Abstract:
Errors in data haunt practitioners in statistics and other fields. A practitioner who comes across a data value that is highly unlikely to be correct can either drop the value (or the whole observation) or replace the value by one or several imputed values. If there is an original recording, such as a paper questionnaire, it is worthwhile to check against this source in case there has been a processing error. It is sometimes possible to go further back and contact the individual or, in general terms, the object again, and make another observation. Then it is usually necessary to prioritise those objects that are most cost-effective to observe again. For that reason a prediction of error sizes of each variable of each unit is required to select a subsample of objects to observe again.
One approach is to predict the true value of one observation using available information and use the absolute value of the difference between the predicted and reported value. This is a number referred to as an *item score*. Usually we want to verify all observations on the same unit rather than each item separately. We discuss ways of forming a *unit score* out of a generic set of p item scores. Based on the unit score several subsample designs are available, e.g. cut-off sampling above a threshold. The problem of prioritising manual statistical editing of business survey data is our motivating example.