Comparison between Clark and Kokic and Bell approaches in winsorization

Romain Lesauvage

Insee, France

Statistical Data Editing
October 6th, 2022

Outline

Table of Contents

Introduction

1. Economic variables with higly skewed distribution very usual in business survey
2. **Influential units problems**
3. it a way to limit the impact of these values in estimators ?
4. Main issue : determination of the atypical units

$\Rightarrow$ **Winsorization**

Table of Contents

Winsorization

- **Winsorization** : transformation of the a variable of interest Y into another Y^* defined as :

$$Y^* = \begin{cases} Y & \text{if } Y \leq K_h \\ \frac{n_h}{N_h} Y + (1 - \frac{n_h}{N_h}) K_h & \text{if } Y > K_h \end{cases}$$

- We have to fix a value for K_h: this is where different approaches come.

Table of Contents

Kokic and Bell approach

We suppose that we have a stratified sample and note h the quantity depending of the strata h.

$$K_h = -\frac{B}{\frac{N_h}{n_h} - 1} + \mu_h$$

1. Using this K_h, the winsorized estimator extend the HT estimator.
2. Winsorized estimator biased but has the smallest error in estimator the total of Y on average of all possible samples.
3. B is the bias of the minimum winsorized estimator, n_h is the number of units sampled in the stratum h, N_h is the size of population in stratum h and μ_h is the expectation of Y in the stratum h.

How to calculate the bias B?

The bias B is calculated as a zero of the function:

$$F(B) = -B[1 + \sum_h n_h E_h(J_h^*)] - \sum_h n_h E_h(Y_h^* J_h^*)$$

- E_h is the expectation in the stratum h
- $Y_h^* = (\frac{N_h}{n_h} - 1)(Y_h - \mu_h)$
- $J_h^* = 1$ if and only if $Y_h \geq K_h$

The function can be rewritten as a function of $L = -B$ and computed as a piecewise affine function.

Clark method

The Clark method works not only for stratified samples, we need auxliaries variables. It's a generalization of Kokic and Bell method.

1. Hypothesis: in each stratum, $Y_h = \mu_h + \epsilon_h$ (same as Kokic and Bell)
2. $K_h = -\frac{B}{\frac{N_h}{n_h} - 1} + \mu_h^*$ with $\mu_h^* = E[\min(Y, K_h)]$, difficult to calculate so we need to estimate it by $\hat{\mu}_h$
3. Find the zero of the function $L - E[\sum_{i \in s} \max(\hat{D}_i - L, 0)]$ with $\hat{D}_i = (Y_i - \hat{\mu}_i)(\omega_i - 1)$, ω_i being the weigth of unit i.

Connecting the two approaches

- The two functions used in the two methods can be connected with some hypothesis, so it seems to be the same method...
- ... But there is a main difference : calculation of μ_h, Kokic and Bell propose to use an independant survey/a previous edition of the survey to compute a value that estimate μ_h whereas Clark proposes to find it using a regression.
- Is there a big difference between the two ways of calculate μ_h ?

Table of Contents

Application to real data : a French survey, ESA-EAP

- The ESANE system makes it possible to produce structural business statistics in France. This is done through an annual survey, ESA-EAP, of approximately 160,000 companies.
- We used the data of the 2020 survey to compare the impact of winsorization with the two methods : Kokic and Bell (KB) and Clark.
- In the survey, we make a difference between the companies with only one legal unit (called independant unit) and those with several legal units (called profiled companies).

Results

	KB	Clark - independant data	Clark - sample	Clark - corr. factor
Ind. units	283	35	1616	1448
Other	158	28	459	340
Total	441	63	2075	1788

- $\times 7$ using KB instead of Clark - independant data (preconised solution)
- but $\times 4$ using Clark with other ways of calculate μ_h

Table of Contents

Conclusion

1. The two methods can be reunited by rewritting the functions we have to use...
2. ... but one main difference : estimation of μ_h
3. In real data, we see that the method leads to very different results.
4. Which method is the best ? Simulations to do to try to see which method has the least RMSE.

Thank you!

Contact: romain.lesauvage@insee.fr