Bascula : Current Status and Future Developments

Jelke Bethlehem, Statistics Netherlands

1. Introduction

1.1. About errors in surveys

Our complex society experiences an ever growing demand for statistical information relating to social, demographic, industrial, economic, financial, political, and cultural situation of the country. Such information enables policy makers and others to take informed decisions for a better future. Sometimes, statistical information can be retrieved from administrative sources. More often there is a lack of such sources. In that case, the sample survey is a powerful instrument to collect new statistical information.

A sample survey collects information on only a small part of the population. In principle, this sample only provides information about the selected elements of the population. Nevertheless, if the sample is selected using a proper sampling design, it is also possible to make inference about the population as a whole. Data about the sample elements can be used to compute estimates of population characteristics.

Estimates will never exactly equal to the population characteristics to be estimated. There will always be some error. This error can have many causes. Two broad categories can be distinguished: sampling errors and non-sampling errors.

Sampling errors are introduced by the sampling design. They are due to the fact that estimates are based on a sample and not on a complete enumeration of the population. The sample is selected by means of a random selection procedure. Every new selection of a sample will result in different elements, and thus in a different value of the estimator. The magnitude of the sampling error can be controlled through the sampling design. For example, by increasing the sample size, or by taking selection probabilities proportional to some well chosen auxiliary variable, the error in the estimate can be reduced.

Non-sampling errors occur even if the whole population is investigated. Non-sampling errors are errors made during the process of recording the answers to the questions. An important source of non-sampling errors is

non-response. It is the phenomenon that sample elements do not provide the required information. There may be various reasons for this: refusal to co-operate, not at home at the time of the visit of the interviewer, or not able to co-operate due to illness or other circumstances.

Due to non-response the sample size will be smaller than planned. This may result in increased variances of population estimates. Another, more serious, consequence is that non-response may be selective. This happens if non-response causes some groups in the population to be over- or under-represented in the sample, and these groups behave differently with respect to the population characteristics to be investigated. For example, if people with high incomes refuse to co-operate in the survey, the estimate of the mean income in the population will be too low.

1.2. Weighting sample surveys

There is ample evidence that non-response often causes estimates to be biased. This means that something has to be done to correct for this bias. A frequently used technique is *weighting*. Weighting is based on the use of *auxiliary information*. Auxiliary information is defined as a set of variables that have been measured in the survey, and for which information on the population distribution is available. By comparing the population distribution of an auxiliary variable with its sample distribution, it can be assessed whether or not the sample is representative for the population (with respect to this variable). If both distributions differ considerably, one must conclude that non-response has resulted in a selective sample.

The auxiliary information can also be used to compute adjustment weights. Weights are assigned to all records of observations. Estimates of population characteristics can now obtained by using the weighted values instead of the unweighted values. The weights are defined in such a way that population characteristics for the auxiliary variables can be computed without error. Then the weighted sample is said to be *representative* with respect to the auxiliary variables used.

If it is possible to make the sample representative with respect to several auxiliary variables, and if these variables have a strong relationship with the phenomena to be investigated, then the (weighted) sample will also be (approximately) representative with respect to these phenomena, and hence estimates of population characteristics will be more accurate.

The Blaise environment contains a tool for computing adjustment weights. It is the program Bascula. Using the file with sample data and a file with the population distribution of auxiliary variables, Bascula can compute these adjustment weights using a number of different techniques. The most simple and straightforward one is *post-stratification*. Due to lack of a sufficient amount of population information, or empty or small sample cells, post-stratification is not always applicable. For such a situation, Bascula offers two alternatives: *linear weighting* and *multiplicative weighting*.

The purpose of this paper is to describe the current status and the future developments of Bascula. Section 2 gives an overview of the theoretical

background. Section 3 describes how Bascula can be used within the Blaise environment. Section 4 gives discusses some future developments that will lead to a new, enhanced version of Bascula.

Bascula is being developed by two departments of Statistics Netherlands. The Department of Statistical Methods is responsible for the theoretical framework, and specific weighting algorithms. The Statistical Informatics Department is in charge of the user-interface and the integration in the Blaise system.

2. Weighting techniques

2.1. Basic concepts of sampling

In order to be able to explain the theoretical background of weighting, first some basic definitions and notations of sampling theory are introduced.

Let the finite *target population* U consist of a set of N identifiable elements, which may be labelled 1,2,...,N. Associated with each element i is a unknown value Y_i of the *target variable*. The vector of all values of the target variable is denoted by

$$Y = (Y_1, Y_2, ..., Y_N)' \qquad (2.1.1)$$

Objective of the sample survey is assumed to be estimation of the population total :

$$Y_T = \sum_{i \in U} Y_k \qquad (2.1.2)$$

To estimate this population parameter, a sample of size n is selected. It is assumed throughout this paper that samples are selected without replacement. In this case, the sample can be represented by a subset s of U ($s \subset U$). The sample size is equal to the number of elements in s. The *first order inclusion probability* of element i in the target population is defined as :

$$\pi_i = P(i \in s) \qquad (2.1.3)$$

It is the probability that the sample s contains element i. In case of a simple random sample, all inclusion probabilities are equal to n/N.

A straightforward way to estimate the population total of the target variable is to apply the estimator defined by Horvitz-Thompson (1952) :

$$y_{HT} = \sum_{i \in S} \frac{Y_i}{\pi_i} \qquad (2.1.4)$$

The Horvitz-Thompson estimator (2.1.4) is an unbiased estimator of the population total. This estimator does not make any use of auxiliary information. However, such information can be used to improve

estimators. Bascula offers three ways of incorporating auxiliary information in the estimation procedure. These three approaches are post-stratification, linear weighting and multiplicative weighting. They will be discussed in the following subsections.

2.2. Post-stratification

Post-stratification is a well known and often used weighting method. To be able to carry out post-stratification, one or more qualitative auxiliary variables are needed.

Suppose, we have an auxiliary variable X having H categories. So it divides the population into H strata. The strata are denoted by the subsets $U_1,U_2,...,U_H$ of the population U. The number of population elements in stratum U_h is denoted by N_h, for h=1,2,...,H. The population size N is equal to $N=N_1+N_2+...+N_H$.

A sample of size n is selected from the population. The set of sample elements in stratum h is denoted by s_h. If n_h denotes the number of sample elements in the sub-sample s_h (for h=1,2,...,H), then $n=n_1+n_2+...+n_H$. Note that the values of the n_h are the result of a random selection process. So, they are random variables.

Post-stratification assigns identical correction weights to all elements in the same stratum. The correction weight d_i for an element i in stratum U_h is in its most general form defined by

$$d_i = \frac{N_h}{\sum_{j \in s_h} \frac{1}{\pi_j}} \qquad (2.2.1)$$

where the sum is taken over all sample elements j in the stratum U_h, i.e. all elements in sub-sample s_h. In case of a simple random sample, all inclusion probabilities π_i are equal to n/N, and the correction weight d_i for an element i in stratum U_h reduces to

$$d_i = \frac{N_h n}{N n_h} \qquad (2.2.2)$$

If the values of the inclusion probabilities and correction weights are imputed in expression (2.2.1), the result is the well known post-stratification estimator

$$y_p = \sum_{h=1}^{H} N_h \bar{y}_h \qquad (2.2.3)$$

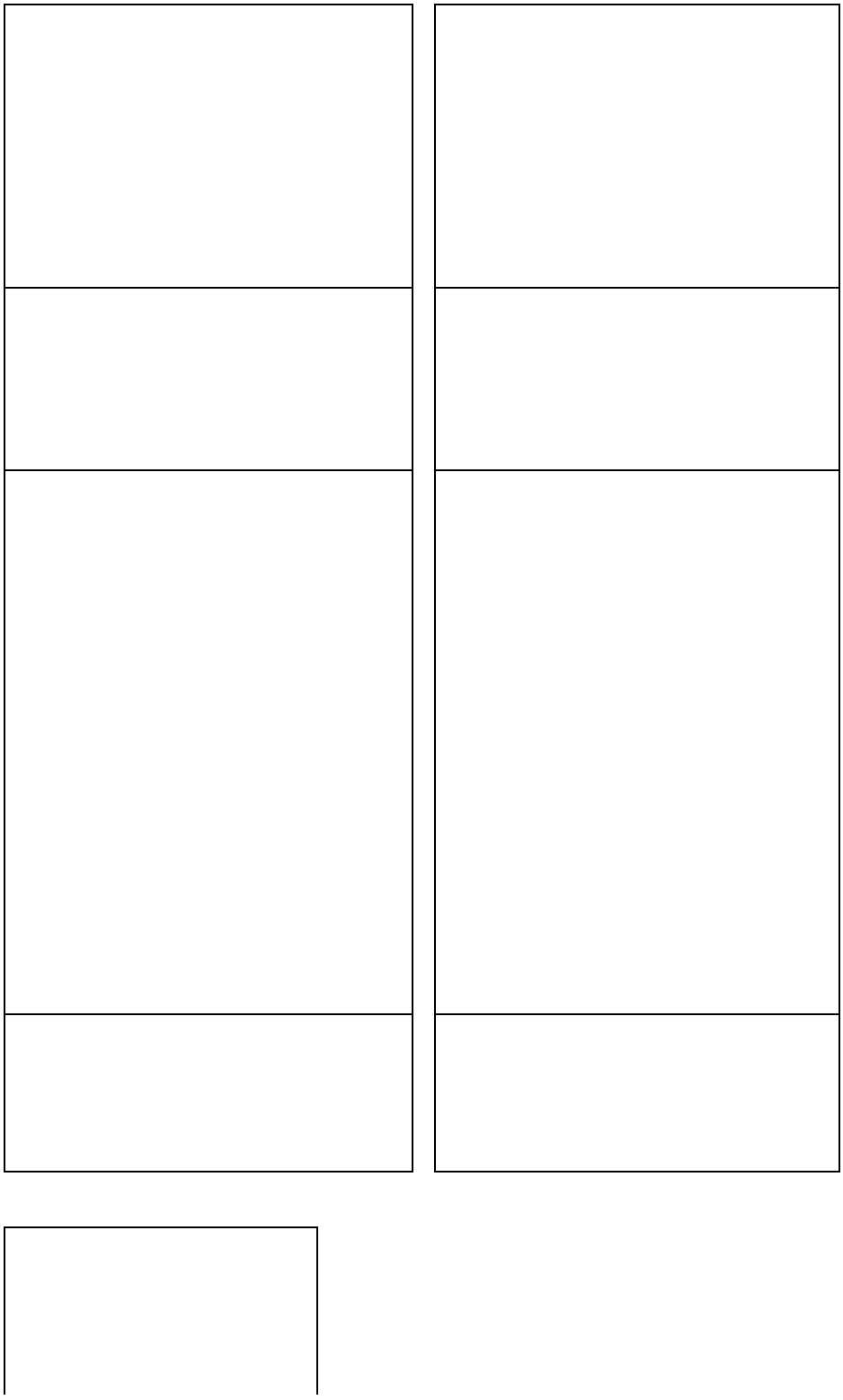
where

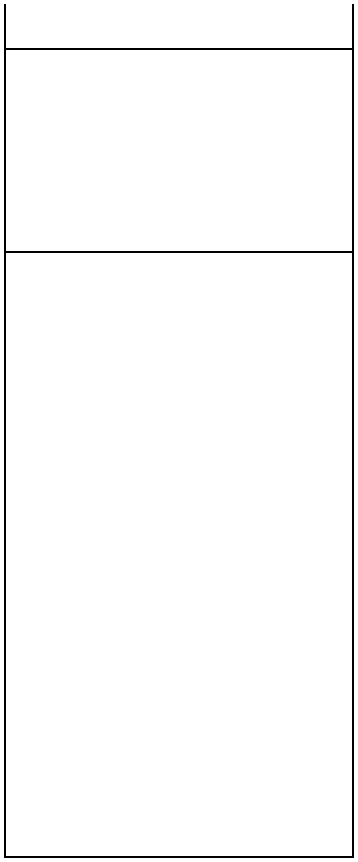
$$\bar{y}_h = \frac{1}{n_h} \sum_{i \in s_h} Y_i \qquad (2.2.4)$$

is the sample mean of the observations in stratum U_h. So, the post-stratification estimator is equal to a weighted sum of sample stratum means.

A simple example shows how weighting works in the case of a simple random sample. From a population of size 1000 a sample of size 100 is selected. There are two auxiliary variables: Sex (with two categories Male and Female), and AgeClass (with three categories Young, Middle and Old). Figure 2.2.1 contains both the population and the sample distribution of these variables.

