Dati mancanti

Breve introduzione

- In un'indagine campionaria le quantità d'interesse possono essere soggette a due tipi di errori:
 - campionari
 - non campionari.
- L'errore campionario è inversamente proporzionale alla dimensione del campione (e quindi nullo se l'indagine coinvolge l'intera popolazione obiettivo) ed è misurabile e controllabile a priori attraverso la teoria della probabilità.
- L'errore non campionario, invece, è direttamente proporzionale alla dimensione campionaria ed è inoltre positivamente influenzato dalla complessità dell'indagine, ma, anche per la sua origine diversa, non è misurabile.

- Gli errori di mancata risposta sono gli errori **non** campionari più frequenti.
- Essi sono legati all'idoneità cognitiva degli intervistati e alla loro volontà di fornire risposte vere, ma possono anche essere specifici della tecnica di rilevazione adottata ed infine possono essere riconducibili al comportamento degli intervistatori, laddove è previsto il loro intervento.
- L'errore di mancata risposta riduce in modo non prevedibile la numerosità campionaria e, qualora i non rispondenti siano diversi dai rispondenti, può portare a distorsioni nelle stime delle quantità di interesse.

- Gli errori di mancata risposta sono configurabili come:
 - Errori di mancata copertura
 - Errori di non risposta totale
 - Errori di non risposta parziale

- Gli errori di mancata copertura consistono nell'esclusione dalla lista di campionamento di alcune unità appartenenti alla popolazione obiettivo, che quindi hanno probabilità nulla di entrare a far parte del campione.

- Questa tipologia di errore è dovuta all'inadeguatezza qualitativa delle liste di campionamento, che possono risultare incomplete, non aggiornate o ridondanti (se sono presenti unità che non rientrano nella lista). Per esempio, in un'indagine telefonica che si avvale dell'elenco degli abbonati come lista di campionamento e che coinvolge le sole famiglie, risultano assenti tutte le famiglie non utenti (lista incompleta), possono essere erroneamente intervistate utenze d'affari (lista ridondante) o alcune famiglie possono mantenere il nominativo di un intestatario dell'abbonamento deceduto (lista non aggiornata).

- Gli errori di non risposta totale consistono nel non rispondere all'intero questionario e sono causati dal rifiuto a collaborare da parte dell'intervistato, dall'impossibilità di rintracciarlo anche dopo ripetute visite, dall'impossibilità di comunicare per problemi di lingua o malattia e dalla perdita del questionario da parte dell'intervistatore.
- In questo caso le sole informazioni a disposizione sul soggetto di interesse sono quelle usate per il campionamento e i motivi della mancata risposta.

- Pertanto gli effetti della mancata risposta totale sono piuttosto gravi soprattutto nel caso in cui il non rispondente differisca molto dai soggetti intervistati perché questo può causare distorsioni nelle stime dei parametri delle quantità di interesse.
- Gli effetti sono meno gravi nel caso di indagini longitudinali in cui il soggetto ha risposto ad almeno un'intervista, perché si dispone di informazioni sull'intervistato che possono essere utilizzate per compensare la mancata risposta.

- Gli errori di mancata risposta parziale, infine, consistono nel non rispondere ad uno o più quesiti del questionario.
- Le cause sono il rifiuto o l'incapacità di rispondere da parte dell'intervistato, la dimenticanza di porre una domanda o registrare una risposta da parte dell'intervistatore, la presenza di risposte inconsistenti eliminate durante la fase di editing ed errori in fase di registrazione delle risposte su supporto informatico.

- Gli effetti della mancata risposta parziale non sono gravi come quelli della totale perché in questo caso si hanno a disposizione delle informazioni sull'intervistato; tuttavia non si dispone di un dataset rettangolare e pertanto non è possibile applicare ai dati le analisi statistiche tradizionali.
- Inoltre la mancata risposta parziale riduce la numerosità campionaria, rendendo le stime delle quantità d'interesse meno efficienti, e, poiché si dispone di informazioni solo su un segmento della popolazione obiettivo, sono possibili distorsioni nei parametri d'interesse.

- È chiaro che il contenuto informativo sulle unità non rispondenti nei due casi di mancata risposta è molto differente dal momento che in quella totale si conoscono solo le informazioni di campionamento mentre in quella parziale si dispone anche di una parte dell'intervista.

- A questa differenza sulle informazioni disponibili corrispondono due metodi di imputazione diversi:
 - nel caso di mancata risposta totale si utilizzano metodi di ponderazione basati su informazioni ricavate dal disegno campionario, ad esempio quelle ricavate dalla lista di campionamento o dall'appartenenza ad un certo strato della popolazione;
 - nel caso di non risposta parziale si usano invece metodi di imputazione, che consistono nella sostituzione del valore mancante con uno scelto opportunamente, usando le informazioni disponibili sull'unità statistica in oggetto.

- Un'analisi dei dati mancanti richiede alcune assunzioni sulla distribuzione dei dati mancanti e in particolar modo su come le distribuzioni dei dati mancanti e di quelli osservati sono correlate.
- Si consideri allora una popolazione di numerosità N da cui si estrae un campione di numerosità n . Sul campione si osserva una variabile risposta Y che presenta m dati mancanti ed un insieme di variabili esplicative X osservate per tutte le unità.

