Discussion paper

Variances of Census Tables after Mass Imputation

Sander Scholtus

December 2018

Content

Summary

In this report we consider variance estimation for frequency tables of cross-classifications that occur in the Dutch virtual population Census, when mass imputation is used to predict all missing values of educational attainment in the population. We develop two alternative approaches: an analytical variance approximation and a bootstrap method. Both approaches are tested in a small simulation study.

Keywords

statistical methods, imputation, censuses, education level

1. Introduction

In this report we consider variance estimation for frequency tables of cross-classifications that involve a variable with missing values that have been predicted by mass imputation. The motivation for studying this problem comes from the next Dutch virtual population Census. Most variables in the Census are available from administrative sources with (near-)complete population coverage. An exception occurs for educational attainment, which is observed partly in education registers and partly in the Labour Force Survey. For about 7 million Dutch persons (of a total of 17 million), educational attainment is not observed. For the next Census, an approach has been developed that imputes all missing values of educational attainment in the population (De Waal and Daalmans, 2018). This naturally leads to the question how to determine the precision of estimated entries in frequency tables based on the imputed data.

In Section 2 we derive an analytical approximation to the variance of an estimated count in a cross-classification table based on data after mass imputation. The resulting formula consists of three terms, two of which are generic and straightforward to compute. The third term depends on the precise form of the imputation model. In Sections 3 and 4 we derive an explicit algorithm to compute this term for the model that has been developed to impute educational attainment in the Dutch virtual Census (a so-called continuation-ratio logistic regression model). In Section 5, an alternative variance estimation method is proposed based on a finite-population bootstrap algorithm that was developed previously by Kuijvenhoven and Scholtus (2011). Both approaches are tested in a small simulation study on synthetic data in Section 6. Some concluding remarks follow in Section 7.

It should be noted that, throughout this report, we use the variable educational attainment as a running example, but the underlying ideas can be applied more generally. The analytical approximations in Section 2 are readily applicable to other categorical variables. The expressions in Sections 3 and 4 are more specific but could also be applied to other variables that are measured on an ordinal scale. The bootstrap method in Section 5 is very general and could be applied to ‘any’ statistic based on mass-imputed data.[1]

[1] The author would like to thank Jacco Daalmans, Jeroen Pannekoek, Eric Schulte Nordholt, and Ton de Waal for helpful discussions about this topic. Special thanks are due to Ton de Waal and Arnout van Delden for reviewing this paper.

2. Analytical variance estimation after mass imputation - general case

2.1 Introduction

We consider the estimation of frequency tables of cross-classifications that include the variable educational attainment. The true count of units in a particular cell of such a frequency table can be written as $\theta_{hc} = \sum_{i \in U} h_i y_{ci}$. Here, U denotes the target population; y_c is an indicator such that $y_{ci} = 1$ if person i has education level c and $y_{ci} = 0$ otherwise ($c = 1, \dots, C$); finally, h is a similar 0-1-indicator for the cross-classification of all other variables in the table. All variables other than educational attainment are supposed to be completely observed for all units in the population. Educational attainment is partially observed – for some units in an administrative source and for other units in a sample survey – and missing values on this variable are imputed throughout the population (mass imputation).

From the imputed data set, we obtain the following estimator for θ_{hc}:

$$\hat{\theta}_{hc} = \theta_{hc1} + \hat{\theta}_{hc2} = \sum_{i \in U_1} h_i y_{ci} + \left(\sum_{i \in S} h_i y_{ci} + \sum_{i \in U_2 \backslash S} h_i \tilde{y}_{ci} \right). \tag{1}$$

Here, U_1 consists of all persons in U with an education level that is observed in an administrative source. From the remaining persons in the population, $U_2 = U \backslash U_1$, a probability sample S is available with observed education levels. Finally, for all persons $i \in U_2 \backslash S$, the education level y_{ci} is unknown and replaced by an imputation $\tilde{y}_{ci}$ in (1). The aim of this section is to obtain a variance formula for the estimator $\hat{\theta}_{hc}$.

In general, we suppose that the missing values of y_{ci} are imputed by drawing – independently for each person $i \in U_2 \backslash S$ – a vector $(\tilde{y}_{1i}, \dots, \tilde{y}_{Ci})$ from a multinomial distribution with estimated probabilities $(\hat{p}_{1i}, \dots, \hat{p}_{Ci})$, so that for each i exactly one of the values $\tilde{y}_{ci}$ is equal to 1 and the other values are equal to 0. The estimated probabilities $\hat{p}_{ci}$ are based on the observed distribution of y_c in the sample S. In Section 3, we will introduce a particular model for these probabilities that is used in the Dutch virtual Census.

2.2 Derivation of the variance

The uncertainty in the estimator $\hat{\theta}_{hc}$ is determined both by the imputations themselves (which are stochastic) and by the uncertainty in the estimated probabilities $\hat{p}_{ci}$ on which these imputations are based. Conditioning on a realisation of the random sample S and using a standard decomposition formula for conditional variances, we can write the variance of $\hat{\theta}_{hc}$ as follows:

$$\operatorname{var}(\hat{\theta}_{hc}) = E_S\{\operatorname{var}(\hat{\theta}_{hc}|S)\} + \operatorname{var}_S\{E(\hat{\theta}_{hc}|S)\}.$$

From (1) and using the fact that imputations for different persons are independent, we obtain:

$$\begin{aligned}
\operatorname{var}(\hat{\theta}_{hc}) &= E_S\{\operatorname{var}(\theta_{hc1} + \hat{\theta}_{hc2}|S)\} + \operatorname{var}_S\{E(\theta_{hc1} + \hat{\theta}_{hc2}|S)\} \\
&= E_S\{\operatorname{var}(\hat{\theta}_{hc2}|S)\} + \operatorname{var}_S\{\theta_{hc1} + E(\hat{\theta}_{hc2}|S)\} \\
&= E_S\{\operatorname{var}(\textstyle\sum_{i\in S} h_i y_{ci} + \sum_{i\in U_2\setminus S} h_i \tilde{y}_{ci}\,|S)\} \\
&\qquad + \operatorname{var}_S\{E(\textstyle\sum_{i\in S} h_i y_{ci} + \sum_{i\in U_2\setminus S} h_i \tilde{y}_{ci}\,|S)\} \\
&= E_S\left\{\sum_{i\in U_2\setminus S} h_i \operatorname{var}(\tilde{y}_{ci}|S)\right\} + \operatorname{var}_S\left\{\sum_{i\in S} h_i y_{ci} + \sum_{i\in U_2\setminus S} h_i E(\tilde{y}_{ci}|S)\right\} \\
&= E_S\left\{\sum_{i\in U_2\setminus S} h_i \hat{p}_{ci}(1-\hat{p}_{ci})\right\} + \operatorname{var}_S\left(\sum_{i\in S} h_i y_{ci} + \sum_{i\in U_2\setminus S} h_i \hat{p}_{ci}\right).
\end{aligned}$$

In the fourth line we used that $h_i^2 = h_i$ and in the last line we used that $E(\tilde{y}_{ci}|S) = \hat{p}_{ci}$ and $\operatorname{var}(\tilde{y}_{ci}|S) = \hat{p}_{ci}(1-\hat{p}_{ci})$.

For all persons in U_2, define a sample inclusion indicator by $a_{Si} = 1$ if $i \in S$ and $a_{Si} = 0$ if $i \notin S$. The associated first- and second-order inclusion probabilities are denoted here by $\tau_{Si} = E(a_{Si}) = P(a_{Si} = 1)$ and $\tau_{Sii} = \tau_{Si}$ and, for $i \neq j$, $\tau_{Sij} = E(a_{Si}a_{Sj}) = P(a_{Si} = a_{Sj} = 1)$. It follows that:

$$\begin{aligned}
\operatorname{var}(\hat{\theta}_{hc}) &= E_S\left\{\sum_{i\in U_2} h_i(1-a_{Si})\hat{p}_{ci}(1-\hat{p}_{ci})\right\} \\
&\qquad + \operatorname{var}_S\left[\sum_{i\in U_2} h_i\{a_{Si}y_{ci} + (1-a_{Si})\hat{p}_{ci}\}\right] \\
&\equiv V_1 + V_2.
\end{aligned}$$

It should be noted that $\hat{p}_{ci}$, the estimated probability that person i has education level c, is obtained from a model that is estimated on the observed data from sample S. Hence, $\hat{p}_{ci}$ is a random variable that depends on S. This complicates the derivation of the variance. To simplify matters, we introduce the assumption that a_{Si} and $\hat{p}_{cj}$ are stochastically independent for all combinations of i, j and c (including the case $i = j$). That is to say, we assume that the sample is drawn independently of the expected education level of the persons involved. In particular, this implies that $E(a_{Si}\hat{p}_{ci}) = E(a_{Si})E(\hat{p}_{ci})$.

Under this assumption, we can reduce the first term in the above expression for $\operatorname{var}(\hat{\theta}_{hc})$ to:

$$\begin{aligned} V_1 &= \sum_{i \in U_2} h_i E\{(1 - a_{Si})\hat{p}_{ci}(1 - \hat{p}_{ci})\} \\ &= \sum_{i \in U_2} h_i E(1 - a_{Si}) E\{\hat{p}_{ci}(1 - \hat{p}_{ci})\} \\ &= \sum_{i \in U_2} h_i (1 - \tau_{Si})\{p_{ci}(1 - p_{ci}) - \mathrm{var}(\hat{p}_{ci})\}. \end{aligned} \tag{2}$$

In the last line, it was used that

$$\hat{p}_{ci}(1 - \hat{p}_{ci}) = p_{ci}(1 - \hat{p}_{ci}) + (\hat{p}_{ci} - p_{ci})(1 - \hat{p}_{ci})$$

and therefore (assuming that $E(\hat{p}_{ci} - p_{ci}) = 0$):

$$\begin{aligned} E\{\hat{p}_{ci}(1 - \hat{p}_{ci})\} &= p_{ci}(1 - p_{ci}) - E\{(\hat{p}_{ci} - p_{ci})\hat{p}_{ci}\} \\ &= p_{ci}(1 - p_{ci}) - E\{(\hat{p}_{ci} - p_{ci})^2\} \\ &= p_{ci}(1 - p_{ci}) - \mathrm{var}(\hat{p}_{ci}). \end{aligned} \tag{3}$$

For the second variance term, we obtain:

$$\begin{aligned} V_2 &= \mathrm{var}_S\left(\sum_{i \in U_2} h_i \{\hat{p}_{ci} + a_{Si}(y_{ci} - \hat{p}_{ci})\} \right) \\ &= \sum_{i \in U_2} \sum_{j \in U_2} \mathrm{cov}\big(h_i\{\hat{p}_{ci} + a_{Si}(y_{ci} - \hat{p}_{ci})\}, h_j\{\hat{p}_{cj} + a_{Sj}(y_{cj} - \hat{p}_{cj})\}\big) \\ &= \sum_{i \in U_2} \sum_{j \in U_2} h_i h_j \big[\mathrm{cov}(\hat{p}_{ci}, \hat{p}_{cj}) + \mathrm{cov}\{a_{Si}(y_{ci} - \hat{p}_{ci}), a_{Sj}(y_{cj} - \hat{p}_{cj})\} \\ &\qquad + \mathrm{cov}\{\hat{p}_{ci}, a_{Sj}(y_{cj} - \hat{p}_{cj})\} + \mathrm{cov}\{a_{Si}(y_{ci} - \hat{p}_{ci}), \hat{p}_{cj}\}\big]. \end{aligned} \tag{4}$$

In what follows, we will use the following lemma.

