività e la valutazione dei livelli di errore nei dati.

In sede di progettazione è auspicabile la conduzione di studi e sperimentazioni o simulazioni di dati con le quali più modalità alternative per la conduzione dell'operazione di codifica siano messe a confronto per valutare l'ipotesi migliore in termini di rapporto costi benefici.

In ogni caso occorre utilizzare, quando possibile e coerentemente con gli scopi della ricerca, le classificazioni standard disponibili, procedendo eventualmente a disaggregazioni o ad aggregazioni di codici in modo da potersi comunque riportare ad una classificazione nota, assicurando così la confrontabilità dei risultati conseguiti con quelli desumibili da altre fonti. E' inoltre opportuno prevedere la figura dei codificatori esperti per risolvere i casi di difficile interpretazione e omogeneizzare il lavoro complessivo.

Le procedure devono essere predisposte avendo cura di prevedere anche i casi per i quali la codifica risulti troppo difficile per gli operatori e si debba ricorrere all'aiuto di codificatori esperti. Per classificazioni gerarchiche di particolare complessità può essere utile compiere uno smistamento preliminare dei dati secondo grandi classi (livelli di codifica più elevati) ed inviarli successivamente a codificatori specializzati.

Ad esempio per la classificazione dell'attività economica delle imprese è comune, nelle indagini Istat, suddividere i questionari secondo grandi classi di attività economica, ed inviarli a codificatori specializzati in settori differenti. Questa pratica, sebbene utile dal punto di vista delle economie di lavorazione, deve tuttavia essere rigorosamente monitorata dal punto di vista della qualità dei dati. Infatti, inviare agli stessi codificatori dati omogenei dal punto di vista dell'attività economica può provocare l'introduzione di distorsioni sistematiche in particolari classi di attività nel caso in cui i corrispondenti codificatori commettano errori. Per ovviare a questo problema si può prevedere l'impiego di due codificatori per ciascuna classe individuata e compiere controlli per identificare eventuali distorsioni.

La formazione dei codificatori deve essere particolarmente accurata e deve trattare in particolare gli aspetti specifici dell'utilizzo del materiale di ausilio alla codifica, avendo cura di predisporre esercitazioni pratiche e sessioni di verifica. E' importante inoltre predisporre carichi di lavoro attesi e livelli minimi desiderati rispetto alla qualità dell'operazione, comunicando tali obiettivi al personale durante la formazione. Si deve sottolineare che i livelli di qualità andrebbero definiti in relazione agli obiettivi dello studio, tenendo anche conto della possibilità di individuare e correggere gli errori nelle fasi successive della lavorazione. Se ad esempio si ha ragione di ritenere che gli errori di classificazione possano essere individuati con certezza confrontando i codici apposti dagli operatori con altre informazioni individuali, ad esempio per mezzo di un programma di correzione automatica, sarà possibile mantenere più basso lo standard di qualità richiesto nella fase di codifica, potendo facilmente operare una correzione nelle fasi successive.

Ogni qualvolta sia possibile è bene ricorrere a strumenti informatizzati per l'aiuto alla codifica. Questi strumenti possono costituire un semplice ausilio al reperimento dei codici da parte dell'operatore o, per una buona parte delle codifiche da effettuare, sostituirsi ad esso, lasciando all'operatore esperto solo i casi più complessi.

E' auspicabile l'istituzione dei revisori, attività che può essere svolta dai codificatori esperti, qualora siano stati previsti. Ai revisori dovrebbero essere demandate le attività di controllo della qualità dell'operazione e la loro documentazione. In generale, per ogni codificatore dovrebbe essere previsto un iniziale controllo esaustivo del materiale codificato che, sulla base dei livelli di errore riscontrati, potrebbe essere ridotto ad un controllo statistico della qualità. In alternativa, se le risorse disponibili lo obbligano, può essere adottata la strategia inversa, partendo da un controllo a campione del lavoro di ciascun codificatore, e passando ad una ispezione esaustiva su quegli operatori per i quali si siano riscontrati tassi di errore campionario eccedenti gli obiettivi prefissati, in modo da riportarli ai livelli accettabili. Seguendo una di queste due modalità di controllo è possibile contenere l'errore di classificazione nei livelli previsti. E' bene osservare come

