Gestione informatica dei dati

Le dimensioni della qualità dei dati

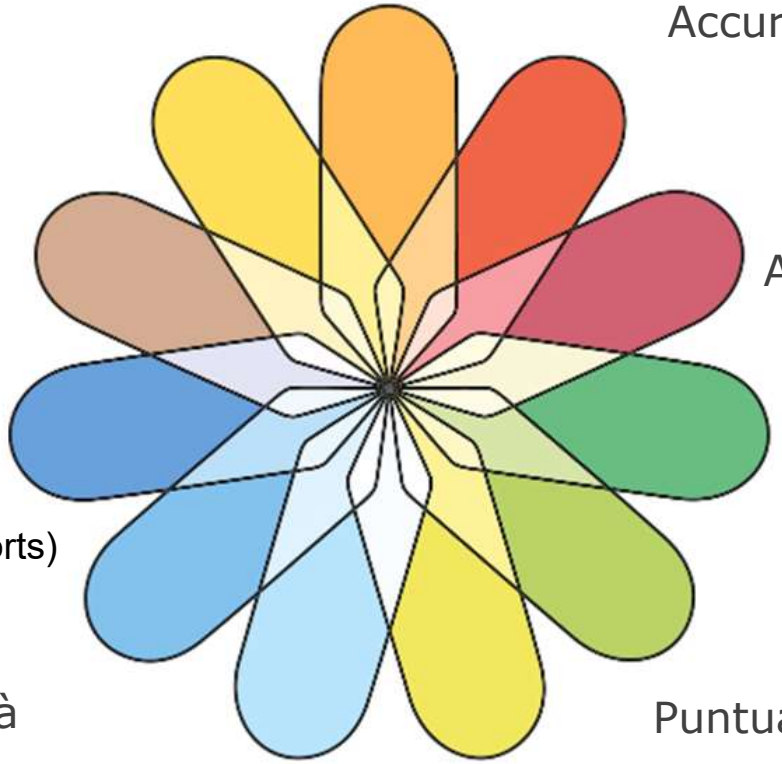
Prof. Roberto Foderà

Cattedra di Gestione informatica dei dati
Dipartimento di Giurisprudenza - Palermo

LUMSA
UNIVERSITÀ

A.A. 2021-2022

Le dimensioni della qualità dei dati



Quality is the degree to which a set of inherent characteristics of an object fulfils requirements.
(ESS Handbook on quality and metadata reports)

Le dimensioni della qualità dei dati

Qualità:

La qualità dei prodotti si misura in base al grado di pertinenza, accuratezza e attendibilità, tempestività e coerenza delle statistiche, in base alla loro comparabilità tra le diverse regioni e i vari paesi e in base alla facilità di accesso per gli utilizzatori, vale a dire in base ai principi di produzione statistica.

Consideriamo la qualità come la base del nostro vantaggio competitivo in un contesto mondiale caratterizzato da una crescente tendenza verso informazioni istantanee di cui spesso non è attestata la qualità.

(Codice delle statistiche europee, 2017)

Le dimensioni della qualità dei dati

PRINCIPIO 4

Impegno a favore della qualità

La qualità è un imperativo per le autorità statistiche, che individuano sistematicamente e regolarmente i punti di forza e di debolezza al fine di migliorare costantemente la qualità dei processi e dei prodotti.

INDICATORE	
4.1	la politica per la qualità è definita e resa pubblica. Esiste una struttura organizzativa e sono disponibili strumenti adeguati per assicurare la gestione della qualità.
4.2	sono in atto procedure per pianificare, monitorare e migliorare la qualità dei processi statistici, compresa l'integrazione di dati provenienti da più fonti.
4.3	la qualità dei prodotti è regolarmente monitorata e valutata tenendo conto dei possibili compromessi; essa è oggetto di relazioni elaborate in base ai criteri di qualità delle statistiche europee.
4.4	è prevista una regolare e approfondita valutazione dei principali prodotti statistici con il ricorso, se del caso, anche a esperti esterni.

https://ec.europa.eu/eurostat/documents/4031688/9394142/KS-02-18-142-IT-N.pdf/2d3874da-4253-4f20-9cfd-304f48a5ed1a

Le dimensioni della qualità dei dati

Pertinenza:
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;

Le dimensioni della qualità dei dati

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;

Le dimensioni della qualità dei dati

Affidabilità:
Il termine non si riferisce al dato ma alla fonte, allo strumento, al metodo, alla procedura, ecc. E' 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 non supera un prestabilito livello.

