Commissione scientifica della S.I.S. su:
"La qualità dei dati statistici"

Glossario dei principali termini su:
"La qualità dei dati statistici"

(*a cura di A. Giommi*)

INDICE DELLE VOCI

PREFAZIONE

L'obiettivo iniziale della Commissione era quello di sviluppare una terminologia standardizzata delle definizioni dei vari concetti ed aspetti della qualità dei dati statistici. Tuttavia, sin dalle prime discussioni in seno alla Commissione emersero subito importanti problemi semantici: nella letteratura corrente lo stesso termine è a volte usato con differenti significati e lo stesso concetto è a volte specificato con differenti termini o definizioni, creando problemi di comunicazione anche tra gli statistici. Per questo motivo, nel predisporre le definizioni da includere nel glossario si potevano seguire due vie: (i) riportare per ogni concetto le definizioni usate nella letteratura, mettendo in evidenza i problemi semantici; (ii) individuare per ogni concetto la definizione più opportuna, da utilizzare in modo coerente nella formulazione di concetti tra loro collegati.

La maggioranza della Commissione ha ritenuto opportuno seguire la seconda alternativa, proprio per cercare di pervenire ad una proposta di standardizzazione delle definizioni dei vari concetti.

L'analisi delle definizioni date in letteratura sull'argomento ha consentito di individuare i termini da inserire nel glossario. Non si tratta di una lista esaustiva, ma soltanto dei principali termini che sono collegati alla qualità dei dati statistici. I termini descritti nel glossario sono scritti in grassetto anche per facilitare collegamenti e rinvii tra le voci.

Il lavoro è stato molto complesso e la Commissione è ben conscia del fatto che le varie definizioni riportate possono essere non completamente soddisfacenti e, in alcuni casi, non condivisibili dal punto di vista linguistico. Lo si considera comunque un primo passo nello sviluppo di un più sistematico approccio alla standardizzazione delle definizioni.

La Commissione si augura, pertanto, che puntuali osservazioni e suggerimenti dei colleghi consentano di migliorare le definizioni e di renderle accettabili ed utilizzabili da parte di tutta la Comunità degli Statistici italiani.

Accuratezza

Il termine accuratezza, nell'accezione più generale, esprime la vicinanza di un valore rilevato al corrispondente **valore vero**. Anche riferito, come spesso avviene, ad una **stima**, il termine conserva lo stesso significato, denotando la ridotta dimensione dell'errore statistico globale (v. **errore globale di stima**). In questo senso l'accuratezza può essere espressa quantitativamente, almeno sul piano teorico, dal reciproco dell'**errore quadratico medio.** Maggiore è l'errore quadratico medio, minore l'accuratezza della stima e vicever-sa. In questa accezione il termine coincide con **attendibilità**.

Adeguatezza

E' la capacità del dato o del sistema di informazione statistico di soddisfare i bisogni conoscitivi dell'utente. Qualità estrinseca del dato precisabile facendo riferimento all'insieme dei criteri che consentono di valutare la soddisfazione dell'utente. Tra i principali criteri vi sono la **pertinenza**, la **tempestività**, la **trasparenza.**

Affidabilità

Nel linguaggio tecnico è affidabile l'impianto, l'apparecchio, ecc., che dà garanzia di buon funzionamento. Nell'indagine statistica, per analogia, il termine non si riferisce al dato ma alla fonte, allo strumento, al metodo, alla procedura, ecc. E' quindi affidabile una procedura dalla quale si ottengono dati di qualità costante o poco variabile in ripetute applicazioni della stessa sotto identiche condizioni.

Nella letteratura specializzata il termine affidabile è utilizzato anche per denotare una stima il cui errore globale (v. **errore globale di stima**) non supera un prestabilito livello.

Aggregazione di dati

L'aggregazione è una qualsiasi funzione dei **dati elementari** registrati in distinti **microdati** utilizzata per riassumere le informazioni in essi contenute, talora anche per renderli pubblicabili evitando il rischio di **riferimento individuale**.

Ammissibilità

La modalità registrata che afferisce alla singola unità statistica è detta ammissibile se non contraddice nessuna delle **regole di compatibilità**.

Arrotondamento casuale

E' l'arrotondamento a caso ad una cifra vicina, terminante per zero o per cinque, delle frequenze basse in una distribuzione di frequenza (v. **segreto statistico attivo**).

Attendibilità

E' espressione della qualità dei dati rilevati con procedure statistiche affidabili. Attendibile, sia in relazione al dato rilevato che alla stima, è il termine comunemente utilizzato per esprimere un livello di qualità valutabile statisticamente.

Campionamento casuale (v. **campionamento probabilistico**)

Campionamento non probabilistico

Sono non probabilistici o, come si dice anche, rappresentativi, i campioni formati senza seguire un criterio probabilistico. Tra i più comuni campioni non probabilistici si individuano i **campioni a scelta ragionata** e i **campioni per quote**. Rispetto a quelli probabilistici, i campioni non probabilistici presentano il duplice svantaggio di rendere incontrollabile il rischio di introdurre distorsioni nel processo di formazione dei dati e di non permettere la valutazione dell'**errore medio di stima**.

Campionamento probabilistico

Si denomina probabilistico o casuale o statistico, il campionamento (v. **campione**) nel quale, ad ogni unità della popolazione, è attribuita una probabilità positiva e nota di fare parte del campione. Dal punto di vista operativo, tale procedura comporta l'utilizzazione in modo appropriato delle tecniche per la selezione casuale del campione. L'applicazione di queste tecniche è finalizzata all'ottenimento dell'indipendenza da qualsiasi variabile interna o esterna all'indagine, compreso, in particolare, l'orientamento di colui che sta predisponendo il campione.

Campione

Data una **popolazione statistica** composta da *N* unità (v. **unità statistica**), con *N* qualsiasi (anche infinito), si denomina campione l'insieme delle *n* unità selezionate fra le *N* che compongono la popolazione.
Un campione può essere formato in base a logiche probabilistiche, oppure in base a criteri soggettivi di rappresentatività Le tecniche campionarie sono utilizzate per stimare statistiche della popolazione, ad esempio: medie, totali, rapporti fra variabili. Le stime ottenute in base a campioni di popolazione sono soggette all'**errore campionario**.

Campione a scelta ragionata

E' a scelta ragionata il campione costruito in modo da assomigliare, per alcune caratteristiche fondamentali, (per esempio, in una indagine su popolazioni umane, per il sesso, l'età, la professione, ecc.) alla **popolazione statistica** cui appartiene. Le informazioni sulla struttura della popolazione sono note a priori.
Tra i tanti metodi di formazione di un campione a scelta ragionata, si annovera il campione di unità tipiche. Questo è formato da unità tratte da certi strati o sottocampioni relativamente omogenei per alcune caratteristiche fondamentali di classificazione. La tecnica si basa sulla logica che le variabili afferenti a una unità statistica sono interdipendenti e che, se una unità è vicina alla media della popolazione per una variabile, in generale differirà di poco dalla media della popolazione anche per diverse altre. Si tratta di un metodo piuttosto arbitrario, da impiegare soprattutto negli studi comparativi di situazioni tipiche non altrimenti rilevabili.

Campione autoponderante

E' autoponderante il campione le cui unità hanno uguale probabilità di appartenervi. L'autoponderazione del campione permette di effettuare le stime corrette operando nel modo più

semplice, ossia come se la rilevazione fosse esaustiva. Il campione può non essere autoponderante: (a) quando le unità sono state selezionate con probabilità diseguali; (b) quando la **copertura** di un eventuale campione autoponderante è incompleta.
Nel caso di mancata autoponderazione, per ottenere stime corrette si dovrà assegnare ad ogni unità statistica un peso variabile e proporzionale all'inverso della probabilità di appartenere al campione, in modo da far pesare di più, nel campione, le unità con probabilità più basse.

Campione per quote

E' per quote il campione per la cui formazione viene fissata a priori la dimensione totale del campione e il numero di unità da assegnare ad ogni intervistatore. La scelta delle unità campionarie è lasciata all'iniziativa degli intervistatori sotto la condizione che rispettino le quote di popolazione che presentano le caratteristiche prefissate.
Le quote vengono determinate in base alla conoscenza della distribuzione marginale o congiunta di alcuni caratteri legati alle variabili sotto studio nella popolazione o nelle sottopopolazioni in cui la stessa è stata preventivamente suddivisa.
Per introdurre una forma di controllo sulla selezione del campione, si può fare in modo che gli intervistatori seguano particolari percorsi, oppure che intervistino il tipo di persone designate nel luogo in cui svolgono la loro attività.

