Les seuils de Kokic et Bell performants dans d'autres cadres que le sondage aléatoire simple et stratifié

Thomas Deroyon, Cyril Favre-Martinoz, Arnaud Fizzala - Insee

Colloque francophone sur les sondages, Lyon, octobre 2018

Plan

Unité atypique

Définition :

- Les unités atypiques sont des unités de la population qui, suivant qu'elles appartiennent ou pas à l'échantillon, changent beaucoup le niveau des estimateurs calculés et diffusés.
- Exemple - dans un sondage stratifié à un degré avec SAS dans les strates (très courant en statistiques d'entreprises), une unité est atypique si ses réponses diffèrent fortement des réponses des autres unités de la strate

Deux types d'unités atypiques

- UNITÉS ATYPIQUES NON REPRÉSENTATIVES : l'unité ne peut représenter qu'elle même
  - Exemple : Chiffre d'affaires paraissant élevé dû à une erreur d'unité lors de la déclaration.
- UNITÉS ATYPIQUES REPRÉSENTATIVES : l'unité diffère des autres unités de sa strate, mais on ne peut pas supposer qu'elle ne représente qu'elle même.
  - Exemple : Chiffre d'affaires élevé mais confirmé par l'entreprise.

Cette présentation concerne le traitement des unités atypiques **représentatives** uniquement.

Causes d'apparition des unités atypiques représentatives

Dans les enquêtes auprès des entreprises, la présence d'unités atypiques est quasiment inévitable :

- Variables d'intérêt très assymétriques ;
- Décalage temporel entre l'information servant à constituer les strates (secteur d'activité, effectif salarié, chiffre d'affaires) et la collecte des données pendant lequel les entreprises ont pu évoluer + Taux de sondage croissant avec la taille des entreprises => Phénomène de "strata jumpers"

Les unités de la partie exhaustive

Les unités de la partie exhaustive (20% à 80% de l'échantillon selon les enquêtes), ne représentent qu'elles-mêmes (poids de 1 !), et ne sont donc pas des unités atypiques représentatives.

- Les méthodes que nous allons voir concernent la partie non-exhaustive de l'échantillon.

Une première solution

Les unités atypiques représentatives ont longtemps été traitées à l'Insee en ramenant leur poids de sondage à 1 et en repondérant les autres unités de la strate de tirage.

Facile à mettre en place, cette solution pose quelques questions :

- Le caractère subjectif de la détection d'une unité : à partir de quel moment une unité est-elle considérée atypique ?
- le passage systématique du poids à 1 : pourquoi pas une autre valeur ?

La winsorisation selon la méthode de Kokic et Bell apporte une réponse "scientifique" à ces questions et est donc de plus en plus utilisée à l'Insee.

Principe

Cadre : sondage stratifié à un degré avec sondage aléatoire simple dans chaque strate

Principe :

- choix d'une variable d'intérêt X
- définition de seuils K_h dans chaque strate
- création d'une variable winsorizée X^w obtenue en rabotant les valeurs de X qui dépassent les seuils dans chaque strate
- deux types de winsorization :

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

Arbitrage biais variance

Estimateur winsorizé : l'estimateur winsorizé du total de X est l'estimateur d'Horvitz-Thompson du total de la variable winsorizée

$$\hat{X}^w = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_h} X_i^w$$

- estimateur biaisé du total de X
- variance plus faible car la variance empirique $S_h^2(X^w)$ de X^w est plus faible que $S_h^2(X)$ dans chaque strate
- $\Rightarrow$ arbitrage **biais - variance**
- **paramètre central** : choix des seuils $K_h \Rightarrow$ méthode de Kokic et Bell

Principe

RÉSULTAT : Seuils qui minimisent l'erreur quadratique moyenne de l'estimateur winsorizé sous

- l'aléa résultant du plan de sondage
- la distribution de X dans chaque strate

HYPOTHÈSES :

- sondage stratifié à un degré avec sondage aléatoire simple dans chaque strate
- variable winsorizée X à valeurs positives ou nulles
- dans une strate, toutes les valeurs de X sont issues de la même loi, d'espérance μ_h
- les seuils K_h sont indépendants de l'échantillon auquel ils sont appliqués

Problèmes pratiques

En pratique, sauf à disposer d'une édition précédente de l'enquête, trouver des valeurs de X sur des unités **indépendantes** de l'échantillon dans chaque strate ne va pas de soi...