Le dimensioni della qualità dei dati

Tempestività:
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.

Puntualità:
Fa riferimento all'esistenza di un calendario di rilascio dell'informazione statistica e misura il grado di aderenza a tale calendario da parte del produttore dell'informazione statistica;

Le dimensioni della qualità dei dati

Accessibilità:
Nota anche col nome di "trasparenza", 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.

Le dimensioni della qualità dei dati

Chiarezza:
La chiarezza delle informazioni statistiche si riferisce alla disponibilità di documentazione appropriata, relativa alle varie caratteristiche e fasi dell'indagine ed, eventualmente, la possibilità di ottenere assistenza nell'utilizzo e all'interpretazione dei dati.

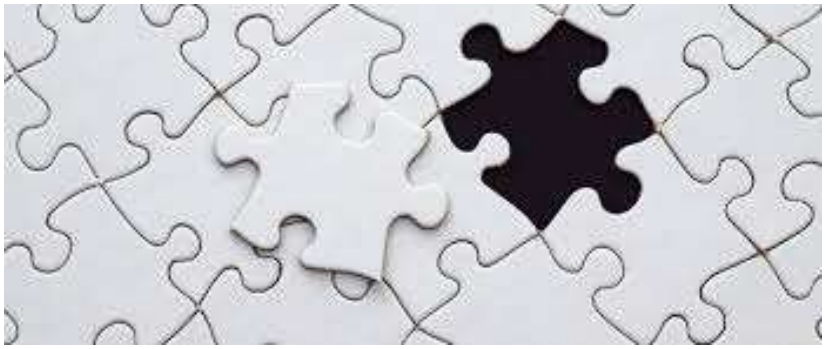
Le dimensioni della qualità dei dati

Confrontabilità (o comparabilità):
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;

Le dimensioni della qualità dei dati

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 siano 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.

Le dimensioni della qualità dei dati



Completezza:
Dal punto di vista dei processi di produzione statistici, si tratta di una caratteristica trasversale e consiste nella capacità delle informazioni di integrarsi per fornire un quadro informativo soddisfacente del dominio di interesse.
Dal punto di vista del dato raccolto su ogni unità statistica che partecipa all'indagine, essa indica l'assenza di dati mancanti nel questionario o scheda di rilevazione ecc. (mancata risposta totale o parziale).

Le dimensioni della qualità dei dati

Il Sistema Informativo sulla Qualità (SIQual) contiene informazioni sulle modalità di esecuzione delle rilevazioni ed elaborazioni condotte dall'Istat e sulle attività svolte per garantire la qualità dell'informazione statistica prodotta.

http://siqual.istat.it/SIQual/

Le dimensioni della qualità dei dati

Il sistema è dedicato alla navigazione dei metadati che descrivono il processo produttivo e le sue caratteristiche: contenuto informativo, scomposizione in fasi e operazioni, attività di prevenzione, controllo e valutazione dell'errore.

Le dimensioni della qualità dei dati

Attraverso il Sistema è possibile conoscere i fenomeni indagati dall'indagine e, quindi, le dimensioni teoricamente raggiungibili per svolgere analisi dei dati.

[R] - Rilevazione sulle forze di lavoro

Fenomeni osservati ()

Dal 31/03/2003

- Caratteristiche dell'attività lavorativa
- Caratteristiche socio-demografiche della popolazione residente
- Disoccupazione
- Forze di lavoro
- Non forze di lavoro
- Occupazione
- Occupazione part-time
- Orari di lavoro
- Ore lavorate
- Precedenti esperienze di lavoro
- Ricerca di lavoro
- Situazione lavorativa
- Studio e formazione

Le dimensioni della qualità dei dati

[R] - Rilevazione sulle forze di lavoro

Unità di Rilevazione( ) e Archivi di Estrazione ( )

