

QUALITA' DEL DATO

Rilevazione statistica come processo di produzione di dati (micro,macro)

Qualità di un prodotto: adeguatezza del prodotto all'uso per il quale è stato realizzato ovvero capacità a soddisfare le proprietà garantite dal produttore

- *di **progettazione*** (es. forma, potenza,...)
- *di **tolleranza*** (limiti entro i quali i requisiti variano: es. durata minima di accensione di una lampadina)

Qualità del dato statistico:

<i>di progettazione</i>	<i>di tolleranza</i>
<ul style="list-style-type: none"> • rilevanza teorica: adeguatezza dell'informazione prodotta alle necessità informative • rilevanza effettiva: quanto dell'informazione prodotta viene effettivamente utilizzato • tempestività: intervallo di tempo fra l'emergere di esigenze informative e la disponibilità dei risultati • trasparenza: possibilità per l'utente di accedere a tutte le informazioni relative agli strumenti di indagine (definizioni, classificazione, questionario, ecc.) 	<ul style="list-style-type: none"> • precisione campionaria • precis. non campionaria (accuratezza): viene definita in termini di distanza fra il valore vero e quello osservato (errore di misura)

Fasi dell'indagine, operazioni e principali fonti di errore non campionario

Fasi	Operazioni	Fonti errore	Tipo errore
Progettazione	scelta unità, variabili	modello concettuale	rilevanza teorica
	redazione questionario	struttura del vocabolario, quesiti, codifica	errori di misura
	diffusione dei dati		rilevanza effettiva trasparenza
Rilevazione sul campo	formazione elenchi di unità	liste supervisorie rilevatori	errori di misura (selez. da mancata copertura)
	raccolta dati	supervisorie rilevatori rispondenti	errori di misura (mancate risposte, incongruenze)
Revisione	registrazione	operatori	errori di misura
	revisione	revisori programmi automatici	errori di misura (possibile selezione)
Elaborazione dati	elaborazione	programmi base statistica	errori di calcolo rilevanza effettiva
	diffusione		tempestività

Il controllo di qualità (accuratezza)

Obiettivi Prevenzione (fase progettazione)
Monitoraggio (fasi di rilevazione e di verifica)

Può essere realizzato mediante:

- 1) sistema di controlli affiancato all'indagine, che fornisce indicatori riferiti ad aspetti particolari dell'indagine (*non dà una misura sintetica dell'errore di misura*)
- 2) indagine sperimentale ad hoc, che consente una misura sintetica dell'errore ma non può affiancare l'indagine in modo continuativo ed esteso (*comporta aggravii economici ed organizzativi*).

In pratica:

la tecnica 2) viene applicata a un sottogruppo di unità dell'indagine principale, una tantum; gli indicatori in 1) sono calcolati con riferimento all'intera indagine, in modo continuativo.

Individuazione e correzione degli errori

- Tipi di controllo quantitativo
 qualitativo
- Revisione manuale (esperti di settore)
 automatica
 rapidità
 uniformità
 controlli complessi
- Correzione sul microdato
 sul macrodato

Il tutto è subordinato alla possibilità di individuazione dell'errore.

Controlli quantitativi

Verifica della coincidenza fra piano di rilevazione teorico ed effettivo

Piano teorico	Piano effettivo
nr. di rilevatori	nr. di rilevatori
nr. comuni (aree)	nr. comuni (aree)
nr. modelli rilev. inviati	nr. modelli ril. pervenuti nr. MTR (Mancanza Totale Risposta) nr. sostituzioni

esempio di controllo quantitativo mediante indagine ad hoc: la stima della copertura del censimento

Per l'indagine censuaria non esiste una numerosità di confronto.

L'Istat svolge un'indagine di controllo mediante un campione di sezioni di censimento nelle quali si individuano tutte le unità contenute. Si stimano quindi i dati della seguente tavola

unità	presenti al censimento	assenti al censimento	Totale
presenti all'indagine	N_{11}	N_{12}	$N_{1\bullet}$
assenti all'indagine	N_{21}	N_{22}	$N_{2\bullet}$
	$N_{1\bullet}$	$N_{2\bullet}$	N

