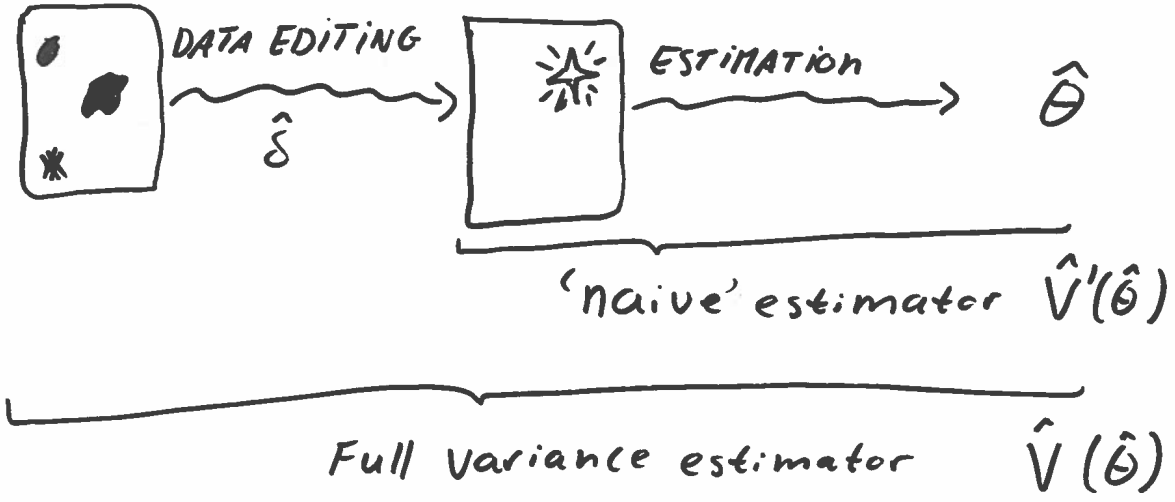
Computational estimates of data editing variance

Mark van der Loo, Jeroen Pannekoek, Lisanne Rijnveld

Statistics Netherlands

Data editing process



Estimation of variance

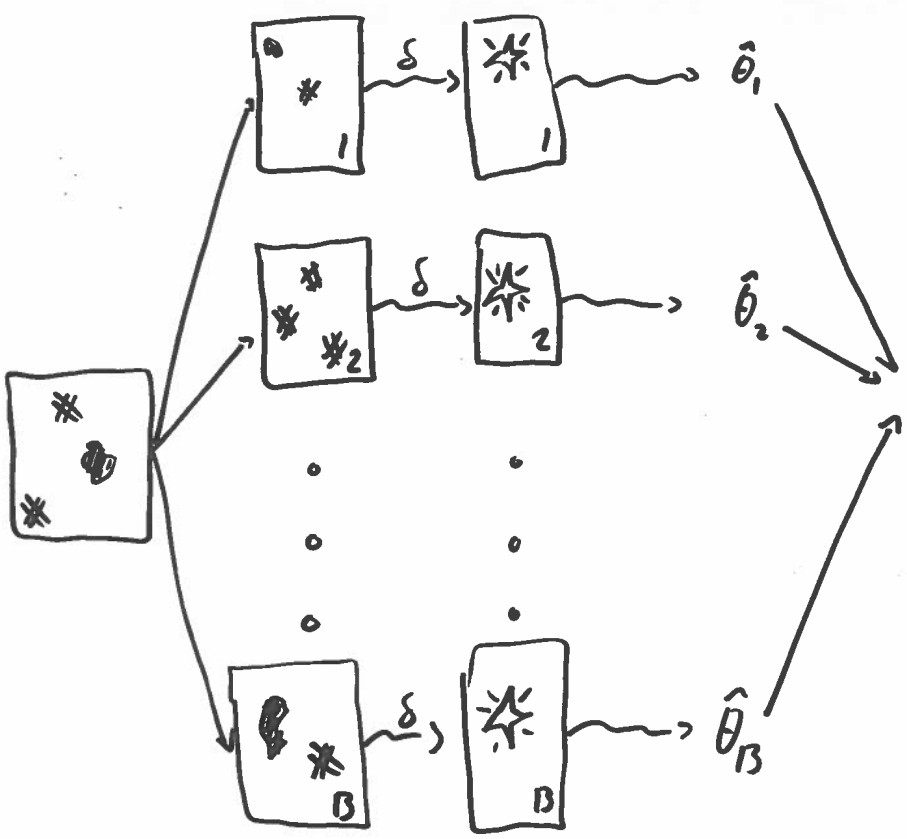
Example: naive estimation of variance of the mean

$$\hat{V}'(\hat{\theta}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

Total variance includes effects of data editing

$$V(\hat{\theta}) = E_\delta V(\hat{\theta}|\delta) + V_\delta E(\hat{\theta}|\delta)$$