Figure 2.2.1. Post-stratification with two auxiliary variables





The sample is not representative for the population. For example, the percentage of young females in the population is 20.9%, whereas in the sample the percentage is 15.0%. The sample contains too few young females. The third table in figure 2.2.1 contains the correction weights as computed by means of expression (2.2.2). For example, the weight for young female is equal to (209/1000) x (100/15) = 1.393. Young females are under-represented in the sample, and therefore get a weight larger than 1. People in over-represented strata get a weight less than 1.

The adjustment weights w_i are obtained by multiplying the correction weights d_i by the inclusion weights c_i. In this case, all inclusion weights are equal to N/n=10. Suppose, we use the weights to estimate the number of young females in the population. The weighted estimate would be 15 x 10 x 1.393 = 209, and this is the exact population total. Thus, application of weights to the auxiliary variable results in perfect estimates. If there is a strong relationship between the auxiliary variable and the target variable, also estimates for the target variable will be improved if these weights are used.

The weighting model obtained by crossing the variables AgeClass and Sex is denoted by

AgeClass x Sex.

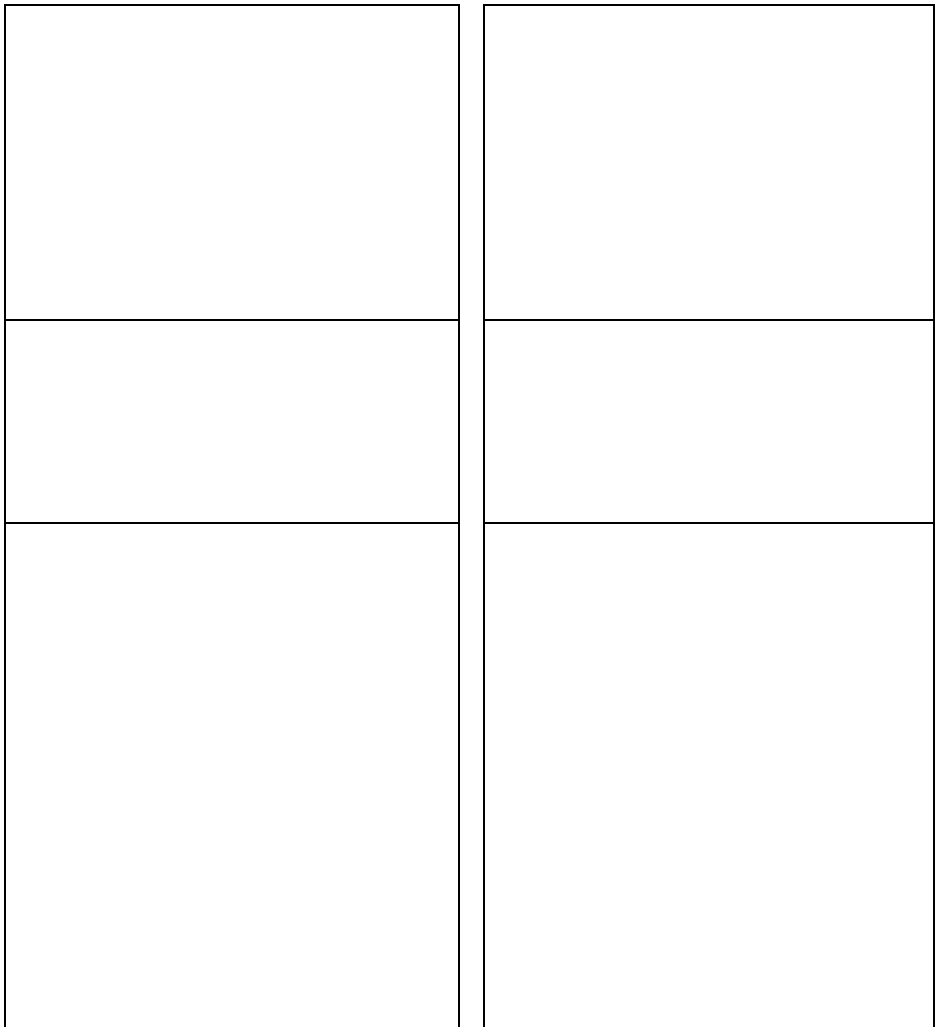
The idea of crossing variables can be extended to more than two variables. As long as the table with population frequencies is available, and all sample frequencies are greater than 0, weights can be computed.

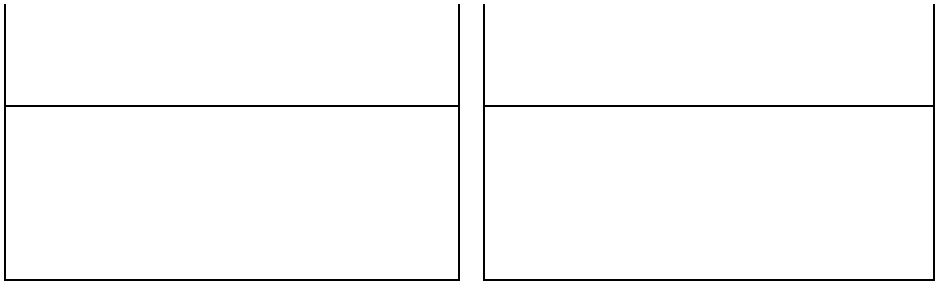
However, if there are no observations in a stratum, the weight can not be computed for that stratum. This leads to incorrect estimates. If the sample frequencies are very small, say less than 5, weights can be computed, but estimates will be unstable.

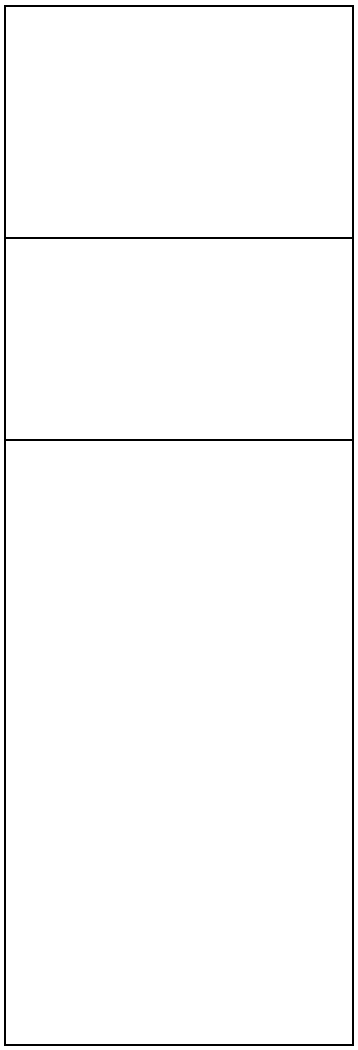
As more variables are used in a weighting scheme, there will be more strata. Therefore the risk of empty strata or strata with too few observations will be larger. There are two solutions for this problem. One is to use less auxiliary variables, but then a lot of auxiliary information is thrown away. Another is to *collapse strata*. This means merging a stratum with too few observations with another stratum. It is important to combine strata that resemble each other as much as possible. Collapsing strata is not a simple job, particularly if the number of auxiliary variables and strata is large. It is often a manual job.

Another problem in the use of several auxiliary variables is the lack of a sufficient amount of population information. This is illustrated in figure 2.2.2. The population distribution of the two variables AgeClass and Sex is known separately, but the distribution in the cross-classification is not known. In this case the post-stratification *AgeClass x Sex* cannot be carried out, because weights cannot be computed for the strata in the cross-classification.

Figure 2.2.2. Lack of population information







One way to solve this problem is to use only one variable, but that would mean ignoring all information with respect to the other variable. What is needed is a weighting technique that uses both marginal frequency distributions simultaneously. There are two weighting techniques that can do that: linear weighting and multiplicative weighting. These two techniques are described in the next two subsections.

2.3. Linear weighting

The technique of linear weighting is based on the theory of general regression estimation. This section only gives an overview of the theory of

linear weighting. More details can be found in Bethlehem and Keller (1987).

Suppose, there are p auxiliary variables available in the survey. For each element i in the population, there is a p-vector $X_i = (X_{i1},X_{i2},...,X_{ip})'$ of values of these auxiliary variables. The p-vector of population totals of the auxiliary variables is denoted by

$$X_T = \left(X_{T1}, X_{T2}, ..., X_{Tp}\right)', \tag{2.3.1}$$

where X_{Tj} denotes the population total of the j-th auxiliary variable :

$$X_{Tj} = \sum_{i \in U} X_{ij}, \tag{2.3.2}$$

where the sum is taken over all elements i in the population U. The Nxp-matrix of all values of the auxiliary variables is denoted by X. The i-th row of this matrix corresponds to the vector X_i.

If the auxiliary variables are correlated with the target variable, then for a suitably chosen p-vector $B = (B_1,B_2,...,B_p)'$ of regression coefficients for a best fit of Y on X, the residuals in the N-vector $E = (E_1,E_2,...,E_N)'$, defined by

$$E = Y - XB, \tag{2.3.3}$$

vary less than the values of the target variable itself. Application of ordinary least squares results in

$$B = (X'X)^{-1}X'Y = \left(\sum_{i \in U} X_i X_i'\right)^{-1}\left(\sum_{i \in U} X_i Y_i\right) \tag{2.3.4}$$

Of course, the vector B cannot be computed in practice, because the required information is not available. Instead, this quantity is estimated using the sample data. Let the p-vector b be the sample estimator for the p-vector B. A good choice for b is

$$b = \left(\sum_{i \in S} \frac{X_i X_i'}{\pi_i}\right)^{-1}\left(\sum_{i \in S} \frac{X_i Y_i}{\pi_i}\right) \tag{2.3.5}$$

The estimator b is an asymptotically unbiased (ADU) estimator of B. Expression (2.3.5) can be used to compute the *general regression estimator*. This estimator is defined as

$$y_R = y_{HT} + (X_T - x_{HT})'b \tag{2.3.6}$$

where y_{HT} is the Horvitz-Thompson estimator for the population mean of the target variable and x_{HT} the p-vector of Horvitz-Thompson estimators for the population totals of the auxiliary variables. The general regression estimator adjusts the simple Horvitz-Thompson estimator for differences

between the true values and estimated values of the totals of the auxiliary variables. The estimator is asymtotically unbiased.

It can be shown, see e.g. Bethlehem and Keller (1987), that the variance of estimator (2.3.7) is small if the residual values in E are small. Hence, the use of auxiliary variables which can explain the behaviour of the target variable, will be rewarded with a precise estimator.

Use of the general regression estimator implies a form of weighting. The estimator can be rewritten as

$$y_R = \sum_{i \in s} \frac{d_i Y_i}{\pi_i}, \qquad (2.3.7)$$

where the correction weight d_i is defined by

$$d_i = X_i' v \qquad (2.3.8)$$

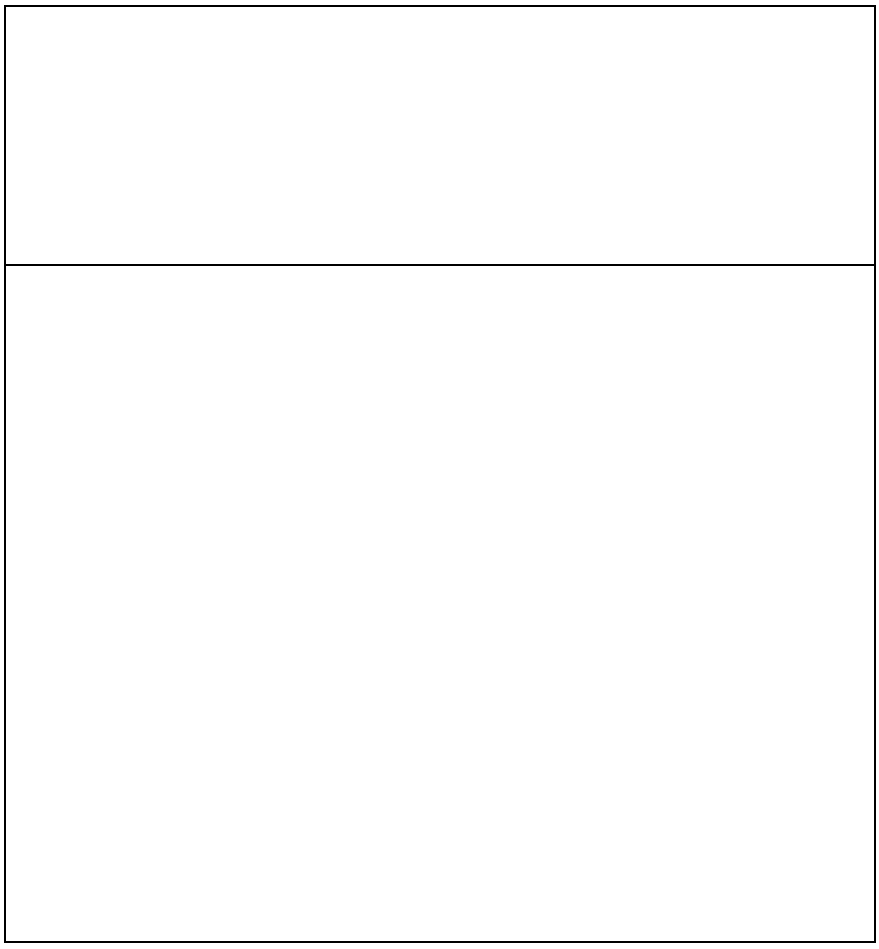
and the p-vector v of weight coefficients is equal to

