ON BALANCED SAMPLING AND CALIBRATION ESTIMATION IN SURVEY SAMPLING

Risto Lehtonen
University of Helsinki, Finland, risto.lehtonen@helsinki.fi

Jaroslav Hajék (1959) introduced an important concept or *representative strategy* as a joint property of an estimator $e(\omega)$ and sampling design $P = \{P(s), s \in S\}$. The pair (e, P) is said representative with respect to numbers x_i if $P(e(s,x) = X) = 1$ i.e. if the estimate applied to the population $(x_1, \dots, x_N)$ estimates the total without error with probability 1 (Hajék 1981, p. 40). In other words, the equation

$$\sum_{i \in s} x_{ki} w_i = \sum_{i=1}^{N} x_{ki} \ (1 \le k \le m), \qquad (1)$$

where $w_i = w_i(s)$ are weights, should hold for any sample s that may be selected under the sampling design *P*. Hajék showed that representativeness is one of the conditions of optimality and simplifies the expression of the mean squared error. I discuss representativeness under the design-based (randomization) framework. The property of strategy representativeness can be obtained if (1) with appropriate weights holds for the sampling design or the estimation design. In balanced probability sampling (Deville and Tillé 2004), inclusion probabilities π_i are derived that fulfill the requirement of $\sum_{i \in s} x_{ki}/\pi_i = \sum_{i=1}^{N} x_{ki}$. In calibration estimation (Deville and Särndal 1992), calibration weights w_i are derived that meet the calibration (balancing) equations (1). Many strategies common today can be reviewed under the framework of representativeness. A (non-calibrated) linear estimator for the total of *y*, such as Horvitz-Thompson estimator $\hat{t}_{HT} = \sum_{i \in s} y_i/\pi_i$, applied to a sampling design *P* balanced on the auxiliary variables x_k, is a representative strategy w.r.t. x_k. A calibration estimator $\hat{t}_{CAL} = \sum_{i \in s} w_i y_i$ of the total of *y* applied to a (non-balanced) sampling design *P* also is a representative strategy. Accuracy benefits from these strategies are expected if the auxiliary x-variables correlate with the target variable *y*. In the strategies touched this far, statistical models only appear implicitly. In penalized balanced sampling (Breidt and Chauvet 2012) and penalized calibration (Guggemos and Tillé 2010), some of the balancing or calibration constraints in (1) are relaxed in a controlled way. In these methods, explicit statistical modelling e.g. with linear mixed models plays an important role Models such as members of the generalized linear mixed models family enter on the scene also in some other recent approaches, e.g. generalized calibration (Deville 2000) and model calibration (Wu and Sitter 2001; Lehtonen and Veijanen 2009). Based on selected literature, we discuss the properties of the various representative strategies and present some numerical examples.

References

Breidt, F.J. and Chauvet, G. (2012) Penalized balanced sampling. Biometrika, 99, 945–958.

Deville, J.-C. (2000) Generalized calibration and application to weighting for non-response. In: Bethlehem J.G. and van der Heijden, P.G.M. (eds) COMPSTAT. Physica, Heidelberg.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87, 376–382.

Deville, J.-C. and Tillé, Y. (2004) Efficient balanced sampling: The cube method. Biometrika, 91, 893–912.

Guggemos, F. and Tillé, Y. (2010) Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. Journal of Statistical Planning and Inference, 140, 3199–3212.

Hajék, J. (1959) Optimum strategy and other problems in probability sampling, Casopis pro Pestováni Matematiky, 84, 387–423.

Hajek, J. (1981) Sampling from a Finite Population. New York: Marcel Dekker.

Lehtonen, R. and Veijanen, A. (2012) Small area poverty estimation by model calibration. Journal of the Indian Society of Agricultural Statistics, 66, 125–133.

Wu, C. and Sitter, R.R. (2001) A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association, 96, 185–193.