- il confronto fra i risultati del censimento e quelli dell'indagine di controllo dà stime di N_{11} , N_{12} e N_{21} ;
- ipotizzando l'indipendenza fra le due indagini si perviene alla stima di N_{22}

$$(N_{22} / N_{12}) = (N_{21} / N_{11}) = (N_{2\bullet} / N_{1\bullet}) \Rightarrow N_{22} = N_{12} N_{21} / N_{11}$$

$$\text{tasso di copertura} = 100 (N_{21} + N_{11}) / N$$

Controlli qualitativi

1) Verifica informazione raccolta

- * *distribuzioni di frequenza*
- * *indici statistici*

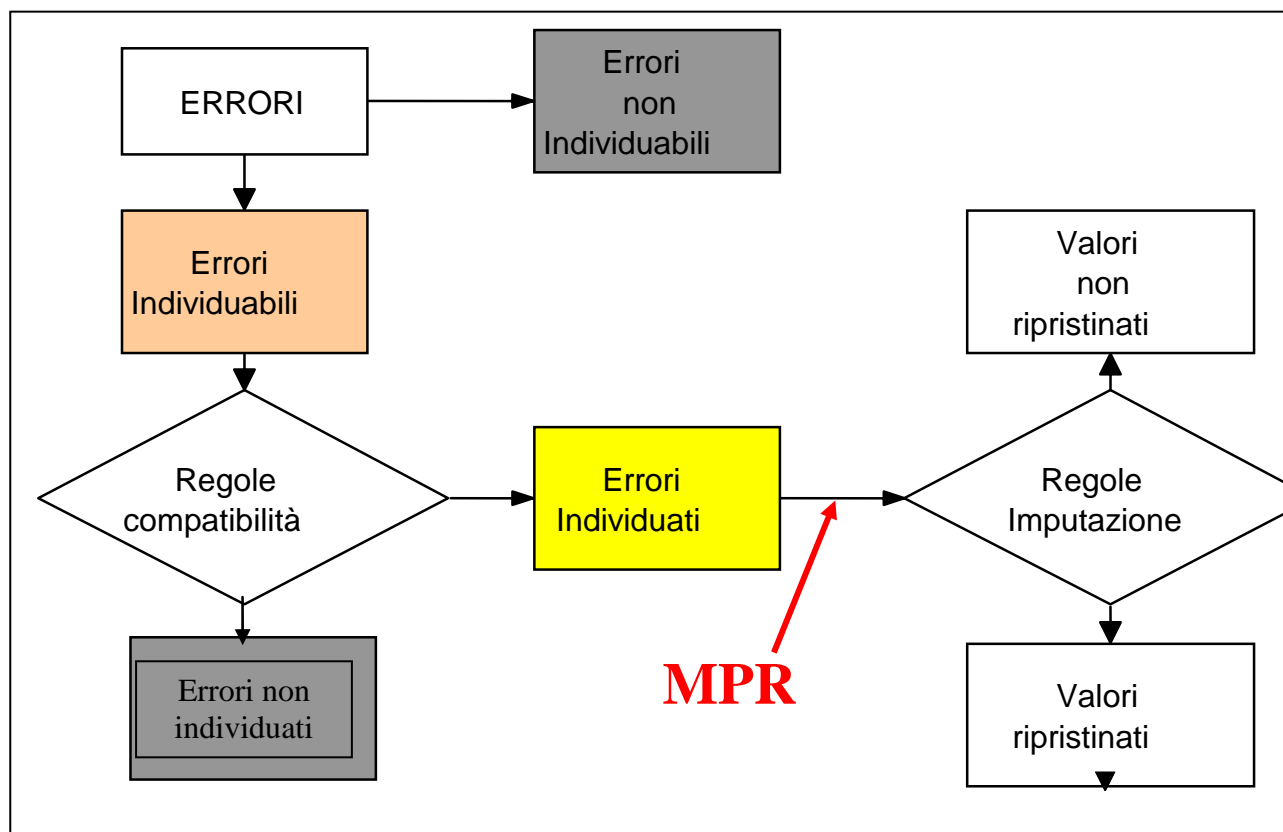
2) Individuazioni di errori sistematici

possono avere varie cause e varie manifestazioni (es. missing, non so)
non esiste una definizione operativa valida in generale che ne permetta l'individuazione

In generale:

si usano piani di compatibilità (programmi automatici)

Programmi di compatibilità e correzione (MPR)



Regole compatibilità

formali (*struttura questionario*)
 sostanziali (*informazioni a priori; es. campo di variazione plausibile per grandezze quantitative*)