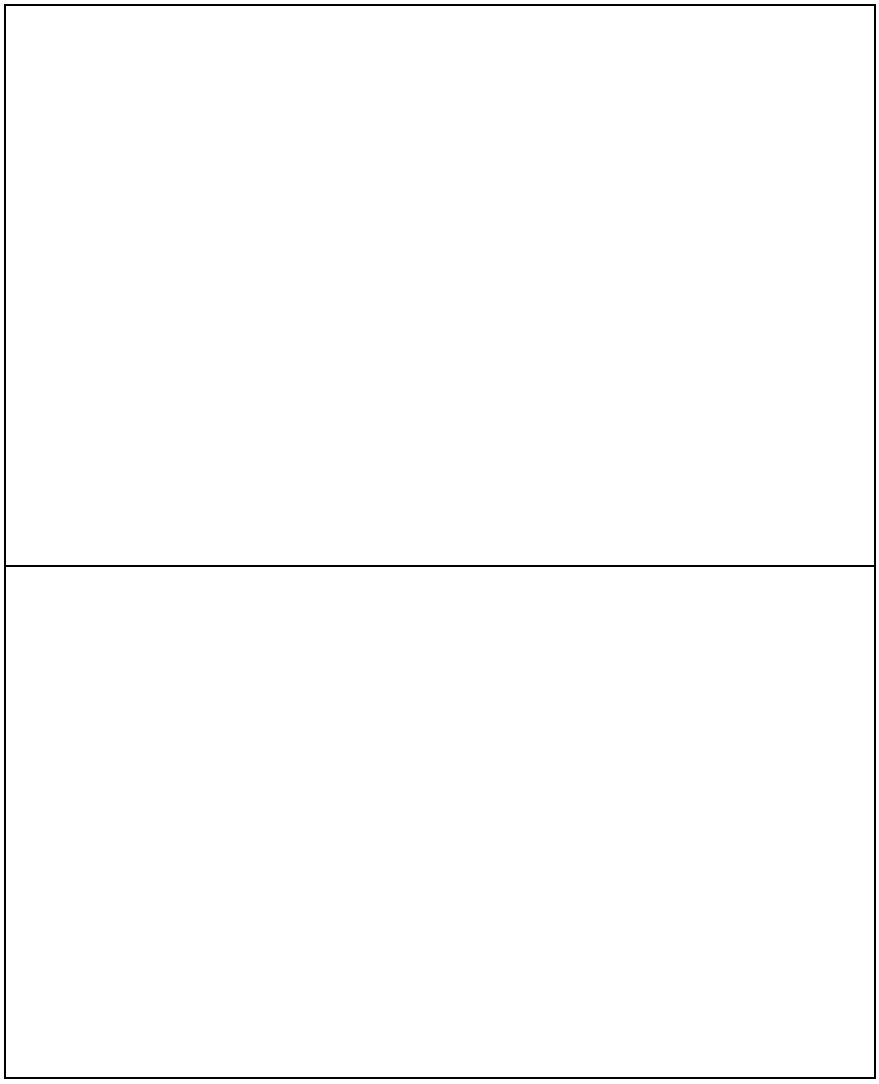
$$v = \left(\sum_{i \in S} \frac{X_i X_i'}{\pi_i} \right)^{-1} X_T \qquad (2.3.9)$$

Post-stratification is a special case of linear weighting. This is illustrated with the example of weighting by Sex and AgeClass. The auxiliary variables are dummy variables. There is a dummy variable for each cell in the table obtained by crossing the variables. The weighting scheme is denoted by *Sex x AgeClass*. The available population information is displayed in figure 2.2.1.

Crossing the two auxiliary variables produces a table with 2 x 3 = 6 cells. So six dummy variables have to be introduced. The possible values of these dummy variables are shown in figure 2.3.1. The bottom line in the table contains the vector of population totals of the auxiliary variables. The values are equal to the population frequencies in the cells of the population table.

Figure 2.3.1. Values of the dummy variables in Sex x AgeClass





Using this information, the value of the vector v of weight coefficients turns out to be equal to

v = (0.983, 0.950, 1.023, 1.393, 0.847, 0.850).

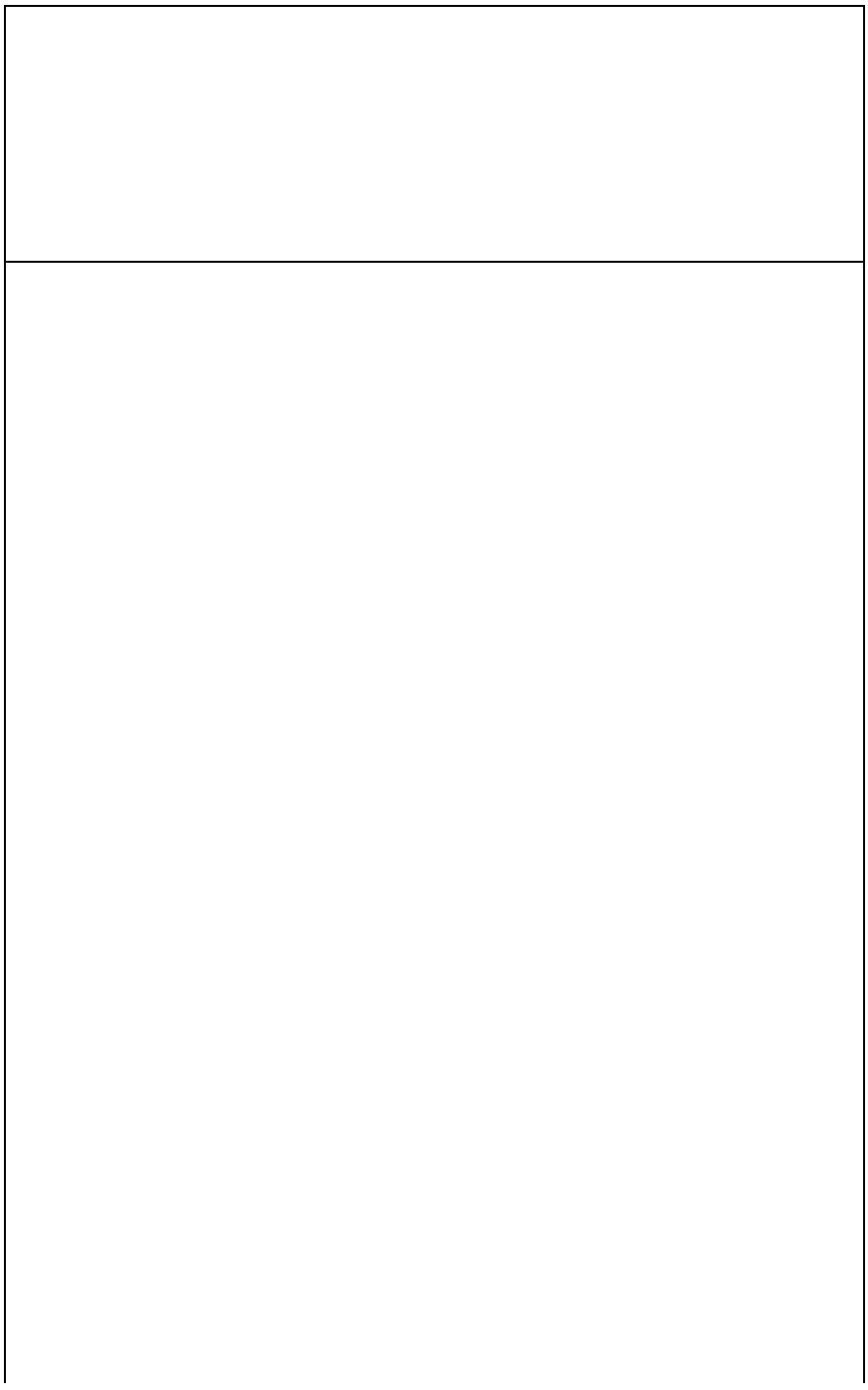
Figure 2.3.2 shows how the corrections weights can be computed using the vector v. The weights are identical to the weights in figure 2.2.1.

Figure 2.3.2. Computation of weights for Sex x AgeClass

Now, it is illustrated how linear weighting can address the problem of the lack of sufficient population information. Linear weighting offers a possibility to include both variables in the weighting scheme without knowing the population frequencies in all cells. The trick is to introduce a different set of dummy variables. Instead of using one set of 2 x 3 = 6 dummy variables for the cells of the tables, use two sets of dummy variables: one set of two dummy variables for the categories of Sex, and another set of 3 dummy variables for the categories of AgeClass. Then

there are 2 + 3 = 5 dummy variables. In each set, always one dummy has the value 1, whereas all other dummies are 0. The possible values of the dummy variables are described in figure 2.3.3. The first dummy variable represents the constant term in the regression model. It always has the value 1. The second and third dummy variable relate to the two sex categories, and the last three dummies represent the three age categories. The vector of population totals is equal to the frequencies of all dummy variables separately. Note that in this weighting scheme always three dummies in a row have the value 1.

Figure 2.3.3. Values of the dummy variables in Sex + AgeClass



Since no information is used about the crossing Sex by AgeClass, but only the marginal distributions, a different notation is introduced. This weighting scheme is denoted by

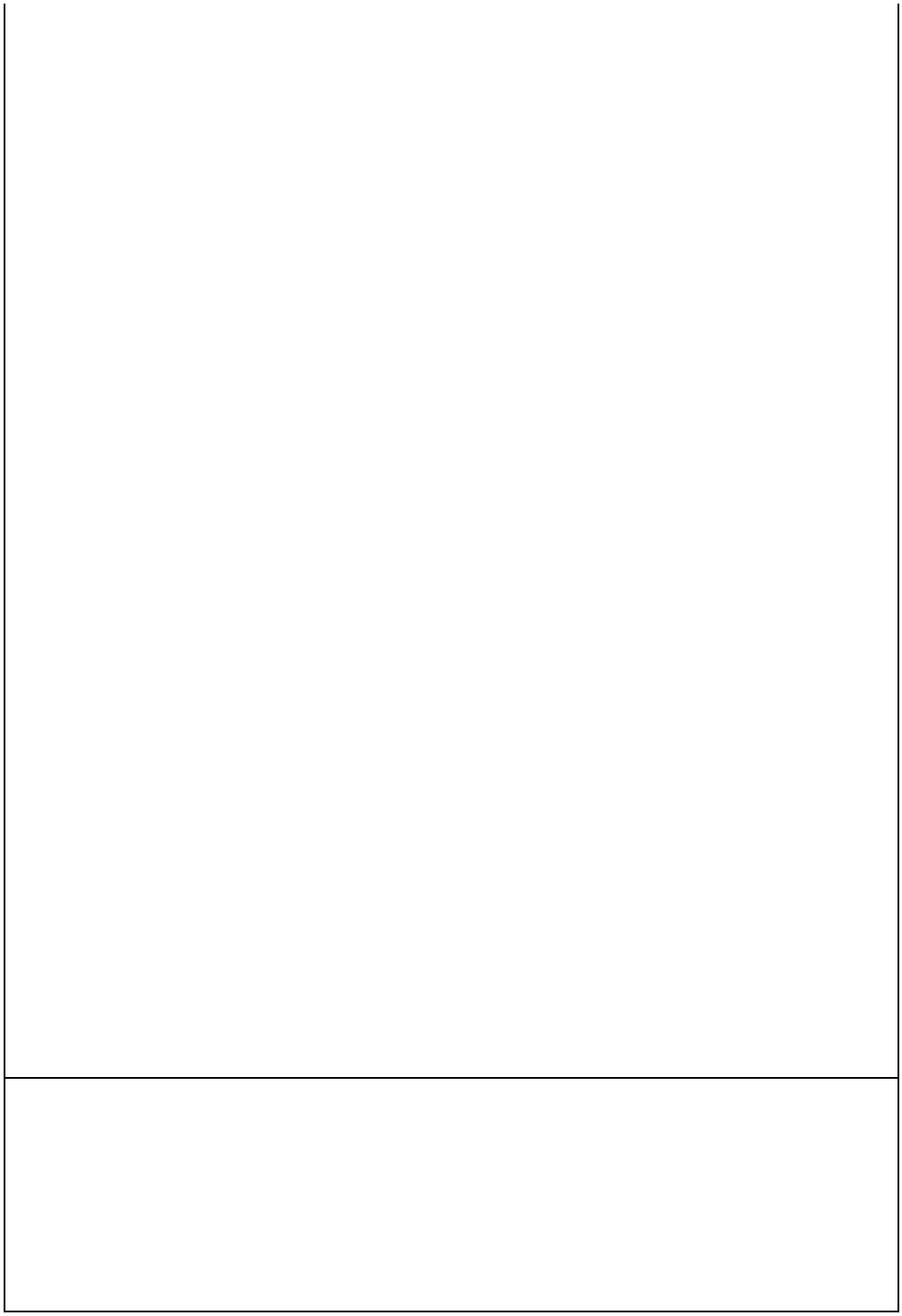
Sex + AgeClass.

