Estimating accuracy for statistics based on register and survey data

Léander Kuijvenhoven and Sander Scholtus

Discussion paper (10007)

Statistics Netherlands

The Hague/Heerlen, 2010

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2008–2009	= 2008 to 2009 inclusive
2008/2009	= average of 2008 up to and including 2009
2008/'09	= crop year, financial year, school year etc. beginning in 2008 and ending in 2009
2006/'07–2008/'09	= crop year, financial year, etc. 2006/'07 to 2008/'09 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

6008310007 X-10

Estimating accuracy for statistics based on register and survey data

Léander Kuijvenhoven and Sander Scholtus

Summary: In this paper, we consider the situation where data are collected from both registers and sample surveys. We show how the bootstrap resampling method can be applied in this situation, in order to obtain insight in the accuracy of statistics based on combined data. The method is applied to the Dutch Educational Attainment File.

Keywords: register, sample, combined data, bootstrap

1 Introduction

In this paper, we shall consider the situation where data are obtained from several different sources, including both administrative registers and probability samples. The combined data are then used to compile statistics on a population of interest, such as totals, averages, and proportions of target variables. Often the estimator is a combination of a register and a survey that is weighted. The problem with such statistics is that it is often difficult to derive measures of accuracy, like variances or confidence intervals.

In this paper, we are primarily interested in developing methodology to assess the accuracy of particular estimates from combined data; see Section 2. Other kinds of estimators based on combined data are described by Smith (2009). We do not discuss the problem of *micro integration* in itself, i.e. how to construct a statistical data base and how to handle inconsistencies between data from different sources. Instead, we assume that a statistical data base has already been constructed. The main topic of this paper is to show how the bootstrap resampling method can be applied in the situation of combined data from administrative sources and sample surveys. Other approaches, such as the jackknife, or multiple imputation of the non-sampled part of the population, are not discussed in this paper.

The outline of this paper is as follows. The setting of this paper is elaborated in Section 2. Section 3 discusses the classical bootstrap method. The method is extended to a combination of register and survey data in Sections 4 and 5. The results of an application to the Dutch Educational Attainment File are described in Section 6. The paper concludes with a discussion.

2 Surveys based on registers and samples

Statistical institutes are making increasing use of existing registers, to reduce costs and to reduce the burden on respondents. Such registers are often primarily used for non-statistical purposes, and therefore not ideal from a statistical perspective. In many cases, an additional sample survey is needed to obtain reliable statistical results. The problem of assessing the accuracy of estimates based on a combination of administrative sources and sample surveys has, therefore, become very relevant to statistical institutes.

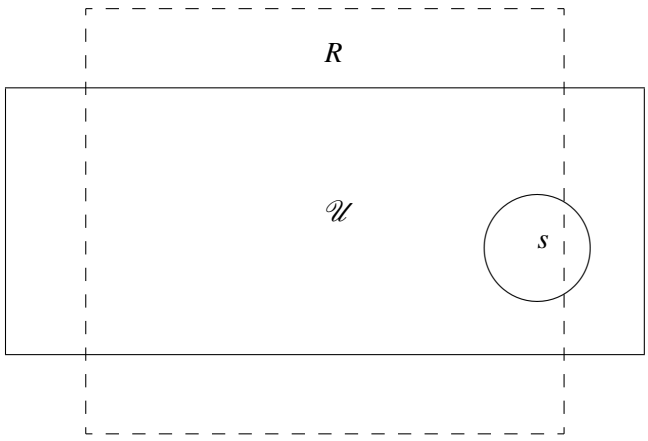
For convenience, we consider the case that a survey is based on one register and one survey sample. However, the register may be assembled from several smaller registers, and the sample may be a combination of several smaller samples. We denote the register by R and the sample by s. Let $\mathscr{U}$ denote the target population. Figure 1 shows the relationship between $\mathscr{U}$, R and s in graphical form.

Conditional on a realisation of the register, we divide $\mathscr{U}$ into the part that can be matched to R,

$$\mathscr{U}_R = \mathscr{U} \cap R,$$

and its complement $\mathscr{U}_{NR} = \mathscr{U} \setminus \mathscr{U}_R$. By construction, information on all elements of $\mathscr{U}_R$ can be found in the register. However, due to overcoverage, the register may also

Figure 1: The target population (rectangle), the register (dashed rectangle), and the sample (circle).



contain information on elements outside the target population. On the other hand, information on $\mathscr{U}_{NR}$ must be inferred from s. Note that s is originally sampled from the target population $\mathscr{U}$, and not from $\mathscr{U}_{NR}$. Therefore, we define

$$s' = s \setminus (s \cap R).$$

This represents the part of the sample that cannot be matched to the register.

