DiPAV, 2007, 20, 51-56

I DATI MANCANTI

Emanuela Chemolli [*]
Margherita Pasini [**]

Sommario

Spesso i ricercatori si trovano di fronte al problema dei dati mancanti. Questo articolo presenta alcune informazioni utili a selezionare e applicare alcune modalità di trattamento dei dati mancanti.

Abstract

Researchers are commonly faced with the problem of missing data. This article presents some information for the selection and application of approaches for handling missing data.

1. Introduzione

«Non sempre i dati che raccogliamo possono essere utilizzati in quanto tali, e spesso dobbiamo "prepararli" in modo da poterli analizzare adeguatamente.» (Luccio, 2005, p. 97)

Ciascun ricercatore sa che ogni fase di una ricerca in psicologia va predisposta e attuata con la maggior cura e il maggior rigore possibili, sia che si tratti dell'analisi della letteratura, della predisposizione del disegno sperimentale, della procedura da seguire per la raccolta dei dati, delle tecniche attraverso cui analizzare i dati raccolti. Un aspetto molto delicato e a volte sottovalutato riguarda la verifica della qualità dei dati raccolti e le decisioni da prendere quando la base di dati sulla quale si intende proseguire con le analisi, come spesso accade, non è completa perché mancano dei dati. Questo breve lavoro si prefigge lo scopo di capire come possono essere

* Dottoranda in Psicologia dell'Organizzazione: processi di integrazione e di differenziazione, Dipartimento Psicologia e Antropologia Culturale, Università degli Studi di Verona. E-mail: emanuela.chemolli@formazione.univr.it; chemanu@hotmail.com
** Professore Associato in Psicometria, Dipartimento di Psicologia e Antropologia Culturale. E-mail: margherita.pasini@univr.it.

"trattati" questi dati mancanti. Ma perchè occuparsene? I dati mancanti sono veramente un problema? Perché perdere tempo sui dei dati che non ci sono? Quasi tutti i metodi statistici standard presumono che ogni caso della nostra matrice di dati abbia l'informazione su tutte le variabili e che solo i casi con tutte le informazioni possano essere inclusi nelle analisi. Per questo motivo il trattamento dei dati mancanti è un aspetto fondamentale dell'analisi dei dati.

Un buon punto di partenza è tentare di ridurre fin dall'origine al minimo i dati mancanti: la rilevazione va predisposta con cura e il processo di raccolta e di inserimento dati va costantemente monitorato. Per quanto tutto ciò richieda molto tempo, in realtà questo sarà ben compensato dalla riduzione del tempo che poi si dedicherà alla predisposizione dei dati per l'analisi. Cohen, Cohen, West e Aiken (2003) ricordano alcune regole base:

- sapere esattamente di quali informazioni si avrà bisogno per ogni variabile che si userà nell'analisi dei dati, prima di iniziare la raccolta dei dati;
- non porre agli intervistati domande alle quali spesso non saranno capaci di rispondere (regola ovvia, ma che viene violata fin troppo spesso);
- rivedere subito i dati raccolti così che non sia troppo tardi per tornare e ottenere le informazioni mancanti.

Seguire tali regole, tuttavia, non risolve completamente il problema. I dati possono essere mancanti per varie ragioni: se si tratta di un questionario autosomministrato, l'intervistato può decidere di non rispondere, oppure può essere nelle condizioni di non sapere cosa rispondere, o ancora può semplicemente dimenticarsi di rispondere. Può succedere anche che intervistatori opportunamente addestrati dimentichino di porre una domanda. Negli studi longitudinali, persone già intervistate una volta potrebbero non essere più reperibili nella seconda tornata di interviste. Anche la fase dell'inserimento dei dati può essere un momento che contribuisce al problema.

2. Domande preliminari al trattamento dei dati mancanti

Per decidere come trattare i dati mancanti è necessario tenere in considerazione alcuni fattori, in particolare quanti sono i dati mancanti, quanto numeroso è il campione e perché mancano. Certamente sarà diverso trattare i dati mancanti se questi sono solo una piccola percentuale (3% o meno) dei dati complessivi rispetto ad una percentuale di dati mancanti molto alta (più del 10%). Spesso si consiglia di fare una analisi di verificare la frequenza

dei dati mancanti sia per "variabile" che per "caso". Se il valore supera il 5% la situazione è critica e solitamente si toglie la variabile/il caso. Se il valore è inferiore al 5% si trova una soluzione. Ci sono situazioni in cui si mantiene la variabile/il caso anche se tale percentuale raggiunge il 6-7%.