Dal 31/03/2003

Famiglie di fatto

Estratte da: Liste Anagrafiche Comunali (LAC) (Dal 02/07/2012)

La situazione della famiglia estratta e quella risultante dall'anagrafe ma la rilevazione indaga sulla famiglia di fatto indipende... [more]

Famiglie di fatto

Estratte da: Anagrafe comunale (dal 31/03/2003 al 02/04/2013)

La situazione della famiglia estratta e quella risultante dall'anagrafe ma la rilevazione indaga sulla famiglia di fatto indipende... [more]

Unità di Analisi ( )

Dal 31/03/2003

Individui

L'analisi dei fenomeni oggetto della rilevazione viene effettuata su base individuale

Unità di Analisi di tipo Subset ( )

Dal 31/03/2003

Individui

L'analisi dei fenomeni oggetto della rilevazione viene effettuata su base individuale

Individui di 15 anni o più (Dal 31/03/2003)

La rilevazione viene fatta sugli individui di 15 anni o più. Per quelli con meno di 15 anni vengono rilevate le sole informazioni ... [more]

Le dimensioni della qualità dei dati

Se il processo statistico rilascia dati on line, attraverso SIQual è possibile consultare tali informazioni.

[R] - Rilevazione sulle forze di lavoro

Elenco dei dati on-line (banche dati e tavole di dati Istat o non Istat) dove è possibile accedere ai dati, cliccando sul nome.

Banche Dati Istat

- **ConIstat: statistiche congiunturali**
  A partire da lunedì 16 aprile 2012 Con'Istat, il database contenente le serie storiche degli indicatori congiunturali prodotti dall'Istat, non è più attivo'Tutte le serie contenute in questo database sono disponibili su I'Stat (http://dati'istat'it)
- **I.Stat: il data warehouse dell'Istat**
- **Sistema di indicatori territoriali**

Tavole Istat

- **Ricostruzione delle serie storiche delle Forze di lavoro**

Il sistema di controllo della qualità

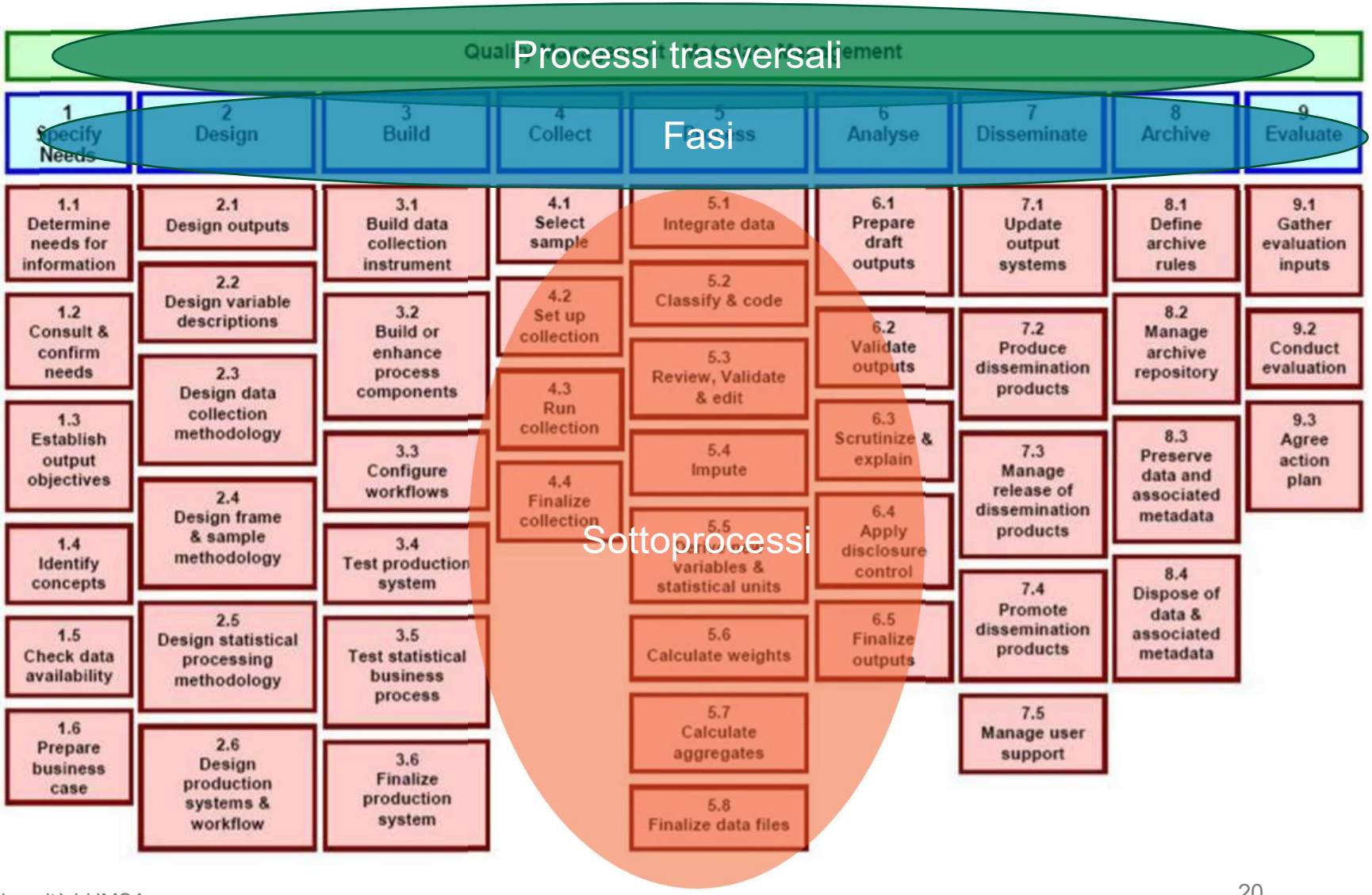
Il **GSBPM** (Generic Statistical Business Process Model) è uno strumento il cui scopo è descrivere le fasi di un processo statistico, indipendentemente dal tipo di processo.