Une pratique courante à l'INSEE est de winsoriser le chiffre d'affaires (disponible pour toutes les unités de la base de sondage et corrélé à la majorité des variables d'intérêt) et de transférer l'effet de la winsorisation au poids des unités selon une règle de trois. De cette façon, la winsorisation du chiffre d'affaires impactera toutes les variables et on conservera les valeurs déclarées dans le fichier.

$$w_i^w = w_i * \frac{X_i^w}{X_i} \text{ puis } \hat{X}^w = \sum_{i=1}^{n} w_i^w X_i$$

Retour sur les hypothèses

Les seuils de Kokic et Bell sont optimaux sous certaines hypothèses :

- le plan de sondage de l'enquête doit être aléatoire simple stratifié
- le paramètre doit être le total d'une variable.

Dans des cas plus complexes, deux solutions sont envisageables :

- proposer des adaptations de la méthode de Kokic et Bell ;
- utiliser les méthodes de biais conditionnel (Beaumont et al., 2013) et (Favre-Martinoz et al., 2016) qui s'appliquent à n'importe quel plan de sondage.

La suite de la présentation concerne deux cas "plus complexes".

Contexte

ECMOSS : Enquête sur le Coût de la Main d'Oeuvre et la Structure des Salaires.

- salaires horaires moyens : comparaison entre les différents pays européens dans un ensemble de domaines d'intérêt (secteurs d'activité, régions,...)

Demande de mise en place de méthode de traitement des valeurs influentes, mais le cadre n'est pas celui de Kokic et Bell :

- Plan de sondage en deux phases
- Paramètre d'intérêt est un Ratio et pas un total

Adaptations proposées

Plan de sondage en 2 phases : on fait comme si l'échantillon de salarié était directement tiré selon un SAS stratifié

- Les écarts de poids au sein d'une même pseudo-strate sont donc négligés.

Ratio : on winsorise la variable linéarisée

- Pour éviter les valeurs négatives, on effectue une translation : $u_k^t = u_k + |\min(u_i)|$
- On winsorise ensuite u_k^t puis on transfère au poids $w_k^w = w_k \frac{u_k^{tw}}{u_k^t}$