queste modalità operative possano comportare, soprattutto all'inizio della loro applicazione, una dilatazione dei tempi di lavorazione. Nel caso tale dilatazione dei tempi sia ritenuta troppo elevata si può adottare una strategia differente, rinunciando alla correzione degli errori di codifica e spostando la valutazione dell'errore in momenti successivi alla fase operativa vera e propria. In questo caso si potrà pur sempre valutare i livelli di errore e predisporre strategie correttive, come una formazione del personale più accurata o migliori strumenti di ausilio, per il miglioramento futuro del processo.

Tecniche di controllo della qualità meno onerose possono prevedere la doppia codifica di un campione di questionari per ciascun codificatore e/o l'individuazione degli errori di codifica in sede di revisione automatica. In questo secondo caso però saranno individuati soltanto gli errori che danno luogo a valori non ammissibili o incongruenti, cioè solitamente i più grossolani. In ogni caso è opportuno che i controlli di qualità siano eseguiti in riferimento ai singoli operatori, avendo cura di predisporre codici identificativi tali che si possa risalire da ogni questionario al codificatore che lo ha lavorato.

I risultati dell'attività di controllo dovrebbero essere documentati in forma standard, riportando le percentuali di errore sostenute da ciascun codificatore sia sul complesso dei dati che per sottoclassi di codici. E' possibile prevedere, se si adotta una classificazione gerarchica, una misura di distanza che tenga conto dell'appartenenza del codice errato alla stessa classe gerarchica di quello corretto o ad una classe differente. La documentazione prodotta dovrebbe essere analizzata studiando sia i valori medi assunti dall'errore, sia la sua variabilità fra gli operatori. In questo modo, come è stato discusso nel paragrafo precedente, è possibile generare ipotesi sulle fonti che hanno agito nella generazione dell'errore, legando a fattori strutturali i livelli medi di errore e a fattori individuali la variabilità rispetto a tali valori medi.

In alcuni casi, come detto sopra, può essere considerata la modalità di codifica presso gli enti territoriali. Tale modalità di lavorazione, a causa della sua bassa qualità attesa, è da adottare soltanto qualora la classificazione non rivesta particolare importanza per lo studio in questione o se le risorse disponibili presso l'ente statistico non permettono assolutamente lo svolgimento in proprio della codifica. In questo secondo caso tuttavia, deve essere devoluta speciale attenzione all'attività di formazione e di controllo della qualità.

La formazione deve essere svolta tenendo conto delle risorse disponibili nelle differenti realtà territoriali e garantendo comunque la possibilità di una assistenza continua, per esempio dedicando personale interno ad una consulenza telefonica. Inoltre sarebbe bene predisporre un calendario di ispezioni nelle quali verificare le condizioni in cui viene condotto il lavoro e l'aderenza alle procedure pianificate. Particolare importanza è inoltre rappresentata dall'identificazione di un referente in ogni ente periferico a cui riferirsi nel caso si riscontrino problemi o cadute di qualità. Anche e soprattutto nel caso della codifica svolta presso gli enti periferici è opportuno prevedere un controllo a posteriori della qualità che può essere condotto da operatori interni con modalità simili a quelle illustrate sopra. Resta anche valido il principio generale di diffondere le misurazioni della qualità presso tutti i livelli coinvolti, a partire dagli operatori che lavorano negli enti periferici fino ad arrivare al responsabile della qualità.

Elaborazioni statistiche (da Statistics Canada, 1987)

Per elaborazioni statistiche si intende il processo di sommarizzazione ed interpretazione dei dati. Tale processo coinvolge uno studio più approfondito di quello richiesto dalla singola produzione di stime conclusive. L'elaborazione (o analisi) statistica è importante per la predisposizione di nuove indagini sulla base dei risultati di studi pilota o precedenti indagini, per la formulazione di obiettivi

realistici riguardanti la qualità, l'identificazione di problemi e di requisiti del processo di produzione.
Anche l'attività di validazione richiede analisi, come nel caso dell'interpretazione delle differenze tra i risultati dell'attività e i dati ad essi correlati. L'analisi può anche richiedere l'esplorazione di questioni sociali e/o economiche mediante l'esame di dati di fonti anche diverse.
Ai fini di garantire la qualità delle elaborazioni statistiche si elencano i seguenti suggerimenti.