Il GSBPM è suddiviso in fasi, sottoprocessi, processi trasversali. È organizzato con una struttura a matrice, che garantisce flessibilità nell'ordinamento e nei passaggi da una fase o da un sottoprocesso all'altro.

L'ordine di produzione dei sottoprocessi non deve essere necessariamente sequenziale e potrebbero non doversi svolgere nella loro totalità.

Il sistema di controllo della qualità

La struttura a matrice del GSBPM



Il sistema di controllo della qualità

L'uso del GSBPM permette di svolgere un controllo continuo sulla produzione dei dati statistici, svolgendo una valutazione ex-ante, ex –post e un monitoraggio in itinere.

Qualità di processo

- Ogni prodotto è il risultato di un processo
- Standardizzare, controllare e migliorare i processi consente di migliorare la qualità del prodotto
- Strumenti e indicatori diversificati

Qualità di prodotto

- Misurazioni della qualità dell'output o del prodotto statistico finale
- Non sempre esistono misure quantitative
- Alcune componenti sono in conflitto tra loro
- L'accuratezza è la componente con la natura più statistica ma anche spesso quella più difficile da calcolare

Il sistema di controllo della qualità

Cosa mettere in valigia prima di partire.

Quando si avvia una rilevazione è molto importante prevedere quali fonti di errori possono esistere.

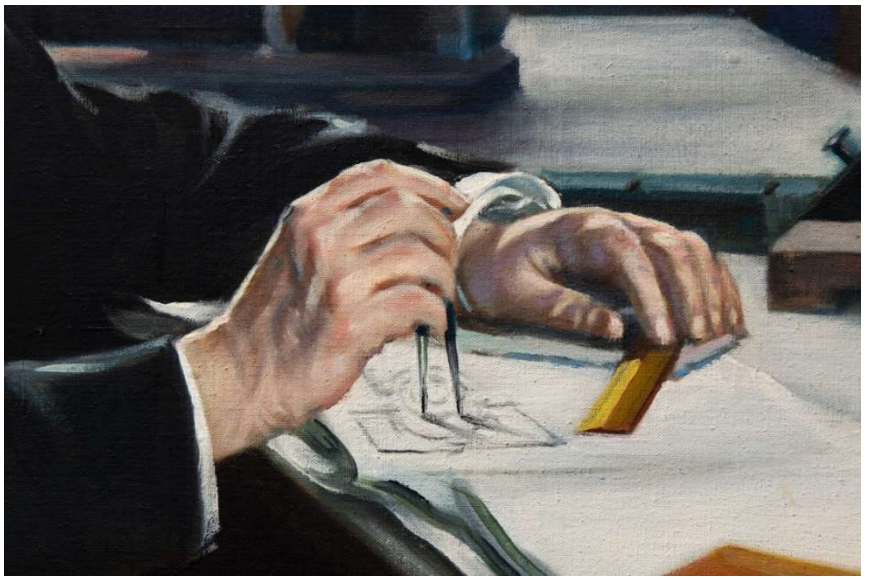
Esempi di errori ex ante possono essere il disporre di una lista non completa della popolazione da rilevare, oppure il costruire delle classificazioni errate per le variabili scelte, o utilizzare domande con una sintassi o una semantica “sbagliate”.

