SOME APPLICATIONS OF WEIGHTING WITH BASCULA

Robert Göttgens
Netherlands Central Bureau of Statistics, Heerlen, The Netherlands

1. Introduction

More and more, statisticians use microcomputers for processing surveys. Since 1987 the Netherlands Central Bureau of Statistics (CBS) is developing the Blaise System for integrated survey processing, see e.g. Bethlehem (1991). Data collection, data editing, tabulation, and analysis can be carried out with this system on a microcomputer or on a network of microcomputers. In 1991 a new tool was added to this system: Bascula, a package for weighting sample survey data.

Survey data, obtained after data collection and data editing, are usually not ready yet for making inference about the population from which the sample has been drawn. The problem is that the data do not constitute a representative sample due to unequal selection probabilities and non-response. In order to correct for these effects, often adjustment weights are computed. Post-stratification is a well-known technique. Every record is assigned some weight, and these weights are computed in such a way that the weighted sample distribution of characteristics like sex, age, marital status, and area reflects the known distribution of these characteristics in the population. Two major problems can make application of post-stratification difficult: empty strata and lack of adequate population information. Research has been carried out at the CBS in order to improve weighting techniques. The result was a new general method for weighting, in which weights are obtained from a linear model which relates the target variables of a survey to auxiliary variables. Post-stratification is a special case of this method. Because of the generality of the method, different weighting schemes can be applied that take advantage of the available population information as much as possible, and at the same time avoid the above mentioned problems (Bethlehem and Keller, 1987).

Bascula is a general weighting package, running on microcomputers under the operating system MS-DOS. It combines several weighting techniques. In the first place, traditional post-stratification can be carried out. And if the number of empty strata is small, one can instruct the program to collapse (i.e. combine) these strata with neighbouring strata. In the case of many empty strata, or lack of sufficient population information, Bascula can carry out the linear weighting technique described above or apply multidimensional iterative proportional fitting (also called multiplicative weighting, or raking ratio estimation).

The user specifies the auxiliary information to be used for the weighting in the form of a model. This model will be confronted with the available sample and population information. If the specified model cannot be applied, Bascula will propose a simpler model. Bascula will carry out a complete post-stratification if possible. If not, the user has to decide either to carry out linear or multiplicative weighting. So if maximum auxiliary information is not available, or if it would result in unstable estimates, one has to choose between linear weighting and multiplicative weighting.

Three different weights are distinguished by Bascula: the inclusion weight, which can be attached to each separate record, the correction weight computed by Bascula and the final adjustment weight, the product of the former two. The resulting final adjustment weights can either be added to the data file, or be stored in a separate file.

It is possible to derive new variables or change the contents of existing variables by making use of the build-in recode facility. The options available in this facility are a subset of the options in a more general file manipulation package called Manipula (Hofman et al., 1990). Manipula can be used if the recoding facility in Bascula is not sufficient.

2. Some applications of weighting with Bascula on CBS-statistics

2.1. The National Voters Survey (NVS)

Before and after the parliamentary elections of 6 September 1989, a survey took place into voting behaviour and backgrounds. This National Voters Survey combines background items with questions scaling opinions on a variety of party-political subjects. To analyse characteristics of the Dutch electorate, which influence party choice, a structural equation model was developed (Schmeets 1991).

The NVS 1989 consisted of two rounds: one interview before and one interview after the election. Eventually 1506 persons were interviewed face-to-face with the aid of a Blaise CAPI program, which had as major advantage the possibility of cleaning the data during the fieldwork. The response being low, a weighting procedure was needed to correct the response pattern for the distribution of voting behaviour, age, sex, marital status, degree of urbanisation and region in order to enhance as much as possible the statistics on voting behaviour and attidudes. So the following characteristics were used to make different weighting models: voting behaviour (8 categories), age (7 categories), sex (2 categories), marital status (4 categories), degree of urbanisation (4 categories) and region (4 categories). By means of these auxiliary information the following weighting models were constructed and evaluated (Schmeets and Molin, 1990):