Missing data

- A questo proposito si possono distinguere tre casi:
 1. il processo non dipende dai valori di X e Y , quindi i dati sono mancanti (MAR) e osservati a caso (OAR, Observed At Random); allora i dati sono detti mancanti completamente a caso (MCAR, Missing Completely At Random) In questo caso il campione incompleto può essere considerato un sottocampione casuale del campione originario;

Missing data

2. il processo che causa dati mancanti (e in particolare la probabilità che un particolare valore di Y sia mancante) non dipende dai valori mancanti di Y (ma solo dai valori osservati di X e Y); allora i dati si dicono mancanti a caso (MAR, Missing At Random);
3. il processo dipende dai valori osservati di Y ; allora i dati non sono mancanti a caso (NMAR).

Missing completely at random (MCAR)

- La probabilità che un certo valore per una data variabile sia missing è indipendente da qualsiasi altro valore per qualsiasi altra variabile, missing o osservata.
- Ovvero i valori missing values sono distribuiti aleatoriamente.
- Esempi
 - Esperimento
 - Avviamo più prove sperimentali con un ordine casuale. Concludiamole dopo 30 minuti.
 - Sondaggi
 - Chiediamo un numero di domande sorteggiate casualmente. Limitiamo al 30% del campione (casuale) un certo item.
 - Data entry
 - Inseriamo gli item in disordine sorteggiandoli dal mucchio. Fermiamo l'inserimento dati a intervalli aleatori.

Missing at random (MAR)

- La probabilità che un valore sia missing è collegata ai valori osservati disponibili
- Esempi
 - Le donne meno spesso indicano l'età
 - Una certa categoria di lavoratori più spesso evade le tasse
- MAR se sesso e categoria lavorativa sono osservati

Not missing at random (NMAR)

- La probabilità che un valore sia mancante è collegata ai valori mancanti
- Esempi
 - I sovrappeso meno spesso indicano il peso.
 - I ricchi il reddito
 - Gli appartenenti all'estrema destra il partito
- NMAR: problema grave

- Bisogna, inoltre, distinguere se il meccanismo generatore dei dati mancanti è **ignorabile** o **non ignorabile**.
- nel primo caso i dati sono di tipo MAR e i parametri del processo dei dati mancanti sono incorrelati con i parametri che devono essere stimati;
- nel secondo caso, invece, i dati non sono MAR.

- Se il meccanismo di non risposta è ignorabile non è necessario modellare il meccanismo dei dati mancanti come parte del processo di stima e i valori osservati possono essere considerati un sottocampione casuale del campione originario; altrimenti è necessario modellare il meccanismo dei dati mancanti per ottenere buone stime dei parametri di interesse.

- Le azioni da perseguire per il trattamento dei dati mancanti comportano spesso costi e tempi di realizzazione elevati: se il numero di dati incompleti o mancanti è poco rilevante, allora l'esiguità del problema non giustifica l'introduzione di complesse procedure correttive; per contro, la frequenza di tassi di mancata risposta più elevati evidenzia una non trascurabilità del fenomeno e delle conseguenze che esso comporta.

- Se è possibile identificare le variabili che spiegano il perché della mancanza (*mechanism variables*) allora si ha il controllo del meccanismo che tale mancanza genera.
- Le variabili che sono collegate più frequentemente al meccanismo di non risposta riguardano il livello d'istruzione, lo status sociale, il bisogno di approvazione o la motivazione: se il ricercatore ha sotto controllo tutte le variabili che possono spiegare il meccanismo di non risposta allora questo è trascurabile, nel senso che i risultati della ricerca non saranno soggetti a distorsioni per la presenza di non risposte

- Nessun procedimento analitico è in grado di affrontare il problema delle non risposte, e dei dati mancanti in generale, nella fase di raccolta dei dati; non essendo gli strumenti a disposizione dei ricercatori sufficienti a tenere sotto controllo il fenomeno delle mancate risposte, è necessario ricorrere a forme d'intervento successivo.
- **Obiettivo** non è quello di eliminare completamente il fenomeno, non potendo far affidamento su alcun metodo statistico che corregga tutte le differenze, ma ottenere una riduzione della non risposta.

- Le metodologie utilizzate per l'analisi dei dati parzialmente mancanti sono molteplici, e la varietà dei metodi proposti in letteratura è cruciale per operare la scelta più adeguata: ad ogni modo, tutti i metodi per il trattamento dei dati incompleti sono basati sull'assunzione che non è possibile verificarne la funzionalità utilizzando i soli dati a disposizione.

Le principali strategie proposte per il trattamento dei dati mancanti

- **Procedure basate sull'analisi delle unità completamente registrate:** l'espedito utilizzato è quello di non considerare nell'analisi le unità per le quali manca, in alcuni campi, la registrazione dei valori; le metodologie basate su tale procedura sono facili da attuare nel caso in cui si abbiano piccole quantità di dati mancanti, ma possono condurre a forti distorsioni per il fatto che alcune unità sono state eliminate.

Le principali strategie proposte per il trattamento dei dati mancanti

- **Procedure basate sull'imputazione:** quando i dati memorizzati sono mancanti o errati è possibile adottare delle tecniche d'imposizione di codici plausibili in modo da creare un set di dati completo che può poi essere analizzato con le tecniche standard; talvolta è possibile che tale procedimento, nel singolo caso, produca a sua volta un errore (poiché il codice "forzato" è plausibile ma non necessariamente vero), anche se i codici usati non minano l'accuratezza delle statistiche di sintesi.

Le principali strategie proposte per il trattamento dei dati mancanti

- **Procedure basate su modelli:** questa classe generale di procedure prevede che sia generato un modello per i dati parzialmente mancanti, il procedimento inferenziale è quindi basato sulla verosimiglianza sotto quel particolare modello.