Per contenere questo tipo di errori è bene mettere in pratica alcune azioni **preventive** agendo sulle fonti (statistiche).

Il sistema di controllo della qualità

Correzioni cammin facendo.

Gli errori che possono emergere durante la rilevazione possono essere molti ed eterogenei, come l'errata comprensione della domanda da parte del rispondente o del rilevatore, o mancata risposta per distrazione o per volontà.



Per contenere questo tipo di errori è bene mettere in pratica alcune azioni di **monitoraggio** e di **correzione** degli errori nel corso del processo produttivo. Bisogna quindi sviluppare strumenti che cercano di individuare gli errori al momento in cui si verificano così da limitarne gli effetti sulle stime finali.

Il sistema di controllo della qualità

Una volta rotte le uova … devi fare la frittata!

Alla fine della rilevazione gli errori dei presentano spesso difficoltà ad essere recuperati. Per esempio se osserviamo molte mancate risposte, o se la classificazione risulta distorta o se il processo di elaborazione ha preso tanto tempo da diffondere dati ormai troppo “vecchi”.

Per contenere questo tipo di errori bisogna mettere in pratica delle azioni di **valutazione a posteriori**. Spesso queste azioni di valutazione dei dati statistici traggono forza essi stessi da analisi statistiche come la stima della varianza o i test di significatività che forniscono misure della (ignota) **distorsione** dovuta agli errori sulle stime finali.

Il sistema di controllo della qualità

Errore di campionamento: errore che deriva dal fatto di disporre di un campione di una popolazione. Più il campione differisce dalla popolazione da rappresentare, più le stime da esso prodotte potranno essere lontane dal parametro.

La teoria statistica del campionamento permette sia di contenere che di fornire una misura di tale errore.

Gli errori non campionari possono derivare da molteplici cause e sono più difficili da valutare e misurare. Inoltre si presentano anche utilizzando fonti amministrative (ovvero su popolazione teoricamente totali).

Si presenta nel seguito una elenco di tali errori.

Il sistema di controllo della qualità

Errori di specificazione:
Definizione: noto anche come validità di costrutto, si riferisce alla corrispondenza tra il piano della ricerca e la teoria di riferimento. Una ricerca è valida se si possono ragionevolmente escludere spiegazioni alternative dei dati rispetto alla teoria di riferimento.
Perché ciò sia possibile è necessario che il riferimento teorico sia chiaro ed univoco.

Correzioni:
Preventive:
a definire chiaramente il costrutto astratto che si vuole analizzare;
b verifica delle correlazioni tra i dati delle variabili che stiamo studiando e le variabili concettualmente simili;
c controllo della manipolazione (manipulation chek), che consiste nel verificare se la manipolazione sperimentale è effettivamente rappresentativa del costrutto ipotizzato.

Il sistema di controllo della qualità

Errori di lista o di copertura:
Definizione: in sovracopertura derivano da errate inclusioni di enti nella popolazione, o duplicazioni delle unità nella lista. in sottocopertura derivano da omissioni nell'inclusione

Correzioni:
Preventive : a) integrazione tra liste, b) aggiornamento della lista

Ex post: a) stima del tasso di sotto-copertura attraverso il confronto con altre fonti o attraverso Indagini di copertura

Il sistema di controllo della qualità

Errori di mancata risposta o non osservazione totale:

Definizione: tutte le informazioni riferite a una unità statistica risultano mancanti o insoddisfacenti. In caso di archivio amministrativo l'unità non vi risulta presente.