Consider a target parameter θ. For instance, suppose that we are interested in estimating the total value of a certain variable y:

$$\theta = \sum_{k \in \mathscr{U}} y_k = \sum_{k \in \mathscr{U}_R} y_k + \sum_{k \in \mathscr{U}_{NR}} y_k = \theta_R + \theta_{NR}. \tag{1}$$

Using the available information in the survey, an estimate $\hat{\theta}$ of θ can be computed. For (1) we have:

$$\hat{\theta} = \sum_{k \in R} y_k + \sum_{k \in s'} w_k y_k = \hat{\theta}_R + \hat{\theta}_{NR}, \tag{2}$$

with w_k the final weight of each element in s'. Often, these final weights are calibrated on known or previously estimated marginals of $\mathscr{U}_{NR}$, so that they cannot be determined until after micro integration of the different sources.

It is natural to ask about the accuracy of the estimate $\hat{\theta}$. In particular, we are interested in its bias and variance. Due to the complexity of combining registers and sample surveys, it is not clear how these measures should be evaluated analytically. For complex situations like this, an approach based on *resampling* can be useful; cf. Shao and Tu (1995). In this paper, we discuss how the well-known bootstrap resampling method could be adapted to our problem.

Instead of weighting the elements of s' to represent $\mathscr{U}_{NR}$, it is also possible to impute the records of the non-sampled elements of $\mathscr{U}_{NR}$. In that case, multiple imputation (see e.g. Rubin (1987)) could be considered as an alternative method for assessing the bias and variance of $\hat{\theta}$. We did not explore this possibility, because we originally developed our method with the application to the Dutch Educational Attainment File,

to be discussed in Section 6, in mind. In this application, the users prefer working with weights rather than imputations.

We make the following assumptions:

- The impact of errors in the register on $\hat{\theta}$ is considered negligable compared to the sampling variance.
- The subpopulation $\mathscr{U}_R$ is considered fixed. Thus, the sample survey is the only source of variance in $\hat{\theta}$.

Under these assumptions, we can (and shall) ignore the difference between θ_R and $\hat{\theta}_R$ in (1) and (2). Measuring the accuracy of $\hat{\theta}$ as an estimate for θ thus becomes equivalent to measuring the accuracy of $\hat{\theta}_{NR}$ as an estimate for θ_{NR}.

3 Introduction to the bootstrap

In this section, we introduce the classical bootstrap method and briefly discuss adaptations of the bootstrap to finite population sampling. In the next two sections, we shall discuss how to extend the bootstrap methodology to a combination of register data and sample survey data.

The bootstrap idea is to mimic the generating process leading to the originally observed data, by estimating the target population from the sample and then resampling from the estimated population. A large number of bootstrap replicates of the target estimator is found by applying the algorithm that produced the original estimate to samples taken from this estimated population.

Let $\mathscr{U}$ denote the target population, $\hat{\mathscr{U}}$ its estimate based on the observed data and $t(.)$ the algorithm that when applied to the population gives the parameter of interest and when applied to the observed data gives the estimate, that is: $\theta = t(\mathscr{U})$ and $\hat{\theta} = t(\hat{\mathscr{U}})$[1]. Then the classical bootstrap method is to resample from $\hat{\mathscr{U}}$, the empirical distribution function (EDF), to obtain a bootstrap sample $\hat{\mathscr{U}}^*$ and a statistic $\hat{\theta}^* = t(\hat{\mathscr{U}}^*)$. To estimate the bootstrap distribution, this procedure is repeated B times independently, yielding replicates $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$. These are then used to estimate the variance of $\hat{\theta}$ by

$$\begin{aligned} \hat{S}^2_B(\hat{\theta}) &= \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_b - \overline{\hat{\theta}^*_B}\right)^2, \qquad (3) \\ \overline{\hat{\theta}^*_B} &= \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b. \end{aligned}$$

The classical bootstrap is normally applied to a single homogeneous sample taken independently from $\mathscr{U}$. It is then common to take $\hat{\mathscr{U}}$ to be the EDF that puts mass n^{-1} on each element of the original sample (where n denotes the sample size), and to generate $\hat{\mathscr{U}}^*$ by sampling with equal probability and with replacement from the originally observed data.

[1] $t(.)$ is in effect a plug-in estimator.

An important problem with the classical bootstrap arises when it is applied to finite population sampling, namely how to mimic the effect of sampling without replacement. In order to obtain a valid measure of the variance of the estimate, it is crucial to capture the effect of the sampling design. In particular, sampling without replacement leads to a smaller variance than sampling with replacement.

There are a number of methods suggested in the literature to adapt the classical bootstrap to finite population sampling (see e.g. Shao and Tu (1995)). However, these methods tend to be difficult to apply in practice. Here, we follow the approach suggested by Canty and Davison (1999), which is to explicitly construct the estimated population $\hat{\mathscr{U}}$. Suppose that a simple random sample of size n is obtained from a population $\mathscr{U}$ of N elements, and suppose for convenience that N/n is an integer. Then a virtual population $\hat{\mathscr{U}}$ is built by taking N/n copies of each sampled element. In the case of stratified sampling, this approach is applied separately to each stratum. Once the estimated population has been made, we can simply apply the original sampling design to obtain the bootstrap samples.

