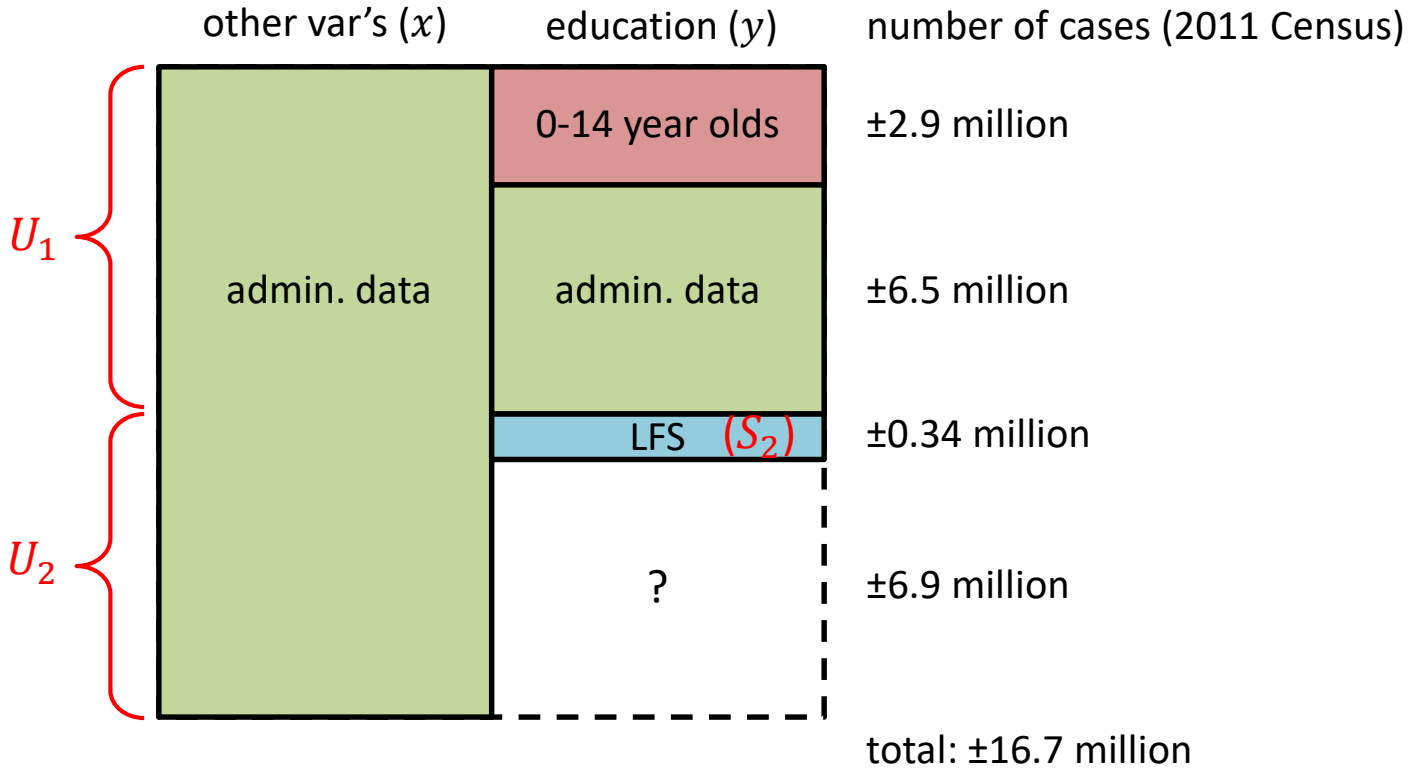
Variance estimation after mass imputation with an application to the Dutch population census

Sander Scholtus and Jacco Daalmans
(Statistics Netherlands)

UN/ECE Workshop on Statistical Data Editing
31 August 2020

Dutch Virtual Census



Dutch Virtual Census

- Goal: estimate frequency tables involving educational attainment
  - Typical element: $\theta_{hc} = \sum_{i \in U} h_i y_{ci} \quad (h_i \in \{0,1\}, y_{ci} \in \{0,1\}, c \in \{1, \ldots, C\})$

other variables	educational attainment (y)				
	level 1	…	level c	…	level C
1					
…					
h			θ_{hc}		
…					
H					