Riguardo alla numerosità del campione, è importante considerare che alcune applicazioni statistiche sono preferibili per campioni grandi (maggiori di 200). Rispondere alla terza questione, ovvero al perché ci siano dati mancanti, è una questione più delicata. Spesso, forse per comodità e semplicità, si ritiene che le persone che hanno risposto ad una determinata domanda non siano diverse da quelle che non hanno risposto. I meccanismi che determinano la presenza di dati mancanti possono essere classificati in tre categorie (Rubin, 1976):

1. *valori mancanti completamente casuali (Missing Completely At Random, MCAR).* La probabilità di dati mancanti su una variabile non è collegata né al valore mancante sulla variabile, né al valore di ogni altra variabile presente nella matrice dati che si sta analizzando;
2. *valori mancanti casuali (Missing At Random, MAR)*. I valori mancanti sono indipendenti dal valore che viene a mancare, ma dipendono da altre variabili, cioè i dati sulla variabile sono mancanti per categorie di intervistati che potrebbero essere identificati da valori su altre variabili;
3. *valori mancanti non ignorabili (Missing Not At Random, MNAR)*. La mancanza di un dato può dipendere sia dal valore del dato stesso che dalle altre variabili. Per esempio, se si studia la salute mentale e le persone depresse riferiscono meno volentieri informazioni riguardanti il loro stato di salute, allora i dati non sono mancanti per caso.

3. La gestione dei dati mancanti

Il passo successivo dopo la definizione dei meccanismi è quello della gestione dei dati mancanti. Sostanzialmente le scelte possibili sono due: l'eliminazione dei casi o la sostituzione dei dati mancanti. Un metodo semplice, indicato solo nel caso in cui l'ammontare dei dati mancanti è limitato e questi sono mancanti completamente a caso (MCAR), è quello di cancellare i casi *(case deletion)*. I modi per eliminare i casi sono due: *listwise deletion* e *pairwise deletion.* Nel primo caso si elimina dal campione ogni caso che ha dati mancanti. Le analisi avverranno quindi solo sui casi che hanno valori validi per tutte le variabili in esame. Si ha una maggiore sem-

plicità di trattazione, tuttavia non si utilizza tutta l'informazione osservata (si riduce la numerosità campionaria e, quindi, l'informazione). Il secondo metodo è la *pairwise deletion*, che utilizza tutti i casi che hanno i dati validi su due variabili volta per volta. In questo modo si riesce a massimizzare la numerosità del campione da utilizzare, ma si tratta comunque di un metodo che presenta dei problemi, per esempio il fatto che con questo approccio i parametri del modello saranno basati su differenti insiemi di dati, con differenti numerosità campionarie e differenti errori standard.

Quando i dati non sono MCAR è opportuno sostituirli con appropriate funzioni dei dati effettivamente osservati (*imputation*). Di seguito sono indicati alcuni metodi.

1. *Mean Imputation*. Il dato mancante viene sostituito con la media della variabile. Questo metodo, utilizzato troppo spesso per la sua semplicità, riducendo la variabilità dei dati, ha invece effetti importanti su molte analisi dei dati e generalmente dovrebbe essere evitato.
2. *Regression Imputation*. Si tratta di un approccio basato sulle informazioni disponibili per le altre variabili. Si stima una equazione di regressione lineare per ogni variabile utilizzando le altre come predittori. «Questo metodo offre il vantaggio di poter utilizzare dei rapporti esistenti tra le variabili per effettuare le valutazioni dei dati mancanti; tuttavia esso è usato raramente, in quanto amplifica i rapporti di correlazione tra le variabili» (Di Nuovo, Di Nuovo & Buono, 2006, p. 182); quindi se le analisi si baseranno su regressioni, questo metodo è sconsigliato.
3. *Hot-Deck Imputation*. I dati mancanti vengono forniti da un "donatore", ovvero un caso privo di dati mancanti, scelto, generalmente entro la stessa base di dati, entro un insieme di casi simili al caso con dati mancanti. Secondo Di Nuovo et al. (2006) «la difficoltà di questo metodo sta nel definire il concetto di "simile"» (p. 182).
4. Howell (2007) scoraggia l'uso delle tecniche di *mean imputation, regression imputation, hot-deck imputation*. Secondo questo autore è importante conoscerle perché è importante scoraggiare il loro uso, ma ancor più importante è riconoscere che queste tecniche hanno condotto agli approcci moderni.
5. *Multiple Imputation*. La tecnica *multiple imputation*, applicabile in caso di MAR, prevede che un dato mancante su una variabile sia sostituito, sulla base dei dati esistenti anche sulle altre variabili, con un valore che però comprende anche una componente di errore ricavata dalla distri-

buzione dei residui della variabile. Forse, come afferma Howell (2007), uno dei problemi principali che ha rallentato l'uso del *multiple imputation* per anni è stata l'assenza di software semplici.

