Statistica Neerlandica (2003) Vol. 57, nr. 1, pp. 1–2

Editorial

Incomplete data occur everywhere. If we draw a sample from a population, data will be missing for the population units that are not part of the sample. If we administer different tests to different subgroups, not a single respondent will have complete data. If we randomise a subject to the control group, the subject's outcome data under the experimental treatment will be unobserved. If we stop testing before all light bulbs have failed, some lamps will have censored lifetime data. If our measurement instruments change over time, data will be missing on both the old and the new instruments.

Over the years, creative statisticians have developed sound and practical solutions to deal with missing data problems. Many of these solutions now belong to the toolbox of the applied statistician and are not recognised (anymore) as incomplete data tools. Classic estimation of population quantities allows us to ignore the units that were not in the sample. Failure time models solve censoring problems by assuming a distribution of the unobserved data. Regression analysis predicts sensible values for units with missing outcomes. Many other standard methods can be viewed as solutions to missing data problems. Missing data are not merely a nuisance during data analysis. Incomplete data problems can and should inspire good scientific thinking.

This special issue of *Statistica Neerlandica* is concerned with *Incomplete data: multiple imputation and model-based analysis*. Significant computational advances over the last decade have helped to establish multiple imputation as a respectable and versatile approach to a broad variety of incomplete data problems. Model-based analysis refers to the situation where a specific model for the missing data is needed, i.e., if the missing data are not missing at random. Multiple imputation still works in this case, but the emphasis shifts to the specification of the model that created the missing data.

Some papers in this issue were prepared for the *10th Symposium on Statistical Software*, devoted to incomplete data, and held in Utrecht, The Netherlands on November 8, 2001. We felt that a selection of the conference proceedings merited publication and additionally invited contributions from other experts working on related topics. The papers were subsequently refereed and the present volume is the result of that review.

The first five papers all deal with aspects of multiple imputation in multivariate data under the assumption that the data are missing at random. **Rubin** deals with an intricate missing data problem in medical expenditures. He suggests applying iterative Bayesian methods to fill up only that part of the missing data that destroys the monotone pattern, and introduces nested imputation as a new method to break

up large imputation problems into manageable pieces. In the second paper, **Schafer** discusses properties of multiple imputation compared with maximum likelihood under various types of mismatches between the imputer's and the analyst's model, and carefully delineates the situations to watch out for. **Brand**, **Van Buuren**, **Groothuis-Oudshoorn** and **Gelsema** describe a practical approach for testing approximate properness of a given imputation method. **Kamakura** and **Wedel** propose an imputation procedure to deal with incomplete transaction databases in marketing, where they assume that the relations between the variables can be modelled by a factor model. **Rässler** deals with a statistical matching problem in marketing by imputation, comparing a new non-iterative method with a number of alternatives.

The next three papers reflect current work on incomplete longitudinal data where the missing data are informative, a very active research area in biostatistics. **Fitzmaurice** presents a review of current methods, covering both simple, inadequate fixes and more principled approaches. **Albert** and **Follmann** work out their transitional model of disease states under the assumption that the missing data are informative. **Molenberghs**, **Thijs**, **Kenward** and **Verbeke** discuss various strategies to perform sensitivity analyses of longitudinal data, using the pattern-mixture model as their point of departure. Finally, **Groeneboom** and **Jongbloed** deal with the problem of estimating a probability density function from data that are corrupted by uniform noise. Only the corrupted data are observable here, whereas the uncorrupted data are missing.

The papers reflect state-of-the-art approaches to the analysis of incomplete data in areas such as marketing, drug research, tobacco litigation, epidemiology and public health. Bringing this work together in a single volume will hopefully stimulate cross-fertilisation across these and other fields. As editors of this special issue, we wish to express our appreciation to the contributors for their effort in preparing the papers and to the reviewers for their sometimes detailed suggestions on improvement. We heartily invite the reader to study the papers.

Stef van Buuren and Rob Eisinga