devono essere

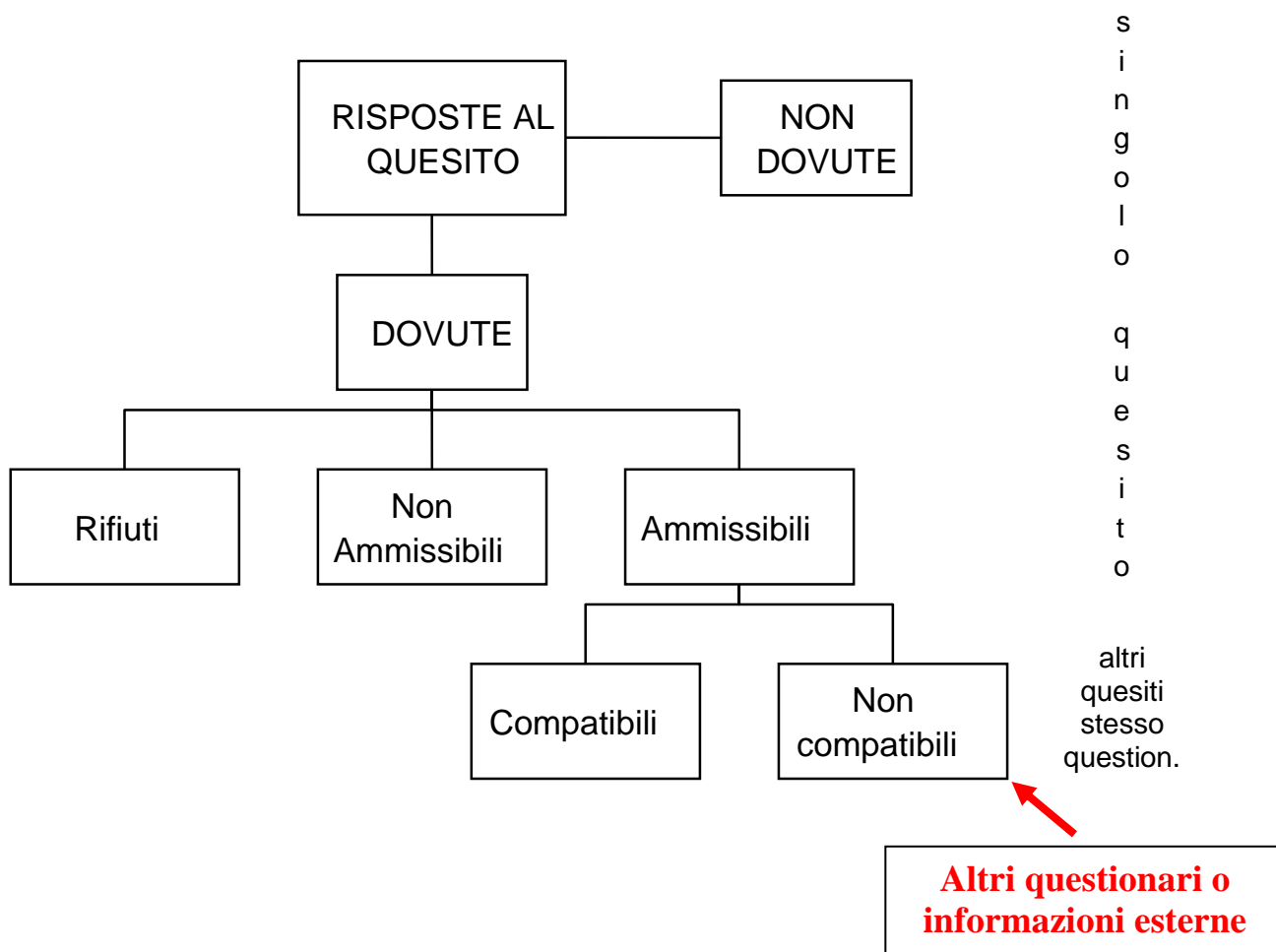
*non ridondanti
 non contraddittorie*

Errori individuati: Mancanza Parziale Risposta

*MPR vera e propria
 valori fuori campo
 incongruenze*

Trattamento della MPR

• Fasi per l'individuazione



• Tecniche di correzione (imputazione)

criteri deterministici *forzature a priori
regressione determ.*

criteri stocastici *da donatore*

Criteri deterministici

Forzature a priori

L'imputazione di valori predeterminati richiede la sequenzialità delle regole (ordinamento gerarchico) secondo l'importanza delle variabili e in modo da realizzare il minimo numero di modifiche

Esempio

SESSO= femmina

ETA' = 67 anni

CONDIZ. PROFES. = servizio di leva

SESSO= femmina

ETA' = 67 anni

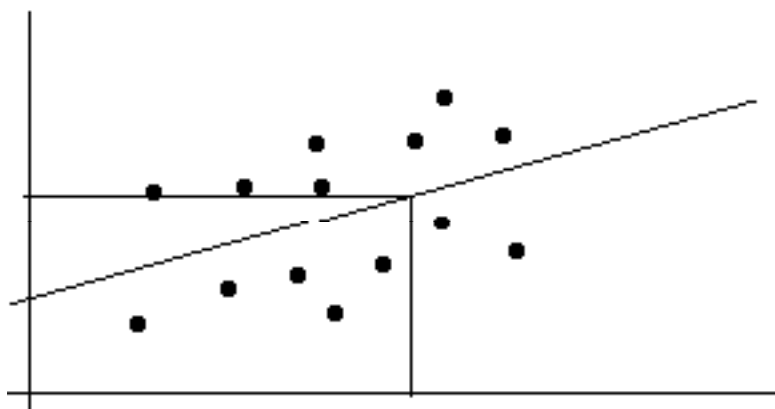
CONDIZ. PROFES. = ritirata dal lavoro (pensionata)

Regressione deterministica (variabili quantit.)