Caso statistico

E' il risultato di un'operazione mediante la quale si realizza una visione schematica della realtà, orientata da obiettivi specifici di conoscenza ed operativi. Ogni entità potenzialmente destinata a costituire una unità statistica è considerabile nell'insieme dei caratteri che la contraddistinguono: il caso statistico nasce da una visione semplificata, la quale isola i caratteri interessanti per il particolare obiettivo, e quindi meritevoli di attenzione, da altri, ritenuti irrilevanti e quindi trascurati. La formazione del caso si completa con una specificazione, per ciascun carattere, dell'elenco esaustivo di modalità mutualmente escludentisi che possono presentarsi nell'unità statistica e sollecitare attenzione mirata. La formazione del caso è un momento creativo, una astrazione che precede la rilevazione di cui costituisce l'essenziale punto di partenza. Un modello di rilevazione ne costituisce la visibile espressione linguistica.

Codificazione

E' l'operazione mediante la quale vengono trasformati in numeri, o in sequenze alfanumeriche, le informazioni espresse in forma verbale contenute in un questionario. L'operazione di codificazione, quando necessaria, precede quella di **registrazione** delle informazioni del questionario sul supporto informatico.

Coefficiente di correlazione intraclasse

E' l'indice che esprime il grado di omogeneità tra unità appartenenti ad una de-

terminata classe. Nell'ambito delle misure dell'errore extracampionario, indica il grado di omogeneità tra dati relativi ad unità rilevate da uno stesso rilevatore (coefficiente di correlazione intra-intervistatore) o tra codici apposti da uno stesso codificatore (coefficiente di correlazione intra-codificatore)
Il coefficiente di correlazione intra-intervistatore è tanto più grande quanto più grandi sono le **distorsioni dei rilevatori**.

Coerenza

In senso lato, vi è coerenza tra dati statistici quando tra di essi non esiste contraddittorietà o incompatibilità. Nell'indagine statistica si parla di controllo di coerenza, sia in rapporto al dato rlativo alla singola unità, sia in rapporto ai risultati complessivi. Coerenza tra dati non implica assenza di errori nei dati.

Cold deck (v. **metodi di imputazione**)

Collegamento esatto

Il collegamento (in inglese *linkage)* è l'operazione mediante la quale si abbinano dati relativi ad un'unica unità statistica ma rilevati in occasioni diverse (due o più indagini, un'indagine e una registrazione amministrativo-contabile, ecc.). L'operazione è possibile a condizione che, nelle varie occasioni, siano stati rilevati caratteri identificatori dell'unità.
Il collegamento consta di due fasi:
a) la ricerca di *records* potenzialmente abbinabili (cioè riferiti alla stessa unità) sulla scorta del confronto dei caratteri di identificazione;
b) la decisione di procedere o meno alla loro riunificazione.
La fase decisionale (b) è conseguente alla eventualità che si presentino discrepanze nel confronto delle modalità dei caratteri di identificazione a causa di errori che possono essere intervenuti in una qualsiasi fase del processo di formazione del dato. Per questo l'operazione di collegamento esatto può dar luogo a due tipi di errore:
i) l'abbinamento di *record* relativi ad unità statistiche simili ma diverse (v. **falso positivo**);
ii) il mancato abbinamento di *record* relativi alla stesa unità statistica (v. **falso negativo**).

Collegamento statistico

Operazione mediante la quale si abbinano con metodi probabilistici, attribuendoli ad un'unica unità statistica, dati relativi ad unità (normalmente due) statistiche provenienti da fonti (indagini) diverse, che si presentano omogenei per le modalità di un insieme di caratteri, generalmente non sufficienti ad identificare l'unità stessa.

Compatibilità (v. **regole di compatibilità**)

Compenetrazione

Si denomina compenetrazione dei subcampioni, o compenetrazione delle assegnazioni degli intervistatori, lo schema di assegnazione di un sottoinsieme casuale del campione selezionato per l'indagine ad ogni intervistatore. Sia, cioè, n la numerosità di un campione selezionato da una popolazione di N unità e siano k gli

intervistatori selezionati per rilevare i dati nell'indagine. La compenetrazione consiste nel suddividere a caso le unità campionarie n *k* subcampioni e nell'associare, in base a un criterio casuale, ogni subcampione a un intervistatore.
Siccome ogni assegnazione è un campione casuale dell'intera popolazione, le differenze tra le medie di differenti assegnazioni possono essere attribuite all'effetto distorsivo degli intervistatori (v. **varianza del rilevatore**). Il confronto fra la variabilità interna ai subcampioni e quella tra le medie degli stessi permette di misurare con approssimazione varie componenti della **varianza extracampionaria** delle stime. Se impiegato anche nella reintervista con gli stessi rilevatori che hanno svolto l'indagine principale, lo schema di compenetrazione delle assegnazioni (che diventa allora una reintervista con doppia compenetrazione delle assegnazioni degli intervistatori), permette di separare con approssimazione la varianza dovuta al campionamento (v. **varianza campionaria**) da quella imputabile agli **errori di rilevazione** e di identificare con maggiore precisione varie componenti d'errore extracampionarie.

Completezza

Il termine, generalmente associato all'informazione raccolta per ogni unità statistica che partecipa all'indagine, indica l'assenza di dati mancanti nel questionario o scheda di rilevazione ecc. Considerando il complesso dei dati raccolti nell'indagine, è detto taso di completezza la quota o percentuale di unità per le quali sia stata raccolta un'informazione completa, sul totale delle unità che hanno partecipato all'indagine.

Confidenzialità (v. **segreto statistico**)

Controllo esterno, indiretto

Viene effettuato in molteplici forme confrontando quanto raccolto in una specifica indagine con quanto risulta da altra fonte, indipendentemente dalla prima, e con riferimento sia a problemi di **copertura** che di **accuratezza** dei risultati. Lo si applica ai **dati elementari** mediante una **indagine successiva** (campionaria) condotta sullo stesso oggetto e in pari condizioni, oppure collegando dati elementari provenienti per la singola unità da un distinto canale di informazione. Lo si applica anche ai **dati aggregati**, secondo la stessa logica. Ad esempio, la stima del consumo nazionale di una certa categoria di prodotti derivata dalle risultanze di indagini presso le famiglie può venir messa a confronto con la disponibilità calcolata per la stessa categoria, nel dato intervallo di tempo, partendo dalla produzione e dal saldo con l'estero.
Talvolta il confronto è fatto con i risultati di un'analisi statistica e dell'applicazione di modelli. Il modello della popolazione stabile può, in certe condizioni, mettere in luce incoerenze tra parametri demografici stimati per una popolazione storica o per una di un Paese con statitiche lacunose o difettose.

Controllo interno, diretto

Si effettua sia sui **microdati** sia sui **macrodati** ottenuti dall'indagine. Nel primo caso rientrano i controlli di

ammissibilità e **compatibilità** che si effettuano, dopo la rilevazione, sui **dati elementari**. Nel secondo si operano analoghi controlli sui **dati aggregati** per evidenziare eventuali distorsioni e carenze. Ad esempio, l'analisi del rapporto di mascolinità dei figli delle decedute può fare emergere l'inattendibilità di quella informazione.