Attività preliminari

1. Studio della documentazione disponibile a riguardo di definizioni, concetti, modalità di rilevazione, disegno campionario, qualità dei dati, ecc. ;
2. Studio della documentazione riguardante i file contenenti i dati. In tale documentazione è infatti sovente raccolta una grossa mole di informazioni che possono modificare in modo sostanziale le interpretazioni delle analisi statistiche condotte;
3. Contatti con il personale responsabile della pianificazione e della implementazione dell'indagine al fine di coprire tutti aspetti poco chiari alla luce della documentazione disponibile;
4. Studio delle procedure di editing imputation e valutazione sull'inclusione in analisi dei dati sottoposti a correzione automatica. Eliminazione di tutti i record non adatti all'elaborazione statistica e loro conservazione in un apposito file archivio.

Analisi dei dati

1. Conduzione di analisi preliminari semplici mediante statistiche descrittive quali indici di posizione delle distribuzioni e istogrammi. Conduzione di analisi esplorative per l'individuazione di assunzioni plausibili sui dati. Test di adattamento finalizzati a valutare l'appropriatezza di distribuzioni teoriche nell'adattamento ai dati. Uso di metodi di rappresentazione grafica;
2. Uso di metodi robusti per la stima dei parametri. Applicazione di tecniche diagnostiche della regressione. Valutare la bontà di adattamento del modello ai dati:
3. Considerare nell'analisi i disegni di campionamento complesso;
4. Applicazione di studi tipo cross-validation dei dati per analizzare se i risultati conseguiti con l'analisi possono essere considerati sufficientemente generalizzabili;
5. Ricorrere ad esperti nell'applicazione dei singoli metodi statistici utilizzati e condividere i risultati preliminari con lo staff di ricerca per eliminare la probabilità degli errori ed imprecisioni più comuni nelle interpretazioni delle analisi.

Dimensioni della qualità

Dal punto di vista della qualità, l'informazione statistica può utilmente essere considerata alla stregua di un qualsiasi bene o servizio in modo da potervi applicare i concetti sviluppati nel settore della qualità dei beni e servizi prodotti in ambito industriale o terziario. In tale contesto adottiamo la definizione di qualità proposta nelle norme ISO 8402-1984 per un bene o servizio: "Il possesso della totalità delle caratteristiche che portano al soddisfacimento delle esigenze, esplicite o implicite, dell'utente". Questa definizione, ai nostri fini certamente non operativa, evidenzia due punti molto importanti:

1. Il soggetto fruitore della qualità è l'utente al quale è rivolto il bene o il servizio;
2. La qualità del bene o servizio consiste nel possesso di determinate caratteristiche.

È inoltre opportuno introdurre un'ulteriore distinzione tra il bene o servizio prodotti e il processo di produzione che porta alla loro creazione. Questa distinzione ci serve per evidenziare che le caratteristiche di qualità di un prodotto possono essere in buona parte ottenute migliorando il

processo di produzione del bene o servizio in questione. È per questo che nel seguito si farà spesso menzione della qualità di processo e della qualità del prodotto, sempre con l'obiettivo del conseguimento della "soddisfazione dell'utente".
A partire da questi concetti generali possiamo passare a definire quali sono le dimensioni che caratterizzano la qualità nel caso in cui il bene (e servizio) in questione sia rappresentato dall'informazione statistica su un collettivo di interesse. Per introdurre tali concetti ci riferiremo alla documentazione Eurostat in materia di valutazione della qualità delle statistiche prodotte dai paesi membri della Comunità Europea:

1. Rilevanza;
2. Accuratezza;
3. Tempestività e puntualità;
4. Accessibilità e chiarezza (o trasparenza);
5. Confrontabilità;
6. Coerenza;
7. Completezza.