Expression (2.3.9) can be applied for the computation of the weight coefficients. The vector v of weight coefficients turns out to be equal to

v = (0.991, -0.033, 0.033, 0.161, -0.095, -0.066).

The weight for an element is now obtained by summing the appropriate elements of this vector. The first value corresponds to the dummy X_1, which always has the value 1. So there is always a contribution 0.991 to the weight. The next two values correspond to the categories of Sex. Note that their sum equals zero. For males, an amount 0.033 is subtracted, and for females, the same amount is added. The final three values correspond to the categories of AgeClass. Depending on the age category a contribution is added or subtracted. Figure 2.3.4 contains the computation of the correction weights for all combinations of Sex and AgeClass. All figures are rounded to three digits.

Figure 2.3.4. Computation of weights for Sex + AgeClass



The weights in figure 2.3.4 are not equal to the weights obtained by complete post-stratification (see figure 2.2.1). This is not surprising, since the model *Sex + AgeClass* uses less information than the model *Sex x AgeClass*.

The examples used to illustrate the theory of linear weighting were based on two auxiliary variables. Of course, it is possible to use more than two variables.

So far, only the applications of linear weighting with qualitative auxiliary variables were discussed. It is also possible to apply linear weighting if only quantitative auxiliary variables are available. Then, linear weighting becomes generalised regression estimation. Moreover, it is also possible to combine qualitative and quantitative variables in a weighting model. This opens new possibilities for weighting. For more details, see the Bascula Reference Manual.

2.4. Multiplicative weighting

If linear weighting is applied, correction weights are obtained that are computed as the sum of a number of relevant weight coefficients. It is also possible to compute correction weights in a different way, namely as the product of a number of weight factors. This weighting method is called *multiplicative weighting*. Sometimes it is also called *raking* or *iterative proportional fitting*.

Generally, multiplicative weighting can be applied in the same situations as linear weighting as long as only qualitative variables are used. An example is the situation in which there is no population information about the complete cross-classification of all auxiliary variables, but there is information about classifications of subsets of these variables. The method of multiplicative weighting computes correction weights by means of an iterative procedure. The resulting weights are the product of factors contributed by all cross-classifications.

There is no general theoretical framework for multiplicative weighting. It is a step-wise process that is continued until satisfactory results are obtained. The general scheme is as follows :

1) Introduce a weight factor for each stratum in each cross-classification term. Set the initial values of all factors to 1.

2) Adjust the weight factors for the first cross-classification term so that the weighted sample becomes representative with respect to the auxiliary variables included in this cross-classification.

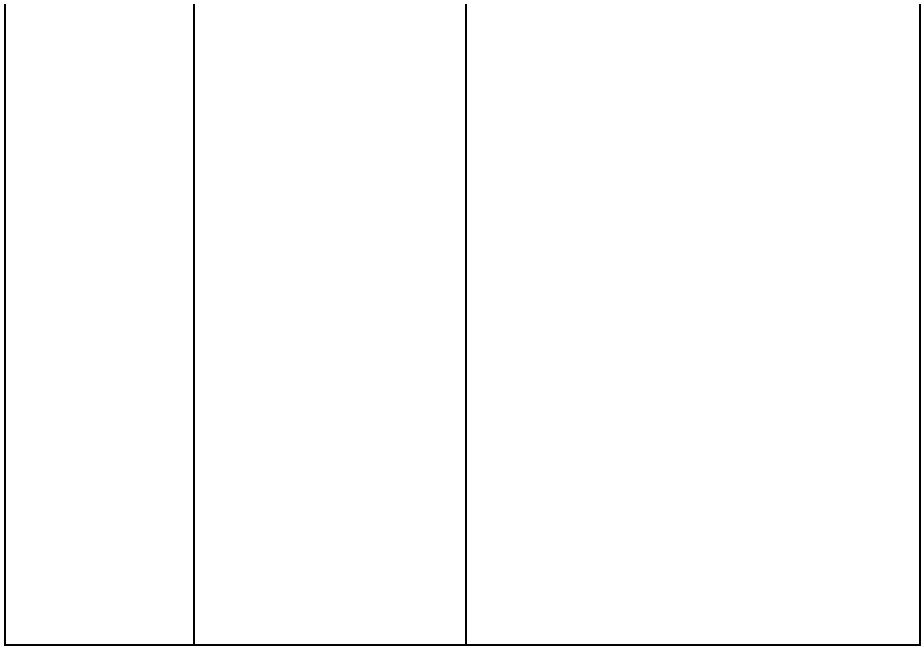
3) Adjust the weight factors for the next cross-classification term so that the weighted sample is representative for the variables involved. Generally, this will disturb representativeness with respect to the other cross-classification terms in the model.

4) Repeat this adjustment process until all cross-classification terms have been dealt with.

5) Repeat steps 2, 3, and 4 until the weight factors do not change any more.

This procedure is illustrated with a simple example. The two qualitative auxiliary variables Sex and AgeClass are used. It is assumed that only the population distribution of Sex (2 categories) and AgeClass (3 categories) separately are available, and not the cross-classification. Figure 2.4.1 contains the starting situation.

Figure 2.4.1. Starting situation for multiplicative weighting

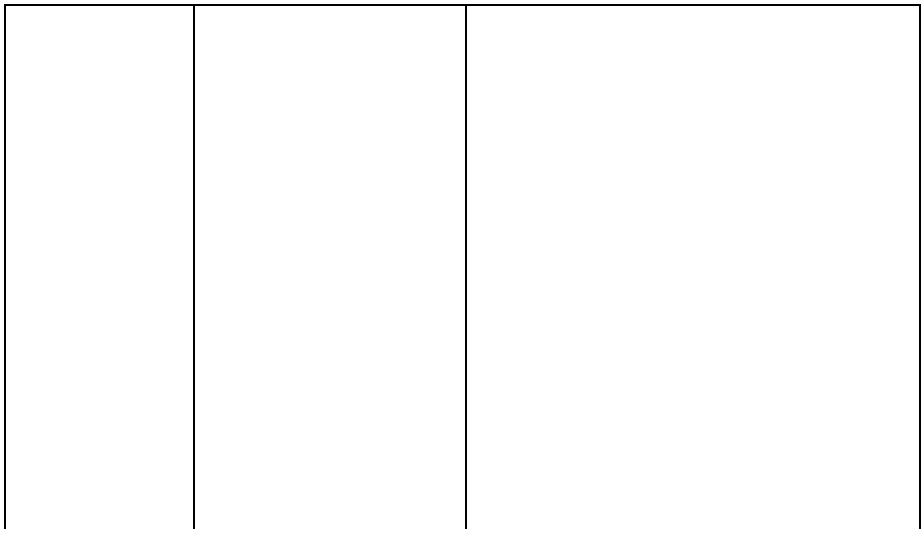


The upper-left part of the table contains the weighted frequencies in the sample for each combination of AgeClass and Sex. They are the Horvitz-Thompson estimates of the corresponding population quantities.

The row and column denoted by ‘Weight factor’ contain the initial values of the weight factors. The values in the row and column denoted by ‘Weighted sum’ are obtained by first computing the weight for each sample cell (by multiplying the relevant row and column factors), and then summing the weighted cell frequencies. Since the initial values of all factors are equal to 1, the weighted sums in figure 2.4.1 are equal to the unweighted sample sums. The row and column denoted by ‘Population distribution’ contain the frequencies of AgeClass and Sex in the population.

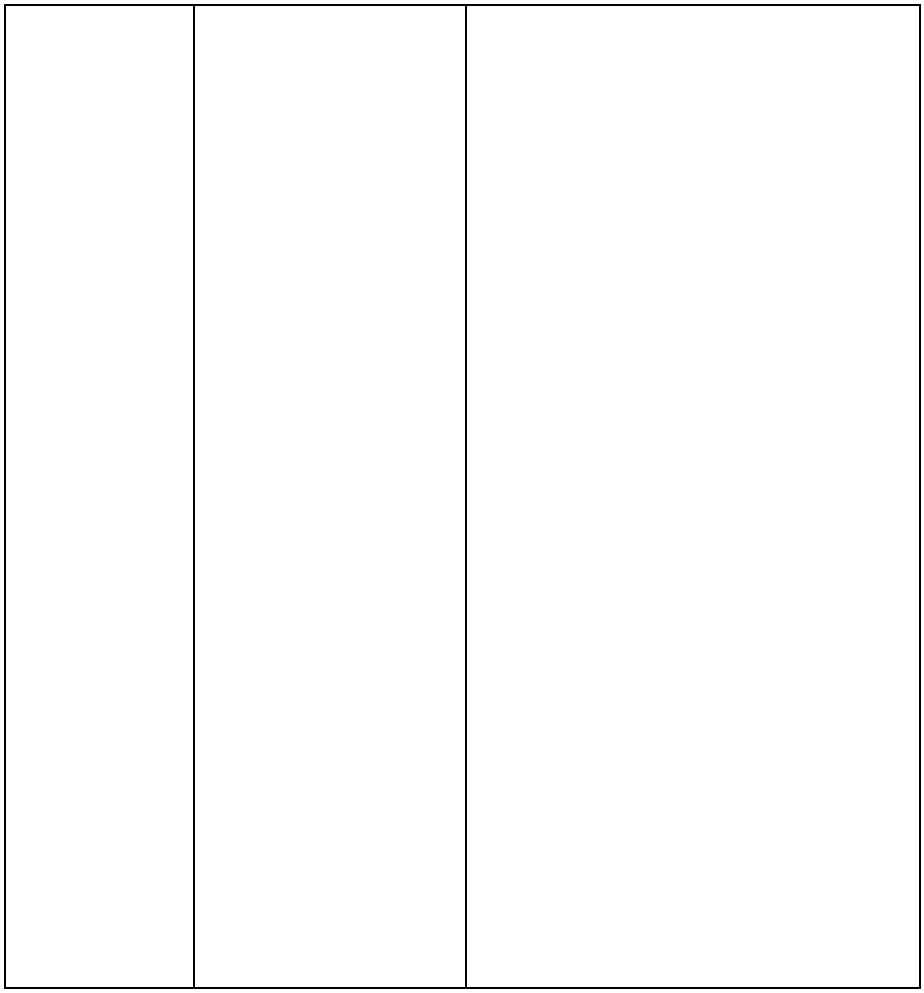
The iterative process must result in row and column factors with such values that the weighted sums match the population distribution. This is clearly not the case in the starting situation. First, the weight factors for the rows are adjusted. The result of that exercise is shown in figure 2.4.2.

Figure 2.4.2. Situation after adjusting for AgeClass



The weighted sums for the rows are now correct, but the weighted sums for the columns still show a discrepancy. The next step will be to adjust the weight factors for the columns such that the weighted column sums match the corresponding population frequencies. The result of this operation is shown in figure 2.4.3.

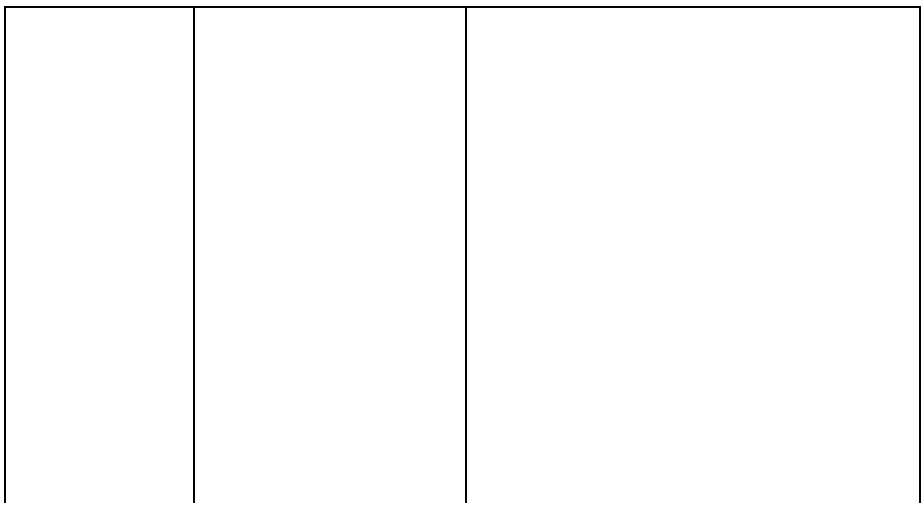
Figure 2.4.3. Situation after adjusting for Sex



Note that the adjustment for Sex has disturbed the adjustment for AgeClass. The weighted sums for the age categories no longer match the relative population frequencies. However, the discrepancy is much less than in the initial situation.