Lemma 1. *For any set of four random variables A, B, C and D such that*
- *A is independent of B and D, and*
- *C is independent of B and D,*

it holds that $\mathrm{cov}(AB, CD) = E(AC)\,\mathrm{cov}(B, D) + E(B)E(D)\,\mathrm{cov}(A, C)$.

Proof. By a standard decomposition, it follows that

$$\begin{aligned} \mathrm{cov}(AB, CD) &= E_{A,C}\{\mathrm{cov}(AB, CD|A, C)\} + \mathrm{cov}_{A,C}\{E(AB|A, C)\,, E(CD|A, C)\} \\ &= E_{A,C}\{AC\,\mathrm{cov}(B, D)\} + \mathrm{cov}_{A,C}\{A\,E(B)\,, C\,E(D)\} \\ &= E(AC)\,\mathrm{cov}(B, D) + E(B)E(D)\,\mathrm{cov}(A, C). \end{aligned}$$

□

For future reference, two corollaries of this result will now be mentioned separately.

Lemma 2. *For any set of three random variables X, Y and Z such that X is independent of Y and Z, it holds that* $\mathrm{cov}(XY, Z) = E(X)\,\mathrm{cov}(Y, Z)$.

Proof. Apply Lemma 1 with $A = X$, $B = Y$, $C = 1$ and $D = Z$. □

Lemma 3. *For any pair of independent random variables X and Y, it holds that* $\operatorname{var}(XY) = \operatorname{var}(X)\operatorname{var}(Y) + E(X)^2\operatorname{var}(Y) + E(Y)^2\operatorname{var}(X)$.

Proof. Apply Lemma 1 with $A = C = X$ and $B = D = Y$. It follows that

$$\begin{aligned}\operatorname{var}(XY) &= E(X^2)\operatorname{var}(Y) + E(Y)^2\operatorname{var}(X)\\ &= \{\operatorname{var}(X) + E(X)^2\}\operatorname{var}(Y) + E(Y)^2\operatorname{var}(X).\end{aligned}$$

□

An application of Lemma 1 with $A = a_{Si}$, $B = y_{ci} - \hat{p}_{ci}$, $C = a_{Sj}$ and $D = y_{cj} - \hat{p}_{cj}$ yields:

$$\begin{aligned}&\operatorname{cov}\{a_{Si}(y_{ci} - \hat{p}_{ci}), a_{Sj}(y_{cj} - \hat{p}_{cj})\}\\ &= E(a_{Si}a_{Sj})\operatorname{cov}(y_{ci} - \hat{p}_{ci}, y_{cj} - \hat{p}_{cj}) + E(y_{ci} - \hat{p}_{ci})E(y_{cj} - \hat{p}_{cj})\operatorname{cov}(a_{Si}, a_{Sj})\\ &= \tau_{Sij}\operatorname{cov}(\hat{p}_{ci}, \hat{p}_{cj}) + (y_{ci} - p_{ci})(y_{cj} - p_{cj})(\tau_{Sij} - \tau_{Si}\tau_{Sj}).\end{aligned}$$

Similarly, it follows from Lemma 2 with $X = a_{Sj}$, $Y = y_{cj} - \hat{p}_{cj}$ and $Z = \hat{p}_{ci}$ that:

$$\begin{aligned}\operatorname{cov}\{\hat{p}_{ci}, a_{Sj}(y_{cj} - \hat{p}_{cj})\} &= \operatorname{cov}\{a_{Sj}(y_{cj} - \hat{p}_{cj}), \hat{p}_{ci}\}\\ &= E(a_{Sj})\operatorname{cov}(y_{cj} - \hat{p}_{cj}, \hat{p}_{ci})\\ &= -\tau_{Sj}\operatorname{cov}(\hat{p}_{ci}, \hat{p}_{cj}).\end{aligned}$$

By reason of symmetry, it also follows that

$$\operatorname{cov}\{a_{Si}(y_{ci} - \hat{p}_{ci}), \hat{p}_{cj}\} = -\tau_{Si}\operatorname{cov}(\hat{p}_{ci}, \hat{p}_{cj}).$$

Substituting these results into expression (4) for V_2 yields:

$$\begin{aligned}V_2 = \sum_{i\in U_2}\sum_{j\in U_2} h_i h_j\big[&(1 + \tau_{Sij} - \tau_{Si} - \tau_{Sj})\operatorname{cov}(\hat{p}_{ci}, \hat{p}_{cj})\\ &+ (y_{ci} - p_{ci})(y_{cj} - p_{cj})(\tau_{Sij} - \tau_{Si}\tau_{Sj})\big].\end{aligned} \tag{5}$$

The second half of expression (5) may be recognised as a standard formula for the sampling variance of the (unweighted) sample total of $h_i(y_{ci} - p_{ci})$:

$$\operatorname{var}_{HT}\left\{\sum_{i\in S} h_i(y_{ci} - p_{ci})\right\} = \sum_{i\in U_2}\sum_{j\in U_2}(\tau_{Sij} - \tau_{Si}\tau_{Sj})h_i h_j(y_{ci} - p_{ci})(y_{cj} - p_{cj});$$

see, e.g., Särndal et al. (1992, p. 43). Furthermore, the contribution to the first half of expression (5) for V_2 for $i = j$ equals

$$h_i^2(1 + \tau_{Sii} - 2\tau_{Si})\operatorname{cov}(\hat{p}_{ci}, \hat{p}_{ci}) = h_i(1 - \tau_{Si})\operatorname{var}(\hat{p}_{ci}).$$

Hence, these variance contributions cancel out against the second term in expression (2) for V_1. Thus, combining the above expressions for V_1 and V_2, we obtain the following variance formula:

$$\begin{aligned}\operatorname{var}(\hat{\theta}_{hc}) = \sum_{i\in U_2} h_i(1-\tau_{Si})p_{ci}(1-p_{ci}) + \operatorname{var}_{HT}\left\{\sum_{i\in S} h_i(y_{ci}-p_{ci})\right\} \\ + \sum_{i\in U_2}\sum_{\substack{j\in U_2\\ j\neq i}} h_i h_j(1+\tau_{Sij}-\tau_{Si}-\tau_{Sj})\operatorname{cov}(\hat{p}_{ci},\hat{p}_{cj}).\end{aligned} \qquad (6)$$

To complete the derivation of the variance, we require an expression for $\operatorname{cov}(\hat{p}_{ci},\hat{p}_{cj})$. This will depend on the precise model that is used for imputation. A particular expression for the model that is used to impute educational attainment for the Dutch Census will be derived in the next two sections.

2.3 Variance estimation

In principle, the variance (6) could be estimated as follows:

$$\begin{aligned}\widehat{\operatorname{var}}(\hat{\theta}_{hc}) = \sum_{i\in U_2} h_i(1-\tau_{Si})\{\hat{p}_{ci}(1-\hat{p}_{ci}) + \widehat{\operatorname{var}}(\hat{p}_{ci})\} \\ + \sum_{i\in S}\sum_{j\in S}\frac{\tau_{Sij}-\tau_{Si}\tau_{Sj}}{\tau_{Sij}} h_i h_j(y_{ci}-\hat{p}_{ci})(y_{cj}-\hat{p}_{cj}) \\ + \sum_{i\in U_2}\sum_{\substack{j\in U_2\\ j\neq i}} h_i h_j(1+\tau_{Sij}-\tau_{Si}-\tau_{Sj})\widehat{\operatorname{cov}}(\hat{p}_{ci},\hat{p}_{cj}).\end{aligned} \qquad (7)$$

Here, the second term corresponds to $\widehat{\operatorname{var}}_{HT}\{\sum_{i\in S} h_i(y_{ci}-p_{ci})\}$, an estimator for the variance of an unweighted sample total (cf. Särndal et al., 1992). Furthermore, $\widehat{\operatorname{cov}}(\hat{p}_{ci},\hat{p}_{cj})$ denotes an unbiased estimator of the covariance of $\hat{p}_{ci}$ and $\hat{p}_{cj}$ (to be derived below) and $\widehat{\operatorname{var}}(\hat{p}_{ci}) = \widehat{\operatorname{cov}}(\hat{p}_{ci},\hat{p}_{ci})$. Finally, the addition of $\widehat{\operatorname{var}}(\hat{p}_{ci})$ in the first term of (7) is necessary to avoid bias since it follows from (3) that

$$E\{\hat{p}_{ci}(1-\hat{p}_{ci}) + \widehat{\operatorname{var}}(\hat{p}_{ci})\} = E\{\hat{p}_{ci}(1-\hat{p}_{ci})\} + \operatorname{var}(\hat{p}_{ci}) = p_{ci}(1-p_{ci}).$$

Variance estimator (7) requires that the first- and second-order inclusion probabilities τ_{Si} and τ_{Sij} are known for all (pairs of) persons in U_2. In practice, for some sample designs the second-order probabilities may not be known exactly. Moreover, in some cases the first-order probabilities may not have been stored for persons outside the realised sample S, and reconstructing these probabilities afterwards may be difficult. In particular, this latter problem occurs for the Dutch Census. In these situations, a different (approximate) variance estimator is needed. We will discuss two alternative approaches.

For the first approximation, we assume that the sample S resembles a simple random sample without replacement of size $|S| = n$ from a population of size $|U_2| = N$. In this case $\tau_{Si} = n/N$ for all $i \in U_2$ and $\tau_{Sij} = n(n-1)/\{N(N-1)\}$ for all $i, j \in U_2$ with $i \neq j$. Hence, in the last term of expressions (6) and (7) we find, for $i \neq j$,

$$1 + \tau_{Sij} - \tau_{Si} - \tau_{Sj} = 1 + \frac{n(n-1)}{N(N-1)} - \frac{2n}{N} = \left(1 - \frac{n}{N}\right)\left(1 - \frac{n}{N-1}\right).$$