$y = b_0 + b_1 x$ ← modello stimato sui dati corretti

y_k : variabile da correggere

x : variabile nota per tutte le unità



Se Y MPR e $X=x_k \implies Y=y_k = b_0 + b_1 x_k$

Oss.: Se $b_1=0$, si imputa la media dei valori di y

Criteri stocastici

Da donatore

Il valore (modalità) errato viene corretto in base a quello di un'altra unità, scelta secondo certe regole.

Individuazione dell'unità donatrice:

si considera la distanza fra l'unità k da correggere e le possibili unità donatrici (insieme dei dati completi) in base ai valori di alcune variabili che devono essere noti per tutte le unità

distanza vettore di codici (var. qualit.)

misura sintetica (var. quantit.)
es. distanza euclidea

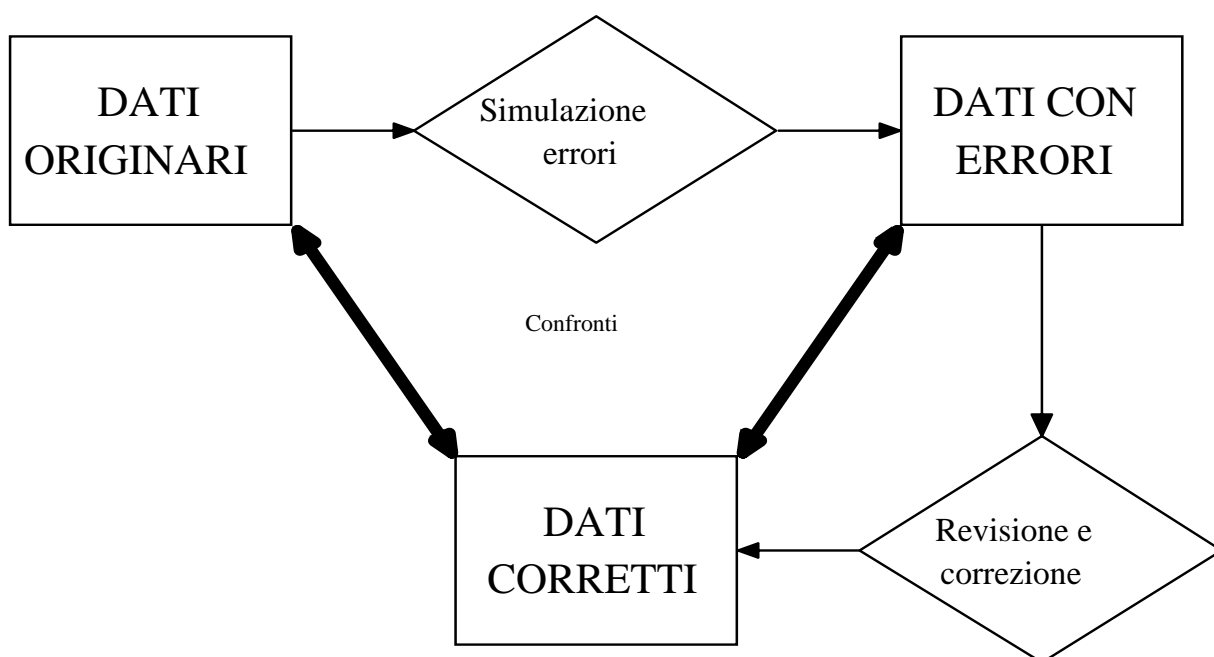
valore imputato codice del donatore
(grandezze qualitative)

valore del donatore "perturbato"
(grandezze quantitative)

NB: basarsi solo sui rispondenti equivale ad assumere che la mancata risposta è *random* (ad es. nel calcolo di una media, equivale ad imputare il valor medio ai non rispondenti)

Controllo del piano di compatibilità e di correzione

Prima di applicare un piano di revisione e di correzione è opportuno testarlo mediante simulazioni secondo il seguente schema



Confronti:

dati originari \longleftrightarrow dati corretti:

effetti degli errori e dei criteri di correzione

dati con errori \longleftrightarrow dati corretti:

effetti dei criteri di correzione

Per testare un piano di compatibilità è necessario simulare più tipi di errore