Bootstrap approach



$$\hat{\theta}^* = \frac{1}{B}\sum_{j=1}^{B} \hat{\theta}_j$$

$$\hat{V}^*(\hat{\theta}^*) = \frac{1}{B}\sum_{j=1}^{B} (\hat{\theta}_j - \hat{\theta}^*)^2$$

The applied data editing procedure

Unit of measure error correction

↓

Correct typing errors

↓

Correct rounding errors

↓

Error localization

↓

Deductive imputation

↓

Ratio imputation

↓

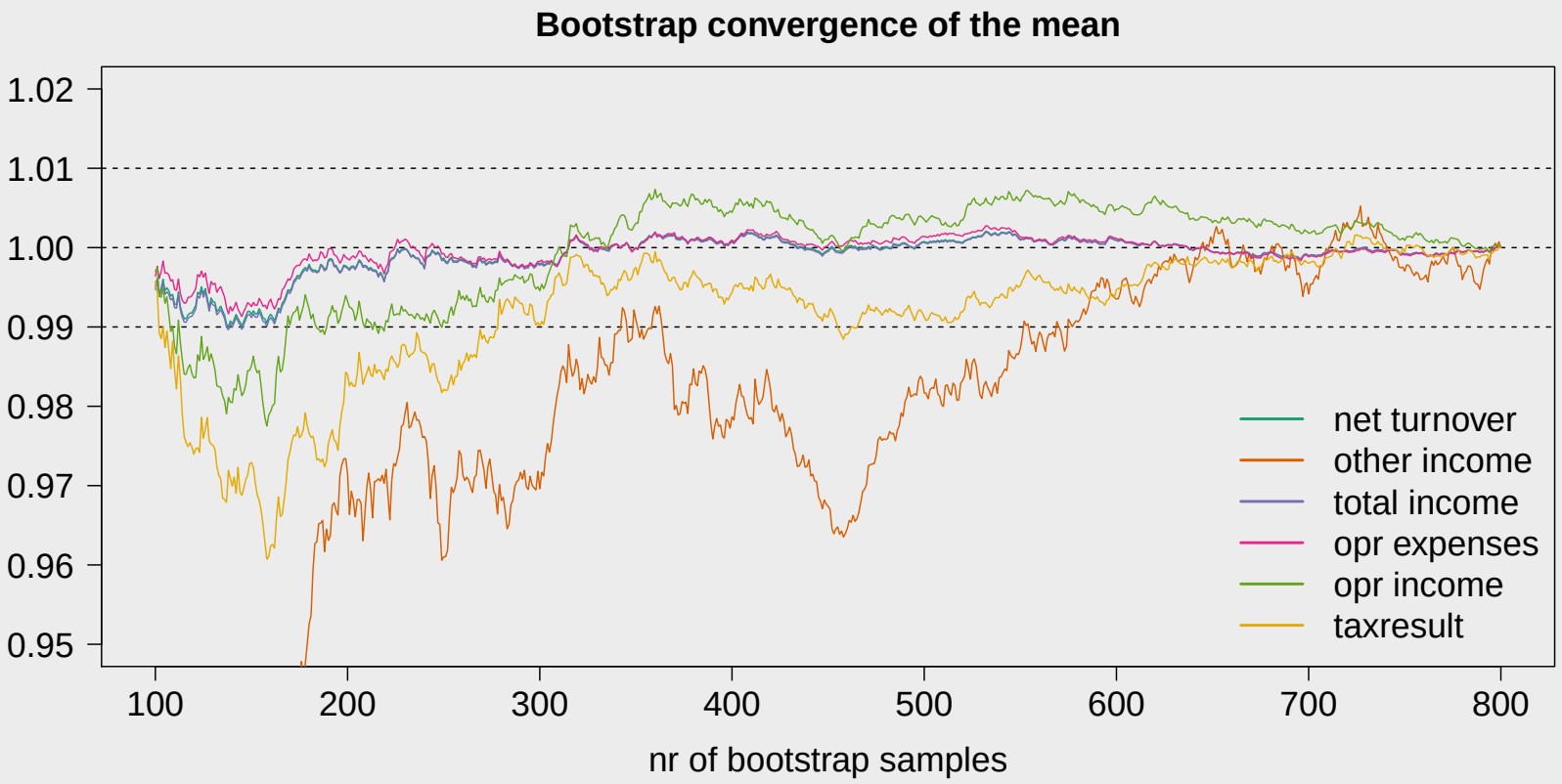
Median imputation

↓

Adjust to edits

} In bootstrap

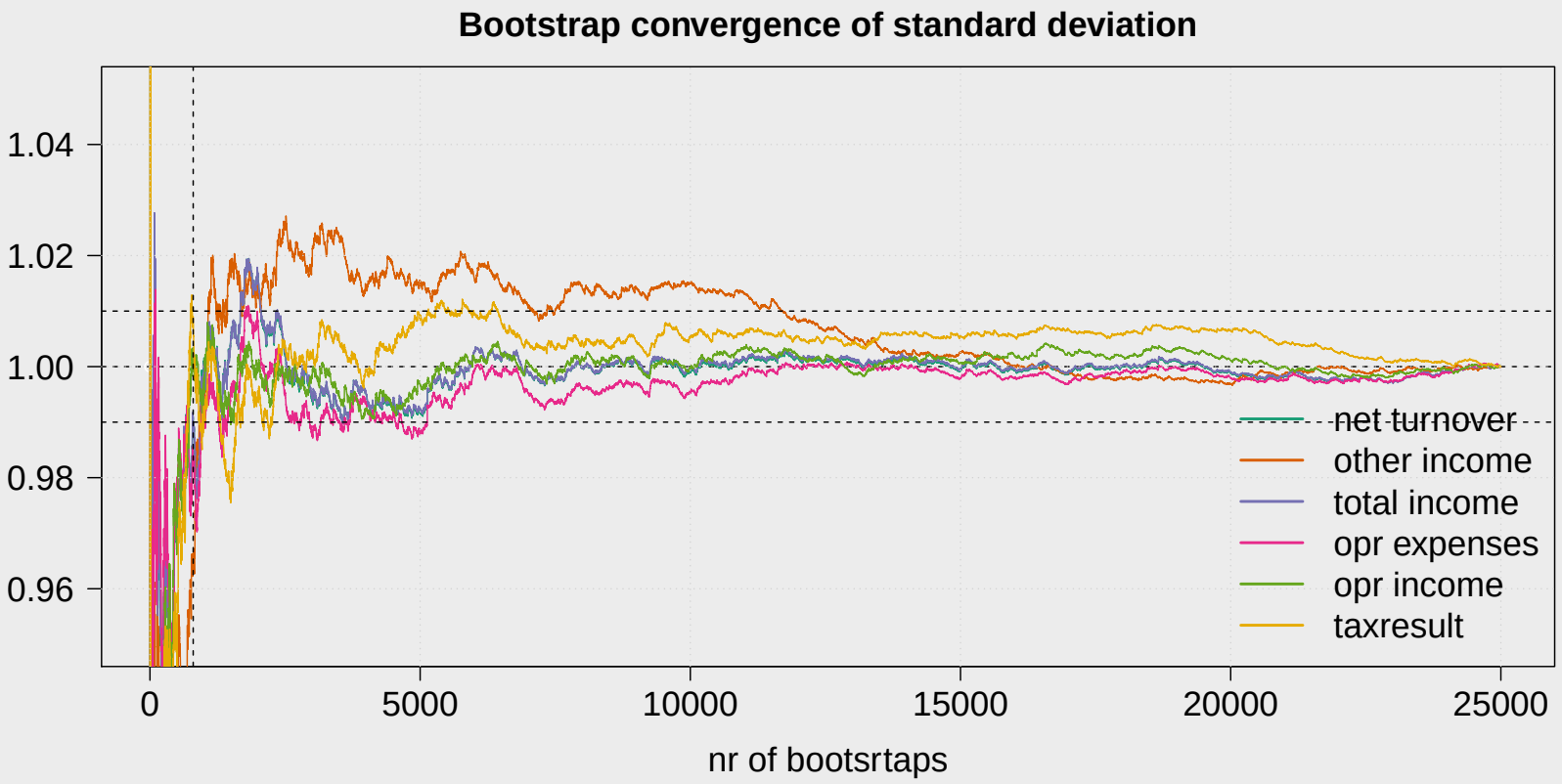
Convergence of the mean



Results

Variable	bootstr mean	percent bias	bootstr stdev	naive stdev	percent diff.
net turnover	27181.0	0.03	2371.5	2364.3	0.6
other income	130.8	2.83	75.1	75.4	-0.7
total income	27328.0	0.05	2388.8	2378.1	0.9
opr expenses	21938.0	-0.09	2058.4	2096.7	-3.8
opr income	758.4	1.66	144.7	134.2	13.9
taxresult	725.8	2.35	176.2	162.5	14.9

Convergence of variance



Results – interpreted

Variable	bootstr mean	percent bias	bootstr stdev	naive stdev	percent diff.
net turnover	27181.0	*0.03*	2371.5	2364.3	*0.6*
other income	130.8	2.83	75.1	75.4	*-0.7*
total income	27328.0	*0.05*	2388.8	2378.1	*0.9*
opr expenses	21938.0	*-0.09*	2058.4	2096.7	*-3.8*
opr income	758.4	1.66	144.7	134.2	13.9
taxresult	725.8	2.35	176.2	162.5	14.9

blue = 'significant' considering bootstrap convergence.

Summary / lookout

Conclusions

- Effect of variance including DE can be estimated with bootstrap
- For skewly distributed data, $B \gg n$ but by tracking convergence, conclusions can be drawn.
- We found larger DE variance for variables of lower quality

To improve

- Simple bootstrap not accurate for complex survey design