4 A generalisation of the bootstrap

In this section and the next, we discuss how to extend the method of Canty and Davison (1999) to the situation where data from administrative sources and sample surveys are combined, as described in Section 2. The current section deals with the construction of bootstrap samples, while Section 5 discusses how these bootstrap samples may be used to estimate the accuracy of statistics.

4.1 Overview

When trying to apply the bootstrap method in our setting, there are a number of complicating factors that need to be addressed:

1. The original sample design refers to the sample s obtained from $\mathscr{U}$, and not to s' from $\mathscr{U}_{NR}$. Note that this means that, although R is considered fixed, the overlap between s and R is stochastic, and thus a source of variance. In order to take this source of variance into account in our bootstrap method, we use s, and the original inclusion probabilities used for sampling s from $\mathscr{U}$, to construct the virtual population $\hat{\mathscr{U}}$. Thus, the virtual population $\hat{\mathscr{U}}$ is constructed without any reference to the register.

   We use the following notation: $\mathscr{U}$ contains $M+N$ elements, of which M belong to $\mathscr{U}_R$ and N belong to $\mathscr{U}_{NR}$. The original sample s contains $m+n$ elements of $\mathscr{U}$, of which m belong to $\mathscr{U}_R$ and n belong to $\mathscr{U}_{NR}$. (Thus s' contains n elements.) To each $k \in \mathscr{U}$ there is associated an inclusion probability π_k, with $0 < \pi_k \leq 1$ and

$$\sum_{k \in \mathscr{U}} \pi_k = m+n. \tag{4}$$

For $k \in s$ we also define an inclusion weight d_k proportional to $1/\pi_k$. The inclusion weights are scaled to have the following property:

$$\sum_{k \in s \cap R} d_k = M \text{ and } \sum_{k \in s'} d_k = N. \tag{5}$$

Note that we do not make any further assumptions about the inclusion probabilities. Thus, we consider a general unequal probability sampling design here.

In analogy to Canty and Davison (1999), a virtual population $\hat{\mathscr{U}}$ is constructed by taking d_k copies of the k-th sampled element. Next, bootstrap samples are obtained by sampling from $\hat{\mathscr{U}}$. To this end, each element of $\hat{\mathscr{U}}$ is given an inclusion probability, as we will discuss further in Section 4.3.

2. Due to nonresponse only a part of the original sample is observed. In our notation, s refers to the set of $m+n$ respondents, so nonrespondents have already been removed from the sample. Since $\mathscr{U}$ refers to the whole population of $M+N$ elements (and not to a "population of respondents"), the inclusion probabilities π_k have been adjusted for the amount of nonresponse (see (4)). Implicitly, we use the so-called Fixed Response Model, cf. Bethlehem (2009): the population is divided into a stratum of respondents (who always respond when sampled) and a stratum of nonrespondents (who never respond when sampled). Data are observed for the intersection of the sample with the stratum of respondents only.[2]

   Since respondents and nonrespondents may differ in some respects, we cannot expect the virtual population $\hat{\mathscr{U}}$ constructed above to fully represent the original population $\mathscr{U}$. Canty and Davison (1999) argue that the bootstrap method is valid provided that the same model that was used to correct the weights for nonresponse in the original sample, is also applied to the weights for each bootstrap resample, under the assumption that the weighting model indeed explains nonresponse behaviour. By this approach, each weighted bootstrap resample will correctly represent the original population. We return to this point in Section 4.4.

3. Since it is not expected that all d_k are integers, it is not possible to just put d_k copies of the k-th sampled element in $\hat{\mathscr{U}}$. Section 4.2 describes a solution to this minor problem.

4.2 Rounding the inclusion weights

To construct a virtual population $\hat{\mathscr{U}}$ of $M+N$ elements from s, we want to take d_k copies of each respondent, exploiting the fact that (5) holds. As d_k need not be integer-valued in general, we apply Fellegi's method for consistent rounding to round these weights to integer-valued "blow up weights" δ_k that satisfy

$$\sum_{k \in s \cap R} \delta_k = M \text{ and } \sum_{k \in s'} \delta_k = N. \tag{6}$$

[2] In principle, the method could be modified to assume the more common Random Response Model, where each element of the population has an associated probability to respond when sampled; see Canty and Davison (1999). Again, we did not explore this, because the method was originally developed for the Dutch Educational Attainment File (see Section 6), where response probabilities could not be estimated.

A description of Fellegi's method is given in Appendix A.

Thus, $\hat{\mathscr{U}}$ is obtained by taking exactly δ_k copies of the k-th element from s.

