5

I trattamenti preliminari dei dati

Prima di effettuare le analisi statistiche occorre procedere ad un controllo accurato dei dati disponibili allo scopo di verificare che essi possiedano le caratteristiche che li rendono idonei per le successive elaborazioni. Il problema della pulizia dei dati (in inglese *data cleaning*) riguarda (Zani and Cerioli, 2007):

1. i dati mancanti ed il loro trattamento;

2. valori anomali ed il loro trattamento.

5.1 I dati mancanti e strategie per il loro trattamento

Nelle analisi statistiche si verifica molto spesso di non rilevare alcune unità statistiche fra quelle originariamente programmate oppure di non possedere le modalità di alcuni individui tra i fenomeni rilevati. I dati mancanti (*missing values*) possono essere dispersi in tutta la matrice dei dati oppure possono comparire soltanto in una o poche variabili. Per superare le difficoltà generate dai dati mancanti si procedere in vari modi (Allison, 2001):

- condurre l'analisi solo sulle unità per le quali sono note tutte le modalità di tutti i fenomeni (*criterio listwise*);

- effettuare l'analisi (univariata) di ciascun fenomeno su tutte le unità per le quali si conoscono i dati dello stesso (*criterio columnwise*). Similmente si possono studiare le relazioni tra coppie di variabili con riferimento alle unità di cui sono noti i dati di entrambe (*crirerio pairwise*);

- ponderare i dati disponibili in modo tale da rappresentare i dati mancanti e/o stimare i valori mancanti.

Spesso è utile applicare una combinazione tra i diversi criteri. Ai fini delle analisi multivariate si possono fornire i seguenti suggerimenti per il trattamento dei dati mancanti:

1. quando il numero di unità statistiche con dati mancanti è modesto (minore del 5%) conviene adottare il criterio *listwise*, effettuando le analisi