1. Sex + Age + MariStat + DegOfUrb + Region
2. Sex x MariStat + Sex x Age + DegOfUrb + Region
3. Sex x MariStat + Sex x Age + DegOfUrb x Region
4. VotBehav + Age + Sex + MariStat + DegOfUrb + Region
5. VotBehav
6. VotBehav + Age

The notation used in these models has some analogy with analysis of variance. An 'x' means that variables are crossed, i.e. the population distribution in the table obtained by crossing the variables is used for the computation of weights. A '+' means that only marginal distributions are used. For example, the model 'Sex x MariStat + Sex x Age' means that the two population tables of Sex by Marital status ans Sex by Age are used, but not the three-way table Sex by Marital status by Age.

For all these weighting models the weights were calculated using the additive weighting method. The different results of the models were compared and after extensive analysis and discussion one decided finally to use the fourth model. It was very evident that one of the major advantages of Bascula during the weighting of this sample turned out to be the flexibility and fastness of selecting and calculating the different weighting models. In earlier years of the NVS one had to specify all desired weighting models in writing and offer this to the statistical methods department. So tailor-made software had to be made for each weighting model. Bascula needs only a few minutes to calculate a weighting model as described above (n = 1506, on a 386 machine), so it saves both time and much programming efforts.

2.2. The Family Expenditure Survey (FES)

The CBS has conducted an annual Family Expenditure Survey ever since 1978, representative for all households in the Netherlands. In the FES 1991 eventually 2859 households were interviewed face-to-face with the laptop computers using a Blaise CAPI program and paper and pencil questionnaires. Like in the NVS described above, to attain a good representative sample of the population one needs to apply some weighting method. In this survey the following characteristics were used as auxiliary information: the income of the household, the number of persons in the household, the socio-economic category and sex of the head of the household and the share of households living in owner-occupied dwellings. By means of these characteristics some 'new' auxiliaries were defined: a combination of income and the size of the household IncSize with 13 categories and a combination of income of the household and socio-economic category of the head of the household IncSec with 37 categories. With this auxiliary variables the following weighting scheme was used:

HireOwn + IncSize + Sex + IncSec

In this case, for two reasons the chosen weighting method concerned multiplicative weighting: first the final adjustment weights should always be positive, which is guaranteed in the multiplicative weighting method. The second reason concerned the possibility of defining lower and upper weight limits in the multiplicative weighting module: by doing so it is guaranteed that the final adjustment weights lie within the range defined by the lower and upper weights.

References

Bethlehem, J.G. and W.J. Keller, 1987, Linear Weighting of Sample Survey Data. Journal of Official Statistics 3, pp. 141-154.

Bethlehem, J.G., A.J. Hundepool, M.H. Schuerhoff and L.F.M. Vermeulen, 1989, Blaise Blaise 2.0 / Language Reference Manual. CBS report, Netherlands Central Bureau of Statistics, Voorburg.

Bethlehem, J.G. and W.J. Keller, 1991, The Blaise System for Integrated Survey Processing. Survey Methodology Journal 17, pp. 43-56.

Göttgens, Robert, Bert Vellen, Marly Odekerken and Lon Hofman, 1991, Bascula, version 1.0 / A weighting package under MS-DOS. User Manual. Netherlands Central Bureau of Statistics, Voorburg, The Netherlands.

Hofman, L.P.M.B., H.W.J.M. Vellen and J.H.A.M. Meertens, 1990, Manipula, Version 1.0 / File manipulation under MS-DOS, User Manual. CBS report, Netherlands Central Bureau of Statistics.

Schmeets, J.J.G., 1991, Party choice in the Netherlands: Model, weighting and political interest, Proceedings of the twelfth meeting of the 'Working group structural equation modelling' at the Netherlands Central Bureau of Statistics in Heerlen, 4 and 5 april 1991.

Schmeets, J.J.G. and E.J.E. Molin (1990): Constructie en evaluatie van weegmodellen in kiezersonderzoek, CBS report, Netherlands Central Bureau of Statistics, Heerlen.