4.3 Determining inclusion probabilities for $\hat{\mathscr{U}}$

We shall use the index $j = 1, \ldots, M+N$ to denote the elements of the virtual population $\hat{\mathscr{U}}$. In order to obtain bootstrap samples of size $m+n$ from $\hat{\mathscr{U}}$, each element j must be given an inclusion probability π_j^*, such that

$$\sum_{j \in \hat{\mathscr{U}}} \pi_j^* = m+n. \tag{7}$$

An obvious choice is to give each copy of $k \in s$ an inclusion probability π_k' proportional to π_k, using a scaling factor to satisfy (7). Since there are δ_k copies of the k-th original element in $\hat{\mathscr{U}}$, (7) may then be written as

$$\sum_{k \in s} \delta_k \pi_k' = m+n. \tag{8}$$

Therefore it suffices to determine $\pi_k' \propto \pi_k$ for $k \in s$ that satisfy (8), and then define, for $j \in \hat{\mathscr{U}}$, $\pi_j^* = \pi_k'$ if the j-th element of $\hat{\mathscr{U}}$ is a copy of the k-th element of s. We describe a simple method to determine π_k' in Appendix B.

4.4 Drawing and weighting bootstrap samples

We have discussed how to construct a virtual population $\hat{\mathscr{U}}$ from the original sample s, and how to assign inclusion probabilities π_j^* to the elements of $\hat{\mathscr{U}}$. A large number, say B, of bootstrap samples may now be obtained by repeatedly sampling without replacement from $\hat{\mathscr{U}}$ using π_j^*. The bootstrap samples are drawn independently of each other.

We recall that the original sample was used to construct the virtual population, in order to take the stochastic nature of the overlap between the sample and the register into account. Once we have obtained a bootstrap sample, all copies of overlapping elements may be removed, since the sample is only used to make inferences about $\mathscr{U}_{NR}$. By removing the elements that are also present in R, we are left with a set of bootstrap samples containing copies of elements of s'. By (6), we know that $\hat{\mathscr{U}}$ contains N such elements, and for simplicity we assume that they are numbered $j = 1, \ldots, N$.

Let $a_{b,j}^*$ denote the sample inclusion indicator of the j-th virtual element in the b-th bootstrap sample:

$$a_{b,j}^* = \begin{cases} 1 & \text{if } j \text{ is included in the } b\text{-th bootstrap sample} \\ 0 & \text{else} \end{cases}$$

for $j = 1, \ldots, N$ and $b = 1, \ldots, B$.

For each bootstrap sample, we calibrate the inclusion weights $d_{b,j}^* = 1/\pi_j^*$ by applying the same weighting model that was used to correct for nonresponse in the original sample. Note that known marginals of $\mathscr{U}_{NR}$ are used in this weighting model. We denote the final weights for the b-th bootstrap sample by $w_{b,j}^*$, $j = 1, \ldots, N$, where by definition $w_{b,j}^* = 0$ if $a_{b,j}^* = 0$.

5 Estimating the accuracy of statistics

In this section, we discuss how the weighted bootstrap samples that were obtained in Section 4 may be used to estimate the accuracy of statistics based on a combination of register data and survey data. Section 5.1 deals with variance estimates and Section 5.2 deals with confidence intervals.

5.1 Estimating the variance

In Section 2 we considered the estimator (2) for (1). Note that the first term in (1) depends only on $\mathscr{U}_R$ and the second term depends only on $\mathscr{U}_{NR}$. Thus, $\theta = t(\mathscr{U})$ is decomposed into $\theta_R = t(\mathscr{U}_R)$ and $\theta_{NR} = t(\mathscr{U}_{NR})$. The original data provide the following estimate $\hat{\theta}$ of θ:

$$\hat{\theta} = \theta_R + \hat{\theta}_{NR}, \tag{9}$$

where $\hat{\theta}_{NR} = t(\hat{\mathscr{U}}_{NR})$.

Each weighted bootstrap sample gives rise to a bootstrap replicate

$$\hat{\theta}^*_b = \theta_R + \hat{\theta}^*_{NR,b} = \sum_{k\in R} y_k + \sum_{j=1}^{N} w^*_{b,j} y^*_j, \qquad b = 1,\ldots,B,$$

where $y^*_j = y_k$ if the j-th element of $\hat{\mathscr{U}}_{NR}$ is a copy of the k-th element of s', and the weights $w^*_{b,j}$ were defined in Section 4.4.

From the bootstrap replicates, the variance of $\hat{\theta}$ is estimated by $\hat{S}^2_B(\hat{\theta})$ from (3). However, since θ_R is considered fixed, the variance of $\hat{\theta}$ is just the variance of the second term in (9). Instead of (3) we may therefore use

$$\begin{aligned} \hat{S}^2_B(\hat{\theta}) &= \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_{NR,b} - \overline{\hat{\theta}^*_{NR,B}}\right)^2, \\ \overline{\hat{\theta}^*_{NR,B}} &= \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_{NR,b}. \end{aligned} \tag{10}$$

To compare the accuracy of different estimates of the form (9), for instance the cell counts in a cross-tabulation, we suggest to use the estimated relative inaccuracy:

$$\hat{\mathrm{rel}}_B(\hat{\theta}) = \frac{\sqrt{\hat{S}^2_B(\hat{\theta})}}{\hat{\theta}}.$$

The relative inaccuracy is a dimensionless quantity which measures the estimated standard deviation of $\hat{\theta}$ as a proportion of the estimate itself. To obtain a very simple publication protocol, we can declare that an estimate is accurate enough to be published, if its relative inaccuracy is sufficiently small. That is: $\hat{\theta}$ is publishable if

$$\hat{\mathrm{rel}}_B(\hat{\theta}) < \beta,$$

with β some predetermined upper bound.