Correzioni:

Preventive : a) lettera di presentazione dell'indagine, garanzia sul segreto statistico, contatti preliminari con i rispondenti; b) rapporti e accordi con i fornitori dei dati amministrativi.

Monitoraggio: a) monitoraggio di indicatori di qualità di MRT per motivo; b) riponderazione per i non rispondenti

Ex post: a) stima della distorsione da MRT (es. studi di identificazione, indagini ad hoc su un sotto-campione di non rispondenti)

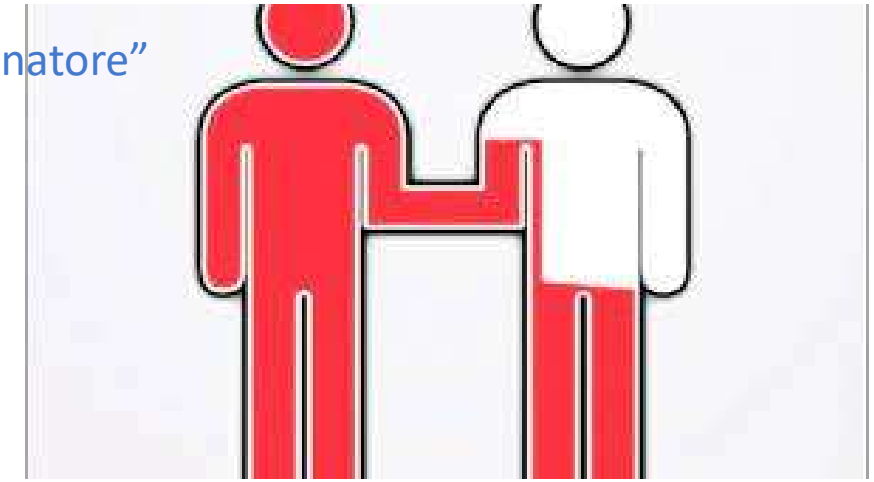
Il sistema di controllo della qualità

Errori di mancata risposta parziale:
Definizione: si ha quando un rispondente fornisce solo alcune, ma non tutte, le informazioni richieste. Da fonte amministrativa si potrebbe avere un dataset con valori mancanti.

Correzioni:
Preventive: a) lettera di presentazione dell'indagine, garanzia sul segreto statistico, contatti preliminari con i rispondenti.