Controllo statistico

L'espressione richiama il controllo statistico della qualità attuato nella produzione in serie. In questa, mantenendo costanti i fattori co-agenti in una linea di produzione, ci si assicura che il complesso causale messo in atto garantisca una relativa uniformità - entro i limiti accettati e prefissati - dei "pezzi" che ne escono: relativa, nel senso che al complesso mantenuto stabile nel tempo si sovrappone un insieme di fattori accidentali sì che le caratteristiche delle unità prodotte possono differire tra loro per ragioni casuali, con una dispersione di risultati che non disturba, in quanto compatibile con i limiti di tolleranza che ci si è proposti di rispettare.
In una indagine statistica, il prodotto è un *unicum* e la definizione di limiti di tolleranza rimane ambigua: infatti, in ogni caso, si mira a rappresentare al meglio, nel dato, la realtà effettuale cui esso si riferisce, e l'errore totale può essere misurato - quando la cosa è fattibile - solo a posteriori. Gli scostamenti osservati in esperienze ripetute non aiutano tanto a fissare gamme di risultati futuri per una realtà in continuo divenire quanto a immaginare che i valori osservati in una indagine statistica siano una tra le tante manifestazioni empiriche possibili di una distribuzione generata da leggi probabilistiche centrata su un valore che è la media dei valori osservabili.. Tenere sotto controllo statistico l'intero insieme di operazioni in cui un'indagine si articola significa mettere in atto provvedimenti atti a garantire in modo efficiente che ogni persona o istituzione coinvolta in una qualsiasi sua fase rispetti sempre e uniformemente regole di comportamento e modalità di azioni quali sono state predisposte nel piano fissato.

Copertura

Copertura è un termine impiegato nelle indagini statistiche per denotare:
(a) il rapporto esistente fra il numero di unità che hanno collaborato all'indagine e quelle designate a parteciparvi (v.**mancata rilevazione**). Il concetto è diverso da quello di **frazione di campionamento**;
(b) la frazione sottoposta a indagine ma in questo senso è preferibile la locuzione frazione di campionamento.

Correttezza

Nella teoria della stima è la proprietà per la quale il valore atteso dello stimatore coincide con quello del parametro stimato.
In riferimento al dato rilevato sulla singola unità statistica esprime una misurazione non affetta da errore; in altri termini il dato è corretto se il valore rilevato coincide col **valore vero**.

Credibilità (v. **attendibilità**)

Dato aggregato, macrodato

E' l'aggregato che si ottiene dalla sintesi di **dati elementari**. In realtà, non vi è

distinzione precisa fra macro e microdato, esistendo piuttosto fra i due un *continuum*. Il consumo familiare procapite è un insieme di consumi individuali distinti e di altri indivisi. A sua volta, può essere preso come dato di base per una distribuzione di famiglie secondo il livello di consumo pro-capite. La dimensione di un comune in termini di numero di abitanti è un aggregato costruibile attraverso informazioni individuali sulla residenza o presenza. E' anche un dato elementare per una distribuzione di comuni secondo la loro popolazione.

Dato definitivo

E' quello diffuso dal produttore di statistiche quando egli ha completato il relativo **piano di rilevazione** su tutti gli elementi in gioco e i cui risultati sono stati sottoposti ad un processo di messa a punto.

Dato elementare

E' elementare il dato che si riferisce al singolo carattere osservato su ciascuna **unità di analisi.** Se l'unità di analisi è un aggregato di unità statistiche, come ad esempio una famiglia, il dato relativo al reddito familiare totale, o procapite, è da considerarsi elementare anche se risulta da una funzione di dati relativi ad unità di livello inferiore. Ciò evidenzia l'impossibilità di distinguere il dato elementare dal **dato aggregato** se non in rapporto all'unità statistica di riferimento.

Dato provvisorio, preliminare

E' quello che il produttore di statistiche rende disponibile in via preliminare, per ragioni di **tempestività** nell'informazione, anche se ottenuto in forma approssimata o incompleta e non ancora sottoposto ad un processo di **revisione**.

Disegno di campionamento (v. **piano di campionamento**)

Disegno di rilevazione (v. **piano di rilevazione**)

Distorsione (v. **errore sistematico**)

Distorsione del rilevatore

In una indagine svolta mediante rilevatori, si denomina distorsione del rilevatore il risultato del condizionamento esercitato dal rilevatore sui dati ottenuti. La distorsione può derivare dall'incapacità del rilevatore di stabilire una relazione adeguata con il **rispondente**, dalla incapacità di porre correttamente le domande e di ottenere le giuste risposte, da errori commessi nella registrazione delle risposte.
La distorsione è quantificabile con lo scarto tra il valore medio ottenuto dal rilevatore e il valor medio atteso. In un **disegno di rilevazione** nel quale sia stata effettuata la **compenetrazione** dei subcampioni, la distorsione di un rilevatore è misurata dallo scarto tra la media del subcampione a lui assegnato e la media campionaria globale

Efficienza

Il termine, proprio della teoria della stima, è utilizzato normalmente col si-

gnificato di **precisione** relativa di uno **stimatore**. Dati due stimatori *T* e *T'* appartenenti ad una prestabilita classe di stimatori, si dice che *T* è più efficiente di *T'* se *V(T)<V(T').*
Se, inoltre, in una classe di stimatori ne esiste uno che ha varianza minima, questo può essere utilmente assunto come termine di confronto per valutare l'efficienza degli altri, mediante le operazioni di differenza e/o rapporto tra le varianze.

Errore accidentale (v. **errore casuale**)

Errore campionario, di campionamento

E' la differenza tra la **stima** e il corrispondente valore che si sarebbe ottenuto esaminando la totalità delle **unità statistiche** della popolazione. Mediamente, l'errore diminuisce in valore all'aumentare della numerosità campionaria, ed è nullo quando il campione è composto dalla totalità delle unità che compongono la popolazione.
L'errore dovuto al campionamento va tenuto distinto da quello non campionario (v. **errore extracampionario**), che si manifesta anche se la rilevazione è esaustiva (v. **indagine esaustiva**).

Errore casuale

E' casuale l'**errore statistico** che si manifesta in modo differente (in dimensione e in segno) nelle unità esaminate. Si distingue dall'**errore sistematico** perché si compensa in media sulle osservazioni effettuate o su quelle teoricamente effettuabili.

Un errore casuale è concepibile come la combinazione di innumerevoli fattori di errore, di peso e segno diversi, e tra loro indipendenti. La casualità implicita nel risultato della combinazione implica la normalità della sua distribuzione.

Errore dell'intervistatore (v. **errore del rilevatore**)

Errore del rilevatore

Errore che, nei valori rilevati, è imputabile all'azione dei rilevatori.

Errore di imputazione

E' l'errore commesso sostituendo, mediante i **programmi di imposizione automatica,** un valore diverso da quello vero (v. **valore vero**)

Errore di rilevazione

Quella parte dell'errore non campionario dovuta a difformità tra il **valore vero** e il **valore rilevato** presso le unità statistiche osservate in un'indagine.

Errore di risposta

Errore non campionario dovuto a difformità tra la risposta data e il valore vero (v. **errore di rilevazione**).

Errore extracampionario

E' l'**errore statistico** risultante dalla somma di tutti gli errori che possono essere commessi in una qualsiasi fase del processo di indagine, nonché delle loro possibili interazioni, con l'esclusione dell'**errore campionario.**

Errore globale di stima

Errore delle stime dovuto all'effetto dell'incompletezza della rilevazione (v. **errore campionario**), degli **errori di rilevazione** e di ogni altro **errore extracampionario,** rilevabile e non, e della eventuale interazione tra l'errore campionario e gli errori extracampionari.
Sulla distribuzione delle stime si manifesta come una varianza per la parte degli errori variabili (v. **errore casuale**) e come una **distorsione** per gli **errori sistematici**. Si misura con l'**errore quadratico medio**.

Errore medio di stima

E' l'indice di variabilità di uno **stimatore**, formalmente espresso dalla radice quadrata della sua varianza campionaria. L'errore medio di stima è tanto minore quanto maggiore è la numerosità del campione, annullandosi quando l'indagine investe la totalità delle unità della popolazione.
L'errore medio si può stimare solo quando il campione è stato selezionato con criteri probabilistici. Sulla base della stima e del suo errore medio, è possibile fare affermazioni probabilistiche definendo, fra l'altro, l'intervallo fiduciario della stima stessa.

Errore quadratico medio

E' definito dal **valore atteso** dello scarto quadratico tra **stimatore** e vero valore della statistica che si desidera stimare nella popolazione. Se lo stimatore è corretto (v. **correttezza**), l'errore quadratico medio si identifica con la varianza dello stimatore, altrimenti è maggiore di tale varianza per una quantità pari al quadrato della **distorsione** dello stesso stimatore.