Formula (3) can also be used to estimate the variance of more general estimates, by defining $\hat{\theta}^*_b = t(\hat{\mathscr{U}}^*_b)$, where $\hat{\mathscr{U}}^*_b$ denotes the union of the register population and the

b-th bootstrap sample from $\hat{\mathscr{U}}_{NR}$. For instance, suppose that we are interested in the ratio of two population totals,

$$\theta = \frac{\sum_{k\in\mathscr{U}} y_k}{\sum_{k\in\mathscr{U}} z_k},$$

which is estimated by

$$\hat{\theta} = \frac{\sum_{k\in R} y_k + \sum_{k\in s'} w_k y_k}{\sum_{k\in R} z_k + \sum_{k\in s'} w_k z_k}.$$

A variance estimate for this ratio estimate is obtained from the weighted bootstrap samples by formula (3), with

$$\hat{\theta}_b^* = \frac{\sum_{k\in R} y_k + \sum_{j=1}^N w_{b,j}^* y_j^*}{\sum_{k\in R} z_k + \sum_{j=1}^N w_{b,j}^* z_j^*}, \qquad b = 1,\ldots,B,$$

where z_j^* is defined in the same way as y_j^*. For ratios and proportions, it seems more natural to base a publication protocol directly on $\hat{S}_B^2(\hat{\theta})$, instead of $\hat{\mathrm{rel}}_B(\hat{\theta})$.

5.2 Estimating confidence intervals

Under the assumption that the bootstrap replicates $\hat{\theta}_1^*,\ldots,\hat{\theta}_B^*$ are normally distributed, the interval

$$\hat{\theta} - z^{(1-\alpha/2)}\sqrt{\hat{S}_B^2(\hat{\theta})} \le \theta \le \hat{\theta} + z^{(1-\alpha/2)}\sqrt{\hat{S}_B^2(\hat{\theta})}, \tag{11}$$

gives an estimated $100(1-\alpha)\%$-confidence interval for $\hat{\theta}$. Here, $z^{(1-\alpha/2)}$ is the $100(1-\alpha/2)$-th percentile of the standard normal distribution: $z^{(1-\alpha/2)} = \Phi^{-1}(1-\alpha/2)$, with Φ the cumulative distribution function of the standard normal distribution.

If the bootstrap replicates are not normally distributed it is better to apply the percentile method. This method estimates the $100(1-\alpha)\%$-confidence interval as follows:

$$\hat{G}^{-1}(\alpha/2) \le \theta \le \hat{G}^{-1}(1-\alpha/2), \tag{12}$$

where $\hat{G}^{-1}(\alpha/2)$ and $\hat{G}^{-1}(1-\alpha/2)$ are the $100\alpha/2$-th and $100(1-\alpha/2)$-th percentile of the bootstrap distribution. That is, if we order the bootstrap replicates in the following way,

$$\hat{\theta}_{(1)}^* \le \hat{\theta}_{(2)}^* \le \cdots \le \hat{\theta}_{(B)}^*,$$

then the $100(1-\alpha)\%$-confidence interval is estimated by:

$$\hat{\theta}_{(B\alpha/2)}^* \le \theta \le \hat{\theta}_{(B(1-\alpha/2))}^*,$$

with the assumption that $B\alpha/2$ is an integer.

There exist more sophisticated methods for estimating a confidence interval. Efron (1987) describes the so-called *BC*-interval. The $100(1-\alpha)\%$-*BC*-confidence interval of $\hat{\theta}$ equals

$$\hat{G}^{-1}\left(\Phi\left(2z_0 + z^{(\alpha/2)}\right)\right) \le \theta \le \hat{G}^{-1}\left(\Phi\left(2z_0 + z^{(1-\alpha/2)}\right)\right). \tag{13}$$

Here z_0 is a correction for the bias. This unknown parameter can be estimated from the bootstrap distribution (cf. Efron (1987)) by

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}_b^* < \hat{\theta}\}}{B}\right),$$

The standard confidence interval (11) sees the estimator $\hat{\theta}$ as a random variable with a $N(\theta, \hat{S}_B^2(\hat{\theta}))$-distribution. For the *BC*-interval (13), it is sufficient that there exist a transformation $\zeta = g(\theta)$, such that $\hat{\theta}$ becomes a random variable $\hat{\zeta}$ which is $N(\zeta - z_0, \hat{S}_B^2(\hat{\zeta}))$-distributed, cf. Efron (1987). This assumption is less severe. The *BC*-interval accounts for a possible bias in the bootstrap distribution with respect to the original estimator, by using the term z_0. Note that if $z_0 = 0$, then the *BC*-interval equals the confidence interval constructed with the percentile method.