Furthermore, for the second term in expression (6), we have:

$$\mathrm{var}_{HT}\left\{\sum_{i \in U_2} a_{Si} h_i (y_{ci} - p_{ci})\right\} = \mathrm{var}_{HT}\left\{\frac{N}{n} \sum_{i \in S} \frac{h_i (y_{ci} - p_{ci})}{N/n}\right\}$$

and for simple random sampling without replacement the following standard formula is available for this variance (see, e.g., Särndal et al., 1992):

$$\begin{aligned}
&\frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\left[\sum_{i \in U_2} h_i \left(\frac{y_{ci} - p_{ci}}{\frac{N}{n}}\right)^2 - \frac{1}{N}\left\{\sum_{i \in U_2} \frac{h_i (y_{ci} - p_{ci})}{\frac{N}{n}}\right\}^2\right] \\
&= \left(1 - \frac{n}{N}\right)\frac{n}{N-1}\left[\sum_{i \in U_2} h_i (y_{ci} - p_{ci})^2 - \frac{1}{N}\left\{\sum_{i \in U_2} h_i (y_{ci} - p_{ci})\right\}^2\right] \\
&\equiv \left(1 - \frac{n}{N}\right) n S^2\{h(y_c - p_c)\},
\end{aligned}$$

where $S^2\{z\}$ denotes the so-called adjusted population variance of variable z. In expression (7), this variance may be estimated from the sample S by:

$$\begin{aligned}
&\widehat{\mathrm{var}}_{HT}\left\{\sum_{i \in U_2} a_{Si} h_i (y_{ci} - p_{ci})\right\} \\
&= \left(1 - \frac{n}{N}\right)\frac{n}{n-1}\left[\sum_{i \in S} h_i (y_{ci} - \hat{p}_{ci})^2 - \frac{1}{n}\left\{\sum_{i \in S} h_i (y_{ci} - \hat{p}_{ci})\right\}^2\right] \\
&\equiv \left(1 - \frac{n}{N}\right) n s^2\{h(y_c - \hat{p}_c)\},
\end{aligned}$$

where $s^2\{z\}$ denotes the sample equivalent of $S^2\{z\}$. In summary, we obtain the following simplified expressions for $\mathrm{var}(\hat{\theta}_{hc})$ and $\widehat{\mathrm{var}}(\hat{\theta}_{hc})$ in this case:

$$\begin{aligned}
\mathrm{var}(\hat{\theta}_{hc}) = &\left(1 - \frac{n}{N}\right)\sum_{i \in U_2} h_i p_{ci}(1 - p_{ci}) + \left(1 - \frac{n}{N}\right) n S^2\{h(y_c - p_c)\} \\
&+ \left(1 - \frac{n}{N}\right)\left(1 - \frac{n}{N-1}\right)\sum_{i \in U_2}\sum_{\substack{j \in U_2 \\ j \neq i}} h_i h_j \,\mathrm{cov}(\hat{p}_{ci}, \hat{p}_{cj})
\end{aligned} \qquad (8)$$

and

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{\theta}_{hc}) = \left(1-\frac{n}{N}\right)\sum_{i\in U_2} h_i\{\hat{p}_{ci}(1-\hat{p}_{ci}) + \widehat{\mathrm{var}}(\hat{p}_{ci})\} \\ + \left(1-\frac{n}{N}\right) ns^2\{h(y_c-\hat{p}_c)\} \\ + \left(1-\frac{n}{N}\right)\left(1-\frac{n}{N-1}\right)\sum_{i\in U_2}\sum_{\substack{j\in U_2\\ j\neq i}} h_i h_j \,\widehat{\mathrm{cov}}(\hat{p}_{ci},\hat{p}_{cj}).\end{aligned} \tag{9}$$

For the second approximation, we avoid second-order inclusion probabilities by assuming that S resembles a with-replacement sample from U_2. More precisely, we consider a with-replacement sampling design that consists of n independent draws from U_2 with drawing probabilities τ_{Si}/n. For $1 \ll n \ll N$, it holds under this sampling design that[2] $\tau_{Sij} \approx \tau_{Si}\tau_{Sj}$ $(i \neq j)$, so

$$1 + \tau_{Sij} - \tau_{Si} - \tau_{Sj} \approx (1-\tau_{Si})(1-\tau_{Sj}).$$

From (7), we now find the following simplified expression for $\widehat{\mathrm{var}}(\hat{\theta}_{hc})$:

$$\begin{aligned}\widehat{\mathrm{var}}_{alt}(\hat{\theta}_{hc}) = \sum_{i\in U_2} h_i(1-\tau_{Si})\{\hat{p}_{ci}(1-\hat{p}_{ci}) + \widehat{\mathrm{var}}(\hat{p}_{ci})\} + ns^2\{h(y_c-\hat{p}_c)\} \\ + \sum_{i\in U_2}\sum_{\substack{j\in U_2\\ j\neq i}} h_i h_j (1-\tau_{Si})(1-\tau_{Sj})\,\widehat{\mathrm{cov}}(\hat{p}_{ci},\hat{p}_{cj}).\end{aligned} \tag{10}$$

Here, the second term is obtained by replacing $\widehat{\mathrm{var}}_{HT}\{\sum_{i\in S} h_i(y_{ci}-p_{ci})\}$ in (7) by the estimated variance of the associated Hansen-Hurwitz estimator for with-replacement sampling, $\widehat{\mathrm{var}}_{HH}\{\sum_{i\in S} h_i(y_{ci}-p_{ci})\}$. Accounting, as before, for the fact that here the sample total is unweighted rather than weighted, we find that this variance estimator is given by:

$$\begin{aligned}&\widehat{\mathrm{var}}_{HH}\left\{\sum_{i\in S}\frac{\tau_{Si}h_i(y_{ci}-p_{ci})}{\tau_{Si}}\right\} \\ &= \frac{n}{n-1}\left[\sum_{i\in S}\left\{\frac{\tau_{Si}h_i(y_{ci}-\hat{p}_{ci})}{\tau_{Si}}\right\}^2 - \frac{1}{n}\left\{\sum_{i\in S}\frac{\tau_{Si}h_i(y_{ci}-\hat{p}_{ci})}{\tau_{Si}}\right\}^2\right] \\ &= \frac{n}{n-1}\left[\sum_{i\in S} h_i(y_{ci}-\hat{p}_{ci})^2 - \frac{1}{n}\left\{\sum_{i\in S} h_i(y_{ci}-\hat{p}_{ci})\right\}^2\right] \\ &= ns^2\{h(y_c-\hat{p}_c)\}.\end{aligned}$$

For more details, we refer to Särndal et al. (1992, Section 11.2) and in particular to Appendix A of Knottnerus and van Duin (2006), who applied the same idea in the context of repeated weighting.

[2] In fact, for this with-replacement sampling design it follows from Särndal et al. (1992, pp. 51 and 60) that:

$$\tau_{Sij} - \tau_{Si}\tau_{Sj} = \left(1-\frac{\tau_{Si}}{n}-\frac{\tau_{Sj}}{n}\right)^n - \left(1-\frac{\tau_{Si}}{n}\right)^n\left(1-\frac{\tau_{Sj}}{n}\right)^n \approx e^{-(\tau_{Si}+\tau_{Sj})} - e^{-\tau_{Si}}e^{-\tau_{Sj}} = 0.$$

As noted above, in practice the first-order inclusion probabilities may not be known (any more) for persons in $U_2 \backslash S$. In the absence of any knowledge of these probabilities, one might approximate the unknown τ_{Si} in (10) by their average value $\bar{\tau}_{Si} = |S|/|U_2| = n/N$.

Treating a without-replacement sample as if it were a with-replacement sample will usually lead to an overestimate of the variance. However, for $n \ll N$, the error that is incurred by this approximation should be negligible. If in addition the true inclusion probabilities τ_{Si} differ strongly from those of a simple random sample, then it may be expected that variance estimator (10) will yield more accurate results than variance estimator (9).

In both cases, we have obtained a variance estimator that consists of three components. The first two components may be computed directly from the data that were used for mass imputation (apart from the term involving $\widehat{\mathrm{var}}(\hat{p}_{ci})$). The final component requires an expression for $\widehat{\mathrm{cov}}(\hat{p}_{ci}, \hat{p}_{cj})$. This part will depend on the precise model that was used for imputation. In the remainder of this report we will derive a particular expression for the model that was used to impute educational attainment in the Dutch Census, based on logistic regression.

3. The continuation-ratio logistic regression model - covariances of conditional probabilities

3.1 The continuation-ratio logistic regression model

Imputation of missing values on educational attainment in the Dutch Census is based on the so-called continuation-ratio logistic regression model, estimated on the sample S (Scholtus and Pannekoek, 2015; De Waal and Daalmans, 2018). This model consists of a sequence of ordinary logistic regression models of the form[3]

$$\log\left(\frac{\pi_{ci}}{1-\pi_{ci}}\right) = \boldsymbol{\beta}_c^T \mathbf{x}_i, \quad (c = 1, \dots, C-1), \tag{11}$$

where π_{ci} denotes the conditional probability that person i, with characteristics $\mathbf{x}_i$, has education level c, given that the education level of this person is not lower than c:

$$\begin{aligned}\pi_{1i} &= P(y_{1i} = 1|\,\mathbf{x} = \mathbf{x}_i),\\ \pi_{ci} &= P\big(y_{ci} = 1\big|y_{1i} = \cdots = y_{(c-1)i} = 0, \mathbf{x} = \mathbf{x}_i\big), \quad (c = 2, \dots, C-1).\end{aligned}$$

From these conditional probabilities, the marginal probabilities $p_{ci} = P(y_{ci} = 1|\,\mathbf{x} = \mathbf{x}_i)$ that were used in Section 2 can be derived by the following recursion:

$$\begin{aligned}p_{1i} &= \pi_{1i},\\ p_{ci} &= \pi_{ci}\left(1 - \sum_{k=1}^{c-1} p_{ki}\right), \quad (c = 2, \dots, C-1),\\ p_{Ci} &= 1 - \sum_{c=1}^{C-1} p_{ci}.\end{aligned} \tag{12}$$

Agresti (2013) shows that maximum-likelihood estimation for the continuation-ratio model is simplified by means of the following factorisation:

$$\begin{aligned}&P(y_{1i}, \dots, y_{Ci}|\mathbf{x}_i)\\ &= P(y_{1i}|\mathbf{x}_i)P(y_{2i}|y_{1i}, \mathbf{x}_i) \cdots P\big(y_{ci}\big|y_{1i}, \dots, y_{(c-1)i}, \mathbf{x}_i\big) \cdots P\big(y_{Ci}\big|y_{1i}, \dots, y_{(C-1)i}, \mathbf{x}_i\big).\end{aligned}$$

In terms of the conditional probabilities $\pi_{1i}, \dots, \pi_{(C-1)i}$, this yields:

[3] An intercept term may be included in the vector of coefficients $\boldsymbol{\beta}_c$ but we do not denote this explicitly. This simplifies the notation slightly in comparison to Scholtus and Pannekoek (2015).

$$P(y_{1i}, \dots, y_{Ci}|\mathbf{x}_i)$$
$$= \begin{cases} \pi_{1i}^{y_{1i}}(1-\pi_{1i})^{1-y_{1i}} \prod_{c=2}^{C-1} \left\{\pi_{ci}^{y_{ci}}(1-\pi_{ci})^{1-y_{ci}}\right\}^{\delta(y_{1i}=\cdots=y_{(c-1)i}=0)} & \text{if } \sum_{c=1}^{C} y_{ci} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here, $\delta(A)$ is an indicator with $\delta(A) = 1$ if A is true and $\delta(A) = 0$ otherwise. We can ignore the last factor $P\left(y_{Ci}\middle|y_{1i}, \dots, y_{(C-1)i}, \mathbf{x}_i\right)$, since it corresponds to a degenerate distribution with

$$P\left(y_{Ci} = 1\middle|y_{1i}, \dots, y_{(C-1)i}, \mathbf{x}_i\right) = \delta\left(y_{1i} = \cdots = y_{(C-1)i} = 0\right).$$

For a sample S of independent observations from the model distribution, the log likelihood function can be written as:

$$\begin{aligned} \ell &= \sum_{i \in S} \log P(y_{1i}, \dots, y_{Ci}|\mathbf{x}_i) \\ &= \sum_{i \in S} \sum_{c=1}^{C-1} \delta\left(y_{1i} = \cdots = y_{(c-1)i} = 0\right)\{y_{ci} \log \pi_{ci} + (1 - y_{ci}) \log(1 - \pi_{ci})\} \\ &= \sum_{i \in S} \sum_{c=1}^{C-1} \delta\left(y_{1i} = \cdots = y_{(c-1)i} = 0\right)\left\{y_{ci} \log\left(\frac{\pi_{ci}}{1 - \pi_{ci}}\right) + \log(1 - \pi_{ci})\right\} \\ &= \sum_{i \in S} \sum_{c=1}^{C-1} \delta\left(y_{1i} = \cdots = y_{(c-1)i} = 0\right)\{y_{ci}(\boldsymbol{\beta}_c^T \mathbf{x}_i) - \log[1 + \exp(\boldsymbol{\beta}_c^T \mathbf{x}_i)]\} \end{aligned}$$

with the convention that $\delta\left(y_{1i} = \cdots = y_{(c-1)i} = 0\right) = 1$ for $c = 1$.