Non esplicitamente compresa tra le caratteristiche richieste da Eurostat, ma tuttavia parametro importante e frequentemente citato, si ritiene utile includere la caratteristica di tutela della riservatezza dei rispondenti.
A seguire vengono date le definizioni per le caratteristiche citate:

- Rilevanza: capacità dell'informazione di soddisfare le esigenze conoscitive degli utenti. Nell'accezione di utente si deve intendere anche i committenti preposti ad organi di governo centrali o locali. È appena il caso di precisare che la caratteristica di rilevanza è strettamente collegata con gli obiettivi di indagine considerati in fase di progettazione;
- Accuratezza: grado di corrispondenza fra la stima ottenuta dall'indagine e il vero (ma ignoto) valore della caratteristica in oggetto nella popolazione obiettivo. I motivi che possono causare delle cadute nell'accuratezza dell'informazione sono denominate fonti dell'errore mentre una sua misura viene fornita dall'errore totale;
- Tempestività e puntualità: intervallo di tempo intercorrente fra il momento della diffusione dell'informazione prodotta e l'epoca di riferimento della stessa. Tempi e costi di un processo di produzione sono strettamente in relazione fra loro;
- Accessibilità e chiarezza: nota anche col nome di "trasparenza", questa caratteristica corrisponde alla semplicità per l'utente di reperire, acquisire e comprendere l'informazione disponibile in relazione alle proprie finalità. Queste caratteristiche sono influenzate dal formato e dai mezzi di diffusione dell'informazione rilasciata nonché dalla disponibilità di meta-informazioni a suo corredo;
- Confrontabilità: possibilità di paragonare nel tempo e nello spazio le statistiche riguardanti il fenomeno di interesse. Il grado di confrontabilità è influenzato, oltre che dalle modificazioni concettuali che possono intervenire nel tempo e nello spazio, anche da cambiamenti intervenuti nelle definizioni e/o nelle caratteristiche operative adottate dal processo di produzione dell'informazione. È ovviamente sul controllo di queste ultime che occorre concentrarsi per aumentare al massimo la confrontabilità dell'informazione prodotta;
- Coerenza: per le statistiche derivanti da una singola fonte la coerenza corrisponde alla possibilità di combinare le inferenze semplici in induzioni più complesse. Qualora derivanti da fonti diverse, ed in particolare per informazioni prodotte con diversa periodicità, le statistiche possono essere considerate coerenti fintantoché basate su definizioni, classificazioni e standard

metodologici comuni. In tal caso le inferenze possibili all'utente saranno più facilmente interrelate o, perlomeno, non risulteranno in contrasto fra loro.

- Completezza: si tratta di una caratteristica trasversale ai singoli processi e consiste nella capacità di questi integrarsi per fornire un quadro informativo soddisfacente del dominio di interesse. A loro volta i domini per i quali sono rese disponibili statistiche dovrebbero riflettere le necessità e le priorità espresse dagli utenti del Sistema Statistico Nazionale (SISTAN);
- Tutela della riservatezza: corrisponde alla garanzia dell'anonimato per ciascuno dei soggetti (individui, famiglie, imprese, ...) che hanno fornito le informazioni utili alla conduzione dell'indagine. La mancata garanzia di questa caratteristica, sebbene non immediatamente collegata alla qualità dell'informazione, si ripercuote negativamente sull'immagine di credibilità dell'ente statistico e, diminuendo la fiducia dei nell'ente da parte dei rispondenti, pregiudica la sua possibilità di rilevare dati affidabili.

Le fonti dell'errore

L'errore totale, misura dell'accuratezza, può essere generato da numerose cause che chiameremo nel seguito fonti dell'errore. Una prima distinzione viene fatta tra l'errore campionario e l'errore non campionario. Con il primo si indica l'influenza indotta dall'operazione di campionamento sulla varianza e sulla distorsione delle stime. Va da se che le indagini totali, come il censimento ad esempio, non sono affette da questo tipo di errore.