For variance estimates and confidence interval (11) one usually takes 250 to 500 bootstrap replicates (Chernick (1999)). For the confidence intervals (12) and (13) one normally uses at least 1000 bootstrap replicates.

6 Application to the Dutch Educational Attainment File

Data on educational attainment are important for socio-economic research and government policy, because studies have shown that education is highly correlated with various social phenomena. For this reason, it was decided at Statistics Netherlands to combine available information on education from different sources at a given reference date into one Educational Attainment File (EAF). See Linder and van Roon (2009) for more information on the background of the EAF.

Statistics Netherlands uses two kinds of sources for the construction of the EAF. Firstly, there are a number of education registers, such as the Exam Results Register and the Central Register for Enrolment in Higher Education. However, as there is no central education register in The Netherlands, there is still a substantial part of the Dutch population without available register data. In particular, since most of the available partial registers have only come into existence in recent years, they do not contain information on persons who completed their education before that time. Therefore it is necessary to supplement the register data with a sample survey. For this the Labour Force Survey (LFS) is used.

After a complex process of integrating the different sources on the micro level (cf. Linder and van Roon (2009)), an EAF is constructed. The EAF contains variables on educational attainment for each person present in a register, and for each respondent of the LFS. Moreover, the LFS respondents are weighted to represent the part of the Dutch population that is not found in any register. These weights have been calibrated on various known marginals of the non-registered population.

By linking the EAF to other files from the Dutch Social Statistical Database (SSD), the variables on educational attainment can be combined with many different background variables. Thus, many different cross-tabulations can be performed. It is natural to ask about the accuracy of such cross-tabulations. Since the EAF is an example of the situation we described in Section 2, we decided to apply the bootstrap resampling method of Sections 4 and 5 to obtain measures of accuracy.

To illustrate the possibilities, Table 1 shows an example of a cross-tabulation obtained from the EAF. Eight groups of young immigrants in The Netherlands are differentiated by highest completed education level. The education levels are numbered from 1 to 5, 1 being the lowest level (only primary education) and 5 being the highest (final stages

Table 1: Estimates of the number of persons in eight small subpopulations, by highest completed education level; the estimated standard deviation and relative inaccuracy from 500 bootstraps are shown in brackets.

subpopulation	1	2	3	4	5	total
Moroccan men	1539	8151	7874	670	223	18457
18–24 yrs	(310)	(463)	(503)	(163)	(129)	(692)
	(20%)	(6%)	(6%)	(24%)	(58%)	(4%)
Moroccan women	2060	7812	10686	954	56	21568
18–24 yrs	(322)	(431)	(547)	(145)	(−)	(629)
	(16%)	(6%)	(5%)	(15%)	(−)	(3%)
Moroccan men	2940	4886	7022	1489	936	17273
25–30 yrs	(435)	(531)	(590)	(214)	(236)	(721)
	(15%)	(11%)	(8%)	(14%)	(25%)	(4%)
Moroccan women	3470	5204	8302	1584	778	19338
25–30 yrs	(439)	(476)	(551)	(137)	(163)	(706)
	(13%)	(9%)	(7%)	(9%)	(21%)	(4%)
Turkish men	2491	8654	7737	744	168	19794
18–24 yrs	(364)	(507)	(499)	(195)	(53)	(679)
	(15%)	(6%)	(7%)	(26%)	(32%)	(3%)
Turkish women	2489	8221	9869	961	215	21755
18–24 yrs	(375)	(481)	(485)	(175)	(91)	(641)
	(15%)	(6%)	(5%)	(18%)	(42%)	(3%)
Turkish men	3842	6520	8671	1350	701	21084
25–30 yrs	(444)	(588)	(587)	(134)	(142)	(720)
	(12%)	(9%)	(7%)	(10%)	(20%)	(3%)
Turkish women	4177	5994	8802	1255	688	20916
25–30 yrs	(421)	(450)	(524)	(104)	(128)	(694)
	(10%)	(8%)	(6%)	(8%)	(19%)	(3%)

of tertiary education). Using $B = 500$ bootstrap replicates, we estimated the standard deviation for each cell count in the cross-tabulation. We also computed the relative inaccuracy. The results are displayed in Table 1. Note that the total number of persons in each subpopulation is also estimated and therefore has a non-zero variance.

We found that 500 bootstrap replicates were sufficient to obtain reliable estimates of the standard deviation. As an example, Figure 2 shows the convergence of the estimated standard deviation in one of the cells of Table 1.