The process of adjusting for AgeClass and Sex is repeated until the weight factors do not change any more. The final situation is reached after a few iterations. Figure 2.4.4 shows the final results.

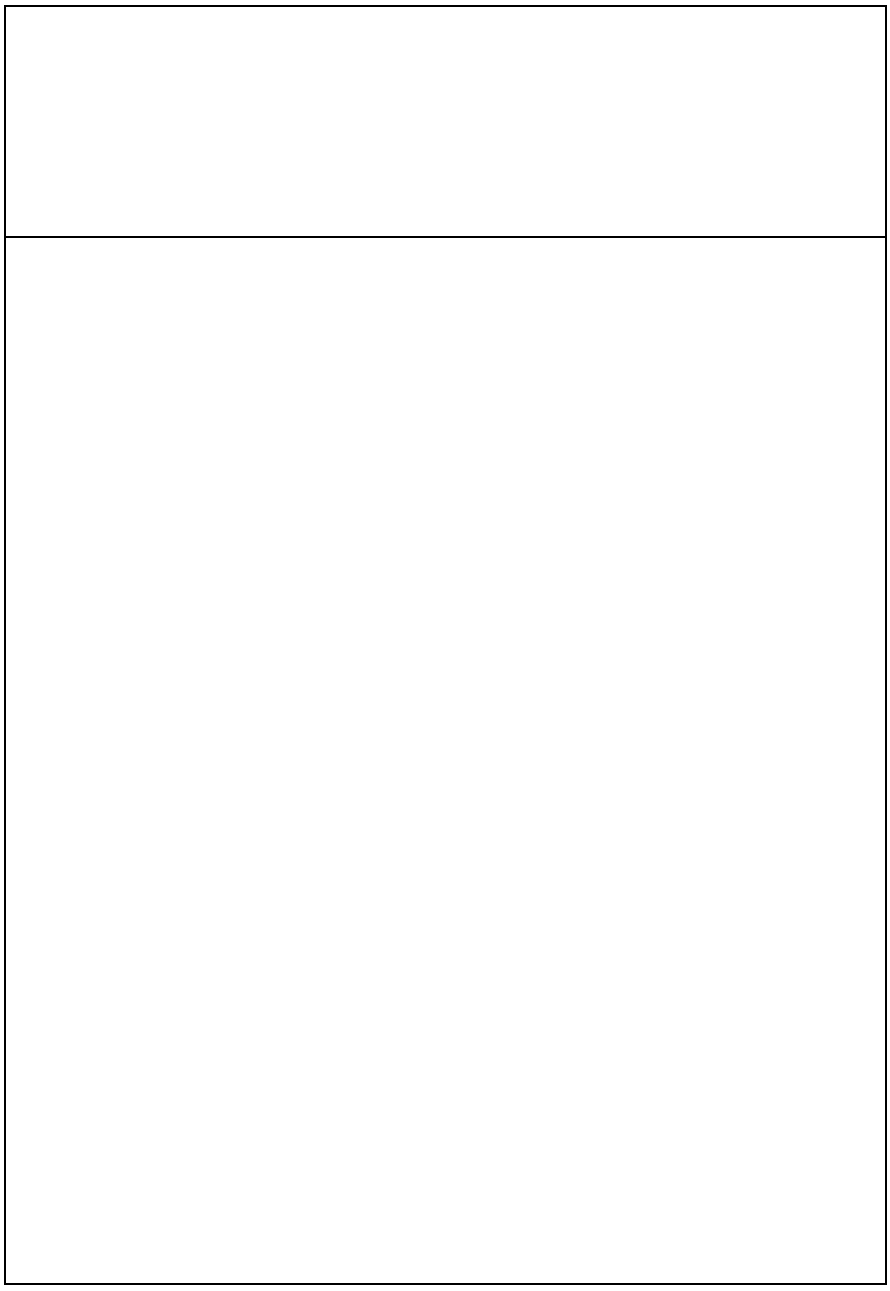
Figure 2.4.4. Situation after convergence



The correction weights for a specific sample element are now obtained by multiplying the relevant weight factors. Figure 2.4.5 contains the resulting correction weights. If the correction weights are multiplied by the inclusion weights, the adjustment weights are obtained. Since the inclusion weights are equal to N/n=10 in the example, the adjustment weights are simply 10 times the correction weights.

For this example, the adjustment weights differ only slightly from those obtained by linear weighting, see figure 2.3.4.

Figure 2.4.5. Computation of weights for Sex + AgeClass



In situations where both linear and multiplicative weighting are possible, a choice must be made. To help making this choice, some of the advantages and disadvantages of both techniques are pointed out.

Linear weighting has the advantage that it is based on a model that explains in which situations weighting will work. It is also possible to describe the effect of linear weighting on the variances of estimates.

Linear weighting may result in negative weights. Such weights are not wrong, but simply a consequence of the theory. Usually, negative weights are an indication that the model does not fit too good. Unfortunately, there are some analysis packages (e.g. SPSS) that do not accept negative weights. That might be a reason not to use linear weighting.

Multiplicative weightings lacks a proper model describing the properties of estimators based on these weights. It is also not very easy to compute variance estimators.

Multiplicative weighting always produces positive weights. If this is necessary requirement, this weighting technique should be chosen.

Although linear and multiplicative weighting seem to be very different weighting techniques, it is possible to create a general framework of which both techniques are special cases. This work was done by Deville and Särndal (1992). They proved that estimators based on linear or multiplicative weighting approximately behave in the same way. So, one could use linear weighting in selecting the proper variables, and then apply multiplicative weighting in order to ensure weights to be positive.

3. Use of Bascula in the Blaise environment

Bascula can be used in two ways: as a stand-alone package, or as a tool within the Blaise environment. To use Bascula as a stand-alone package, there must be a sample data file in Ascii format. Furthermore, a description of the variables in the file must be provided. This can be done interactively from within Bascula. This section describes how Bascula can be used as a tool within the Blaise environment. This is the most efficient way to use Bascula, because both data and meta-data are provided by Blaise.

To show how Bascula can be used, a simple example is used. A very small country was constructed. It was called Samplonia. It is a small island with 1000 inhabitants. The country is divided into two provinces: Agria and Induston. The province of Agria consists of three districts: Wheaton, Greenham, and Newbay. Induston has four districts: Oakdale, Crowdon, Mudwater, and Smokeley.

3.1. Data collection with Blaise

Statistics Samplonia has conducted a survey based on a sample of 100 inhabitants. Information was collected about place of residence (province and district), sex, age, employment status, and income. Figure 3.1.1 contains the Blaise data model for this survey.

The data model should contain all information required in the cause of processing the survey data. Therefore, it is important to realise at the design stage which variable will be needed in the ultimate data file. The data model in figure 3.1.1 contains two fields that are not filled during data collection, but afterwards.

The first field is *AgeClass*. This field represents a derived variable. It is a discrete variable, and its values are derived from the continuous variable Age. Survey publications often do not contain statistics on continuous variables. One reasons may be protection of confidentiality. Another could be that the values of the continuous variable are not very accurate. Often, it is already known at the design stage of the survey which tables will be published. In such a case, it is a matter of good documentation practice to include required derived variables in the data model. The variable *AgeClass* is a good example of a publication variable that is derived from an interview variable Figure 3.1.1 shows that *AgeClass* will divide the Age in three classes.

To be able to compute unbiased estimates of population characteristics, the sample design must be know. More specifically, the first order inclusion probabilities are required. For a simple random sample without replacement, all inclusion probabilities are equal to n/N, where n is the sample size and N is the population size. To have the inclusion probabilities available, the field *IncWeights* is included in the data model of figure 3.1.1. *IncWeight* represents the inclusion weight, which is defined as the reciprocal of the inclusion probability. The inclusion probabilities in the example are all equal to 0.1, so the inclusion weights are all equal to 10.