Errore sistematico, distorsione

Sistematica è la deviazione tra un valore empirico e il suo **valore atteso** che si manifesta sempre nella stessa direzione e misura. Si distingue dall'**errore casuale** che varia tra prove campionarie e può, quindi,bilanciarsi in media.
Per introdurre, in una misura di variabilità di una stima (v. **errore medio di stima**), la componente dovuta alla **distorsione**, si ricorre all'**errore quadratico medio**, dato dall'unione della **varianza campionaria**, di quella dovuta agli errori variabili di rilevazione (v. **varianza extracampionaria**), di eventuale covarianza tra questi e quelli e del quadrato della distorsione.

Errore statistico

E' la discrepanza tra **valore vero** e valore disponibile dall'indagine statistica. La definizione, di carattere generale, ha contenuto diverso a seconda che si riferisca alla singola **unità statistica** oppure ad una statistica di sintesi dei dati rilevati mediante indagine campionaria. Nel primo caso, infatti, la discrepanza è dovuta al complesso degli **errori extracampionari** che si commettono sull'unità statistica. Nel secondo, agli errori extracampionari si aggiunge quello campionario a causa del quale, anche nell'ipotesi puramente teorica che i valori di ogni unità fossero rilevati con esattezza, la statistica campionaria differirebbe comunque dal valore osservabile in un'indagine completa.

Errore variabile (v. **errore casuale**)

Falso negativo

In una classificazione di **unità statistiche** in base ad un attributo del quale possono essere portatrici, si dice falso negativo l'attribuzione (erronea) al gruppo delle unità che non possiedono l'attributo di un'unità che lo possieda. Il grado in cui lo strumento di classificazione è capace di evitare falsi negativi è detto **sensibilità**.

Falso positivo

In una classificazione di **unità statistiche** in base ad un attributo del quale possono essere portatrici, si dice falso positivo l'attribuzione (erronea) di una unità non portatrice dell'attributo al gruppo di unità che lo possiede. Il grado in cui lo strumento di classificazione è capace di evitare tale tipo di errore è detto **specificità.**

Forzatura (v. **metodi di imputazione**)

Frazione di campionamento

Sia *N* il numero di unità statistiche di una popolazione (v. **popolazione statistica**) e *n* (*n<=N)* il numero di unità campionarie (v. **campione**) tratte dalla popolazione. Si denomina frazione di campionamento la proporzione *n/N.*
Se la selezione del campione è effettuata da più strati, da ogni strato può essere selezionata una frazione di campionamento differente. Se la frazione di campionamento è costante negli strati, il campione si dice autoponderante (v. **campione autoponderante**).

Generalizzabilità

La voce, riferita allo strumento di misurazione (o tecnica o procedura), esprime la possibilità di adeguata utilizzazione al di fuori del campo di applicazione per il quale lo strumento è stato concepito.
Riferita ai risultati di un'indagine statistica, denota la possibilità di estendere gli stessi a una realtà fenomenica più ampia di quella per la quale i risultati sono stati ottenuti.

Hot-deck (v. **metodi di imputazione**)

Imputazione

Operazione che consiste nel sostituire i valori di una o più variabili dopo l'analisi di compatibilità, sulla base di regole e metodi prestabiliti; il valore sostitutivo è detto forzatura (v. **metodi di imputazione**).

Indagine campionaria (v. **indagine parziale**)

Indagine di controllo

Indagine svolta per effettuare il **controllo statistico** della qualità dei dati. Le indagini di controllo possono essere svolte in concomitanza con la rilevazione principale (v. **compenetrazione**), o successivamente a questa.

Indagine esaustiva

Esaustiva, o totale, è l'indagine statistica svolta sulla totalità delle unità che compongono la popolazione (v. **popolazione statistica**). Questo tipo di indagine si contrappone a quella parziale (v. **indagine parziale**).

Indagine parziale

Parziale è l'indagine svolta su un sottoinsieme delle unità che compongono la popolazione. Se l'insieme da rilevare è selezionato con l'intento di stimare caratteristiche della popolazione, l'indagine parziale si dice campionaria (v. **campione**).

Indagine pilota, **preliminare**

Si denomina pilota, o preliminare, l'indagine svolta prima di quella principale, con l'intento di assumere informazioni che permettono di rendere più efficiente lo svolgimento dell'indagine principale.
L'indagine pilota è condotta, in genere, su piccola scala e su sottoinsiemi mirati della popolazione. Per esempio, può essere utilizzata per sottoporre a verifica un questionario, per avere un'idea del tempo necessario per lo svolgimento dell'intervista presso certi sottoinsiemi di unità, per conoscere la variabilità dei fenomeni che interessano la ricerca e determinare conseguentemente la numerosità del campione sufficiente ad ottenere stime che abbiano un'**attendibilità** prefissata.

Indagine preliminare (v. **indagine pilota**)

Indagine successiva

Si dice successiva l'indagine svolta sulla stessa popolazione per verificare la qualità dei dati rilevati nell'indagine principale.
Il caso più tipico di indagine successiva è la **reintervista**, condotta, in genere, su una parte limitata delle unità della popolazione che hanno collaborato all'indagine principale. L'indagine successiva può, naturalmente, essere svolta anche con una tecnica di indagine e su unità di rilevazione diverse da quelle dell'indagine principale. Per esempio, in seguito a una indagine sullo stato di salute della popolazione, svolta con la tecnica del questionario, può essere svolta una limitata **indagine di controllo** su dati o con tecniche diagnostiche più affidabili delle dichiarazioni dei soggetti intervistati; oppure, per confrontare l'**accuratezza** delle risposte sul reddito, sul risparmio e sul patrimonio, fornite dalle famiglie, si possono rilevare a parte dati ufficiali sulla consistenza degli aggregati.

Iniezione di errori casuali

Consiste nell'aggiungere algebricamente alle frequenze di una casella, quando esse sono basse, un numero casuale - fatto pari, ad esempio, a un '1', a uno '0', a un '-1' -preservando, in una tabella a plurima entrata, i totali marginali.

Intervista

E' uno dei metodi con i quali si effettua la **rilevazione** dei dati.

Lista della popolazione

E' un elenco delle unità della popolazione con il corrispondente numero d'ordine o, meglio, di un'etichetta che consenta di identificare univocamente l'unità. La lista ha in genere un riferimento fisico: per esempio un elenco di indirizzi oppure un elenco di *record* individuali su un supporto magnetico.
Per svolgere un'indagine statistica è necessario conoscere la lista delle unità che compongono la popolazione per effettuare l'analisi della **copertura** e verificare la **completezza** della rilevazione. Il conoscere la lista della popolazione è fondamentale per formare un campione probabilistico della popolazione (v. **campione**), anche se ci sono casi in cui non è necessario conoscere l'etichetta di tutte le unità della popolazione per individuare il campione da sottoporre a rilevazione. Si pensi, ad es., ad una **indagine campionaria** sui clienti di un negozio, dove il campione è individuato prendendo un cliente ogni mille tra quelli che si presentano alla cassa: è ovvio che non occorre conoscere anche il nome dei clienti perché il campione sia probabilistico.

Macrodato (v. **dato aggregato**)

Mancata intervista (v. **mancata rilevazione**)

Mancata rilevazione

In una indagine sulla popolazione, per mancata rilevazione si intende l'insuccesso nella richiesta di intervista o di compilazione del questionario. La mancata rilevazione può essere causata dal mancato contatto delle unità designate a partecipare all'indagine (per trasferimento, per assenza ripetuta dal domicilio, per morte) o dal rifiuto esplicito a collaborare.
Le mancate interviste di unità che non appartengono all'insieme di riferimento (è il caso delle unità decedute, ma può essere anche quello delle unità trasferite, se escono dall'area oggetto di osservazione) possono essere ignorate per qualsiasi finalità inferenziale. Se però non vengono intervistate unità perché sono assenti o sono dichiaratamente non collaborative anche dopo ripetuti tentativi di contatto, è cruciale che si riconoscano, per lo meno, le loro caratteristiche, per valutare in quale misura è plausibile utilizzare solo i dati rilevati tra i rispondenti, oppure si debbano trovare misure correttive dei dati.
Vari studi hanno mostrato che chi non collabora per scelta o per pigrizia non è assimilabile, in media, a chi collabora e quindi, in presenza di mancate interviste, si pone il problema di inferire sul valore dell'insieme di unità che i mancati rispondenti rappresentano (se l'indagine è campionaria).
Il problema delle mancate interviste è esiziale soprattutto nelle indagini postali.