Let $S_{\geq 1} = S$ and $S_{\geq c} = \left\{i \in S\middle|y_{1i} = \cdots = y_{(c-1)i} = 0\right\}$ $(c = 2, \dots, C - 1)$ denote nested subsamples of the original sample. The log likelihood can be written as a sum where each term depends only on parameters $\boldsymbol{\beta}_c$ for one particular c:

$$\ell = \sum_{c=1}^{C-1} \ell_c = \sum_{c=1}^{C-1} \sum_{i \in S_{\geq c}} \{y_{ci}(\boldsymbol{\beta}_c^T \mathbf{x}_i) - \log[1 + \exp(\boldsymbol{\beta}_c^T \mathbf{x}_i)]\}.$$

Since the individual terms ℓ_c have no parameters in common, maximising the log likelihood ℓ is equivalent to maximising each term ℓ_c separately. Hence, it follows that maximum-likelihood estimates of all parameters in the continuation-ratio model can be obtained by estimating the binary logistic regression models in (11) separately (Agresti, 2013). This involves setting the first-order partial derivatives of ℓ_c with respect to $\boldsymbol{\beta}_c$ equal to zero. The resulting likelihood equations are:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}_c} = \frac{\partial \ell_c}{\partial \boldsymbol{\beta}_c} = \sum_{i \in S_{\geq c}} \left\{y_{ci} \mathbf{x}_i - \frac{\exp(\boldsymbol{\beta}_c^T \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}_c^T \mathbf{x}_i)} \mathbf{x}_i\right\} = \sum_{i \in S_{\geq c}} (y_{ci} - \pi_{ci})\mathbf{x}_i.$$

For the second-order derivatives, we find:

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}_c \partial \boldsymbol{\beta}_c^T} = \frac{\partial^2 \ell_c}{\partial \boldsymbol{\beta}_c \partial \boldsymbol{\beta}_c^T} = -\sum_{i \in S_{\geq c}} \frac{\exp(\boldsymbol{\beta}_c^T \mathbf{x}_i)}{\{1 + \exp(\boldsymbol{\beta}_c^T \mathbf{x}_i)\}^2} \mathbf{x}_i \mathbf{x}_i^T = -\sum_{i \in S_{\geq c}} \pi_{ci}(1 - \pi_{ci}) \mathbf{x}_i \mathbf{x}_i^T.$$

Hence, the information matrix of sub-model ℓ_c is given by:

$$\mathbf{I}_c = -\frac{\partial^2 \ell_c}{\partial \boldsymbol{\beta}_c \partial \boldsymbol{\beta}_c^T} = \sum_{i \in S_{\geq c}} \pi_{ci}(1 - \pi_{ci}) \mathbf{x}_i \mathbf{x}_i^T = \mathbf{X}^T \boldsymbol{\Delta}_{\geq c} \mathbf{X}, \tag{13}$$

where $\mathbf{X}$ denotes an $n \times (K + 1)$ matrix that contains $\mathbf{x}_i^T$ as rows and $\boldsymbol{\Delta}_{\geq c}$ denotes an $n \times n$ diagonal matrix with

$$\begin{aligned}
&(\boldsymbol{\Delta}_{\geq 1})_{ii} = (\boldsymbol{\Delta})_{ii} = \pi_{ci}(1 - \pi_{ci}), \\
&(\boldsymbol{\Delta}_{\geq c})_{ii} = \delta\big(y_{1i} = \cdots = y_{(c-1)i} = 0\big) \pi_{ci}(1 - \pi_{ci}), \quad c = 2, \ldots, C - 1.
\end{aligned}$$

The information matrix of the full model is a block diagonal matrix with $\mathbf{I}_c$ as blocks. By a general property of maximum likelihood estimation (see, e.g., Van der Vaart, 1998), the inverse of this matrix can be used as an asymptotic variance-covariance matrix of the parameter estimates:

$$\begin{aligned}
\operatorname{cov}\big(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_{C-1}\big) &= \begin{bmatrix} \mathbf{I}_1^{-1} & & & \\ & \mathbf{I}_2^{-1} & & \\ & & \ddots & \\ & & & \mathbf{I}_{C-1}^{-1} \end{bmatrix} \\
&= \begin{bmatrix} \{\mathbf{X}^T \boldsymbol{\Delta}_{\geq 1} \mathbf{X}\}^{-1} & & & \\ & \{\mathbf{X}^T \boldsymbol{\Delta}_{\geq 2} \mathbf{X}\}^{-1} & & \\ & & \ddots & \\ & & & \{\mathbf{X}^T \boldsymbol{\Delta}_{\geq C-1} \mathbf{X}\}^{-1} \end{bmatrix}.
\end{aligned}$$

From this derivation, it follows in particular that the estimated parameters in different binary logistic regression models are asymptotically uncorrelated (i.e., for large enough sample sizes), even though they are partly based on the same observations. In fact, they are asymptotically independent, since maximum-likelihood estimators follow a normal distribution for sufficiently large samples.

Once the model parameters have been estimated, the conditional probability that a person with characteristics $\mathbf{x}_i$ has education level c is estimated by

$$\hat{\pi}_{ci} = \frac{\exp\big(\widehat{\boldsymbol{\beta}}_c^T \mathbf{x}_i\big)}{1 + \exp\big(\widehat{\boldsymbol{\beta}}_c^T \mathbf{x}_i\big)}.$$

The variance of $\hat{\pi}_{ci}$ may be approximated by using a first-order Taylor series expansion of the function $f(z) = \exp(z)\,/[1 + \exp(z)]$, with $f'(z) = \exp(z)\,/[1 + \exp(z)]^2$:

$$\begin{aligned}
\hat{\pi}_{ci} &\approx \frac{\exp(\boldsymbol{\beta}_c^T \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}_c^T \mathbf{x}_i)} + \frac{\exp(\boldsymbol{\beta}_c^T \mathbf{x}_i)}{\{1 + \exp(\boldsymbol{\beta}_c^T \mathbf{x}_i)\}^2} \mathbf{x}_i^T \big(\widehat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}_c\big) \\
&= \pi_{ci} + \pi_{ci}(1 - \pi_{ci}) \mathbf{x}_i^T \big(\widehat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}_c\big).
\end{aligned}$$

It follows that:

$$\mathrm{var}(\hat{\pi}_{ci}) \approx \{\pi_{ci}(1-\pi_{ci})\}^2 \,\mathrm{var}\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c\right) = \{\pi_{ci}(1-\pi_{ci})\}^2 \mathbf{x}_i^T \{\mathbf{X}^T \boldsymbol{\Delta}_{\geq c} \mathbf{X}\}^{-1} \mathbf{x}_i.$$

Moreover, as we have seen previously that the parameter estimates from different binary logistic regression models are asymptotically independent, it holds asymptotically that $\mathrm{cov}(\hat{\pi}_{ci}, \hat{\pi}_{di}) = 0$ for all $c \neq d$.

By a similar derivation, the covariance between estimated probabilities for different persons i and j is found to be approximately equal to:

$$\begin{aligned} \mathrm{cov}\left(\hat{\pi}_{ci}, \hat{\pi}_{cj}\right) &\approx \pi_{ci}(1-\pi_{ci})\pi_{cj}\left(1-\pi_{cj}\right) \mathrm{cov}\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \mathbf{x}_j^T \hat{\boldsymbol{\beta}}_c\right) \\ &= \pi_{ci}(1-\pi_{ci})\pi_{cj}\left(1-\pi_{cj}\right) \mathbf{x}_i^T \{\mathbf{X}^T \boldsymbol{\Delta}_{\geq c} \mathbf{X}\}^{-1} \mathbf{x}_j, \end{aligned} \tag{14}$$

whereas asymptotically $\mathrm{cov}\left(\hat{\pi}_{ci}, \hat{\pi}_{dj}\right) = 0$ for all $c \neq d$. It is seen that the full $N \times N$ variance-covariance matrix of $\hat{\pi}_{ci}$ for all persons in a population of size N can be approximated as follows. Construct an $N \times (K+1)$ matrix $\mathbf{X}_{pop}$ which contains $\mathbf{x}_i^T$ as rows for all i in the population and construct an $N \times N$ diagonal matrix $\boldsymbol{\Delta}_{pop}$ with $(\boldsymbol{\Delta}_{pop})_{ii} = \pi_{ci}(1-\pi_{ci})$. The asymptotic variance-covariance matrix of $\hat{\pi}_{c1}, \dots, \hat{\pi}_{cN}$ is given by:

$$\mathbf{V}_{\pi c,pop} = \boldsymbol{\Delta}_{pop} \mathbf{X}_{pop} \{\mathbf{X}^T \boldsymbol{\Delta}_{\geq c} \mathbf{X}\}^{-1} \mathbf{X}_{pop}^T \boldsymbol{\Delta}_{pop}.$$

In fact, in the special case that the logistic regression model for π_{ci} involves one or more categorical predictors that are used as stratifying variables, the above computations can be simplified further. Let $\mathbf{x}_i = \mathbf{x}_{1i} \otimes \mathbf{x}_{2i}$, where $\mathbf{x}_1$ denotes the stratifying variables and $\mathbf{x}_2$ denotes the other predictor variables in the model; also $\otimes$ denotes a Kronecker product. Using some standard properties of the Kronecker product, we obtain from (13):

$$\begin{aligned} \mathbf{I}_c &= \sum_{i \in S_{\geq c}} \pi_{ci}(1-\pi_{ci})(\mathbf{x}_{1i} \otimes \mathbf{x}_{2i})(\mathbf{x}_{1i} \otimes \mathbf{x}_{2i})^T \\ &= \sum_{i \in S_{\geq c}} \pi_{ci}(1-\pi_{ci})(\mathbf{x}_{1i} \otimes \mathbf{x}_{2i})\left(\mathbf{x}_{1i}^T \otimes \mathbf{x}_{2i}^T\right) \\ &= \sum_{i \in S_{\geq c}} \pi_{ci}(1-\pi_{ci})(\mathbf{x}_{1i}\mathbf{x}_{1i}^T \otimes \mathbf{x}_{2i}\mathbf{x}_{2i}^T). \end{aligned}$$

Since $\mathbf{x}_{1i}\mathbf{x}_{1i}^T$ is a matrix with exactly one element on the main diagonal equal to one and all other elements equal to zero, this implies that $\mathbf{I}_c$ is a block diagonal matrix, with each block associated to one stratum of the subsample $S_{\geq c}$. If we denote the unique strata with respect to $\mathbf{x}_1$ as $g = 1, \dots, G$, this yields:

$$\mathbf{I}_c = \begin{bmatrix} \mathbf{X}_{2,1}^T \boldsymbol{\Delta}_{\geq c,1} \mathbf{X}_{2,1} & & & \\ & \mathbf{X}_{2,2}^T \boldsymbol{\Delta}_{\geq c,2} \mathbf{X}_{2,2} & & \\ & & \ddots & \\ & & & \mathbf{X}_{2,G}^T \boldsymbol{\Delta}_{\geq c,G} \mathbf{X}_{2,G} \end{bmatrix}.$$

Here, $\mathbf{X}_{2,g}$ is an $n_g \times K_2$ matrix that contains $\mathbf{x}_{2i}^T$ as rows for all i in stratum g of the sample and $\boldsymbol{\Delta}_{\geq c,g}$ is the $n_g \times n_g$ block in $\boldsymbol{\Delta}_{\geq c}$ associated with these sample units, so that

$$\mathbf{X}_{2,g}^T \boldsymbol{\Delta}_{\geq c,g} \mathbf{X}_{2,g} = \sum_{i \in S_{\geq c,g}} \pi_{ci}(1 - \pi_{ci}) \mathbf{x}_{2i} \mathbf{x}_{2i}^T,$$

where $S_{\geq c,g}$ denotes stratum g within $S_{\geq c}$.