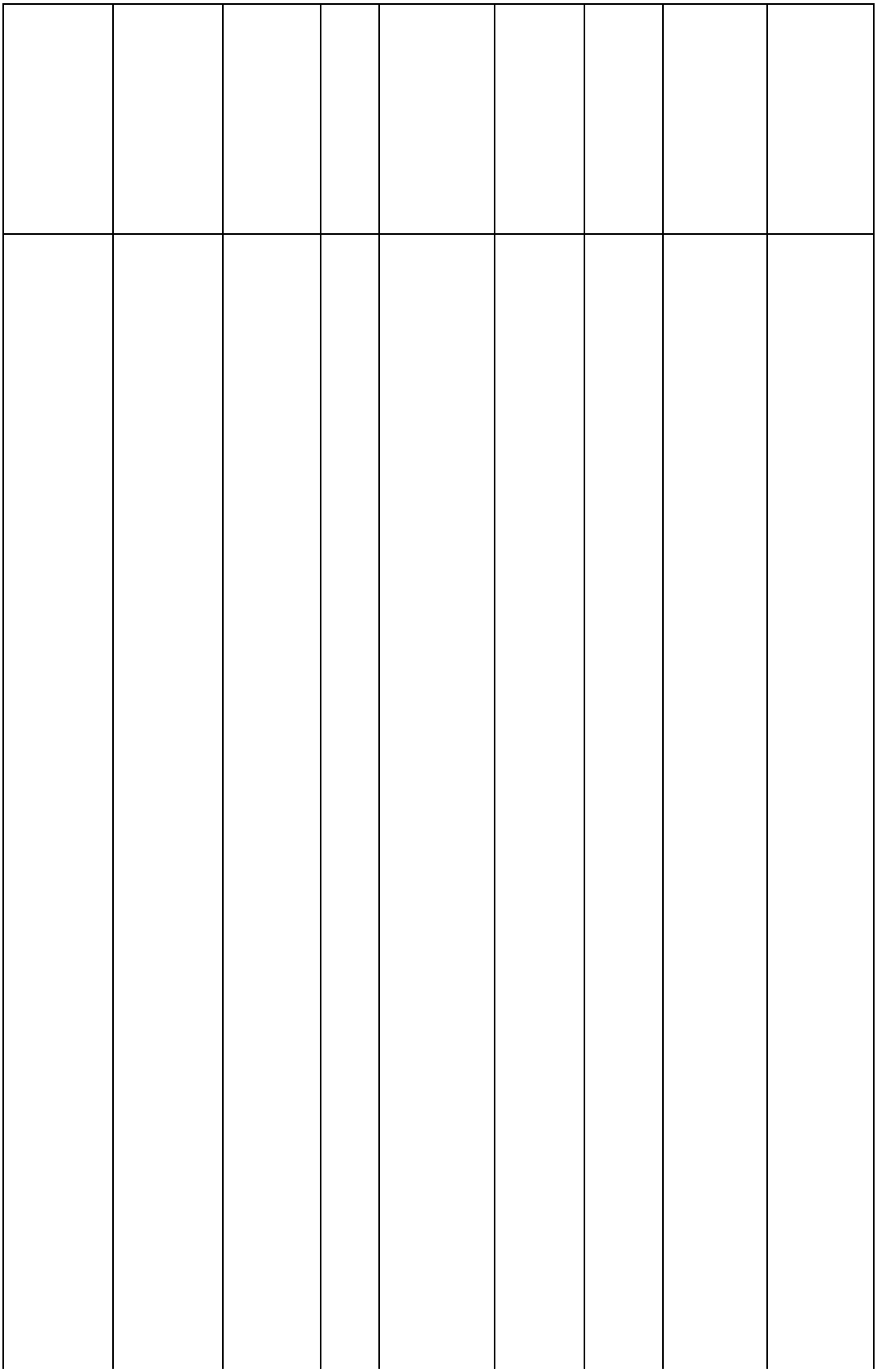
Figure 3.1.1. The Blaise data model

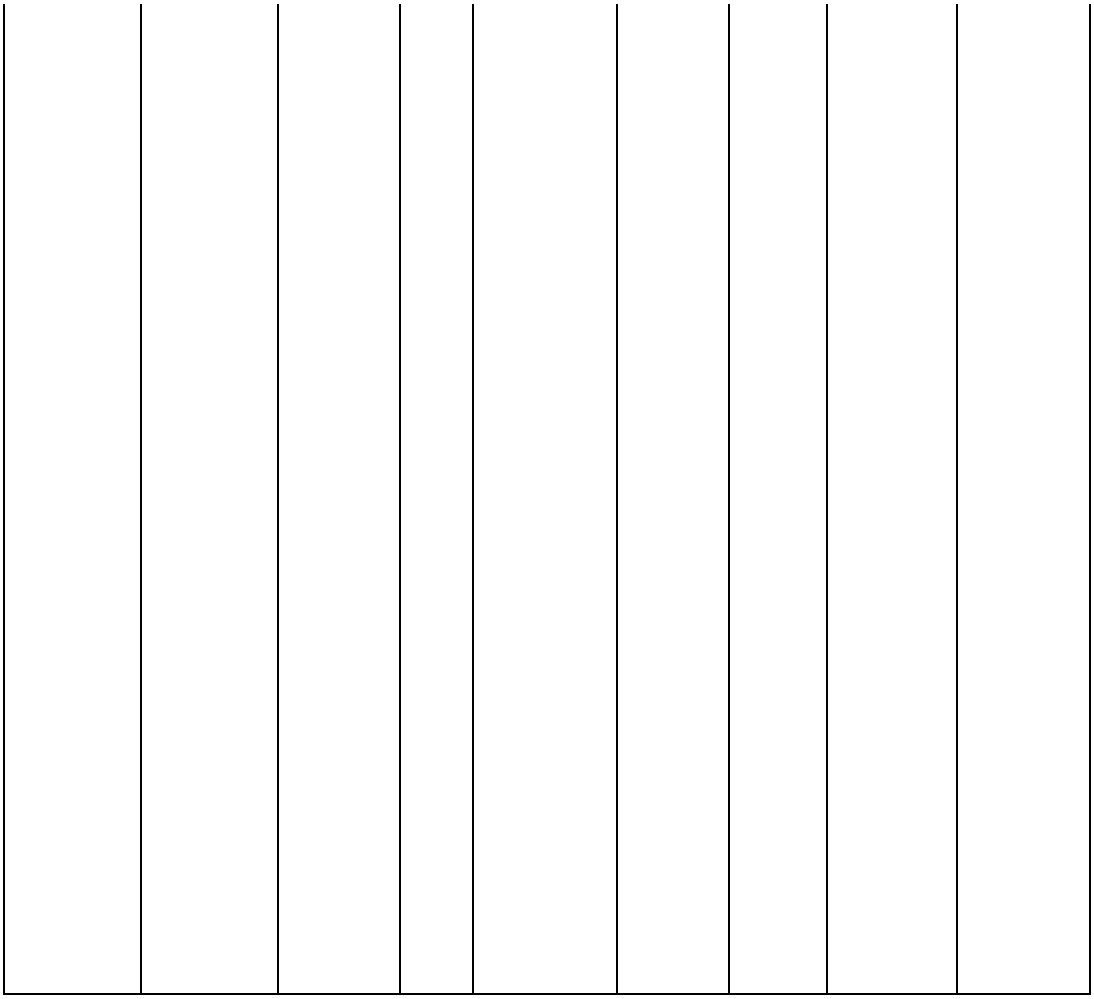
```
DATAMODEL   Samplon1   "Samplonian   Population
Survey"



FIELDS

   District   "District of residence": (Wheaton,
Greenham, Newbay,

                                        Oakdale,
Crowdon,

                                        Smokeley,
Mudwater)

   Province    "Province  of  residence":  (Agria,
Induston)

   Sex          "Sex  of  respondent"     :  (Male,
Female)

   Age       "Age of respondent"    : 0..99

   AgeClass   "Age class"               : (Young,
Middle, Elderly)

   Employ      "Employment  status"      :  (Job,
NoJob)

   Income    "Monthly net income"   : 0..6000

   IncWeight   "Inclusion   weight"             :
0.000..1000.000

   AdjWeight   "Adjustment   weight"           :
0.000..1000.000



RULES

   Province District Sex Age

   IF Age <= 30 THEN

      AgeClass:= Young

   ELSEIF AGE <= 55 THEN

      AgeClass:= Middle

   ELSE

      AgeClass:= Elderly

   ENDIF

   Employ Income

   IncWeight:= 10.000



ENDMODEL
```

Note that for the case of equal inclusion weights, it is not really necessary to specify these weights. If no inclusion weights are given, Bascula will assume them all to be equal, and can compute adjustment weights without knowing the inclusion weights. So the inclusion weights could have been omitted in the example. However, it is considered to be good documentation practice to include this information.

Most surveys suffer from non-response. Therefore, it is likely that some kind of weighting has to be carried out. The computed inclusion weights must be added to the data file. Again, good documentation practice suggests to account for weight variables in the data model. To that end, the field *Weight* has been included in the data model of figure 3.1.1. Note that initially this field has no value assigned to it. In a later stage, Bascula will replace this value by the true value of the weight. Figure 3.1.2 contains a view at the data file just after the fieldwork has been completed.

Figure 3.1.2. The first 10 records of the sample data file before weighting



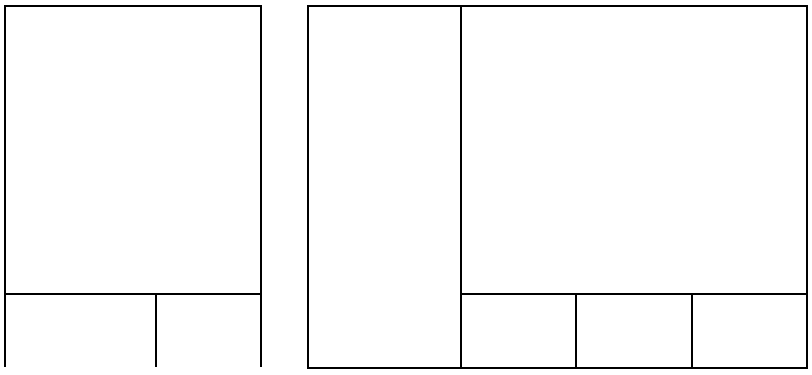


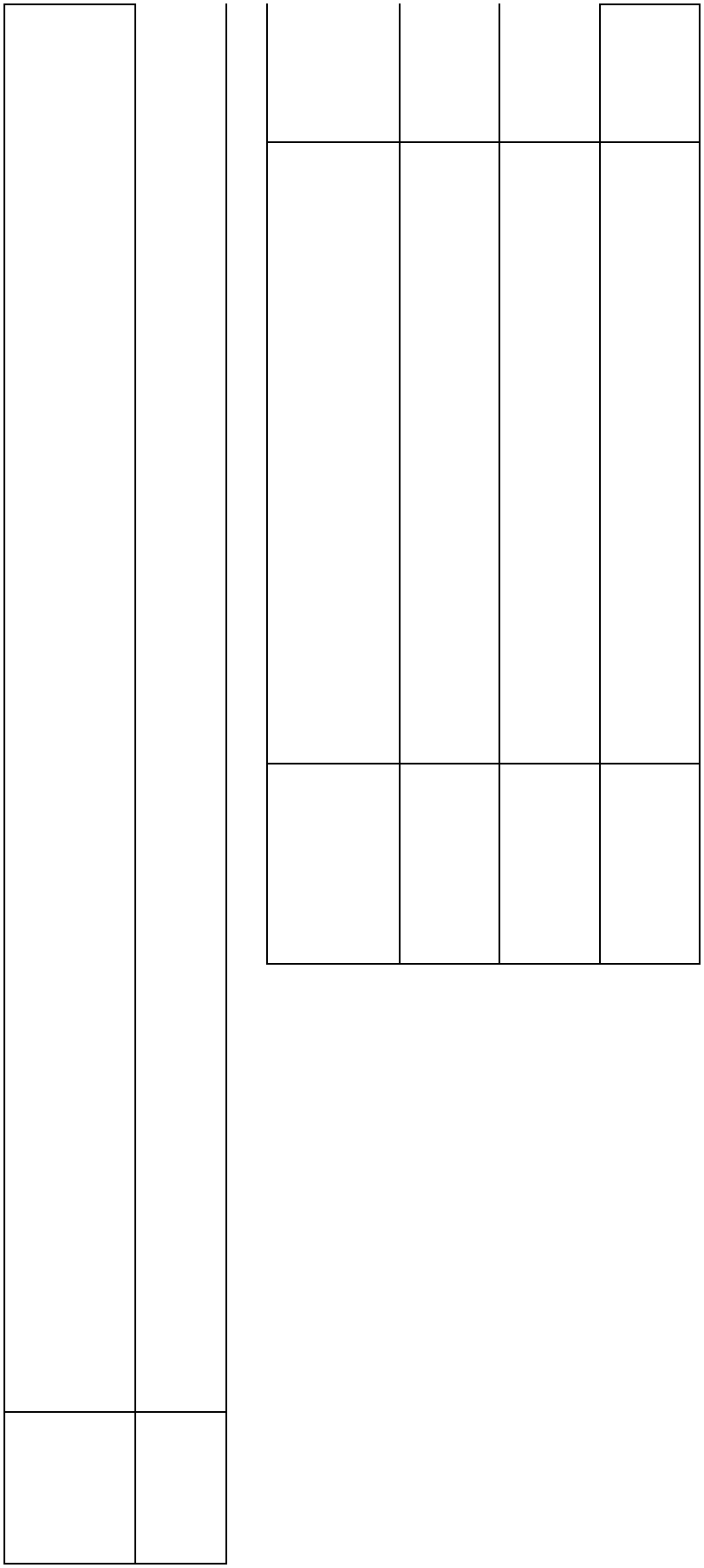
After all data has been collected, and as much as possible detected errors have been corrected, a weighting procedure can be carried out. This is the topic of the next section.

3.2. Weighting with Bascula

To be able to compute adjustment weights, auxiliary information must be available. It is assumed that three auxiliary variables can be used: *District*, *Sex* and *AgeClass*. The available population information for these three variables is displayed in figure 3.2.1. There are two tables. The first one contains the population counts for the districts, and the second one has the counts for age class by sex.

Figure 3.2.1. The available population information





If there is only one population table available, simple post-stratification can be applied. This is not the case here. Therefore, a choice has to be made between linear weighting and multiplicative weighting. Linear weighting will be applied.

If Bascula is installed in the Blaise directory, the weighting package can be run from within the Blaise Control Centre. Bascula is one of the options in the *Tools menu*.

The first thing to do is to inform Bascula about the sample data. The *Sample menu* has options for this. The file type is set to Blaise with the option *File type*, and the name of the data file is selected with the option *Sample file*. Note that no meta-data has to be specified . Once Bascula knows the name of the Blaise data file, it can locate the corresponding meta-data file, and extract all required information about the variables from this meta-data file.

Next, Bascula must be instructed which variables to use for weighting. This can be done with the option *Assign variables* in the *Weighting menu*. The *Select button* produces a list of all variables in the Blaise data file. The relevant variables are selected in this list. For the example at hand, it comes down to selecting the auxiliary variables *District*, *AgeClass* and *Sex*, the inclusion weight *IncWeight*, and the final weight *AdjWeight*. After these variables have been selected, Bascula must know what roles they play in the weighting proces. The *Edit* button is used for this. The variables *District*, *AgeClass* and *Sex* are assigned the role of auxiliary variable, *IncWeight* is the inclusion weight, and *AdjWeight* is to contain the final weight.

Three auxiliary variables will be used in the weighting model. For these variables the relevant population tables must be entered. This is done with the option *Population tables* in the *Weighting menu*. To be able to use the population information in figure 3.2.1, two tables must be defined. The first table only contains the variable *District*. The second table is obtained by crossing *AgeClass* and *Sex*. This table is denoted by *AgeClass ´ Sex*.

Tables are defined with the *Define button*. The defined tables are filled with the population distribution by pressing the *Data* button.