Mancata risposta, **non risposta**

Per mancata risposta si intende l'assenza di risposta a una o più domande poste nel questionario, in altre parti compilato.
Il problema delle mancate risposte ad una domanda può essere affrontato da diversi punti di vista. Una soluzione può essere quella della determinazione statistica delle risposte ottenibili utilizzando informazioni provenienti da

fonti esterne all'indagine (registri, censimenti, ecc.), o tramite un supplemento di indagine (v. **indagine successiva**).
Un'altra via perseguibile è quella della utilizzazione di tutte le informazioni rilevate nell'indagine per imputare (v. **imputazione**) il valore più probabile alle unità che non hanno espresso una risposta valida. Se si tratta di una variabile quantitativa, la forma di imputazione più elementare consiste nell'attribuire alle unità con risposta assente un valore medio di risposte valide.
Procedure appropriate per qualsiasi tipo di variabili, di frequente utilizzazione, sono quelle denominate *hot deck* e *cold deck* (v. **metodi di imputazione**). Si può anche decidere di escludere dall'analisi il dato mancante. L'esclusione va comunque effettuata con la consapevolezza che l'analisi delle sole risposte valide conduce normalmente a risultati distorti (v. **errore sistematico**).

Memorizzazione (v. **registrazione**)

Metadato

E' ogni informazione che può in qualche modo far luce sul significato e/o sulla qualità dei dati. Non si tratta quindi solo di quelle espressamente elaborate a questo fine rientrandovi anche quelle notizie che consentono di ripercorrere le fasi di lavoro nelle quali si è sviluppata l'indagine statistica. E' la disponibilità di metadati che sostanzia il requisito della **trasparenza**.

Metodi di imputazione

Esistono numerosi metodi per imputare valori mancanti e per la maggior parte di essi sono disponibili diverse varianti. Considerando i criteri di base che li ispirano è possibile classificare tali metodi in due gruppi, anche se nella pratica è comune un uso combinato degli stessi:

i) Deterministici; ;possono essere assimilati ad uno schema decisionale ad "albero", in cui le variabili e le loro relazioni sono analizzate secondo una gerarchia specificata a priori; le variabili sono quindi corrette secondo un ordine e valori predeterminati (se x_i non soddisfa la relazione con x_{i+1}, allora si ponga x_i = valore).

ii) Stocastici; consistenti nell'estrarre casualmente i valori correttivi da distribuzioni di valori precostituite utilizzando informazioni provenienti dalla stessa indagine per la quale si operano le imputazioni o già disponibili prima dell'indagine. Tra questi metodi due rivestono particolare importanza.

a) *Cold-deck*; il metodo comporta in primo luogo la costruzione di classi di imputazione, ottenibili raggruppando i *record* relativi alle unità statistiche rilevate in base a una o più variabili di controllo disponibili dall'indagine. Quindi si procede all'assegnazione a ciascuna classe di un certo numero di valori per ciascuna variabile imputabile. Tali valori provengono da conoscenze sulla popolazione disponibili prima dello svolgimento dell'indagine, o, se si sta eseguendo un'indagine periodica, da una occasione precedente a quella nella quale si operano le imputazioni. Una volta

predisposto questo insieme di valori, che prende appunto il nome di *cold-deck*, si esaminano i *record* dell'indagine corrente e in corrispondenza di un valore mancante per una determinata variabile lo si estrae dal *cold-deck.* Il metodo *cold-deck* è importante per essere stato uno dei primi metodi di imputazione automatica; attualmente ad esso viene preferito il metodo *hot-deck*.

b) *Hot-deck*; il metodo comporta il raggruppamento dei *record* dell'indagine in classi di imputazione del tutto analoghe a quelle descritte per il metodo cold-deck. Quindi, per ciascuna classe, i *record* dell'indagine corrente vengono esaminati sequenzialmente secondo un ordine prestabilito Al *record* che, per una variabile imputabile, presenti un dato mancante, viene attribuito il valore che aveva per la stessa variabile il *record* esaminato in precedenza. Per ogni classe di imputazione viene anche individuato, (per ciascuna variabile imputabile) un valore cold-deck da utilizzarsi nel caso che il primo *record* esaminato presenti dati mancanti.

Microdato

E' l'insieme dei dati elementari rilevati sull'**unità di analisi.** Se come spesso avviene tali dati sono registrati su supporto informatico, è invalso l'uso di indicare il microdato con il termine inglese *record* individuale (difficilmente traducibile) o,più semplicemente, *record*. Ma tale termine si riferisce alla forma della **registrazione** piuttosto che al suo contenuto.

Non risposta (v. **mancata risposta**)

Pertinenza

E' un aspetto della qualità dei dati relativo all'oggetto dell'indagine. La pertinenza denota la rispondenza tra informazione prodotta e necessità informative dell'utente.

Piano di campionamento

Il piano o, come si dice con un neologismo, il disegno di campionamento è l'insieme delle decisioni prese nel formare un campione. In alcuni casi il termine viene impiegato per comprendere anche il metodo di stima.
In genere, il piano campionario comprende la struttura del campione - stratificazione, stadi o livelli su cui è selezionato, regole seguite per la selezione - con o senza reimmissione delle unità estratte, sistematica, con probabilità costanti o variabili - con rotazione delle unità, **compenetrazione** delle assegnazioni dei rilevatori - la numerosità del campione ai vari stadi, la probabilità di selezione delle unità, **la frazione di campionamento**

Piano di rilevazione

Si dice piano, o disegno, di rilevazione l'insieme delle fasi elementari per l'espletamento della **rilevazione** dei dati in una indagine statistica. Il piano di rilevazione comprende, dunque, sia il **piano di campionamento** (se l'indagine è per campione), sia le scelte per la rilevazione dei dati (diretto, per in-

tervista o con questionari da autocompilare, o indiretto, per osservazioni o su fonti già esistenti), sia la predisposizione degli strumenti (questionario, altro) e la formazione del personale da adibire alla rilevazione delle informazioni.

Plausibilità

La plausibilità di un dato statistico, per chi ne viene a conoscenza, dipende da una sua ragionevole **compatibilità** con altre informazioni, esterne al dato stesso, delle quali costui è in possesso e alle quali accorda fiducia.

Popolazione statistica

Popolazione statistica è denominato ogni insieme finito o infinito di unità, le quali non sono necessariamente organismi viventi. Il termine ha ormai sostituito "universo", usato in altri tempi per designare l'analogo concetto e derivante dall' "universo del discorso" della logica. Un altro termine praticamente sinonimo è "aggregato".
In una indagine statistica si possono individuare diverse popolazioni:
a) quella obiettivo o ideale, che è la popolazione sulla quale si vuole eseguire l'indagine;
b) quella raggiungibile (si intende con i mezzi a disposizione); per esempio, in una indagine postale, è l'insieme delle unità delle quali si possiedono gli indirizzi esatti, in una indagine telefonica è l'insieme delle unità raggiungibili per telefono;
c) quella raggiunta o rilevata o osservata nell'indagine o nell'esperimento;
d) quella di riferimento, ossia l'aggregato al quale si riportano le stime e si estendono le verifiche effettuate sulle ipotesi di ricerca saggiate con l'indagine o l'esperimento. La popolazione di riferimento può essere quella osservata, oppure quella ideale, eventualmente corretta per tener conto della parte non raggiunta.

Precisione

Il termine assume significati diversi in rapporto all'oggetto cui è associato.
Se riferito allo stimatore (o alla stima) è espressione del grado di dispersione di una classe di misure (stime) ottenibili da ipotetiche replicazioni, in identiche condizioni, della procedura di campionamento, intorno al loro valore atteso. La misura della precisione è data dal reciproco dell'**errore medio di stima**. Maggiore l'errore medio, minore la precisione e viceversa .
Se riferito a singole misure, il termine esprime l'entità dell'approssimazione della misurazione stessa. Si dice precisa una misura, (il termine è riferibile anche allo strumento di misurazione), il cui grado di approssimazione non supera un prestabilito limite di tolleranza.