It is seen that, within the block diagonal matrix $\mathrm{cov}(\widehat{\boldsymbol{\beta}}_1, \dots, \widehat{\boldsymbol{\beta}}_{C-1})$ that was found previously, in this special case each block $\mathbf{I}_c^{-1} = \{\mathbf{X}^T \boldsymbol{\Delta}_{\geq c} \mathbf{X}\}^{-1}$ is itself a block diagonal matrix. The smaller blocks in these matrices can be computed per stratum. Moreover, it follows from

$$\mathrm{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj}) \approx \pi_{ci}(1 - \pi_{ci}) \pi_{cj}(1 - \pi_{cj}) (\mathbf{x}_{1i} \otimes \mathbf{x}_{2i})^T \mathbf{I}_c^{-1} (\mathbf{x}_{1j} \otimes \mathbf{x}_{2j})$$

that in this case $\mathrm{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj}) \approx \pi_{ci}(1 - \pi_{ci}) \pi_{cj}(1 - \pi_{cj}) \mathbf{x}_{2i}^T \{\mathbf{X}_{2,g}^T \boldsymbol{\Delta}_{\geq c,g} \mathbf{X}_{2,g}\}^{-1} \mathbf{x}_{2j}$ if i and j belong to the same stratum and $\mathrm{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj}) \approx 0$ otherwise. That is to say, for units in different strata the estimated probabilities $\hat{\pi}_{ci}$ and $\hat{\pi}_{cj}$ are asymptotically independent.

3.2 Application to variance estimation for Census tables

For the first (lowest) level of educational attainment ($c = 1$), it follows from (12) that $\hat{p}_{1i} = \hat{\pi}_{1i}$. Therefore, for $c = 1$ the results of Section 3.1 can be applied directly to approximate the last term in variance formula (8) when imputation is based on the continuation-ratio logistic regression model. Denote the sub-population of U_2 that belongs to stratum g of $\mathbf{x}_1$ by U_{2g} ($g = 1, \dots, G$). We find:

$$\begin{aligned}\sum_{i \in U_2} \sum_{\substack{j \in U_2 \\ j \neq i}} h_i h_j \,\mathrm{cov}(\hat{p}_{1i}, \hat{p}_{1j}) &= \sum_{i \in U_2} \sum_{\substack{j \in U_2 \\ j \neq i}} h_i h_j \,\mathrm{cov}(\hat{\pi}_{1i}, \hat{\pi}_{1j}) \\ &\approx \sum_{g=1}^{G} \sum_{i \in U_{2g}} \sum_{\substack{j \in U_{2g} \\ j \neq i}} \Big[h_i h_j \pi_{1i}(1 - \pi_{1i}) \pi_{1j}(1 - \pi_{1j}) \\ &\quad \times \mathbf{x}_{2i}^T \{\mathbf{X}_{2,g}^T \boldsymbol{\Delta}_g \mathbf{X}_{2,g}\}^{-1} \mathbf{x}_{2j} \Big],\end{aligned}$$

with $\boldsymbol{\Delta}_g = \boldsymbol{\Delta}_{\geq 1,g}$. We can eliminate the summation over j in this expression by defining an auxiliary variable $\mathbf{z}_{h1i} = h_i \pi_{1i}(1 - \pi_{1i}) \mathbf{x}_{2i}$, with $\mathbf{z}_{gh1,pop} = \sum_{i \in U_{2g}} \mathbf{z}_{h1i}$. It holds that:

$$\sum_{i \in U_2} \sum_{\substack{j \in U_2 \\ j \neq i}} h_i h_j \,\mathrm{cov}(\hat{p}_{1i}, \hat{p}_{1j}) \approx \sum_{g=1}^{G} \sum_{i \in U_{2g}} \mathbf{z}_{h1i}^T \{\mathbf{X}_{2,g}^T \boldsymbol{\Delta}_g \mathbf{X}_{2,g}\}^{-1} \sum_{\substack{j \in U_{2g} \\ j \neq i}} \mathbf{z}_{h1j}$$

$$= \sum_{g=1}^{G} \sum_{i \in U_{2g}} \mathbf{z}_{h1i}^T \left\{ \mathbf{X}_{2,g}^T \boldsymbol{\Delta}_g \mathbf{X}_{2,g} \right\}^{-1} (\mathbf{z}_{gh1,pop} - \mathbf{z}_{h1i})$$

$$= \sum_{g=1}^{G} \left[\mathbf{z}_{gh1,pop}^T \left\{ \mathbf{X}_{2,g}^T \boldsymbol{\Delta}_g \mathbf{X}_{2,g} \right\}^{-1} \mathbf{z}_{gh1,pop} - \sum_{i \in U_{2g}} \mathbf{z}_{h1i}^T \left\{ \mathbf{X}_{2,g}^T \boldsymbol{\Delta}_g \mathbf{X}_{2,g} \right\}^{-1} \mathbf{z}_{h1i} \right].$$

Similarly, the corresponding term in the variance estimator (9) can be approximated by:

$$\sum_{i \in U_2} \sum_{\substack{j \in U_2 \\ j \neq i}} h_i h_j \, \widehat{\mathrm{cov}}(\hat{p}_{1i}, \hat{p}_{1j}) \approx \sum_{g=1}^{G} \left[\hat{\mathbf{z}}_{gh1,pop}^T \left\{ \mathbf{X}_{2,g}^T \widehat{\boldsymbol{\Delta}}_g \mathbf{X}_{2,g} \right\}^{-1} \hat{\mathbf{z}}_{gh1,pop} - \sum_{i \in U_{2g}} \hat{\mathbf{z}}_{h1i}^T \left\{ \mathbf{X}_{2,g}^T \widehat{\boldsymbol{\Delta}}_g \mathbf{X}_{2,g} \right\}^{-1} \hat{\mathbf{z}}_{h1i} \right],$$

with $\widehat{\boldsymbol{\Delta}}_g = \widehat{\boldsymbol{\Delta}}_{\geq 1g}$ an $n_g \times n_g$ diagonal matrix with $(\widehat{\boldsymbol{\Delta}}_g)_{ii} = \hat{\pi}_{1i}(1 - \hat{\pi}_{1i})$ for $i \in U_{2g} \cap S$, $\hat{\mathbf{z}}_{h1i} = h_i \hat{\pi}_{1i}(1 - \hat{\pi}_{1i}) \mathbf{x}_{2i}$ and $\hat{\mathbf{z}}_{gh1,pop} = \sum_{i \in U_{2g}} \hat{\mathbf{z}}_{h1i}$. A similar expression can be obtained for the last term of the alternative variance estimator (10), but in this case we have to replace $\hat{\mathbf{z}}_{h1i}$ by $\hat{\mathbf{z}}_{h1i,alt} = h_i(1 - \tau_{Si}) \hat{\pi}_{1i}(1 - \hat{\pi}_{1i}) \mathbf{x}_{2i}$.

If educational attainment were defined as a dichotomous variable ($C = 2$), we would have been done now. The remaining expressions for $c = 2$ are in that case identical to those for $c = 1$, since it holds that

$$\mathrm{cov}(\hat{p}_{2i}, \hat{p}_{2j}) = \mathrm{cov}(1 - \hat{\pi}_{1i}, 1 - \hat{\pi}_{1j}) = \mathrm{cov}(\hat{\pi}_{1i}, \hat{\pi}_{1j}).$$

However, in general educational attainment is defined as a variable with at least three levels ($C > 2$). This means that the recursive expressions in (12) are non-trivial, and we cannot use the above results about the conditional probabilities $\hat{\pi}_{ci}$ directly to evaluate the variances for education levels $c \geq 2$. (We can always use the above expressions for the case $c = 1$.) Instead, we have to consider the marginal probabilities $\hat{p}_{ci}$. The next section takes up this point.

4. The continuation-ratio logistic regression model - covariances of marginal probabilities

4.1 Covariance approximations for marginal probabilities

For $C > 2$, estimated marginal probabilities $\hat{p}_{ci}$ are derived from the estimated conditional probabilities $\hat{\pi}_{ci}$ found in Section 3, analogously to (12):

$$\begin{aligned} \hat{p}_{1i} &= \hat{\pi}_{1i}, \\ \hat{p}_{ci} &= \hat{\pi}_{ci}\left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki}\right), \quad (c = 2, \dots, C-1), \\ \hat{p}_{Ci} &= 1 - \sum_{c=1}^{C-1} \hat{p}_{ci}. \end{aligned} \tag{15}$$

We will now derive expressions to compute the asymptotic covariances of these marginal probabilities, given that we know (from Section 3) how to obtain the asymptotic covariances of the conditional probabilities. We assume that the auxiliary information in the logistic regression models involves both stratifying variables $\mathbf{x}_1$ and other variables $\mathbf{x}_2$, which means that the asymptotic covariances of the conditional probabilities can be evaluated by the expression at the end of Section 3.1; i.e., for $i, j \in U_{2g}$ it holds that

$$\operatorname{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj}) \approx \pi_{ci}(1 - \pi_{ci})\pi_{cj}(1 - \pi_{cj})\mathbf{x}_{2i}^T\{\mathbf{X}_{2,g}^T\mathbf{\Delta}_{\geq c,g}\mathbf{X}_{2,g}\}^{-1}\mathbf{x}_{2j} \tag{16}$$

and $\operatorname{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj}) \approx 0$ if i and j belong to different strata with respect to $\mathbf{x}_1$.

Denote $\operatorname{cov}(\hat{p}_{ci}, \hat{p}_{dj}) = C_{cdij}$, for $i, j \in U_2$ and $1 \leq c, d \leq C$. To evaluate the variance formulas in Section 2, we only need the 'diagonal' terms C_{ccij}. However, to derive expressions for these diagonal terms below we will also need to consider, along the way, the terms C_{cdij} with $c \neq d$.