Dutch Virtual Census

- Goal: estimate frequency tables involving educational attainment
  - Typical element: $\theta_{hc} = \sum_{i \in U} h_i y_{ci} \quad (h_i \in \{0,1\}, y_{ci} \in \{0,1\}, c \in \{1, \ldots, C\})$

- Proposal (De Waal and Daalmans, 2018): use mass imputation
  - Estimate (e.g., logistic regression) model for $y_1, \ldots, y_C$ based on S_2
  - For $i \in U_2 \backslash S_2$, obtain predicted probabilities $\hat{p}_{1i}, \ldots, \hat{p}_{Ci}$ from estimated model
  - For $i \in U_2 \backslash S_2$, draw imputed 0-1-values $\tilde{y}_{1i}, \ldots, \tilde{y}_{Ci}$ based on $\hat{p}_{1i}, \ldots, \hat{p}_{Ci}$
  - Estimator for θ_{hc}:

$$\hat{\theta}_{hc} = \theta_{hc1} + \hat{\theta}_{hc2} = \sum_{i \in U_1} h_i y_{ci} + \left(\sum_{i \in S_2} h_i y_{ci} + \sum_{i \in U_2 \backslash S_2} h_i \tilde{y}_{ci} \right)$$

How to evaluate the variance of $\hat{\theta}_{hc}$?

- Analytical variance formula
  - Simplifying assumption: suppose that

$$\sum_{i \in S_2} h_i (\hat{p}_{ci} - y_{ci}) = 0$$

  - Then the variance of $\hat{\theta}_{hc}$ can be estimated by:

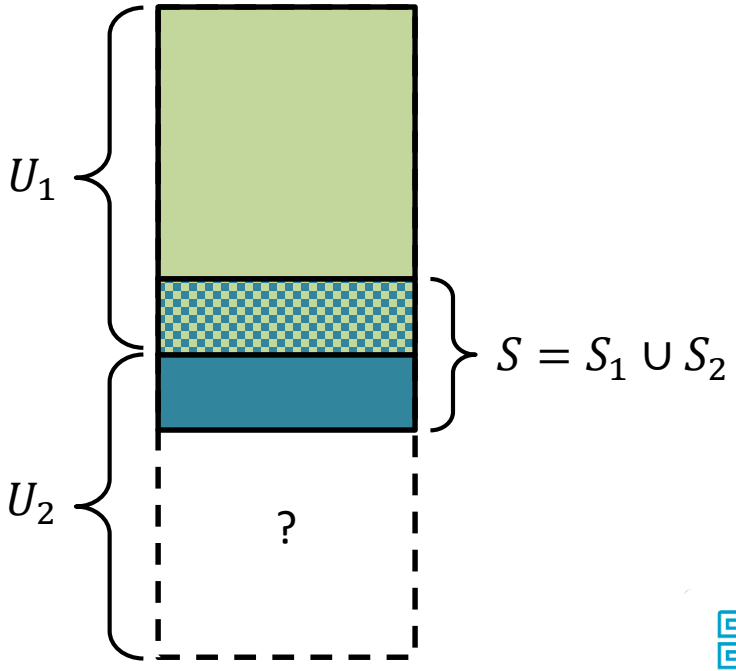
$$\widehat{\mathrm{var}}(\hat{\theta}_{hc} - \theta_{hc}) = \sum_{i \in U_2 \setminus S_2} h_i \hat{p}_{ci}(1 - \hat{p}_{ci}) + \sum_{i \in U_2} \sum_{j \in U_2} h_i h_j \, \widehat{\mathrm{cov}}(\hat{p}_{ci}, \hat{p}_{cj})$$

  - Expression for $\widehat{\mathrm{cov}}(\hat{p}_{ci}, \hat{p}_{cj})$ depends on imputation model (see paper)

- Bootstrap procedure

General set-up

- Target population $U = U_1 \cup U_2$
- Subpopulation U_1:
  - Admin. data available
  - Considered fixed (no variance)
- Probability sample S:
  - May have overlap with admin. data
  - $S_1 = S \cap U_1; \quad S_2 = S \cap U_2$
  - Inclusion probabilities π_i known
- Estimator of interest: $\hat{\theta} = t(U_1, S)$