Profilo degli errori

Il profilo degli errori (in inglese *error profile*) è un resoconto schematico delle operazioni che hanno condotto ai risultati dell'indagine e, idealmente, dell'impatto che le singole operazioni hanno avuto sull'**errore globale delle stime**. Anche quando non sia possibile quantificare ogni componente di errore, un buon profilo degli errori rende evidenti per lo meno la mappa degli errori possibili e la loro incidenza qualitativa sulle stime.

Programmi di imposizione automatica

L'insieme delle procedure che hanno il compito di correggere le informazioni registrate su supporto informatico, prima della elaborazione finale dei dati.
Essi procedono attraverso due fasi:
i) rilevazione, sulla base delle regole fornite, dell'incoerenza sulla singola variabile o tra variabili;
ii) scelta della variabile da cambiare e del nuovo valore da assegnare (v. **metodi di imputazione**)
La fase di compatibilità e correzione si configura quindi come un "filtro" degli errori accumulatisi nelle fasi precedenti; è da notare che è possibile individuare e correggere solo la parte di essi che contravviene alle **regole di compatibilità**.

Qualità del dato statistico

Per qualità dei dati si intende la rispondenza degli stessi alla realtà fattuale che con essi si intende quantificare. Naturalmente la realtà fattuale è semplicemente quella che, in rapporto all'indagine, viene definita convenzionalmente attraverso concetti e classificazioni astratte.
L'esecutore dell'indagine, più o meno esplicitamente, attribuisce al risultato della rilevazione un preciso significato. Assimilando il processo produttivo dei dati statistici ad un qualsiasi processo produttivo, si può anche affermare che la qualità del "dato-prodotto" è la sua capacità di soddisfare le proprietà garantite dal "produttore".
Queste proprietà riguardano due ambiti distinti, quello dell'**attendibilità** che concerne i livelli di **accuratezza** delle stime e quello dell'**adeguatezza** che concerne particolari aspetti del piano di una rilevazione e più precisamente quelli connessi, da un lato, alla definizione degli obiettivi e, dall'altro, alla diffusione dei risultati. Si parla in quest'ultimo caso di **tempestività** dell'informazione, di **trasparenza** del dato, ecc..

Qualità del sistema informativo statistico

Per sistema informativo statistico possiamo intendere l'insieme delle indagini statistiche che sono dirette verso gli stessi "oggetti" o che comunque esplorano singoli aspetti di uno stesso fenomeno o ambiti riconducibili ad una ben individuata problematica. Ciò premesso, si può affermare che la qualità di un sistema informativo statistico se è per certi versi legata alla qualità dell'informazione statistica che deriva dalle singole indagini, da altri punti di vista ne prescinde nel senso che va ben oltre: se per una singola rilevazione buona qualità vuol dire anche possedere una ricca portata informativa sul particolare fenomeno investigato, per un sistema informativo statistico buona qualità vuol dire anche capacità complessiva da parte del corpo di indagini di soddisfare esaurientemente le molteplici esigenze conoscitive che si pongono con riferimento a realtà composite. Può quindi accadere che su ogni singola indagine possa esprimersi un giudizio positivo e che, allo stesso tempo, si debbano esprimere riserve sulla validità del sistema informativo statistico al quale le singole indagini vanno ricondotte. E' evidente che se ciò si verifica è perché sussistono delle carenze nel processo di **integrazione** fra le varie indagini, sotto forma di ridon-

danze e/o di vuoti da colmare. Inconvenienti di questa natura dipendono generalmente dall'assenza, a monte, di un progetto unico che elimini a priori il rischio di eventuali disarmonie.

Registrazione, **memorizzazione**

E' l'operazione mediante la quale l'informazione contenuta nel questionario - rappresentata dai codici corrispondenti alle risposte date a domande chiuse o apposti con l'operazione di codificazione (v. **codificazione**) in corrispondenza delle risposte a domande aperte - viene trasferita sul supporto elettromagnetico per eseguire tramite elaboratore le successive operazioni di revisione, correzione (v. **programmi di imposizione automatica** e correzione) e elaborazione.

Regole di compatibilità

Relazioni tra i valori che possono essere assunti da due o più variabili o tra la singola variabile ed il suo campo di variazione. Si basano su:
i) definizioni di aggregati;
ii) disposizioni formali per la compilazione del questionario (ad esempio la regola di "salto" di uno o più quesiti in presenza di specifica risposta a domanda precedente);
iii) il piano di codifica;
iv) informazioni a priori sui fenomeni oggetto di rilevazione.
Possono essere distinte in regole "esplicite" (ovvero quelle esplicitate dall'esperto) ed in regole "implicite" (ovvero derivate dalle esplicite mediante operazioni logiche). Se ci si limita a considerare solo le regole "esplicite" possono verificarsi contraddizioni nel sistema di regole; per evitare tali situazioni è necessario costruire l'insieme minimo non ridondante e non contraddittorio di regole (esplicite ed implicite).

Reintervista

La reintervista è svolta a una certa distanza e con personale diverso da quello dell'indagine principale per ottenere dalle persone già intervistate risposte alle stesse domande poste nell'indagine principale, con l'obiettivo di misurare (singolarmente o in media) la concordanza di dati ottenibili da rilevatori diversi (v.**indagine successiva**). Gli obiettivi della reintervista sono, dunque, differenti sia da quelli delle indagini svolte in occasioni successive con l'obiettivo di rilevare dati inerenti a punti temporali diversi, sia da quelli dei tentativi di ottenimento della intervista non ancora concessa, tipici delle indagini postali o telefoniche.
La reintervista è ordinariamente svolta da personale più specializzato nella rilevazione di quello che ha svolto l'indagine principale, su un campione di minori dimensioni e su una parte limitata delle informazioni utili alla ricerca (di solito solo sulle domande cruciali e su certe caratteristiche degli individui, fondamentali per l'analisi statistica dei dati): Se si assegna maggiore fiducia ai risultati della reintervista rispetto a quelli dell'indagine principale, la differenza fra le statistiche di sintesi delle due indagini può essere considerata una misura della **distorsione**.
Sulla base del confronto fra le risposte ottenute presso le stesse persone, si può anche misurare in modo approssimato la **varianza extracampionaria** delle stime.

Revisione

Con il termine revisione si indica il processo di identificazione di errori o lacune rappresentati da incoerenze, omissioni e valori fuori campo, nel *record* di ciascuna **unità statistica** partecipante all'indagine. Se l'**unità di rilevazione** comprende più di una unità statistica è talvolta possibile individuare incompatibilità anche tra i *record* che costituiscono l'unità di rilevazione. Ne è un esempio la presenza all'interno di una famiglia di più di un "capofamiglia"
Con accezione più ampia, si fanno rientrare tra le operazioni di revisione anche quelle di correzione delle suddette incompatibilità quando queste ultime avvengono sulla base di informazioni disponibili all'interno dei singoli *record* soggetti a revisione o all'interno di *record* costituenti l'unità di rilevazione. Sia le operazioni di individuazione degli errori che quelle di correzione vengono di norma effettuate con programmi informatici su elaboratore elettronico (v. **programmi di imposizione automatica**)

Riferimento individuale

Indica la possibilità, attraverso il dato pubblicato, di risalire all'unità statistica specifica (individuo, impresa,.ecc.) cui lo stesso si riferisce. La pubblicazione di dati individuali può non consentire di individuarne l'origine. Vedi, ad es., i campioni di dati individuali, resi disponibili in qualche Paese in forme particolari, che escludono quella possibilità: scelta a caso in una popolazione numerosa, aggregazione per caratteristiche rare, ecc. D'altro canto, quando si rendono disponibili dati aggregati (v. **dato aggregato**, **macrodato**), non si è certi di coprire col segreto il dato individuale: è il caso in cui nasca, ad esempio, una coalizione fra i rispondenti che reciprocamente si informano sulla risposta data per individuare, mediante differenza col totale pubblicato, quella fornita da una singola unità non rientrante nella coalizione.

Rilevazione

E' l'accertamento della presenza o meno del carattere o fenomeno (o caratteri e fenomeni) che interessa nella **unità statistica** e della modalità sotto cui, nella stessa, esso si presenta.