Il secondo tipo di errore è provocato da tutte le possibili imprecisioni e inaccuratezze commesse o subite durante l'indagine. A questa seconda classe di errori appartengono ad esempio i rifiuti a rispondere o le risposte errate da parte delle unità statistiche interpellate. Allo stesso modo gli errori generati durante le fasi operative dell'indagine successive alla rilevazione dei dati, come gli errori di registrazione su supporto magnetico, gli errori di codifica o gli errori commessi in fase di revisione del materiale, appartengono a questa categoria.

Gli errori campionari e non campionari concorrono nel determinare l'entità dell'errore totale. Sia la distorsione che la varianza componenti l'errore totale possono essere scomposte additivamente in relazione al peso dovuto a ciascuna fonte d'errore. La stima delle componenti dell'errore totale attribuibile a ciascuna singola fonte d'errore prende il nome di profilo dell'errore e rende possibile l'attività di validazione dell'informazione prodotta. Nel seguito viene proposta una ulteriore classificazione per gli errori non campionari comunemente accettata in ambito scientifico:

- Errori campionari
- Errori non campionari
  - Copertura
  - Mancate risposte
    - Totali
    - Parziali
  - Misurazione

Errori di copertura: errori dovuti ad imperfezioni nella corrispondenza fra la lista utilizzata per selezionare e contattare le unità statistiche (archivi di base) e la popolazione oggetto di indagine. Gli errori di copertura possibili sono di due tipi: l'inclusione nell'indagine di unità non appartenenti alla popolazione oggetto di interesse (sovracopertura); l'impossibilità di selezionare o coinvolgere nell'indagine unità appartenenti alla popolazione oggetto (sottocopertura). Gli errori di sovracopertura sono meno gravi in quanto possono essere scoperti in fase di indagine

predisponendo appositi quesiti per le unità statistiche contattate. Più gravi sono gli errori di sottocopertura i quali non possono essere scoperti se non svolgendo apposite indagini di controllo.

Errori di mancata risposta: errori dovuti al rifiuto o all'impossibilità a rispondere da parte delle unità statistiche contattate. Sono suddivisi in totali, se l'unità non partecipa affatto all'indagine, e parziali, quando l'unità fornisce solo alcune particolari risposte.

Errori di misurazione: errori costituiti dalla differenza fra il vero valore della caratteristica da misurare su una data unità statistica e il valore effettivamente osservato dall'indagine. Tali differenze possono essere introdotte dal rispondente stesso (per dimenticanza, imprecisione o dolo) oppure dallo svolgimento delle fasi di elaborazione successive alla raccolta del dato. Esempi di questo secondo caso sono tutti gli errori introdotti dalle operazioni di registrazione su supporto informatico o di codifica dei quesiti aperti.