Bootstrap

- Classical bootstrap (Efron, 1979) does not account for
  - Finite-population sampling
  - Complex survey design
- Different extensions of the bootstrap available
  - Overview: Mashreghi, Haziza and Léger (2016)
- Here: extension based on pseudo-populations
  - Theory: Booth, Butler and Hall (1994), Chauvet (2007)
  - Previous application: Kuijvenhoven and Scholtus (2011)

Bootstrap

- First assume: $w_i = 1/\pi_i$ is always integer-valued

Bootstrap algorithm:

1. Create a pseudo-population $\widehat{U}^*$ by taking w_i copies of each unit $i \in S$.

2. For each $b = 1, \ldots, B$ do the following:
   - Draw sample S_b^* from $\widehat{U}^*$ analogous to design used to draw S from U.
   - Analogously to $\hat{\theta} = t(S, U_1)$, construct replicate $\hat{\theta}_b^* = t(S_b^*, U_1)$.
3. Compute the variance estimate for $\hat{\theta}$ based on pseudo-population $\widehat{U}^*$ as:
   $\widehat{\mathrm{var}}_{\mathrm{boot}}(\hat{\theta} - \theta) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2$, with $\overline{\hat{\theta}^*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$.

Bootstrap

- General case: $w_i = 1/\pi_i = \lfloor w_i \rfloor + \varphi_i$, with $\lfloor w_i \rfloor \in \mathbb{N}$, $\varphi_i \in [0,1)$

Bootstrap algorithm:

1. Create a pseudo-population $\hat{U}^*$ by taking ω_i copies of each unit $i \in S$.
   Random inflation weight: $\omega_i = \lfloor w_i \rfloor$ with probability $1 - \varphi_i$,
   $\omega_i = \lfloor w_i \rfloor + 1$ with probability φ_i.
2. For each $b = 1, \ldots, B$ do the following:
   - Draw sample S_b^* from $\hat{U}^*$ analogous to design used to draw S from U.
   - Analogously to $\hat{\theta} = t(S, U_1)$, construct replicate $\hat{\theta}_b^* = t(S_b^*, U_1)$.
3. Compute the variance estimate for $\hat{\theta}$ based on pseudo-population $\hat{U}^*$ as:
   $\widehat{\mathrm{var}}_{\mathrm{boot}}(\hat{\theta} - \theta) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2$, with $\overline{\hat{\theta}^*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$.

Bootstrap

- General case: $w_i = 1/\pi_i = \lfloor w_i \rfloor + \varphi_i$, with $\lfloor w_i \rfloor \in \mathbb{N}$, $\varphi_i \in [0,1)$

Bootstrap algorithm:

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

1. Create a pseudo-population $\widehat{U}_a^*$ by taking ω_i copies of each unit $i \in S$.
   Random inflation weight: $\omega_i = \lfloor w_i \rfloor$ with probability $1 - \varphi_i$,
   $\omega_i = \lfloor w_i \rfloor + 1$ with probability φ_i.
2. For each $b = 1, \ldots, B$ do the following:
   - Draw sample S_{ab}^* from $\widehat{U}_a^*$ analogous to design used to draw S from U.
   - Analogously to $\hat{\theta} = t(S, U_1)$, construct replicate $\hat{\theta}_{ab}^* = t(S_{ab}^*, U_1)$.
3. Compute the variance estimate for $\hat{\theta}$ based on pseudo-population $\widehat{U}_a^*$ as:
   $v_a(\hat{\theta} - \theta) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_{ab}^* - \overline{\hat{\theta}_a^*} \right)^2$, with $\overline{\hat{\theta}_a^*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_{ab}^*$.