Riservatezza

E' la condizione di ciò che è affidato alla discrezione e al rispetto del segreto. La condizione riguarda normalmente il **dato elementare** (talora il **microdato**). E' riservato il dato per il quale esiste un prestabilito limite di diffusione e notorietà. Ciò è in conflitto con la necessità di diffusione e pubblicità del dato, che può tuttavia essere realizzata previa adeguata manipolazione (arrotondamento, iniezioni di errori, aggregazione di più dati, ecc.) del dato stesso, in modo che da esso non sia possibile risalire all'unità statistica cui è associato.

Rispondente

E' il soggetto cui è richiesto di fornire i dati che ci si propone di raccogliere con una rilevazione statistica. Non sempre coincide con l'**unità di rilevazione** e

ciò anche quando si tratta di indagini sulle famiglie o sugli individui. E' opportuno che sia individuato con estrema precisione. Ciò risulta indispensabile in quei rari casi per i quali sono previste delle sanzioni a carico di chi, pur obbligato da disposizioni di legge, si rifiuta di prestare la propria collaborazione.

Segreto statistico

Per il produttore di statistiche ufficiali è l'esclusione della possibilità di individuare l'unità statistica alla quale si riferisce un dato reso pubblicamente disponibile. In qualche caso il segreto non vale *erga omnes*, nel senso che quella informazione è ammessa per categorie privilegiate di utenti. La legge italiana del 1929 per l'Istat prevedeva esplicitamente, ad esempio, il magistrato.
Talvolta si vede usato il termine confidenzialità, brutta traduzione di *confidentiality*, di cui non rende il senso preciso. Confidenziale, in Italiano, è fatto o detto in confidenza, non l'azione dello statistico ufficiale intesa a dare la massima diffusione possibile delle informazioni che possiede, salva la garanzia del rispetto del segreto statistico. Altri confonde la *privacy* con il segreto statistico. Certamente violando il segreto si offende la *privacy*, ma questa va presa in considerazione prima di iniziare un'indagine e la connessa rilevazione per evitare non necessarie intrusioni nella sfera intima del soggetto. Così, per rispetto della *privacy*, l'Italia aderisce alla convenzione europea che esclude per le statistiche ufficiali domande intorno a razza, religione, opinioni politiche, ecc.

Segreto statistico attivo

In senso lato indica una procedura per impedire - o rendere estremamente difficoltoso - che si possa, dal dato pubblicato, identificare l'unità cui lo stesso si riferisce. Ciò può essere fatto in vari modi: a) non pubblicando il dato anche quando è disponibile, b) facendo aggregazioni di dati, c) iniettando **errori casuali**, d) adottando **arrotondamenti casuali**
In senso stretto, segreto statistico attivo è espressione usata per certe statistiche economiche. Ad es., nel nostro Annuario di Statistiche Industriali, ed. 1986, nella Tavola 24 sulla distribuzione delle unità rilevate per classi di addetti, incrociate con alcuni rami o classi di attività economica, vengono raggruppate le classi di addetti in cui, per qualche casella, il numero di unità rilevate è inferiore a tre. Nei volumi dell'Eurostat che riportano per voci della NACE, secondo classi di addetti per imprese, informazioni su numero di imprese, personale occupato, spese per lo stesso, cifra d'affari, valore aggiunto al lordo, numerose righe, corrispondenti ad altrettanti Paesi, risultano completamente vuote, altre riportano solo il totale, altre aggiungono a questo il dato per una o due caselle. Spesso poi si adotta un'altra cautela, cioè di non pubblicare un **dato aggregato** riferito a un gruppo di aziende quando una di queste è responsabile per più di una quota elevata del totale complessivo.
L'espressione quindi designa l'applicazione della regola, cosiddetta dell' *n* x *k*, ove *n* è un numero pari a 3, 4, 5 e *k* una percentuale del 70, 75, 80%, poniamo, a seconda dei casi. Talora, a maggior garanzia di tutela del segreto, si preferisce tacere anche sui valori di *n* e *k* cui ci si attiene.

Segreto statistico passivo

Può capitare che il produttore di dati statistici non sia in grado di sapere quali informazioni analitiche, se pubblicate, possano aprire la strada alla individuazione della fonte originaria del reperto; oppure che, per arrivare a saperlo, debba sostenere costi assai elevati sia in termini economici sia rendendo troppo tardiva la pubblicazione. E' il caso soprattutto delle statistiche del commercio con l'estero quando riportino dettagliati incroci di categorie di prodotti e Paesi di origine o, rispettivamente, di destinazione. In tali circostanze, l'istituzione responsabile pubblica i dati analitici e, se riceve proteste da parte di chi si senta con questo offeso dalla possibilità di divenire oggetto di **riferimento individuale**, procede ad adatte aggregazioni con categorie affini (o rimandando ad una generica voce "altri") nelle edizioni successive della stessa statistica. Questo è il modo di operare che va sotto la dizione di segreto statistico passivo.

Sensibilità

Proprietà di uno strumento di misura, funzione della capacità di reagire agli stimoli per la cui rilevazione è predisposto Se il carattere rilevato è un attributo, la sensibilità è misurata dalla frazione di unità statistiche che possiedono l'attributo e sono classificate correttamente (in rapporto al totale di unità che possiedono l'attributo). I casi che, possedendo l'attributo, sono classificati erroneamente si dicono **falsi negativi**

Specificità

Proprietà di uno strumento di misura valutata in relazione alla capacità di discriminare dalle altre le unità che possiedono un dato attributo. Si misura con la frazione di unità che, possedendo l'attributo, sono classificate correttamente (in rapporto al totale delle unità nelle quali l'attributo è assente). I casi che, pur non possedendo l'attributo, sono classificati come portatori dello stesso si dicono **falsi positivi**.

Stima

In senso stretto, è il particolare valore ottenuto dall'applicazione di uno **stimatore** in determinate circostanze. Il termine è comunemente usato, in senso più ampio, per indicare anche l'insieme delle regole attraverso le quali è stato ottenuto quel particolare valore, ovverosia lo stimatore stesso.

Stimatore

Si denomina stimatore una regola o un metodo per stimare una costante della popolazione. E' di solito espresso come una funzione dei valori campionari ed è quindi una variabile statistica la cui distribuzione è di grande importanza nell'accertamento dell'**attendibilità** della stima che dallo stimatore deriva

Tempestività

E' un aspetto della qualità dei dati relativo al lasso di tempo che intercorre tra la loro rilevazione e la loro disponibilità. Questi possono essere considerati tempestivi se, in rapporto a determinati

obiettivi, sono diffusi in tempi congrui al loro raggiungimento.

Trasparenza

Con riferimento ai risultati di un'indagine statistica, o più in particolare ai passaggi operativi di una singola procedura, il termine sottintende la disponibilità delle informazioni necessarie per esprimere un giudizio sulla qualità dei dati. E' quindi espressione di un particolare atteggiamento del soggetto che, avendo effettuato l'indagine, consente ad altri di valutare in modo approfondito il proprio operato.

Unità campionaria

Unità campionaria è una delle unità che compongono il campione. Conviene distinguere questa denominazione da quella di "unità di campionamento", con la quale si intende una delle unità che compongono un aggregato che deve essere sottoposto a campionamento, che sono individuali e indivisibili nel momento in cui si effettua la selezione campionaria. Le unità da estrarre possono essere definite su basi naturali (persone,famiglie di fatto, animali, ecc.), giuridiche (famiglie anagrafiche), amministrative (comuni, ospedali) o su qualsiasi altra base (aree in cui è suddiviso un territorio da sottoporre a indagine). Nel campionamento su più stadi, le unità di campionamento sono diverse ad ogni stadio: sono più grandi al livello superiore e gradatamente più piccole a mano a mano che si scende di livello.

Unità di analisi

E' l'unità appartenente all'insieme che viene analizzato a fini di stima statistica. Se i dati sono forniti sotto forma di tabelle, si denomina **unità di tabulazione**.Può non coincidere con l'**unità di rilevazione**

Unità d'informazione

E' l'unità che fornisce le informazioni statistiche. Può trattarsi di una persona, e in tal caso lo si denomina **rispondente**, o di uno strumento di rilevazione. Per esempio, le centraline per la rilevazione dell'inquinamento atmosferico, le stazioni pluviometriche, il *meter* per la rilevazione della visione dei programmi televisivi sono le unità di informazione delle indagini.