We already know how to evaluate $C_{11ij} = \operatorname{cov}(\hat{\pi}_{1i}, \hat{\pi}_{1j})$. For $2 \leq c \leq C - 1$:

$$C_{ccij} = \operatorname{cov}\left\{\hat{\pi}_{ci}\left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki}\right), \hat{\pi}_{cj}\left(1 - \sum_{l=1}^{c-1} \hat{p}_{lj}\right)\right\}.$$

According to (15), each $\hat{p}_{ki}$ is constructed from just the probabilities $\hat{\pi}_{1i}, \dots, \hat{\pi}_{ki}$. Since the conditional probabilities $\hat{\pi}_{1i}, \dots, \hat{\pi}_{Ci}$ are asymptotically mutually independent, it follows that $\hat{\pi}_{ci}$ is asymptotically independent of all $\hat{p}_{1i}, \dots, \hat{p}_{(c-1)i}$ and all $\hat{p}_{1j}, \dots, \hat{p}_{(c-1)j}$ for $j \neq i$. Therefore, we can (asymptotically) apply Lemma 1 from Section 2 to the above expression for C_{ccij}, with $A = \hat{\pi}_{ci}$, $B = 1 - \sum_{k=1}^{c-1} \hat{p}_{ki}$, $C = \hat{\pi}_{cj}$ and $D = 1 - \sum_{l=1}^{c-1} \hat{p}_{lj}$. This yields:

$$\begin{aligned} C_{ccij} &\approx E(\hat{\pi}_{ci}\hat{\pi}_{cj}) \operatorname{cov}\left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki}, 1 - \sum_{l=1}^{c-1} \hat{p}_{lj}\right) \\ &\quad + E\left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki}\right) E\left(1 - \sum_{l=1}^{c-1} \hat{p}_{lj}\right) \operatorname{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj}) \\ &= \left\{\operatorname{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj}) + \pi_{ci}\pi_{cj}\right\} \sum_{k=1}^{c-1}\sum_{l=1}^{c-1} \operatorname{cov}(\hat{p}_{ki}, \hat{p}_{lj}) \\ &\quad + \left(1 - \sum_{k=1}^{c-1} p_{ki}\right)\left(1 - \sum_{l=1}^{c-1} p_{lj}\right) \operatorname{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj}) \end{aligned}$$

and therefore (for $c = 2, \dots, C - 1$)

$$\begin{aligned} C_{ccij} &\approx \left\{\operatorname{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj}) + \pi_{ci}\pi_{cj}\right\} T_{c-1,ij} \\ &\quad + \left(1 - \sum_{k=1}^{c-1} p_{ki}\right)\left(1 - \sum_{l=1}^{c-1} p_{lj}\right) \operatorname{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj}), \end{aligned} \tag{17}$$

with the short-hand notation

$$T_{c,ij} = \sum_{k=1}^{c}\sum_{l=1}^{c} C_{klij}, \quad (c = 1, \dots, C - 1).$$

For the remaining case $c = C$, it follows directly from (15) that

$$C_{CCij} = \operatorname{cov}\left(1 - \sum_{k=1}^{C-1} \hat{p}_{ki}, 1 - \sum_{l=1}^{C-1} \hat{p}_{lj}\right) = \sum_{k=1}^{C-1}\sum_{l=1}^{C-1} \operatorname{cov}(\hat{p}_{ki}, \hat{p}_{lj}) = T_{C-1,ij}. \tag{18}$$

It remains to find an expression for $T_{c,ij}$. We do this by means of the following two lemmas. Proofs of both lemmas are given in the appendix of this paper.

Lemma 4. *For $c = 1, \dots, C - 1$ it holds asymptotically that*

$$T_{c,ij} \approx \sum_{k=1}^{c} C_{kkij} - \sum_{k=2}^{c} (\pi_{ki} + \pi_{kj}) T_{k-1,ij} \tag{19}$$

with the convention that the second sum is zero for $c = 1$.

Lemma 5. *For $c = 1, \dots, C - 1$ it holds asymptotically that*

$$T_{c,ij} \approx \sum_{k=1}^{c} C_{kkij} \left\{ \prod_{l=k+1}^{c} (1 - \pi_{li} - \pi_{lj}) \right\}, \tag{20}$$

with the convention that for $k = c$ the empty product $\prod_{l=k+1}^{c}(1-\pi_{li}-\pi_{lj}) = 1$.

We now have all the required ingredients to compute asymptotic approximations to all covariances $\mathrm{cov}(\hat{p}_{ci}, \hat{p}_{cj}) = C_{ccij}$ that occur in formula (8) . The following algorithm can be used:

1. Approximate $\mathrm{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj})$ by (16) for all $c = 1, \dots, C-1$ and define $C_{11ij} = T_{1,ij} = \mathrm{cov}(\hat{\pi}_{1i}, \hat{\pi}_{1j})$.
2. Repeat the following steps for $c = 2, \dots, C-1$:
    a. Approximate C_{ccij} by (17).
    b. Approximate $T_{c,ij}$ by (19) or (20).
3. Finally, define $C_{CCij} = T_{C-1,ij}$ according to (18).

By way of illustration, here are the first few iterations of this algorithm:

- (Step 1) Approximate all covariances $\mathrm{cov}(\hat{\pi}_{ci}, \hat{\pi}_{cj})$ $(c = 1, \dots, C-1)$ by (16).
- (Step 1) Define $C_{11ij} = T_{1,ij} = \mathrm{cov}(\hat{\pi}_{1i}, \hat{\pi}_{1j})$.
- (Step 2a, $c = 2$) Compute the approximation
$$C_{22ij} \approx \{\mathrm{cov}(\hat{\pi}_{2i}, \hat{\pi}_{2j}) + \pi_{2i}\pi_{2j}\}T_{1,ij} + (1-p_{1i})(1-p_{1j})\,\mathrm{cov}(\hat{\pi}_{2i}, \hat{\pi}_{2j})$$
according to (17).
- (Step 2b, $c = 2$) Compute the approximation
$$T_{2,ij} \approx C_{11ij} + C_{22ij} - (\pi_{2i} + \pi_{2j})T_{1,ij}$$
according to (19). Alternatively, compute
$$T_{2,ij} \approx C_{11ij}(1-\pi_{2i}-\pi_{2j}) + C_{22ij}$$
according to (20).
- (Step 2a, $c = 3$) Compute the approximation
$$C_{33ij} \approx \{\mathrm{cov}(\hat{\pi}_{3i}, \hat{\pi}_{3j}) + \pi_{3i}\pi_{3j}\}T_{2,ij} + (1-p_{1i}-p_{2i})(1-p_{1j}-p_{2j})\,\mathrm{cov}(\hat{\pi}_{3i}, \hat{\pi}_{3j})$$
according to (17).
- (Step 2b, $c = 3$) Compute the approximation
$$T_{3,ij} \approx C_{11ij} + C_{22ij} + C_{33ij} - (\pi_{2i} + \pi_{2j})T_{1,ij} - (\pi_{3i} + \pi_{3j})T_{2,ij}$$
according to (19). Alternatively, compute
$$T_{3,ij} \approx C_{11ij}(1-\pi_{2i}-\pi_{2j})(1-\pi_{3i}-\pi_{3j}) + C_{22ij}(1-\pi_{3i}-\pi_{3j}) + C_{33ij}$$
according to (20).
- (Step 2a, $c = 4$) Compute the approximation
$$C_{44ij} \approx \{\mathrm{cov}(\hat{\pi}_{4i}, \hat{\pi}_{4j}) + \pi_{4i}\pi_{4j}\}T_{3,ij} + (1-p_{1i}-p_{2i}-p_{3i})(1-p_{1j}-p_{2j}-p_{3j})\,\mathrm{cov}(\hat{\pi}_{4i}, \hat{\pi}_{4j})$$
according to (17).
- …

Note that the algorithm manages to avoid a circular argument, because C_{ccij} is computed using $T_{c-1,ij}$ and $T_{c,ij}$ is computed using only the covariances $C_{11ij}, \dots, C_{ccij}$.

Using (19) or (20) to approximate $T_{c,ij}$ should yield the same results, as these expressions are equivalent. In practice, expression (19) may lend itself to a slightly faster implementation, because we can recursively build and store the two sums $\sum_{k=1}^{c} C_{kkij}$ and $\sum_{k=2}^{c}(\pi_{ki} + \pi_{kj})T_{k-1,ij}$ over the course of the algorithm to avoid duplicate computations.

4.2 Application to variance estimation for Census tables

Recall from Section 3 that if the imputation model involves stratifying variables, the covariances $\mathrm{cov}(\hat{\pi}_{ci},\hat{\pi}_{cj})$ are asymptotically zero for units i and j in different strata. From (17) and (20), it can be shown easily by induction that the same holds for the covariances $\mathrm{cov}(\hat{p}_{ci},\hat{p}_{cj})$. Hence, in the notation used previously in this section, the final term in (8) can be simplified to:

$$\sum_{i\in U_2}\sum_{\substack{j\in U_2\\ j\neq i}} h_i h_j \,\mathrm{cov}(\hat{p}_{ci},\hat{p}_{cj}) = \sum_{g=1}^{G}\sum_{i\in U_{2g}}\sum_{\substack{j\in U_{2g}\\ j\neq i}} h_i h_j C_{ccij}, \tag{21}$$

In principle, the algorithm defined in Section 4.1 can be used to evaluate all C_{ccij}. In Section 3.2 we were able to reduce the double summation over i and j for $c=1$ to two single sums, which significantly reduces the amount of computational work. Unfortunately, a similar simplification is not possible for $c>1$ when $C\geq 3$.

Suppose that stratum U_{2g} consists of N_g persons, with $\sum_{g=1}^{G} N_g = N$. One could make a one-off effort to compute C_{ccij} for all $\sum_{g=1}^{G} N_g(N_g+1)/2$ possible pairs (i,j) with $i\leq j$ in each stratum, and store all these values for later use. (Note that it suffices to consider pairs with $i\leq j$, since $C_{ccij}=C_{ccji}$.) This would require a considerable amount of initial computational work, but once it is done, the variance of any entry in any estimated frequency table can be determined easily with formula (8). Alternatively, one could compute, for each $\hat{\theta}_{hc}$ of which the variance is to be determined, only those C_{ccij} for pairs (i,j) in U_{2g} with $h_i = h_j = 1$, since only these pairs have a non-zero contribution to (21). This would require only a limited amount of computational work for each entry, but it would involve some repetition and it would also mean that new covariances have to be computed for each new frequency table. Nevertheless, in practice it may be preferable to use this approach rather than the first one. This is certainly the case if it turns out that storing all $C\sum_{g=1}^{G} N_g(N_g+1)/2$ values of C_{ccij} is not practically feasible.

Turning to variance estimators, we can use the following estimate of (21) for the final term in variance estimator (9):

$$\sum_{i\in U_2}\sum_{\substack{j\in U_2\\ j\neq i}} h_i h_j \,\widehat{\mathrm{cov}}(\hat{p}_{ci},\hat{p}_{cj}) = \sum_{g=1}^{G}\sum_{i\in U_{2g}}\sum_{\substack{j\in U_{2g}\\ j\neq i}} h_i h_j \hat{C}_{ccij}. \tag{22}$$

Here, $\hat{C}_{ccij}$ can be determined analogously to C_{ccij}, by replacing all unknown elements π_{ci} and p_{ci} in (17), (18) and (19) or (20) by their estimates $\hat{\pi}_{ci}$ and $\hat{p}_{ci}$. Moreover, the covariances $\mathrm{cov}(\hat{\pi}_{ci},\hat{\pi}_{cj})$ can be estimated analogously to Section 3 by:

$$\widehat{\mathrm{cov}}(\hat{\pi}_{ci},\hat{\pi}_{cj}) = \hat{\pi}_{ci}(1-\hat{\pi}_{ci})\hat{\pi}_{cj}(1-\hat{\pi}_{cj})\mathbf{x}_{2i}^T\left\{\mathbf{X}_{2,g}^T\widehat{\boldsymbol{\Delta}}_{\geq c,g}\mathbf{X}_{2,g}\right\}^{-1}\mathbf{x}_{2j},\quad i,j\in U_{2g},$$

where $\widehat{\mathbf{\Delta}}_{\geq c,g}$ is an $n_g \times n_g$ diagonal matrix with

$$(\widehat{\mathbf{\Delta}}_{\geq c,g})_{ii} = \delta\big(y_{1i} = \cdots = y_{(c-1)i} = 0\big)\hat{\pi}_{ci}(1 - \hat{\pi}_{ci})$$

for all $i \in U_{2g} \cap S$. For the final term in the alternative variance estimator (10), a similar expression can be found. Finally, note that we can use $\hat{C}_{ccii}$ in place of $\widehat{\mathrm{var}}(\hat{p}_{ci})$ in the first term of (9) or (10).