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

Bootstrap

Key step: Analogously to $\hat{\theta} = t(S, U_1)$, construct replicate $\hat{\theta}^*_{ab} = t(S^*_{ab}, U_1)$

- Example: Dutch Virtual Census with mass imputation
- Original estimator:
  $\hat{\theta}_{hc} = \sum_{i \in U_1} h_i y_{ci} + \left(\sum_{i \in S_2} h_i y_{ci} + \sum_{i \in U_2 \setminus S_2} h_i \tilde{y}_{ci} \right)$
- Construction of bootstrap replicate $\hat{\theta}^*_{hc,ab}$:
  - $\hat{U}^*_{2a}$ is the subpopulation of $\hat{U}^*_a$ consisting of copies of units from S_2
  - $S^*_{2ab} = S^*_{ab} \cap \hat{U}^*_{2a}$; note: size of overlap is random
  - Use S^*_{2ab} to re-estimate the imputation model for $y_1, \ldots, y_C$
  - Impute the missing values of $y_1, \ldots, y_C$ in $\hat{U}^*_{2a} \setminus S^*_{2ab}$
  - Compute: $\hat{\theta}^*_{hc,ab} = \sum_{k \in U_1} h_k y_{ck} + \left(\sum_{k \in S^*_{2ab}} h_k y_{ck} + \sum_{k \in \hat{U}^*_{2a} \setminus S^*_{2ab}} h_k \tilde{y}_{ck} \right)$

Simulation study

	true counts				true standard deviations		
	educational attainment				educational attainment		
age (years)	low	medium	high		low	medium	high
young (15–35)	330	795	400		34.5	42.2	36.8
middle (36–55)	115	560	480		22.3	36.8	36.1
old (56+)	120	525	400		22.8	35.6	34.5

- Synthetic target population of $N = 3725$ persons
- Simple random sample of size $n = N/5 = 745$; no admin. data
- Mass imputation based on extension of logistic regression (see paper)
- Imputation model: Gender × (Age + Income)
- True standard deviations:
  estimated by repeating sampling and imputing 20 000 times

Simulation study

	true counts				true standard deviations		
	educational attainment				educational attainment		
age (years)	low	medium	high		low	medium	high
young (15–35)	330	795	400		34.5	42.2	36.8
middle (36–55)	115	560	480		22.3	36.8	36.1
old (56+)	120	525	400		22.8	35.6	34.5

- Analytical variance approximation, repeated 100 times
- Bootstrap procedure with $A = 1$, $B = 200$, repeated 100 times

	estimated analytical st. dev.				estimated bootstrap st. dev.		
	educational attainment				educational attainment		
age (years)	low	medium	high		low	medium	high
young (15–35)	34.1	41.8	36.6		34.1	41.9	36.4
middle (36–55)	22.5	36.8	36.1		22.7	36.6	36.0
old (56+)	22.8	35.4	34.3		22.5	35.2	34.5