The owners of the EAF feel that an estimate should only be published if its relative inaccuracy is less than 20%. Thus, from the estimated relative inaccuracies displayed in Table 1, it can be seen that nearly all cells on lower education levels can be published[3], but that most of the estimated cell counts on the highest education level are too inaccurate. This can be explained by the fact that for these cells, the number of LFS respondents tends to be small. Table 2 displays the number of LFS respondents for each cell from Table 1.

We remark that the cell of Moroccan women aged 18–24 with highest completed ed-

[3]The cell counts in Table 1 should be rounded off before being published, because otherwise a degree of precision is suggested that cannot be fulfilled in this case.

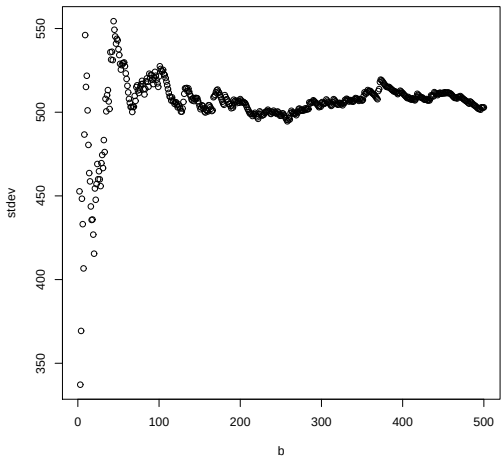


Figure 2: Estimated standard deviation for the estimated number of Moroccan men aged 18–24 with highest completed education level 3, using b bootstraps ($b = 1, \ldots, 500$).

Table 2: Number of LFS respondents for each cell count in Table 1.

subpopulation	1	2	3	4	5	total
M. men (18–24 yrs)	9	33	29	5	2	78
M. women (18–24 yrs)	19	38	55	7	0	119
M. men (25–30 yrs)	38	52	70	9	11	180
M. women (25–30 yrs)	60	99	149	6	9	323
T. men (18–24 yrs)	22	48	40	6	1	117
T. women (18–24 yrs)	29	54	69	8	3	163
T. men (25–30 yrs)	61	106	130	6	9	312
T. women (25–30 yrs)	83	125	207	11	9	435

ucation level 5 does not contain any LFS respondents. This means that the estimated cell count is completely based on the register. Since our bootstrap method only measures the inaccuracy due to sampling, we cannot use it to make any statement on the accuracy of this cell count.

7 Discussion

In this paper, we have shown how the bootstrap resampling method for sampling without replacement from a finite population, as discussed by Canty and Davison (1999), can be extended to a setting where data are collected from both administrative registers and sample surveys. This methodology may be used to obtain insight in the accuracy of statistics based on combined data. The method was successfully applied to the Dutch Educational Attainment File.

In an implementaton of this bootstrap method, to a large extent use can be made of existing applications. In particular, obtaining bootstrap samples from the virtual population, applying a weighting scheme to these resamples, and computing the expressions in Section 5.1 are all very similar to regular statistical production, and every statistical office has software available to perform these steps. We used the Blaise applications Manipula and Bascula for this. The construction of the virtual population itself can be implemented using general statistical software such as SPSS. A large amount of computational work is required to perform the steps described in Section 4. However, once this initial work has been done, the same bootstrap resamples can be used to measure the accuracy of any estimate based on the original data, at virtually no additional costs.

In our discussion, we have considered the register as fixed. In addition, we assumed that errors in the register could be ignored because their effect on the estimates was negligable compared to the sampling variance. Because of these assumptions, the method should not be used if an estimate is solely based on the register, i.e. if the original sample does not contain any contributing elements. We saw an example of this at the end of Section 6. In this case, the method would assume that all population elements that contribute to the estimate are covered by the register. In general, assessing the validity of this assumption is a very difficult problem.

In order to obtain more insight into the quality of the register, the register could be regarded as a sample drawn from a superpopulation. This may be an interesting topic for future research.

Appendix A Fellegi's method

Fellegi's method for consistent rounding is a generic method for rounding real numbers $g_1, \ldots, g_k$ with the property that

$$g_1 + \cdots + g_{k-1} = g_k,$$

in such a way that the rounded numbers $\gamma_1, \ldots, \gamma_k$ still satisfy the above summation. This method is stochastic and has the property that for each rounded number: $E(\gamma_j) = g_j$ and $\lfloor g_j \rfloor \leq \gamma_j \leq \lceil g_j \rceil$. See Fellegi (1975).

In this case, we want to round off the inclusion weights d_k, $k \in s$ such that it continues to hold that:

$$\sum_{k \in s \cap R} d_k = M \text{ and } \sum_{k \in s'} d_k = N.$$

Fellegi's method now consists of the following steps:

1. Compute for each k the remainder $\phi_k = d_k - \lfloor d_k \rfloor$, $\phi_k \in [0, 1)$.
2. Compute the partial sums $s_k = \sum_{l=1}^{k} \phi_l$.
3. Draw a random number ξ from the uniform distribution on $(0, 1]$.
4. Define the rounded weights by $\delta_k = \lceil d_k \rceil$ for all k in

$$\{k : s_{k-1} < \xi + u \leq s_k \text{ for some } u \in \mathbb{N}\},$$

   and $\delta_k = \lfloor d_k \rfloor$ for all other k.