Finally, it may be noted that it follows from the above derivations that $\mathrm{cov}\big(\hat{p}_{ci_1}, \hat{p}_{cj_1}\big) = \mathrm{cov}\big(\hat{p}_{ci_2}, \hat{p}_{cj_2}\big)$ whenever $\mathbf{x}_{i_1} = \mathbf{x}_{i_2}$ and $\mathbf{x}_{j_1} = \mathbf{x}_{j_2}$. Thus, in principle we do not have to evaluate these covariances for all possible pairs of units, but only for all pairs of unique combinations of covariate values $\mathbf{x}$ that occur in the population. In the special case that $\mathbf{x}$ consists solely of categorical variables with a limit number of levels, this could lead to a substantial reduction in computational work. However, when the number of covariates is large, or when $\mathbf{x}$ also contains numerical variables, this reduction may be small.

5. A bootstrap approach

As an alternative to the above analytical variance estimation method, we can consider an approach based on bootstrapping. The classical bootstrap method (see, e.g., Efron and Tibshirani, 1993) uses resampling with replacement from the original sample to approximate the sampling distribution of a target estimator. This method cannot be used directly here, since it does not account for sampling without replacement from a finite population. In fact, mass imputation is only meaningful in the context of a finite population.

Different extensions of the bootstrap to finite-population sampling have been developed. For mass imputation, a particularly useful extension is based on generating pseudo-populations. This methodology was developed by Gross (1980), Booth et al. (1994), Canty and Davison (1999) and Chauvet (2007). At Statistics Netherlands, Kuijvenhoven and Scholtus (2011) applied this type of bootstrap method to data on educational attainment, in a setting that is similar to the present paper but with estimators based on weighting rather than imputation.

We will re-use some of the notation in Section 2. In addition, let $w_{Si} = 1/\tau_{Si}$ denote the sampling weight of person $i \in S$. Let $w_{Si} = \lfloor w_{Si} \rfloor + \varphi_i$, with $\lfloor w_{Si} \rfloor \in \mathbb{N}$ and $\varphi_i \in [0,1)$.[4] The bootstrap algorithm consists of the following steps (Kuijvenhoven and Scholtus, 2011):

For each $a = 1, \dots, A$ do the following:

1. Create a pseudo-population[5] $\widehat{U}_a^* = U_1 \cup \widehat{U}_{2a}^*$ by taking ω_{Si} copies of each unit $i \in S$, where the random inflation weight ω_{Si} is chosen to be $\omega_{Si} = \lfloor w_{Si} \rfloor$ with probability $1 - \varphi_i$ and $\omega_{Si} = \lfloor w_{Si} \rfloor + 1$ with probability φ_i.[6]
2. For each $b = 1, \dots, B$ do the following:
   a. Draw a sample S_{ab}^* from $\widehat{U}_{2a}^*$ according to the same design that was used to draw S from U_2. If $k \in \widehat{U}_{2a}^*$ is a copy of $i \in S$ then its inclusion probability is $\tau_{S_{ab}^* k} \propto \tau_{Si}$, where the proportionality constant is chosen such that $\sum_{k \in \widehat{U}_{2a}^*} \tau_{S_{ab}^* k} = |S|$.
   b. Use S_{ab}^* to re-estimate the imputation model for $y_1, \dots, y_C$ that was used in the original estimation process.
   c. Use the estimated imputation model from Step 2b to impute the missing values of $y_1, \dots, y_C$ in $\widehat{U}_{2a}^* \backslash S_{ab}^*$.
   d. Analogously to the original estimates (1), compute the replicates $\hat{\theta}_{hc,ab}^* = \sum_{k \in U_1} h_k y_{ck} + \left(\sum_{k \in S_{ab}^*} h_k y_{ck} + \sum_{k \in \widehat{U}_{2a}^* \backslash S_{ab}^*} h_k \tilde{y}_{ck} \right)$.
3. Compute the variance estimate for $\hat{\theta}_{hc}$ based on pseudo-population $\widehat{U}_a^*$ as

[4] Here, $\lfloor z \rfloor$ denotes the integer part of a real number z, i.e., the largest integer that is smaller than or equal to z.

[5] In this description it is assumed, as in Section 2, that the sample S is drawn from U_2. Kuijvenhoven and Scholtus (2011) present a slightly different version of the algorithm based on the assumption that S is drawn from U.

[6] In Kuijvenhoven and Scholtus (2011), these random inflation weights are obtained using Fellegi's method for consistent rounding. This has the nice property that $|\widehat{U}_a^*| = |U|$ with certainty.

$$v_a(\hat{\theta}_{hc}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_{hc,ab} - \overline{\hat{\theta}^*_{hc,a}}\right)^2,$$
$$\overline{\hat{\theta}^*_{hc,a}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_{hc,ab}.$$

Compute the final variance estimate for $\hat{\theta}_{hc}$ by averaging over the pseudo-populations:

$$\widehat{\mathrm{var}}_{boot}(\hat{\theta}_{hc}) = \frac{1}{A}\sum_{a=1}^{A} v_a(\hat{\theta}_{hc}).$$

The outer for loop of this algorithm is intended to reduce the noise due to the random assignment of integer-valued inflation weights to units with non-integer sampling weights. Previous results in Kuijvenhoven and Scholtus (2011) and Chauvet (2007) suggest that this additional for loop may have little added value in practice (i.e., choosing $A = 1$ leads to variance estimates of a similar accuracy as choosing $A > 1$). It can certainly be avoided in the special case that all units in the original sample have integer-valued sampling weights. For variance estimation, $B = 200$ replicates is often sufficient.

The bootstrap method is straightforward to implement and can in fact re-use most of the code that was created to compute the original estimates. It is a computationally intensive method, although in this case the same could be said of the analytical approach discussed above. One advantage of the bootstrap method is that the time-consuming parts of the above algorithm (Steps 1 and 2a–c) have to be performed only once. The resulting mass-imputed pseudo-populations can be stored and used to compute a variance estimate for any estimator $\hat{\theta}_{hc}$ by generating the replicates $\hat{\theta}^*_{hc,ab}$ 'on the fly'.

It may be noted that, to store the information from the bootstrap procedure efficiently, we do not need to keep the full pseudo-population(s). Since each pseudo-population consists of copies of units in the original sample S, we can store all relevant information in a matrix of $|S|ABC$ integers, by computing, for each unit $i \in S$ and each c, the values $y^*_{ci,ab} = \sum_{k\in S^*_{ab}} \alpha_{ki} y_{ck} + \sum_{k\in \hat{U}^*_{2a}\setminus S^*_{ab}} \alpha_{ki}\tilde{y}_{ck}$, with $\alpha_{ki} = 1$ if k is a copy of unit i and $\alpha_{ki} = 0$ otherwise. Then for any target estimate, the replicates $\hat{\theta}^*_{hc,ab}$ can be computed as

$$\hat{\theta}^*_{hc,ab} = \sum_{i\in U_1} h_i y_{ci} + \sum_{i\in S} h_i y^*_{ci,ab},$$

since $h_k = h_i$ for all $k \in \hat{U}^*_{2a}$ with $\alpha_{ki} = 1$. In addition, note that the contribution of $\sum_{i\in U_1} h_i y_{ci}$ to $\hat{\theta}^*_{hc,ab}$ is constant and could therefore be ignored when computing the bootstrap variance estimate.

6. A small simulation study

To test the two proposed variance estimation methods in this paper (analytical approximation and bootstrap resampling), a small simulation study was conducted. As a basis for this study, we used the data of the synthetic Samplonia population (see, e.g., Bethlehem, 2009). All computations were done in the R environment for statistical computing.

As our target population, we used all persons over 14 in the Samplonia population ($N = 745$). In this simulation, there are no register data, so $U_1 = \emptyset$ and $U = U_2$. The sample S was drawn according to a simple random sampling design without replacement, with sample size $n = 149$, so that $N/n = 5$. Mass imputation of educational attainment for persons $U_2 \backslash S$ was based on the continuation-ratio logistic regression model of Section 3, with gender (two classes) as stratifying variable and age (three levels) and income (continuous) as additive predictors. For a second round of simulations, a larger population was created by concatenating five copies of the above population ($N = 3725$). The same sample design was used, with $n = N/5 = 745$.

The target frequency table in this simulation study consisted of a cross-classification of age (three levels) and educational attainment (three levels). The table below shows the true population counts in the small (left panel) and large (right panel) target population.

	true counts (small population)			true counts (large population)		
	educational attainment			educational attainment		
age (years)	low	medium	high	low	medium	high
young (15-35)	66	159	80	330	795	400
middle (36-55)	23	112	96	115	560	480
old (56+)	24	105	80	120	525	400

The next table shows the approximate true standard deviations of the associated estimated counts for the above sample design and mass imputation method. These were obtained by drawing 20 000 samples from each target population and for each of them estimating the imputation model, applying mass imputation and tabulating the target estimates.

	true standard deviations (small population)			true standard deviations (large population)		
	educational attainment			educational attainment		
age (years)	low	medium	high	low	medium	high
young (15-35)	15.7	19.3	16.8	34.5	42.2	36.8
middle (36-55)	10.3	16.9	16.6	22.3	36.8	36.1
old (56+)	10.3	16.2	15.6	22.8	35.6	34.5

Next, we simulated 100 samples from each population and applied formulas (9) and (22) to estimate analytical standard deviations. We also simulated 100 samples from each population and applied the bootstrap algorithm from Section 5, with $A = 1$ (there were no rounding issues since $w_{Si} = N/n = 5$ for all units) and $B = 200$.

The table below shows the mean value of the estimated standard deviations for the analytical method and (in brackets) their standard deviation across 100 simulations.

	estimated analytical st. dev. (small population)			estimated analytical st. dev. (large population)		
	educational attainment			educational attainment		
age (years)	low	medium	high	low	medium	high
young (15-35)	14.2	17.6	15.5	32.2	39.5	34.5
	(1.3)	(0.8)	(1.3)	(1.1)	(0.6)	(1.1)
middle (36-55)	9.0	15.0	14.7	20.6	34.0	33.3
	(1.7)	(0.9)	(1.0)	(1.5)	(0.7)	(0.8)
old (56+)	9.2	14.7	14.2	21.1	32.8	31.8
	(1.8)	(0.9)	(1.1)	(1.6)	(0.7)	(0.9)

It is seen that, for both populations, the analytical method underestimates the standard deviation for all entries in the target table. For all classes of educational attainment, the negative relative bias is stable at about 10% for the small population and about 7% for the large population. This negative bias might be due to the fact that the asymptotic approximations used in the derivations in Section 3 and 4 do not work well for small-sized samples. In fact, the relative bias is consistently closer to zero for the large population, which suggests that we may expect a better approximation for populations that are even larger.

The average relative contributions of the three terms in (9) to the total estimated variance are similar for all entries in the table: the first term contributes about 19% of the total estimated variance, the second term about 4%, and the third term about 77%. These fractions hold for both populations. Note that the third term, which is the most demanding to calculate, also has the largest contribution.