Bibliografia

BAILAR, B., A. (1989); Information needs, survey and measurement errors, in Panel Survey, Kasprzyk,
Duncan, Kalton, Singh (eds.), Wiley and Sons, NY, pp. 1-24
BARCAROLI, G., D'AURIZIO, L., LUZI, O. MANZARI, A., PALLARA, S. (1999); Metodi e software per il controllo e la produzione dei dati, Documenti ISTAT, n. 1/1999
BELLINZONA, E. (1997); Excel per la qualità, le carte di controllo, F. Angeli
BRACKSTONE, G., J. (1987); Statistical uses of administrative data: issues and challenges; Proceedings of
Statistics Canada symposium of administrative data, November 1987, pp. 5-16
BRADBURN, N., M., SUDMAN, S. (1991); the current status of questionnaire design, in Measurement
error in surveys, Biemer, Groves, Lyberg, Mattiowetz, Sudman (Eds.), John Wiley and Sons, NY, pp.29-40
BRANCATO, G., FANFONI, L., FORTINI, M., SCANU, M., SIGNORE, M. (2000); Il sistema SIDI: uno strumento generalizzato per il controllo di qualità delle indagini Istat, in corso di pubblicazione su "Scritti di statistica economica".
COCHRAN, W., G. (1977); Sampling techniques, 3rd ed., Wiley, NY
DE ANGELIS, R., MACCHIA, S. (1999); Qualità e praticabilità della codifica automatica di dati censuari:
risultati della sperimentazione sulle variabili Professione, Attività economica e Titolo di studi. Atti del
convegno SIS "Verso i censimenti del 2000", 7-9 giugno 2000, Udine
DENMARK STATISTIK (1995); Statistics on persons in Denmark: a register based statistical system (English ed.), Eurostat - Office for official publications of the European Communities, Luxembourg
DUNCAN, G., J., KALTON, G. (1987); Issues of design and analysis of surveys across time, International
Statistical Review, 55, pp. 97-117
FABBRIS, L., (1989); L'indagine campionaria, Metodi, disegni e tecniche di campionamento, La Nuova
Italia Scientifica, Roma
FORSMAN, G., SHREINER, I. (1991); The design and Analysis of reinterview: an overview, , in
Measurement error in surveys, Biemer, Groves, Lyberg, Mattiowetz, Sudman (Eds.), John Wiley and
Sons, NY, pp. 279-301
FORTINI, M. (1998); Gli indicatori standard di qualità nel sistema infomativo di documentazione delle
indagini, Contributi ISTAT, n. 7/1998
FOWLER, F., J. (1988); Survey research methods, vol. 1, SAGE Publication, Applied social research methods, Beverly Hills
GROVES, R., M. (1989); Survey errors and survey costs, Wiley and Sons, NY
HYMAN, H., H., SHEATSLEY, P., B. (1950); The current status o American public opinion, in J.C. Payne
(Ed.), The teaching of contemporary affairs, National Council of Social Studies
ISTAT (1989); Manuali di tecniche di indagine, voll. 1-6, Istat, collana metodi e norme, Roma
ISTAT (1999) ATECO '91 a cinque cifre, Metodi e norme, serie C, n. 11, 1999
KASPRZYK, K. D., DUNKAN, G., J., KALTON, G., SING, M. P., (1989); Panel Surveys, John Wiley and

Sons, NY
LATOUCHE, M., BERTHELOT, J., M. (1992); Use of score function to rpioritize and limit recontacts in Editing Business Surveys, JOS, vol. 8, n.3 Part II.
LIBERG, L., KASPRZYK, D. (1991); Data collection methods and measurement error: an overview, in Measurement error in surveys, Biemer, Groves, Lyberg, Mattiowetz, Sudman (eds.), John Wiley and
Sons, NY, pp. 237-257
LUZI, O. (1998); L'editing selettivo come strumento per la razionalizzazione del processo di editing: un
primo studio su occupazione, retribuzioni e orari di lavoro nelle grandi imprese; Quaderni di ricerca;
ISTAT, vol. 3/1998 p.143
MACCHIA, S., D'ORAZIO, M. (2000); Impatto delle diverse tecniche di registrazione dei dati sulla codifica
automatica ed analisi di qualità rispetto alla codifica manuale, Atti del convegno della SIS, 26-28 aprile,
Firenze.
MARBACH, G. (1975); Sull'uso di quesiti che tutelano la completezza dell'informazione, Metron, vol.
XXXIII, n. 3-4
ONU (1977); International classification of diseases (ICD). Manual of the international statistical classification of disease, injuries and causes of death. 9
th
revision, vol.1 Geneva, Switzerland, ONU
RICCINI, E., SILVESTRI, F., BARCAROLI, G., CECCARELLI, C., LUZI, O., MANZARI, A. (1995); La metodologia di editing e imputazione per variabili qualitative implementata in SCIA, Documento interno ISTAT, Dicembre 1995
SCHUMAN, H. PRESSER, S. (1981); Questions and answers in attitude surveys, Academic press, NY
SIS (1990); Contributi della statistica alla progettazione di basi dati amministrativi, Riunione satellite della
XXXV riunione scientifica della SIS, Padova, 18 aprile 1990.
STATISTICS CANADA (1987); Statistics Canada quality guidelines, 2nd ed., Minister of Supply and Services Canada, Ottawa
STATISTICS CANADA (1998); Statistics Canada quality guidelines, 3nd ed., Minister of Industry, Ottawa
TAGUCHI, G. (1995); Introduzione alle tecniche per la qualità: progettare qualità nei prodotti e nei processi,
De Agostini, 1995