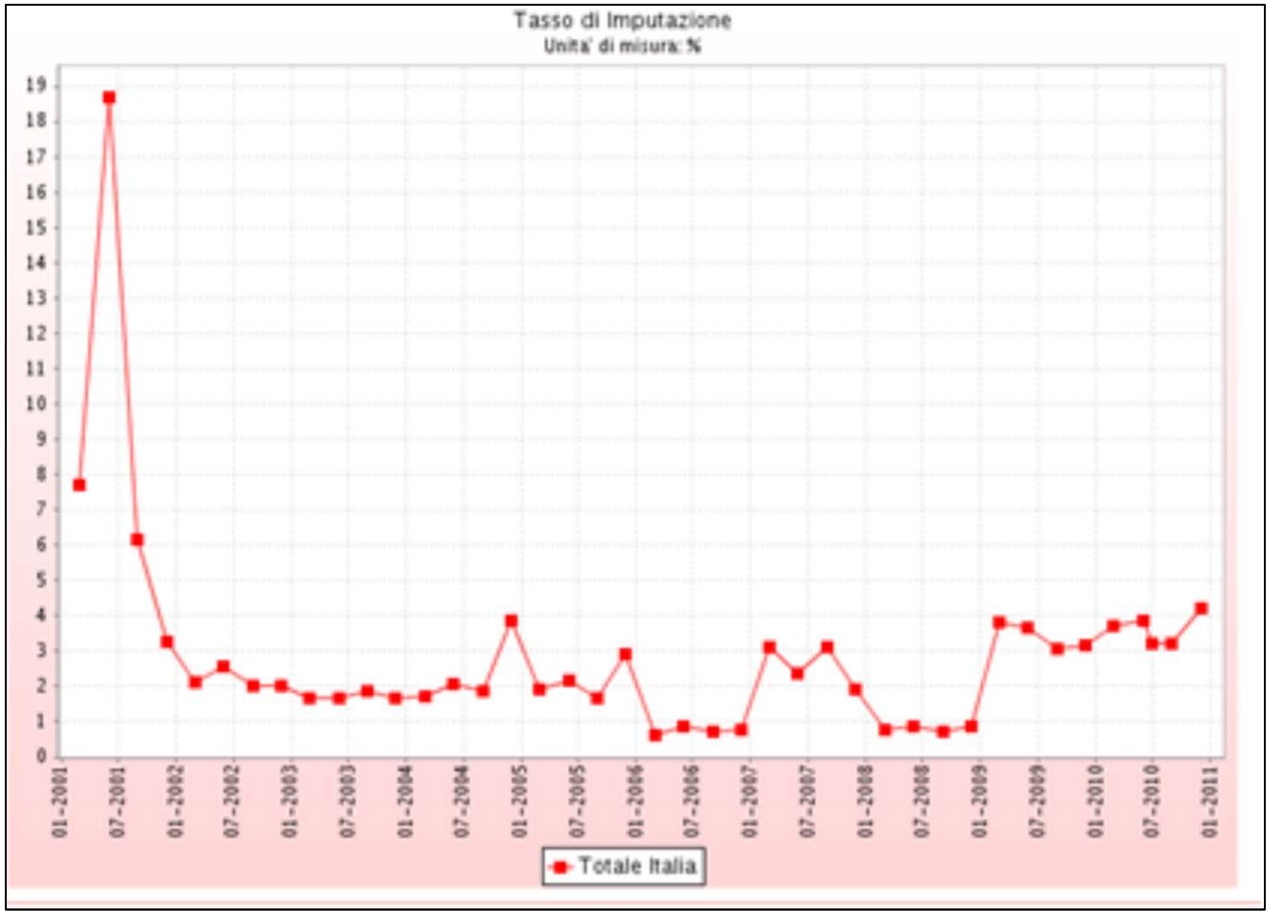
Monitoraggio: a) monitoraggio delle distribuzioni di risposta, delle mancate risposte parziali, dei «non so» e degli indicatori di qualità sul controllo e correzione; b) monitoraggio dei rilevatori

Ex post: imputazione della risposta da un "donatore"



Il sistema di controllo della qualità

Tasso di imputazione in % per una informazione mancante



Il sistema di controllo della qualità

Errori di misurazione:

Definizione: Differenza tra il valore «vero» di una variabile e quello osservato durante l'acquisizione del dato.

Correzioni:

Preventive : a) errata formulazione dei quesiti; b) errata scelta delle modalità di rilevazione.

Monitoraggio: Monitoraggio e supervisione dei rilevatori

Ex post: Controllo del dizionario di codifica

Il sistema di controllo della qualità

Errori di trattamento:

Definizione

Errori introdotti, dopo che il dato è stato acquisito, durante le operazioni per la sua elaborazione, come ad esempio nelle fasi di codifica o registrazione.

Correzioni:

Preventive : a) uso di chiavi di linkage affidabili; b) test degli strumenti (software per la codifica, controllo e correzione)

Monitoraggio: a) monitoraggio e supervisione dello staff; b) monitoraggio di indicatori su controllo e correzione

Ex post: Stimatori di calibrazione per garantire coerenza con totali noti della popolazione

Il sistema di controllo della qualità

Errori di diffusione:

Definizione

Le informazioni diffuse non sono pertinenti con le richieste degli utenti; non risultano chiare, tempestive o puntuali.

Correzioni:
Preventive : Impegno verso gli utenti sulle modalità di diffusione

Esistenza di un calendario delle diffusioni

Metadati a supporto dell'interpretazione e uso dei dati

Monitoraggio: Monitoraggio degli accessi

Analisi dei feedback degli utenti

Ex post: Valutazione della qualità del servizio attraverso consultazioni o indagini sulla soddisfazione degli utenti