This scheme is performed for $k \in s \cap R$ and for $k \in s'$ separately, so that we obtain:

$$\sum_{k \in s \cap R} \delta_k = M \text{ and } \sum_{k \in s'} \delta_k = N.$$

To prove the first identity, let x denote the number of elements in $s \cap R$, and observe that

$$M = \sum_{k \in s \cap R} d_k = \sum_{k \in s \cap R} \lfloor d_k \rfloor + \sum_{k \in s \cap R} \phi_k = \sum_{k \in s \cap R} \lfloor d_k \rfloor + s_x,$$

from which it follows that s_x is an integer. Therefore, there are exactly s_x weights that are rounded up in step 4, namely for those intervals that contain the numbers $\xi, \xi + 1, \ldots, \xi + s_x - 1$. Hence,

$$\sum_{k \in s \cap R} \delta_k = \sum_{k \in s \cap R} \lfloor d_k \rfloor + s_x = M.$$

The proof of the second identity is completely analogous.

Appendix B Determining inclusion probabilities for $\hat{\mathscr{U}}$

In this appendix we describe a simple approach to determine inclusion probabilities π'_k that satisfy (8). Starting with the original inclusion probabilities π_k, we perform the following steps:

1. Compute, for each $k \in s$:

$$\pi_k^{(1)} = (m+n)\frac{\pi_k}{\sum_{l\in s}\delta_l\pi_l}.$$

   By construction, the $\pi_k^{(1)}$ satisfy (8). However, it is possible that some $\pi_k^{(1)} > 1$, so these are not valid inclusion probabilities.

2. Compute, for each $k \in s$:

$$\pi_k^{(2)} = I\left\{\pi_k^{(1)} \geq 1\right\} + I\left\{\pi_k^{(1)} < 1\right\}(m+n-q^{(1)})\frac{\pi_k^{(1)}}{\sum_{l\in s}\delta_l\pi_l^{(1)}I\left\{\pi_l^{(1)} < 1\right\}},$$

   where

$$I\{A\} = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{else} \end{cases}$$

   and $q^{(1)} = \sum_{k\in s}\delta_k I\left\{\pi_k^{(1)} \geq 1\right\}$.

In the second step, all values of $\pi_k^{(1)} > 1$ are set to one, and all values of $\pi_k^{(1)} < 1$ are rescaled such that (8) also holds for $\pi_k^{(2)}$. Because of this rescaling, it may still happen that some $\pi_k^{(2)} > 1$. In that case, the second step has to be applied iteratively. As soon as we find $0 < \pi_k^{(t)} \leq 1$ for all $k \in s$, we define $\pi'_k = \pi_k^{(t)}$. Note that this happens after a finite number of iterations, because once a value has been set to one, it cannot become larger than one anymore.

For the final choice of π'_k, the associated inclusion probabilities π_j^* in the virtual population satisfy (7). This can be seen as follows:

$$\begin{aligned}
\sum_{j\in\hat{\mathscr{U}}}\pi_j^* &= \sum_{k\in s}\delta_k\pi'_k \\
&= \sum_{k:\pi_k^{(t-1)}\geq 1}\delta_k\pi'_k + \sum_{k:\pi_k^{(t-1)}<1}\delta_k\pi'_k \\
&= \sum_{k:\pi_k^{(t-1)}\geq 1}\delta_k + \sum_{k:\pi_k^{(t-1)}<1}\delta_k(m+n-q^{(t-1)})\frac{\pi_k^{(t-1)}}{\sum_{l:\pi_l^{(t-1)}<1}\delta_l\pi_l^{(t-1)}} \\
&= q^{(t-1)} + (m+n-q^{(t-1)}) \\
&= m+n.
\end{aligned}$$

References

Bethlehem, J. (2009), *Applied Survey Methods: a statistical perspective*, Wiley, Hoboken.

Canty, A. J. and Davison, A. C. (1999), 'Resampling-based variance estimation for labour force surveys', *The Statistician* **48**, pp. 379–391.

Chernick, M. R. (1999), *Bootstrap Methods: A Practitioner's Guide*, John Wiley & Sons, New York.

Efron, B. (1987), 'Better bootstrap confidence intervals', *Journal of the American Statistical Association* **82**, pp. 171–185.

Fellegi, I. (1975), 'Controlled random rounding', *Survey Methodology* **1**, pp. 123–133.

Linder, F. and van Roon, D. (2009), 'Deriving educational attainment by combining data from administrative sources and sample surveys; recent developments towards the 2011 census', International Conference Statistics Investment in the Future, Prague, 14-15 September 2009.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer, New York.

Smith, P. (2009), 'Methodological challenges in integrating data collections in business statistics', Paper prepared for NTTS, Brussels.