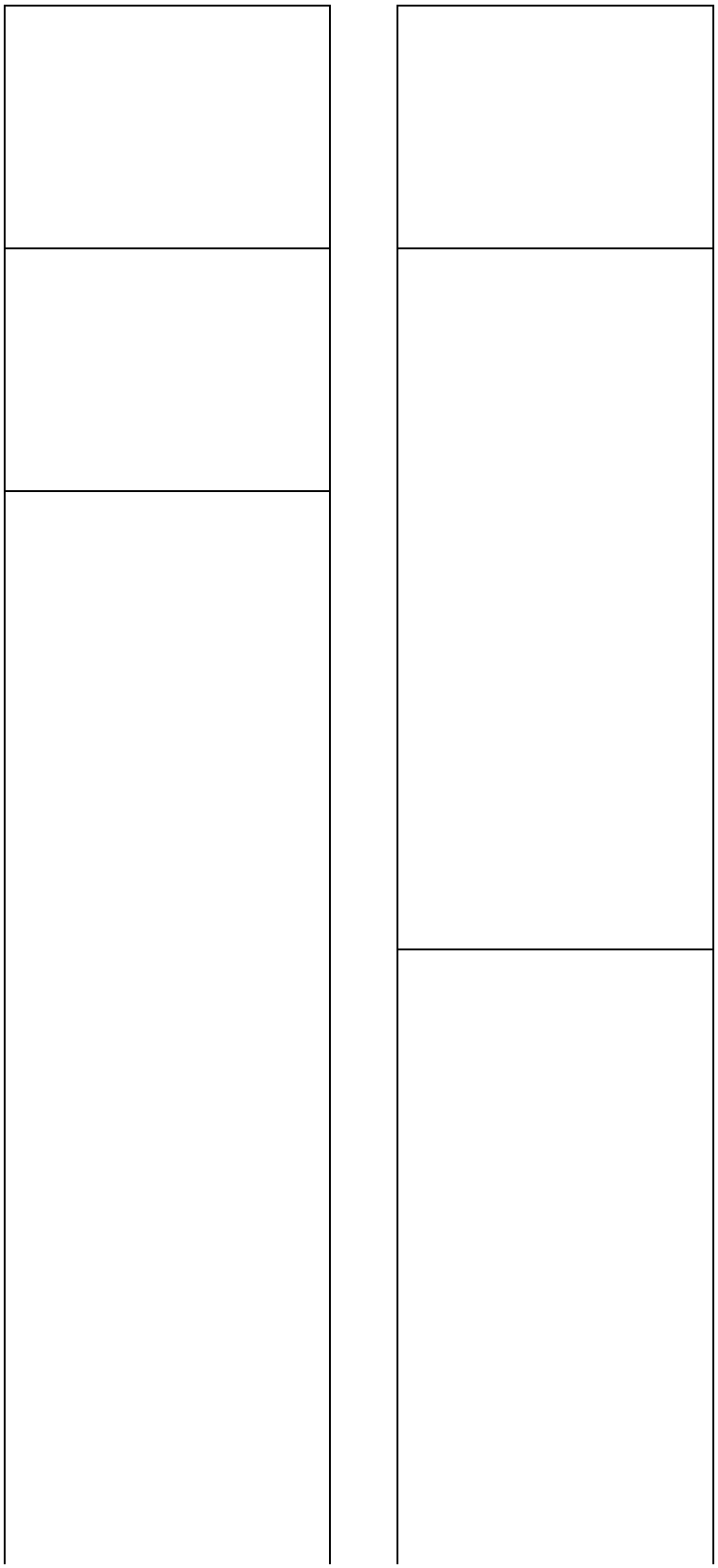
If more than one table is defined, they must all be checked for consistency. For example, the grand totals of all tables must be the same. Consistency is checked with the *Validate* button.

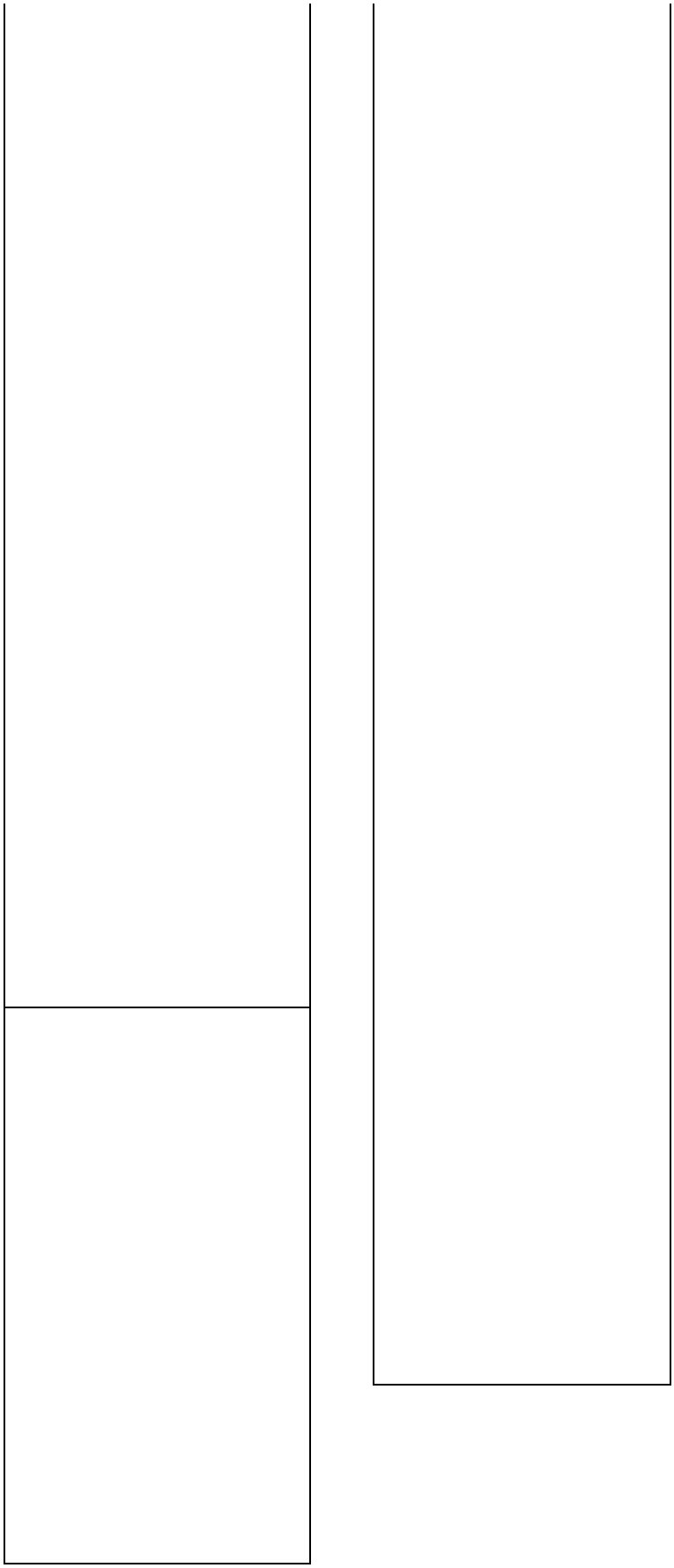
After all required information (sample data and population data) has been specified, the weighting model can be selected. This is done with the option *Select model* in the *Weighting menu*. Usually weighting produces the best results if the maximum possible population information is used. The maximal weighting model is selected by pressing the *Maximal* button. A weighting model must always be checked to see that sufficient sample observations are available in each stratum. This check is carried out by pressing the *Check* button.

If no problems are encountered, the computation of adjustment weights can be initiated through the option *Start weighting* in the *Weighting* menu.

For the example in this paper, two weighting techniques can be applied: linear or multiplicative weighting. Linear weighting is selected. Figure 3.2.2 contains the resulting weight coefficients.

Figure 3.2.2. The weight coefficients





The weight for a specific sample case is computed by adding up the weights in the relevant categories. For example, the weight of and elderly male in Crowdon is equal to

1.089+ 0.432 + 0.053 + 0.343 + 0.166 = 2.083

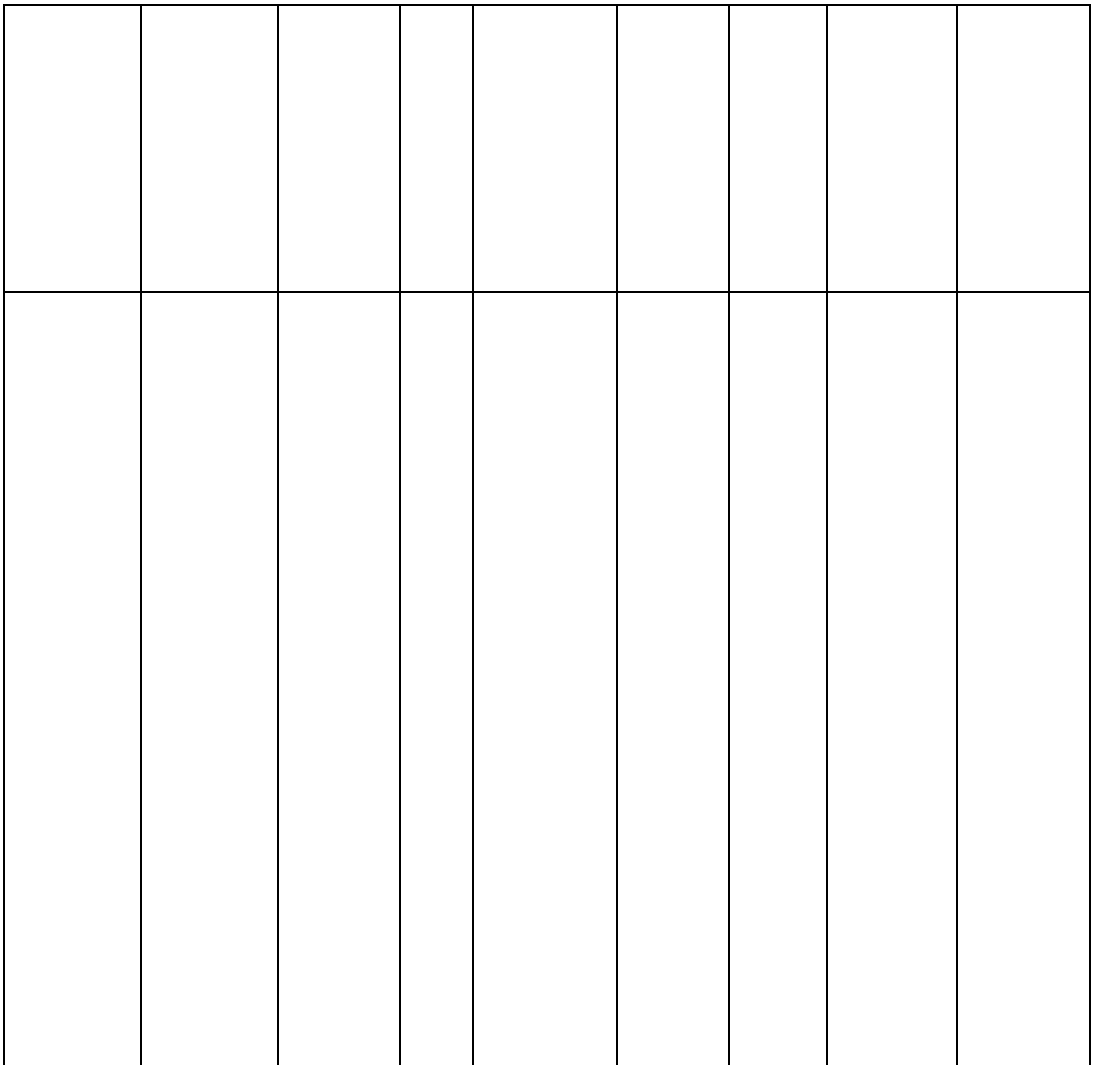
A closer look a the weight coefficients shows which categories are over- or under-represented in the sample. Negative weight coefficients denote under-representation and positive weights over-representation.

