# A comparison of Kokic & Bell winsorisation and Conditional bias methods for outlier treatment

Work Session on Statistical Data Editing, The Hague, 24-26 April 2017

Thomas Deroyon & Cyril Favre-Martinoz - Insee

INSEE - Statistical Methods Division

# Introduction

# Outlier - a definition

DEFINITION:

- a unit is an outlier in a survey sample if estimates **vary strongly** whether **the unit is sampled** or not ([Chambers, 1986])
- Example: in one-stage stratified sampling with simple random sampling in each stratum, outlier = answers very different from the answers of the other units of the stratum

CLASSIFICATION:

- **Non-Representative outliers**: mainly errors, treated at the E & I step
- **Representative outliers**: treatment involves specific methods

$\Rightarrow$ presentation focuses on representative outliers' identification and treatment

# Outliers Origins

Outliers come from:

- Flaws in sampling frames, especially on stratification variables
- Weak correlation between design variables and survey variables of interest
- Multipurpose surveys, with weakly correlated variables of interest
- Highly skewed distributions of variables of interest
- Partial renewal of samples, generating strata jumpers

# Treatment Methods

BASIC PRINCIPLE:

Treatments aim at **limiting outliers' influence** on estimates, by changing answers or estimation weights

$\Rightarrow$ introduces bias but decreases estimation variance

$\Rightarrow$ **bias-variance trade-off**

CONFIGURATION:

Outliers are linked to certain configurations defined by four elements: variable of interest, parameter of interest, sampling design, sample estimate of the parameter

$\Rightarrow$ a unit may be an outlier in a configuration but not in another one

AIM OF THE PRESENTATION:

Comparing two methods on a practical case study

- Kokic & Bell winsorisation: tailored to stratified simple random sampling
- Conditional bias methods: based on a more general measure of influence

# Methods

# Methods

## Kokic and Bell Winsorisation

# Winsorisation

PRINCIPLE: applied to one-stage stratified sampling

- variable of interest $X$, parameter of interest is $\mathbf{X}$, $X$ total in population, estimator is the usual expansion estimator $\hat{\mathbf{X}} = \sum_h \frac{N_h}{n_h} \sum_{i \in s_h} X_i$
- definition of thresholds $K_h$ in each stratum
- computation of a winsorized variable $X^w$ obtained by cutting $X$ values above the threshold in each stratum

$$X^w = \begin{cases} X \text{ si } X < K_h \\ \textbf{Type II Winsorisation : } \frac{n_h}{N_h} X + (1 - \frac{n_h}{N_h}) K_h \text{ if } X > K_h \end{cases}$$

(see [Dalén, 1987] and [Tambay, 1988])

- winsorized estimate of $\mathbf{X}$ in population is $\hat{\mathbf{X}}^w = \sum_h \frac{N_h}{n_h} \sum_{i \in s_h} X_i^w$

# Framework

## Hypothesis:

- one-strage stratified sampling with simple random sampling in each stratum
- winsorized variable $X$ is always superior or equal to 0
- in each stratum, all $X$ variable values are independent realizations of the same random variable
- in each stratum, we have observations of the variable of interest $X$ independent from the survey sample
- thresholds $K_h$ are independent from the sample to which they are applied

## Objective:

computation of thresholds that can identify and treat outliers

- on average on the distribution of the variable of interest $X$
- whatever part of the population is sampled
- with **the objective of minimizing the winsorized estimates' MSE**

# Results

See [Kokic and Bell, 1994]:

- Thresholds minimizing winsorized estimate's MSE are all linear functions of the optimum winsorized estimate's bias
- Bias can be calculed as the point where a piecewise affine function depending on the variable of interest's distribution in each stratum is equal to zero
- Optimum bias and thresholds can be estimated with the values of $X$ available in each stratum and independent from the survey sample

# Methods

## Conditional Bias

# Definition

CONDITIONAL BIAS:

For a sampling design $\Pi$, a sampled unit $i$ and the estimate $\hat{\theta}$ of a parameter $\theta$

$$\mathbb{B}_{1i}(\hat{\theta}) = \mathbb{E}_{\Pi}(\hat{\theta}/i \in S) - \theta$$

See [Moreno-Rebollo et al., 1999], [Moreno-Rebollo et al., 2002] and [Beaumont et al., 2013]

REMARKS:

- direct measure of outlierness and influence
- can be estimated without bias with the sample
- special case: for Poisson sampling and the expansion estimator of a total:

$$\mathbb{B}_{1i}\Big(\sum_{i \in S} d_i X_i\Big) = (d_i - 1) X_i$$

# Conditional Bias outlier robust estimate

A DESIGN-ROBUST ESTIMATE BASED ON CONDITIONAL BIAS:

