Problem of Missing Data in Census - Who Are the Non-Response Respondents?

Jan Hora , Faculty of Nuclear Science and Physical Engineering, Czech Tech. University

Jiří Grim, Institute of Information Theory and Automation, AS CR

Key words: Census information, statistical model, missing data, non-response substitution.

The confidentiality of census data is known to be rather restrictive for economic and social research. To improve the availability of census information we have proposed recently a new method of interactive presentation of census results by means of statistical models. The method is based on estimation of the joint probability distribution of data records in the form of a distribution mixture. The estimated mixture model can be used as a knowledge base of a probabilistic expert system and in this way we can derive the statistical information from the distribution mixture without any further access to the original database. The statistical model does not contain the original data and therefore the final interactive software product can be made freely available via internet without any confidentiality concerns.

The method of interactive statistical model has been applied to the 2001 Czech Census data with the aim to verify its applicability to the next census in 2011 and to analyze the arising new possibilities of treatment of missing data. The considered source database of 10,230,060 records included about 15% of incomplete questionnaires; the total number of missing values was 2,933,427. Since the missing values are identified as non-response, they can be simply included as additional response alternatives. We remark that by estimating the corresponding “extended” statistical model we have a unique possibility to analyze the statistical properties of the “non-response” respondents. Another approach is to modify the EM algorithm to be directly applicable to incomplete data. The “incomplete data” model can be used to estimate the missing values and to compute the statistical model again by using the substituted complete data. In experiments we succeeded to correctly identify about 73% of missing values on the average. In the final tests the accuracy of the “extended” and “substituted” statistical model was comparable; the mean error of displayed histogram columns was about 4%.