Unità di rilevazione

E' l'unità empirica su cui si basa la **rilevazione**. Non coincide necessariamente né con l'unità che fornisce le informazioni (v. **unità di informazione**) né con l'**unità statistica**, cui, in ultima analisi si è interessati, e le cui caratteristiche si vogliono conoscere, ogni volta che essa ne raggruppa più di una.(v. **unità di analisi, unità di tabulazione**) Nel censimento della popolazione, la famiglia di censimento è una delle unità di rilevazione (le altre sono la convivenza e l'ospite di esercizio alberghiero). Se la famiglia è composta di un solo membro, l'unità di rilevazione coincide con l'unità statistica.

Unità di tabulazione

E' l'unità di cui, in una tavola, si presenta la classificazione secondo un assortimento di modalità di caratteri. Il nucleo familiare, in un censimento demografico, non è l'unità statistica (il censito), né una di quelle di rilevazione, ma una derivazione da una di queste (la famiglia di censimento) attraverso il carattere "relazione col capofamiglia", e costituisce oggetto di classificazione e pubblicazione in tavole specifiche.

Unità statistica

E' l'unità elementare della **popolazione statistica**. Può trattarsi di una persona fisica (il censito, il dimesso da un istituto di cura), di una persona giuridica (l'impresa), di un'istituzione (la scuola), di un evento (un matrimonio, una nevicata), ecc.

Validità

Riferito allo strumento di misurazione o alla tecnica, alla procedura, ecc., esprime la rispondenza agli scopi per i quali lo strumento è utilizzato. Il termine mantiene lo stesso significato anche se riferito al dato statistico, poiché validità del dato significa, sostanzialmente, rispondenza dello stesso ai fini della ricerca.

Valore atteso

E' il risultato dell'applicazione dell'operazione matematica "media" ad una statistica dipendente da dati campionari e avente una distribuzione statistica.
Il "valore atteso" non è necessariamente il più frequente; può addirittura non esistere: si pensi ad una variabile che può assumere solo i valori *0* (con probabilità *1-p*) o *1* (con probabilità *p*, $p<1$), il cui valore atteso (*p*) è compreso fra questi due valori.
Il valore atteso di un **dato elementare** è la media dei valori osservabili sotto certe condizioni presso l'unità statistica: in assenza di **errore sistematico** coinciderà con il valore vero dell'unità.

Valore osservato (v. **valore rilevato**)

Valore rilevato, **osservato**

E' la modalità del carattere preso in considerazione accertata con la rilevazione sulla singola unità statistica.

Valore vero

Nel **microdato** può essere designato come la modalità del carattere in considerazione posseduta dall'entità oggetto di rilevazione. Nel **macrodato** può essere inteso come una misura esattamente corrispondente a ciò che nella realtà esiste. Lo si prende in considerazione per definire gli scarti dello stesso dalle singole osservazioni o dalle stime dedotte da queste.
Quando si tratti di caratteri qualitativi, talora, le modalità possono essere delineate con minore ambiguità di quelle di caratteri quantitativi. Per esempio, è poco ambiguo il sesso, ma non lo stato civile (vi sono le unioni consensuali socialmente accettate) né l'attività economica prevalente (con quale criterio è verificata la prevalenza?). In quei casi la modalità registrata rispecchia esattamente o non una condizione reale dell'unità statistica. Ma proprio in quei casi lo scostamento non è definibile che

convenzionalmente in termini quantitativi, trattandosi soltanto di un risultato corretto o errato dell'accertamento effettuato.
Se il carattere è quantitativo, si può sostenere che una valore vero in senso assoluto non esiste. La lunghezza di una barretta d'acciaio è continuamente variabile in funzione, almeno, della temperatura, il rapporto dei sessi nella popolazione presente pure varia, con continuità, nel tempo. Per cui, se di un valore vero si può parlare, esso va riferito a un preciso istante temporale. Ma misure ripetute da cui si ricavi una stima, vengono effettuate in momenti diversi, così come la **registrazione** del sesso dei presenti non avviene per tutti simultaneamente. D'altro canto la misura del supposto valore vero può dipendere dalla **precisione** dello strumento che ho a disposizione per rilevarla. Per chi nega che esista un valore vero, esso può essere concepito come un dato astratto di comodo, scelto in base a considerazioni di opportunità, che si presta a definire un grado di importanza degli scarti.

Varianza di campionamento delle stime

E' la varianza di una qualsiasi statistica del campione calcolata sull'universo dei campioni associato ad un determinato **piano di campionamento**. La radice quadrata della varianza campionaria si denomina errore di campionamento (v. **errore campionario**).

Varianza correlata di risposta

E' la varianza di uno stimatore affetta da errori (variabili) di risposta tra loro linearmente dipendenti. Le varianze correlate di risposta individuate da Fellegi (1964) si identificano con le fonti di errore:

i) la varianza del rilevatore, originata dal condizionamento esercitato dal rilevatore sulle risposte ottenute nell'indagine (v. **varianza del rilevatore**; **distorsione del rilevatore**);
ii) la varianza del supervisore, funzione del condizionamento esercitato dal supervisore o dalle persone preposte all'addestramento sulla qualità del lavoro svolto dai rilevatori.

Varianza del codificatore

E' la misura della variabilità indotta sulle stime dall'azione del codificatore. Ha proprietà e si calcola in modo analogo alla **varianza del rilevatore**. Se rilevatore e codificatore coincidono, la varianza dovuta all'azione del codificatore si confonde con quella del rilevatore.

Varianza di stima

Varianza di uno stimatore attorno al proprio valor medio. E' data dalla somma delle varianze dovute alle singole componenti di variabilità, quella campionaria (v. **varianza campionaria**) e quelle extracampionarie (v. **varianza extracampionaria**) e delle covarianze tra le componenti.

Varianza elementare di risposta

E' la varianza di una stima affetta da errori (variabili) di risposta tra loro indipendenti. Dato un insieme di *N* unità statistiche, la varianza elementare di ri-

sposta è il valore medio degli scarti quadratici dell'errore di risposta atteso per le singole unità. Ciò implica che un'osservazione è idealmente ripetibile sotto identiche condizioni osservazionali, o equivalentemente, che una osservazione su una unità ha una ben precisa, anche se ignota, distribuzione di probabilità dalla quale essa è tratta per l'indagine specifica. Siccome non è possibile ottenere più di un campione di osservazioni per ogni indagine (ossia, ogni nuova misurazione è svolta sotto condizioni osservazionali che hanno almeno un termine di riferimento diverso), il concetto di ripetibilità va inteso in senso lato.

Varianza extracampionaria

E' la **varianza di stima** indotta dalla presenza nei dati di errori extracampionari di varia natura. Comprende la varianza di risposta o di rilevazione (v. **varianza correlata di risposta**, **varianza elementare di risposta**), la varianza dovuta a codifiche di risposte a domande "aperte" (v. **varianza del codificatore**), la varianza per errori commessi nella fase di **memorizzazione** dei dati e nella successiva correzione ed **imputazione** con tecniche casuali (v. **metodi di imputazione**), ed altro ancora.
La varianza campionaria si usa per misurare la parte di inaccuratezza delle stime attribuibile ad errori variabili, (v. **errore casuale**) e che è naturalmente presente anche nelle rilevazioni esaustive.

Varianza dell'intervistatore (v. **varianza del rilevatore**)

Varianza del rilevatore

E' la misura della variabilità indotta sulle stime dall'azione del rilevatore nella fase di raccolta dei dati. Si ottiene calcolando la media (aritmetica ponderata) degli scarti quadratici tra il valor medio ottenuto dai singoli rilevatori e la media globale.
E' una varianza correlata degli errori di risposta (v. **varianza correlata di risposta**). Se si denota, infatti, con σ_R^2 la varianza elementare di risposta e con ρ_i il coefficiente di correlazione intra-intervistatore (v. **coefficiente di correlazione intraclasse**), la varianza tra errori di risposta imputabili agli intervistatori si può anche scrivere:

$$\sigma_R^2 \{1 + \rho_i (n-1)\}$$

dove *n* è il numero medio di unità osservate da ognuno degli intervistatori impiegati per la raccolta dei dati.
Il **disegno di rilevazione** proponibile per misurare la varianza dell'intervistatore è la **compenetrazione** delle assegnazione dei rilevatori