For the expansion estimate of a variable's total in the population, [Beaumont et al., 2013] suggest

$$\hat{\mathbf{X}}^{CB} = \hat{\mathbf{X}} - \sum_{i \in S} \hat{\mathbb{B}}_{1i}(\hat{\mathbf{X}}) + \sum_{i \in S} \Psi_c[\hat{\mathbb{B}}_{1i}(\hat{\mathbf{X}})]$$

with $\Psi_c$ Huber fonction defined by $\Psi_c(t) = \begin{cases} c & \text{if} \quad t \geq c \\ t & \text{if} \quad -c < t < c \\ -c & \text{if} \quad -c \leq t \end{cases}$

$\Rightarrow$ designed to produce an estimate with truncated values of the more extreme conditional biases

$c =$ tuning constant defined by the desired properties of $\hat{\mathbf{X}}^{CB}$ (for instance minimum MSE)

# Beaumont-Haziza-Ruiz Gazen estimate

OBJECTIVE:

[Beaumont et al., 2013] suggest to choose

$$c^* \in \mathit{argmin}_c \, \mathit{argmax}_{i \in S} \mid \hat{B}_{1i}(\hat{\mathbf{X}}) \mid$$

The obtained estimate has a very simple form:

$$\hat{\mathbf{X}}^{BHR} = \hat{\mathbf{X}} - \frac{\mathit{min}_{i \in S} \, \hat{\mathbb{B}}_{1i}(\hat{\mathbf{X}}) + \mathit{max}_{i \in S} \, \hat{\mathbb{B}}_{1i}(\hat{\mathbf{X}})}{2}$$

See [Favre-Martinoz et al., 2015] and [Favre-Martinoz et al., 2016] for generalisations

# Summary

Kokic & Bell winsorization:

- applies to expansion estimates of totals
- specific to stratified simple random sampling
- needs rich auxiliary information
- based on optimization of MSE

Conditional Bias method:

- applies to expansion estimates of totals
- based on a more general approach, that can be applied to almost all sampling designs
- does not need any information outside the sample
- based on minimization of maximum conditional bias

# Application to the Wage Structure and Labor Cost Survey

# The survey in brief

- Annual survey aiming at mainly estimating average hourly labour costs in domains (Nace sections, Nace sections * Nuts 2, Nace sections * number of employees)
- Two stage sampling designs:
  - FIRST STAGE: selection of local unit's sample with stratified simple random sampling
  - SECOND STAGE: selection of employees in the local unit's first stage sample
- Sampled local units detail elements of labour costs and number of worked hours for their sampled employees
- Estimation is based on the employees's sample with their estimation weight (taking into account sampling design, non-reponse treatment and calibration)

# Adaptation of outlier treatment methods

PARAMETER OF INTEREST:

average hourly labour costs in domain $D$, estimated by

$$\hat{R}(D) = \frac{\sum_{i \in S \cap D} w_i E_i}{\sum_{i \in S \cap D} w_i H_i}$$

whose variance is asymptotically

$$\mathbb{V}(\hat{R}(D)) \approx \mathbb{V}\Big(\sum_{i \in S \cap D} w_i L_i\Big) \text{ with } L_i = \big(E_i - \hat{R}(D) H_i\big) / \Big(\sum_{j \in S \cap D} w_j H_j\Big)$$

$\Rightarrow$ applications of outlier treatment methods on the estimate of $L_i$'s total

OTHER ADAPTATIONS:

- Kokic & Bell: employees treated as if selected directly by stratified simple random sampling in local unit's selection strata
- Conditional bias: employees' sampling design described as a Poisson sampling

VALIDATION:

by simulation, on information on labour costs and worked hours available in the sampling frame