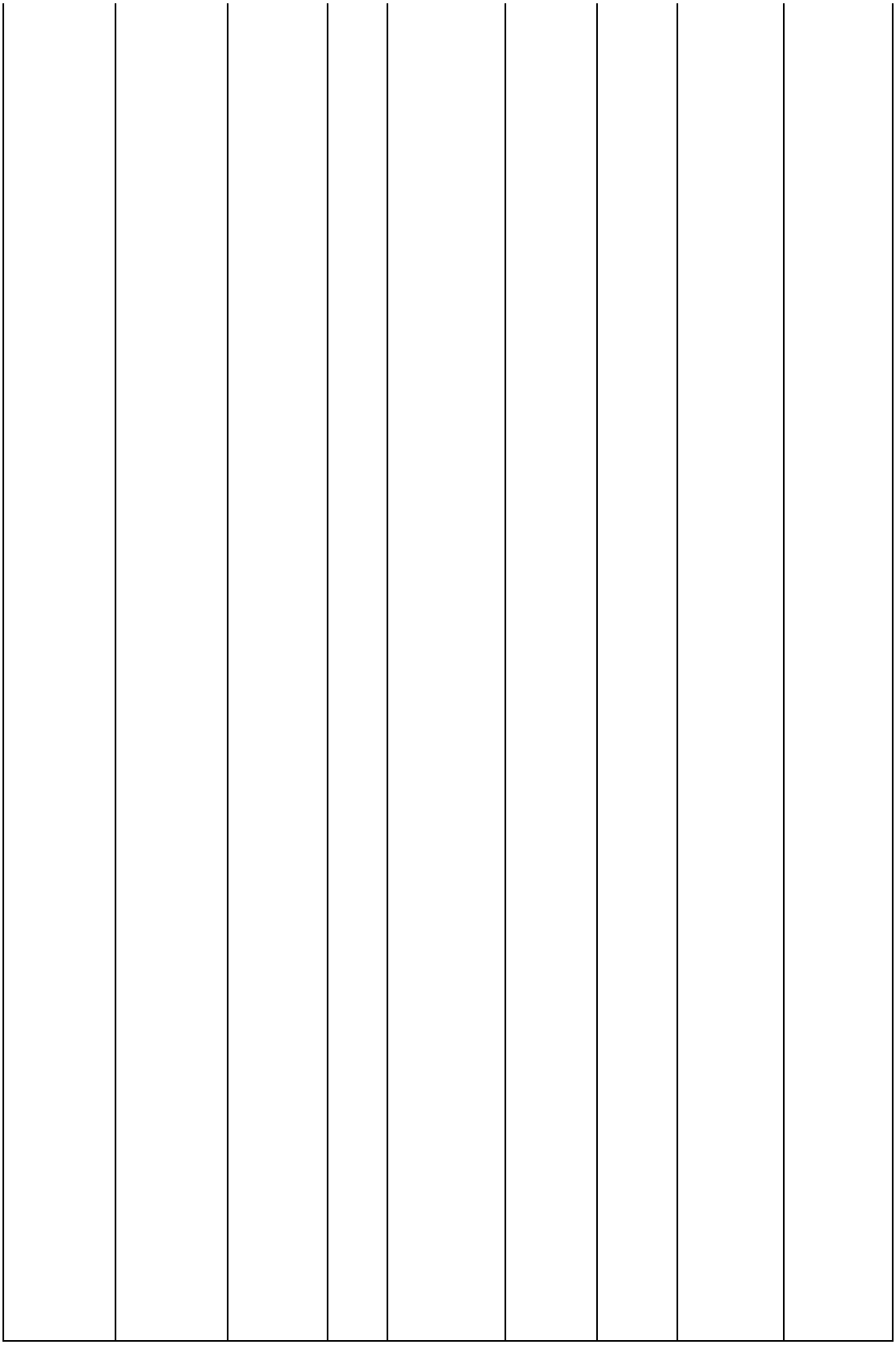
Note that the weights computed in this way, are correction weights. The final weights are obtained by multiplying the inclusion weights with these correction weights.

The final thing to do in Bascula is to write the final weights to the Blaise data file. The *Write button* takes care of this.

After leaving Bascula, it is a good idea to check the data file. You can take a look with the option *View data* in the *Tools menu* of the Blaise Control Centre. Figure 3.2.3 displays the contents of the first 10 records of the data file.

Figure 3.2.3. The first 10 records of the sample data file aftere weighting





The only difference with figure 3.1.2 is the contents of the column AdjWeight. The zeros are replaced by the real values of the adjustment weights.

3.3. Tabulation with Abacus

The next step in the statistical production process could be to make some publication tables. It is now shown that it is very easy to make unweighted and weighted tables with Abacus. For the weighted tables, Abacus can use the weights that were computed by Bascula.

First, a simple unweighted table is constructed. Abacus is accessed through the *Tools menu* of the Blaise Control Centre. All fields are selected for tabulation. A two-dimensional table is defined. The fields *Province* and *District* are used for the rows of the table, where *District* is nested in *Province*. The fields *Employ* and *Sex* are used for the columns, and *Sex* is nested in *Employ*. The resulting table is shown in figure 3.3.1.

Figure 3.3.1. The unweighted table

It is clear that this table contains sample frequencies, since the grand total of the table is equal to the sample size, which is 100.

To make estimates of population quantities, weights have to be incorporated in the table. That can be done in Abacus. The *Edit menu* of Abacus has the option *Calculations*. After activating this option, the weight variable can be selected by pressing the *Weight button*. The field *AdjWeight* is selected for this purpose. If the table is now recomputed, the result will be as displayed in figure 3.3.2.

The counts in the table add up to 1000, and that is the population size. Note that the totals for the field *District* exactly match the frequency distribution in figure 3.2.2. This is consequence of the fact that this field was used in the weighting model.

Figure 2.3.2. The weighted table

This section illustrates the use of Bascula as a tool in the statistical production process. Since the Blaise Control Centre handles all data and meta-data aspects, the use of Bascula becomes simple and straightforward. Bascula 2.0 was used to process the example in this section. Although Bascula is very useful to compute weighted estimates, it still has its limitations. Possible extensions of Bascula are discussed in the next section.

4. Future developments

Statistics Netherlands uses Bascula 2.0 to carry out adjustment weighting for several surveys. Experience has shown that this package is not applicable in every situation. Therefore, there is a demand to enhance the package with new functions. A project has been started to accomplish this. Work is in progress that will lead to a new version 3 of Bascula. In this section, an overview is given of some of the new features.

4.1. Limited weights

There are several reasons why a statistician may want to have some control over the values of the adjustment weights. One reason is that extremely large weights are generally considered undesirable. Large weights usually correspond to population elements with rare characteristics. Use of such weights may lead to unstable estimates of population parameters. To reduce the impact of large weights on estimators, a weighting method is required that keeps the adjustment weights within pre-specified boundaries, and at the same time enables valid inference.

Another reason to have some control over the values of the adjustment weights is that application of linear weighting might produce negative weights. Formally, this is not incorrect. The theory does not prevent negative weights. Usually, negative weights are an indication that the used regression model does not fit the data very well. Negative weights may cause problems in subsequent analysis. For example, if analysis is done on a sub-sample of the data containing a high portion of negative weights, the estimates will be inaccurate. Another problem is caused by some statistical analysis packages that can work with weights, but expect these weights to be positive. These packages are not able to properly process data sets containing negative weights.

To force weights within certain limits, several techniques have been proposed. A technique developed by Deville et al. (1993) comes down to repeating the weighting process (either linear or multiplicative) a number of times. First a lower bound L and an upper bound U are specified. After the

first run, weights smaller than L are set to L and weights larger than U are set to U. Then, the weighting process is repeated, but records from the strata with the fixed weights L and U are excluded. Again, weights may be produced not satisfying the conditions. These weights are also set to either the value L or U. The weighting process is repeated until all computed weights fall within the specified limits. Convergence of this iterative process is not guaranteed. Particularly, if the lower bound L and upper bound U are not far apart, the process may not converge.

In Bascula 3.0, a technique developed by Huang and Fuller (1978) will be implemented. Their algorithm produces weights that are a smooth, continuous, monotone increasing function of the original weights computed from the linear model. The algorithm is iterative. At each step, the weights are checked against a user-supplied criterion value M. This value M is the maximum fraction of the mean weight by which any weight may deviate from the mean weight. For example, if M is set to 0.75, then all weights are forced into the interval with lower bound equal to 0.25 times the mean weight and upper bound equal to 1.75 times the mean weight. Setting the value to 1 implies that all weights are forced to be positive.

One iteration of the algorithm consist of computing weights based on the linear regression model, where an adjustment factor g_i applied to every case i. This adjustment factor is the result of a 'bell' shaped function. The adjustment factor is large for vectors $X_i = (X_{i1},X_{i2},...,X_{ip})'$ close to the mean of the auxiliary variables, and the value of the factor approaches zero as the distance to the mean becomes larger.

The iterative process is repeated until all computed weights satisfy the criterion value. The value of M must be chosen with some care. If the value of M is too small, the algorithm may fail to find a solution. Huang and Fuller (1978) suggest to take a value of M that satisfies the condition

$$0.5 \le M \le 1 - \frac{n}{N}. \qquad (3.1.1)$$

Huang and Fuller (1978) also prove that the asymptotic properties of the regression estimator constructed with their algorithm are asymptotically the same as those of the general regression estimator. So, restricting the weights has (at least asymptotically) no effect on the properties of population estimates computed with the these weights.

4.2. Household weights

Household surveys are usually based on a hierarchical data model. The highest level in the data model is the household. Households are composed of persons and they represent the next level in the model. Such surveys collect information about both levels in the data model. The collected information can be used to make estimates for two populations: the population consisting of all households, and the population consisting of all individual persons.

If the aim of the survey is to make inference on the population of all individual persons, the process is fairly straightforward. The unit of measurement is the individual person. The data file must be approached

as a file of records with data on persons. Available population information on the distribution of personal characteristics can be used to compute adjustment weights, and these weights are assigned to the individual records.

For making inference on the population of households, the same approach can be used. However, there is a problem. Usually there is no information available on the population distribution of household variables. Even information on simple variables like size of the household and household composition is lacking. This makes it impossible to carry out an efficient weighting procedure.

Since it is possible to compute weights for the members of the household, one may wonder whether it is possible to use the person weights in some way to compute a household weight. Possible approaches could be to take (1) the weight of the head of the household, (2) the weight of a randomly selected household member, or (3) to compute some kind of average weight of the household members. Whatever approach is used, there are always problems :

- If the household weights are applied to the members of the households, the weighted estimates of personal characteristics will not match known population frequencies. This discrepancy will not occur if the personal weights are used.

- Inconsistenties may turn up. For example, an estimate of the total income through the households will not be equal to an estimate based on the individual persons.

Generalised regression estimation offers a solution to these problems. Suppose the population consist of N persons distributed over M households. A sample of m households is selected, and these households contain all together n persons.

A new nxm-matrix H is formed defining household membership. The element H_{ij} of H gets the value 1 if person i belongs to household j (for i=1,2,...,n and j=1,2,...,m). Otherwise, the value of H_{ij} is equal to 0. Next, the matrix Z is defined as Z = H'X. This matrix aggregates the values of the auxiliary variables within the households. For example, if there is one auxiliary variable Sex, then there are two dummies (one for male, and one for female). A row of Z represents a household, and the values in the row the number of males and females in that household.

The theory of linear weighting can now be applied, where the sample person data matrix X is replaced by the sample household data matrix Z. The same vector of totals X_T of the auxiliary variables is used in the computation of the weights. For more details about this technique, see Nieuwenbroek (1993).

4.3. Variance estimation

The current version of Bascula can compute adjustment weights. Use of such weights allows for the publication of better quality population statistics. However, to be able to judge the quality of statistics, some

indication of the quality of the published statistics is required. Therefore, it is important to be able to compute variances of estimates.

The theory of linear weighting provides formulae for the computation of the variances of estimators of population characteristics. However, computation of these variances requires the second order inclusion probabilities to be available. Particularly for large samples, this is a considerable computational effort. For multiplicative weighting there are no straightforward variances expressions.

The Department of Statistical Methods of Statistics Netherlands is now in the process of implementing a general variance estimation module that can be applied for both linear and multiplicative weighting, and that does not require second order inclusion probabilities to be known.

The variance estimation technique is based on the method of *Balanced Half Samples*. The general idea is the following. Assume that a stratified sample has been selected. Two elements have been selected (with replacement) from each stratum. Using the sample data, an estimator t for a population characteristic T can be computed.

By selecting one of the sample elements in each of the L strata, a so-called half sample is formed. Such a half sample can also be used to compute an estimate of the population characteristic. Many half samples are possible. In fact, there are 2^L such samples. Denote the estimator based on half sample α by t_α. Now, the variance of the estimator t can be estimated by the quantity

$$\frac{1}{K}\sum_{\alpha=1}^{K}(t_\alpha - t)^2, \qquad (4.3.1)$$

where K is the number of half samples.

One complication is that not every sampling design is a stratified design with two observations per strata. So, for other designs adjustments have to be made to obtain such a situation. Another complication is that for a large number of strata, the number of half samples can be very large. Since an estimate has to be computed for every half sample, the computational effort can be substantial. In order to avoid this, it is possible to work with a subset of half samples. This subset has to be selected very carefully. The theory of Balanced Half Samples provides means to generate such a set of balanced half samples. For more information on the BHS method, see e.g. Wolter (1985).

Implementation of the theory of BHS method will mean that for each half sample a weighting procedure is carried out, resulting in a set of weights. To compute an estimate of the variance of some statistic, the statistic is computed for all sets of weights, and then expression (4.3.1) is applied.

The final tool will not be built into Bascula. Instead, it will be a separate tool that will repeatedly call the Bascula module responsible for computing the weights.

4. References

Bethlehem, J.G and Keller, W.J. (1987), Linear weighting of sample survey data. Journal of Official Statistics 3, pp. 141-153.

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

Deville, J.C., Särndal, C.E. and Sautory, O. (1993) , Generalized raking procedures in survey sampling. Journal of the American Statistical Association 88, pp. 1013-1020.

Huang, E.T. and Fuller, W.A. (1978), Nonnegative regression estimation for sample survey data, Proceedings of the Social Statistics Section, American Statistical Association, pp. 300-305.

Nieuwenbroek, N.J. (1993), An integrated method for weighting characteristics of persons and households using the linear regression estimator. Report 8445-93-M1-1, Statistics Netherlands, Voorburg, The Netherlands.

Wolter, K.M. (1985), Introduction to variance estimatio, Springer Verlag, New York.