6. *Expectation-Maximization.* Un altro approccio moderno del trattamento dei dati mancanti è l'applicazione dell'algoritmo *Expectation-Maximization (EM).* Questo algoritmo, divenuto noto grazie ai lavori di Dempster, Laird e Rubin (1977) è stato via via migliorato (McLachlan e Krishnan, 1997; Meng e van Dick, 1997; Oakes, 1999). La tecnica è quella di stimare i parametri sulla base dei dati osservati, e di stimare poi i dati mancanti sulla base di questi parametri (fase *E*). Poi i parametri vengono nuovamente stimati sulla base della nuova matrice di dati (fase *M*), e così via. Questo processo viene iterato fino a quando i valori stimati convergono (Barbaranelli, 2003).

4. Conclusioni

La presenza di mancate risposte provoca dei problemi in fase di analisi dei dati, in particolare: perdita di efficienza delle stime causata dalla riduzione della dimensione campionaria dei dati completi (gli errori standard sono più elevati, gli intervalli di confidenza sono più ampi e quindi la potenza dei test statistici si riduce); possibile distorsione nelle stime in presenza di una mancata risposta sistematica, quando i rispondenti sono sistematicamente diversi dai non rispondenti; maggiore difficoltà per effettuare le analisi (i data set incompleti richiedono metodi complessi per la stima dei parametri che potrebbero non essere disponibili nei software statistici solitamente utilizzati per l'analisi dei data set completi).

I vari modelli di gestione dei dati mancanti possono produrre soluzioni radicalmente differente, pur sullo stesso campione. Le differenze di questi metodi dipendono da diversi fattori, incluso la quantità di dati mancanti e il tipo di processo che ha causato l'incompletezza dei dati.

I metodi decisamente migliori risultano essere la *Multiple Imputation* e *Expectation-Maximization.* Tuttavia, come sottolinea Allison (2001), entrambi i metodi richiedono l'assunzione i dati siano MAR, cosa non sempre garantita se i dati sono di natura psicologica; comunque si tratta di un assunto più debole rispetto all'assunto che siano MCAR. In via di principio, questi metodi possono anche essere usati per dati mancanti non ignorabili, ma questo richiede un modello corretto del processo per via del quale i dati

sono mancanti, e questo di solito è una cosa che è difficile da ricostruire. Allison (2002) ritiene anche possibile applicare tali metodi anche partendo dall'assunto che i dati non siano MAR, avvertendo però che ottenere buoni risultati è meno sicuro.

Bibliografia

Allison, P. D. (2002). Missing data. *Sage University Papers Series on Quantitative Applications in the Social Sciences*, *7*, 136-150. Thousand Oaks, CA: Sage.

Barbaranelli, C. (2003). *Analisi dei dati. Tecniche multivariate per la ricerca psicologica e sociale.* Milano: LED.

Cohen, J., Cohen, P., West, S. G., Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences. Third edition.* London: LEA.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*, 1-38.

Di Nuovo, A. G., Di Nuovo, S., Buono, S. (2006). Analisi con dati mancanti: confronto tra metodi statistici e un algoritmo di classificazione fuzzy. *TPM, 13*, 179-192.

Howell, D. C. (2007). The Treatment of Missing Data. In W. Outhwaite, P. Turner Stephen (a cura di) (2007), *Handbook of Social Science Methodology*. London: Sage.

Luccio, R. (2005). *Ricerca e analisi dei dati in psicologia. I. La raccolta dei dati.* Bologna: il Mulino.

MacLachlan, G. J., Krishnan, T. (1997). *The EM Algorithm and Extension.* New York: Wiley-Interscience.

Meng, X. L., Van Dyk, D. (1997). The EM Algorithm – an old folk song sung to a fast new tune. *Journal of the Royal Statistical Society*, *59*, 511-567.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society, 61*, 479-482.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *61*, 581-592.