The next table shows the estimated standard deviations for the bootstrap method.

	estimated bootstrap st. dev. (small population)			estimated bootstrap st. dev. (large population)		
	educational attainment			educational attainment		
age (years)	low	medium	high	low	medium	high
young (15-35)	14.8	18.7	16.3	34.1	41.9	36.4
	(1.7)	(1.4)	(1.5)	(2.2)	(2.3)	(2.0)
middle (36-55)	10.1	16.5	16.2	22.7	36.6	36.0
	(2.2)	(1.3)	(1.3)	(2.4)	(2.0)	(2.1)
old (56+)	10.0	15.6	15.0	22.5	35.2	34.5
	(2.1)	(1.1)	(1.3)	(1.9)	(2.1)	(2.2)

It is seen that the bootstrap produces estimated standard deviations that are close to their true values, although in particular for the smaller population a slight negative bias still occurs. In fact, standard bootstrap variance estimates are known to have a negative bias of order $O(1/n)$ which can be non-negligible when the sample size n is small (Efron and Tibshirani, 1993).

The total computation time for the analytical method was about 16 minutes (so about 10 seconds per sample) for the small population and about 454 minutes (so about 4.5 minutes per sample) for the larger population. It should be noted that in this study formula (22) was evaluated by computing $\widehat{\mathrm{cov}}(\hat{p}_{ci}, \hat{p}_{cj})$ once for all unique pairs, rather than only those with $h_i = h_j = 1$ for each separate row in each target table. The latter approach would have been much faster here, given that we are only interested in one frequency table.

For the bootstrap method, the total computation time was about 92 minutes (so about 55 seconds per sample) for the small population and about 237 minutes (so about 2.4 minutes per sample) for the large population. This suggests that, for larger populations, the bootstrap method becomes more efficient than the analytical method.

7. Conclusion

In this report, two different approaches were developed for estimating the variances of frequency tables of cross-classifications that involve educational attainment, where missing values on this variable have been estimated by mass imputation. The first approach involves an analytical approximation, the second approach is based on a bootstrap procedure.

For the regression model that has been developed for the imputation of educational attainment in the Dutch virtual Census, the analytical method requires extensive computations, to such an extent that in practice it may be more time- and memory-consuming than the bootstrap method. In a small simulation study, both approaches yielded reasonable variance estimates. The bootstrap method was somewhat more accurate than the analytical method, which led to a slight underestimation in this example. In fact, the asymptotic approximations used in the derivation of the analytical formulas may be inappropriate when the sample size is small.

For practical applications, the bootstrap method may be more promising than the analytical method, as it is more flexible and more predictable in terms of the amount of computational work involved. For the analytical method, the required minimal amount of computing time and memory may vary widely between different target tables, as it depends on the number of pairs of units that contribute to each cell. To use this method reliably in practice, a further simplification of the formulas presented here may be necessary. Unfortunately, from the application in Section 6 it appears that the term in the variance approximation that is the most computationally complex is also the dominating term. It is not obvious how to find an approximation to this term that is both accurate and easy to compute.

One possible approach could be based on the remark made at the end of Section 4, that the double sum of covariance terms may be reduced by considering only all pairs of unique combinations of covariate values. If the number of unique combinations is still too large for practical computation – in particular, if numerical covariates occur in the imputation model – then variance approximations that are easier to compute but less accurate could be obtained by discretising all numerical covariates into a limited number of classes and/or merging classes of existing categorical covariates. This option could be investigated further to see whether a practical balance can be found between accuracy and the amount of computational work.

References

A. Agresti (2013), *Categorical Data Analysis* (Second Edition). John Wiley & Sons, New York.

J. Bethlehem (2009), *Applied Survey Methods: A Statistical Perspective*. John Wiley & Sons, Hoboken, NJ.

J.G. Booth, R.W. Butler, and P. Hall (1994), Bootstrap Methods for Finite Populations. *Journal of the American Statistical Association* **89**, 1282–1289.

A.J. Canty and A.C. Davison (1999), Resampling-based Variance Estimation for Labour Force Surveys. *The Statistician* **48**, 379–391.

G. Chauvet (2007), *Méthodes de Bootstrap en Population Finie*. PhD Thesis (in French), L'Université de Rennes.

B. Efron and R.J. Tibshirani (1993), *An Introduction to the Bootstrap*. Chapman & Hall/CRC, London.

S. Gross (1980), Median Estimation in Sample Surveys. In: *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 181–184.

P. Knottnerus and C. van Duin (2006), Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey. *Journal of Official Statistics* **22**, 565–584.

L. Kuijvenhoven and S. Scholtus (2011), Bootstrapping Combined Estimators based on Register and Sample Survey Data. Discussion Paper (201123), Statistics Netherlands, The Hague.

C.-E. Särndal, B. Swensson, and J. Wretman (1992), *Model Assisted Survey Sampling*. Springer-Verlag, New York.

S. Scholtus and J. Pannekoek (2015), Massa-imputatie van opleidingsniveaus. Report (PPM-2015-12-11-SSHS-JPNK; in Dutch), Statistics Netherlands, The Hague.

A.W. van der Vaart (1998), *Asymptotic Statistics*. Cambridge University Press, Cambridge.

T. de Waal and J. Daalmans (2018), Mass imputation for Census Estimation: Methodology. Report, Statistics Netherlands, The Hague.

Appendix: Additional proofs

Proof of Lemma 4. Expression (19) is trivially correct for $c = 1$. (In fact $T_{1,ij} = C_{11ij}$ holds exactly by definition.) Therefore, suppose that $2 \leq c \leq C - 1$. We begin by evaluating the 'off-diagonal' terms C_{cdij} with $c \neq d$. First suppose that $d < c$. We can write:

$$C_{cdij} = \operatorname{cov}\left\{\hat{\pi}_{ci}\left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki}\right), \hat{p}_{dj}\right\}.$$

Since $\hat{\pi}_{ci}$ is asymptotically independent of all $\hat{p}_{1i}, \dots, \hat{p}_{(c-1)i}$ and also of $\hat{p}_{dj}$, Lemma 2 can be applied to this expression, with $X = \hat{\pi}_{ci}$, $Y = 1 - \sum_{k=1}^{c-1} \hat{p}_{ki}$ and $Z = \hat{p}_{dj}$. This yields:

$$\begin{aligned} C_{cdij} &\approx E(\hat{\pi}_{ci}) \operatorname{cov}\left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki}, \hat{p}_{dj}\right) \\ &= -\pi_{ci} \sum_{k=1}^{c-1} \operatorname{cov}(\hat{p}_{ki}, \hat{p}_{dj}) \\ &= -\pi_{ci} \sum_{k=1}^{c-1} C_{kdij}, \quad (d < c). \end{aligned}$$

Similarly, we obtain for $d < c$ that

$$C_{dcij} \approx -\pi_{cj} \sum_{l=1}^{c-1} C_{dlij}, \quad (d < c).$$

These expressions can be substituted in the definition of $T_{c,ij}$, to find:

$$\begin{aligned} T_{c,ij} &= \sum_{k=1}^{c} \left\{C_{kkij} + \sum_{l=1}^{k-1} C_{klij} + \sum_{l=k+1}^{c} C_{klij}\right\} \\ &\approx \sum_{k=1}^{c} \left\{C_{kkij} - \sum_{l=1}^{k-1} \left(\pi_{ki} \sum_{m=1}^{k-1} C_{mlij}\right) - \sum_{l=k+1}^{c} \left(\pi_{lj} \sum_{m=1}^{l-1} C_{kmij}\right)\right\} \\ &= \sum_{k=1}^{c} C_{kkij} - \sum_{k=2}^{c} \left(\pi_{ki} \sum_{l=1}^{k-1} \sum_{m=1}^{k-1} C_{mlij}\right) - \sum_{l=2}^{c} \left(\pi_{lj} \sum_{k=1}^{l-1} \sum_{m=1}^{l-1} C_{kmij}\right) \\ &= \sum_{k=1}^{c} C_{kkij} - \sum_{k=2}^{c} \pi_{ki} T_{k-1,ij} - \sum_{l=2}^{c} \pi_{lj} T_{l-1,ij}, \end{aligned}$$

from which the result follows. In the third line, we used that the middle term is empty (hence zero) for $k = 1$ and we re-arranged the summation over k and l in the right-most term. □

Proof of Lemma 5. The statement is trivially correct for $c = 1$. (In fact $T_{1,ij} = C_{11ij}$ holds exactly by definition.) We proceed by induction on c. According to Lemma 4:

$$T_{c,ij} \approx \sum_{k=1}^{c} C_{kkij} - \sum_{k=2}^{c} (\pi_{ki} + \pi_{kj}) T_{k-1,ij}.$$

Using the induction hypothesis, we obtain:

$$\begin{aligned} T_{c,ij} &\approx \sum_{k=1}^{c} C_{kkij} - \sum_{k=2}^{c} (\pi_{ki} + \pi_{kj}) \left[\sum_{m=1}^{k-1} C_{mmij} \left\{ \prod_{l=m+1}^{k-1} (1 - \pi_{li} - \pi_{lj}) \right\} \right] \\ &= \sum_{k=1}^{c} C_{kkij} - \sum_{k=1}^{c-1} C_{kkij} \left[\sum_{m=k+1}^{c} (\pi_{mi} + \pi_{mj}) \left\{ \prod_{l=k+1}^{m-1} (1 - \pi_{li} - \pi_{lj}) \right\} \right] \\ &= C_{ccij} + \sum_{k=1}^{c-1} C_{kkij} \left[1 - \sum_{m=k+1}^{c} (\pi_{mi} + \pi_{mj}) \left\{ \prod_{l=k+1}^{m-1} (1 - \pi_{li} - \pi_{lj}) \right\} \right]. \end{aligned}$$

In the second line, we re-arranged the summation over k and m.

The expression in square brackets can be manipulated as follows (using the convention that the empty product for $m = k + 1$ is equal to 1):

$$\begin{aligned} &1 - \sum_{m=k+1}^{c} (\pi_{mi} + \pi_{mj}) \left\{ \prod_{l=k+1}^{m-1} (1 - \pi_{li} - \pi_{lj}) \right\} \\ &= 1 - \pi_{k+1,i} - \pi_{k+1,j} \\ &\qquad - \sum_{m=k+2}^{c} (\pi_{mi} + \pi_{mj})(1 - \pi_{k+1,i} - \pi_{k+1,j}) \left\{ \prod_{l=k+2}^{m-1} (1 - \pi_{li} - \pi_{lj}) \right\} \\ &= (1 - \pi_{k+1,i} - \pi_{k+1,j}) \left[1 - \sum_{m=k+2}^{c} (\pi_{mi} + \pi_{mj}) \left\{ \prod_{l=k+2}^{m-1} (1 - \pi_{li} - \pi_{lj}) \right\} \right]. \end{aligned}$$

We have extracted a factor $1 - \pi_{k+1,i} - \pi_{k+1,j}$ and are left with a similar expression as before, but with one fewer term. By repeating this procedure as many times as possible, we eventually obtain the identity:

$$1 - \sum_{m=k+1}^{c} (\pi_{mi} + \pi_{mj}) \left\{ \prod_{l=k+1}^{m-1} (1 - \pi_{li} - \pi_{lj}) \right\} = \prod_{l=k+1}^{c} (1 - \pi_{li} - \pi_{lj}).$$

Hence, we find that:

$$T_{c,ij} \approx C_{ccij} + \sum_{k=1}^{c-1} C_{kkij} \left\{ \prod_{l=k+1}^{c} (1 - \pi_{li} - \pi_{lj}) \right\}.$$

Now again using the convention that the empty product is equal to 1, it follows that the statement also holds for c. □

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2017–2018	2017 to 2018 inclusive
2017/2018	Average for 2017 to 2018 inclusive
2017/’18	Crop year, financial year, school year, etc., beginning in 2017 and ending in 2018
2013/’14–2017/’18	Crop year, financial year, etc., 2015/’16 to 2017/’18 inclusive

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

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information