# Results 1/2

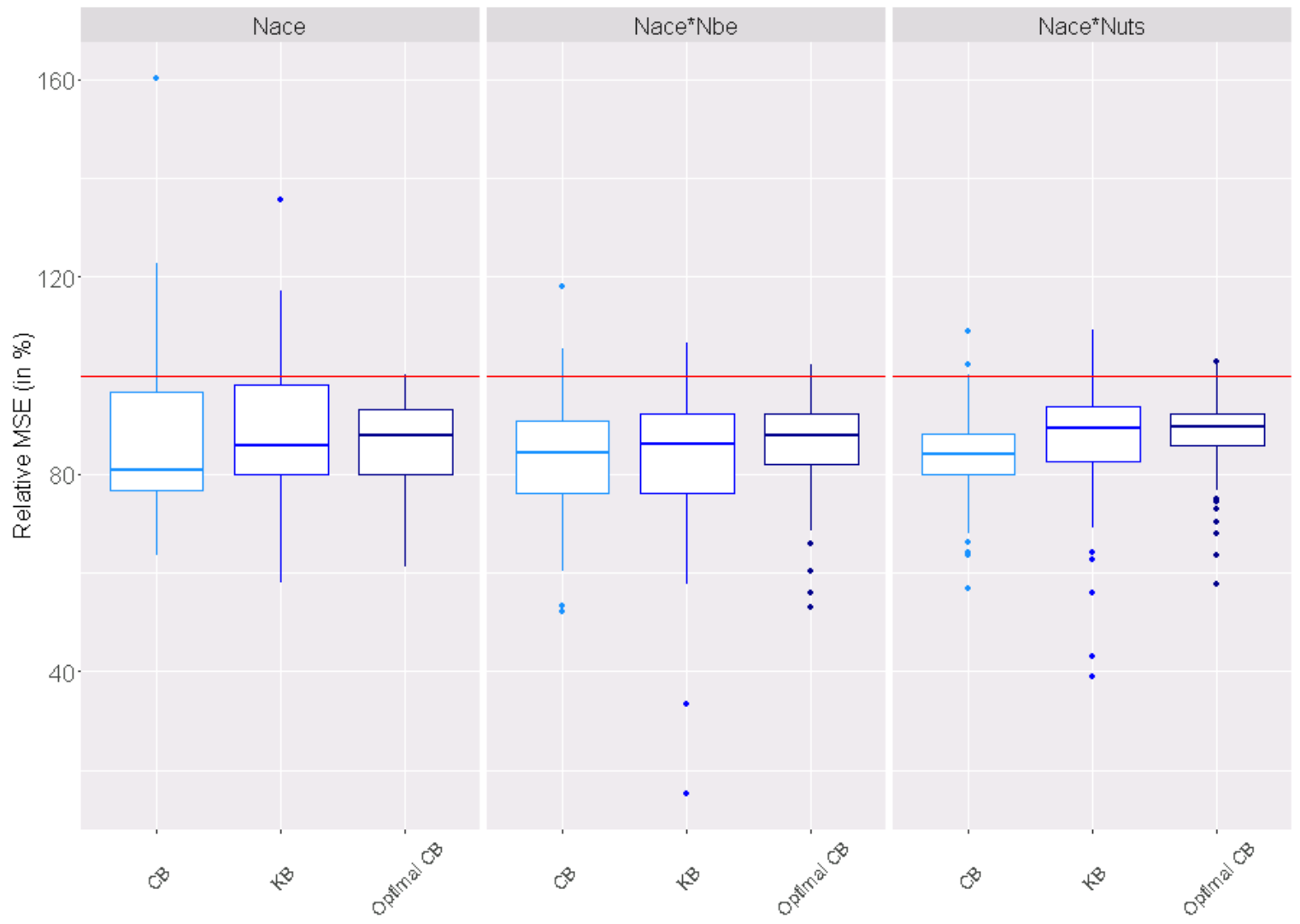


**Figure 1:** MSE's Distribution of average hourly wages' estimates by domains

# Results 2/2

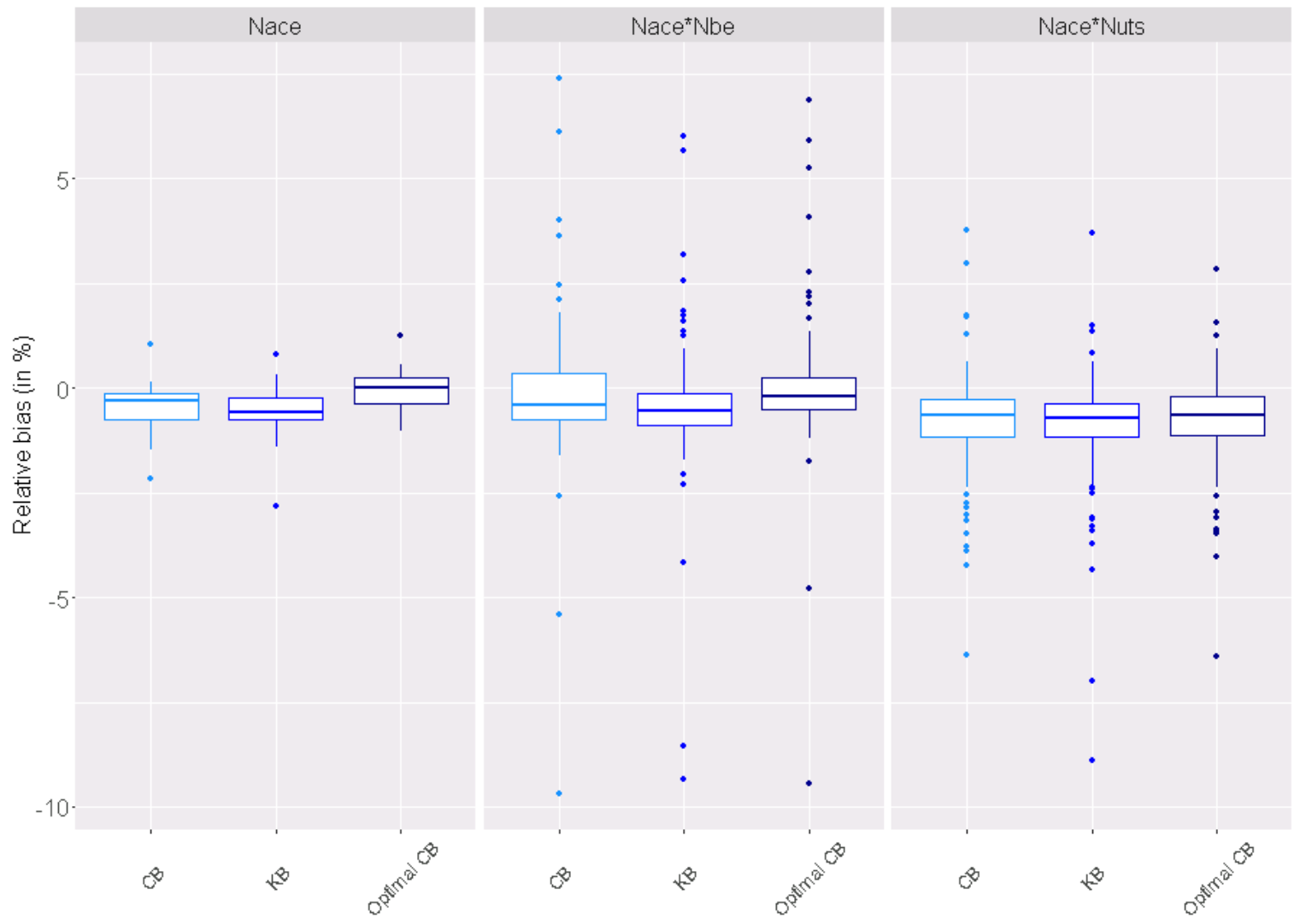


**Figure 2:** Relative bias's Distribution of average hourly wages' estimates by domains

# Conclusion

- Both methods show very similar results
- Both are able to identify and treat a limited number of outliers
- With a significant effect on variance and limited introduced bias
- Methods may decrease precision in some domains
- But these domains are those where the original estimates have a very low variance

$\Rightarrow$ Outlier treatments (Kokic & Bell winsorization) have been integrated to the survey's production process in 2016

Beaumont, J., Haziza, D., and Ruiz-Gazen, A. (2013).
**A unified approach to robust estimation in finite population sampling.**
*Biometrika*, 100:555–569.

Chambers, R. (1986).
**Outlier robust finite population estimation.**
*Journal of the American Statistical Association*, 81:1063–1069.

Dalén, J. (1987).
**Practical estimators of a population total which reduce the impact of large observations.**
R & D Report, Statistics Sweden.

Favre-Martinoz, C., Haziza, D., and Beaumont, J. (2015).
**A method for determining the cut-off points for winsorized estimators with application to domain estimation.**

*Survey Methodology*, 41:51–77.


Favre-Martinoz, C., Haziza, D., and Beaumont, J. (2016).
**Robust inference in two-phase sampling designs with application to unit nonresponse.**
*Scandinavian Journal of Statistics*, 43:1019–1034.


Kokic, P. and Bell, P. (1994).
**Optimal winsorizing cut-offs for a stratified finite population estimation.**
*Journal of Official Statistics*, 10(4):419–435.


Moreno-Rebollo, J., Muñoz Reyez, A., Jimenez-Gamero, J., and Muñoz Pichardo, J. (2002).
**Influence diagnostics in survey sampling: estimating the conditional bias.**
*Metrika*, 55:209–214.

Moreno-Rebollo, J., Muñoz Reyez, A., and Muñoz Pichardo, J. (1999).
**Influence diagnostics in survey sampling: conditional bias.**
*Biometrika*, 86:923–968.

Tambay, J.-L. (1988).
**An integrated approach for the treatment of outliers in sub-annual surveys.**
In *Proceedings og the Survey Research Methods Section*, pages 229–234. American Statistical Association.