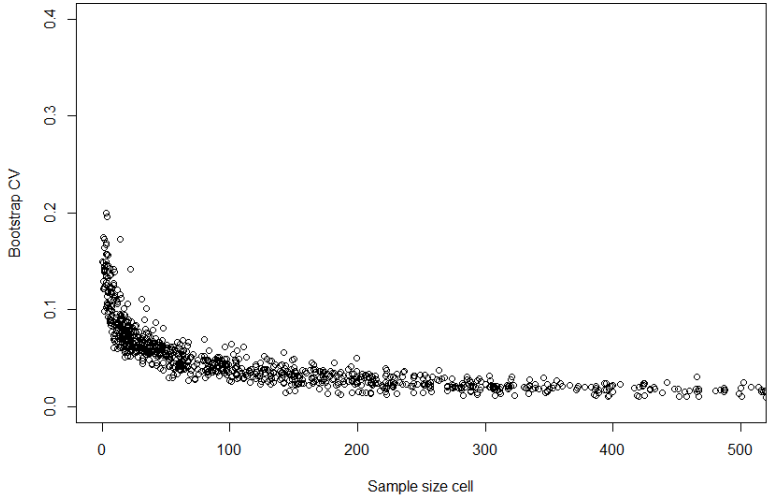
Application to Dutch Census 2011

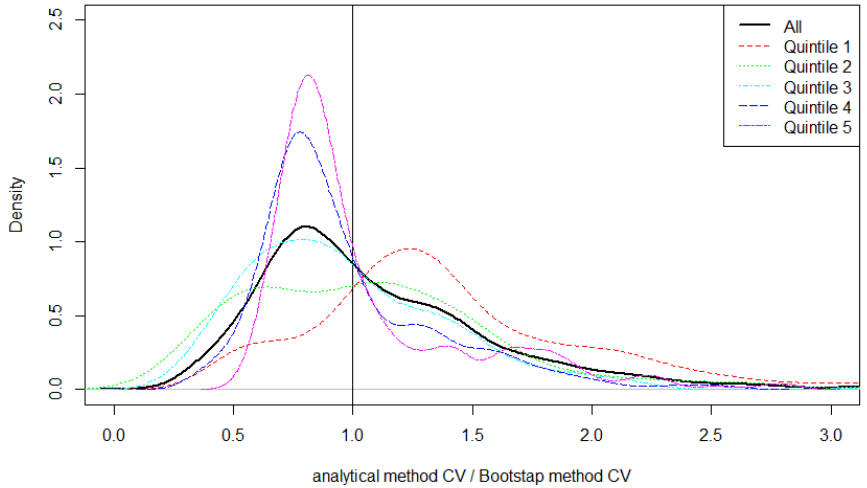
- Target table: Geographic area × Gender × Age × Educational attainment (1152 cells)
- Mass imputation for educational attainment based on extension of logistic regression (see paper)
- Imputation model: Income × (Geographic area + Gender + Age)
- Variance estimation:
  - Analytical variance approximation (computation time: 48.5 minutes)
  - Bootstrap procedure with $A = 1$, $B = 200$ (computation time: 21.5 hours)

Application to Dutch Census 2011

Results for Geographic area × Gender × Age × Educational attainment

Note: assumption $\sum_{i \in S_2} h_i(\hat{p}_{ci} - y_{ci}) = 0$ is *not* satisfied here

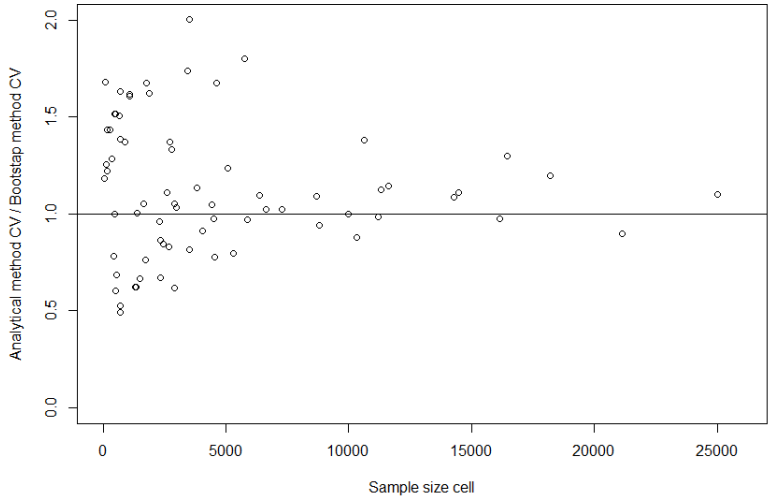




Application to Dutch Census 2011

Results for Geographic area × Educational attainment (72 cells)

Note: assumption $\sum_{i \in S_2} h_i(\hat{p}_{ci} - y_{ci}) = 0$ is satisfied approximately here



Conclusion

- Two approaches for estimating accuracy of frequency tables based on mass imputation: analytical method and bootstrap
  - Analytical method: requires simplifying assumptions, not easily generalisable
  - Bootstrap method: very flexible, fewer assumptions, computationally intensive

- Possible questions for future work:
  - How to extend analytical approach to imputation methods not based on a parametric model (e.g., hot deck imputation)
  - How to extend either approach to account for additional sources of uncertainty:
    - mass imputation for more than one variable
    - measurement error in observed values
    - micro-integration of survey and administrative data in overlapping part

References

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

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

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

B. Efron (1979), Bootstrap methods: another look at the jack-knife. *The Annals of Statistics* **7**, 1–26.

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

Z. Mashreghi, D. Haziza, and C. Léger (2016), A Survey of Bootstrap Methods in Finite Population Sampling. *Statistics Surveys* **10**, 1–52.