Simulations

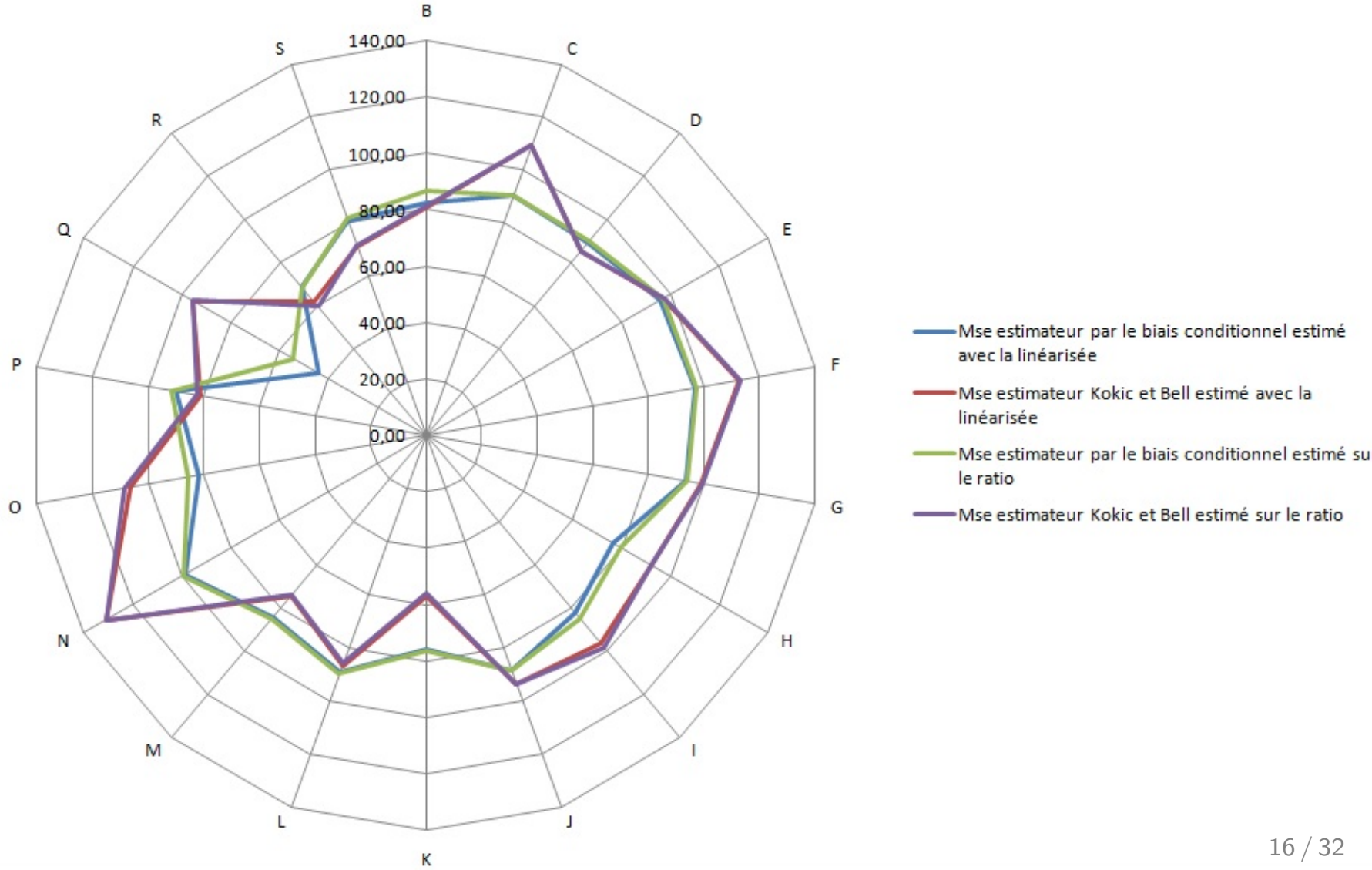
Salaires et nombre d'heures de travail issus des DADS (disponibles pour tout le champ).

1000 réplications de tirage

Comparaison avec des méthodes de Biais conditionnel qui nécessitent des hypothèses moins fortes :

- plan de sondage assimilé à du poissonien
- pas de translation à réaliser sur la variable linéarisée

Ratio des MSE par section



Résultats

- Amélioration des MSE par rapport à une situation sans traitement des valeurs influentes pour la quasi-totalité des domaines de diffusion
- Des résultats avec Kokic et Bell très proches de ceux obtenus avec les biais conditionnels

Cadre et problème

Cadre

Problématique actuelle de statistique d'entreprises : Passage des unités légales aux entreprises... Plus d'éléments dans la présentation **Le traitement des changements de contours des entreprises par partage des poids** à 10h30 !

Problème (ici)

Appliquer une méthode de traitement des valeurs influentes après un partage des poids...
...Mais dont l'échantillon initial est un SAS stratifié !

Adaptations de KB envisagées I

1 : Winsoriser l'échantillon original et effectuer ensuite le partage des poids à partir des poids winsorisés

- Détection des unités dont l'influence serait amplifiée par le partage des poids ?

2 : Faire comme si l'échantillon après partage des poids était issu d'un SAS stratifié et appliqué "directement" la winsorisation

- Conditions de calcul des seuils de Kokic et Bell ne sont pas vérifiées...

Adaptations de KB envisagées II

3 : Winsoriser l'échantillon original avec la variable Z , transformation judicieuse de la variable Y qui permet de retrouver l'estimateur du total de la variable Y après partage de poids, à partir de l'estimateur du total de la variable Z avec les poids issus de l'échantillon original.
On "profite" alors des propriétés du plan de sondage de l'échantillon original.

- Est-ce que winsoriser la variable Z permet de détecter les unités influentes pour la variable Y ?

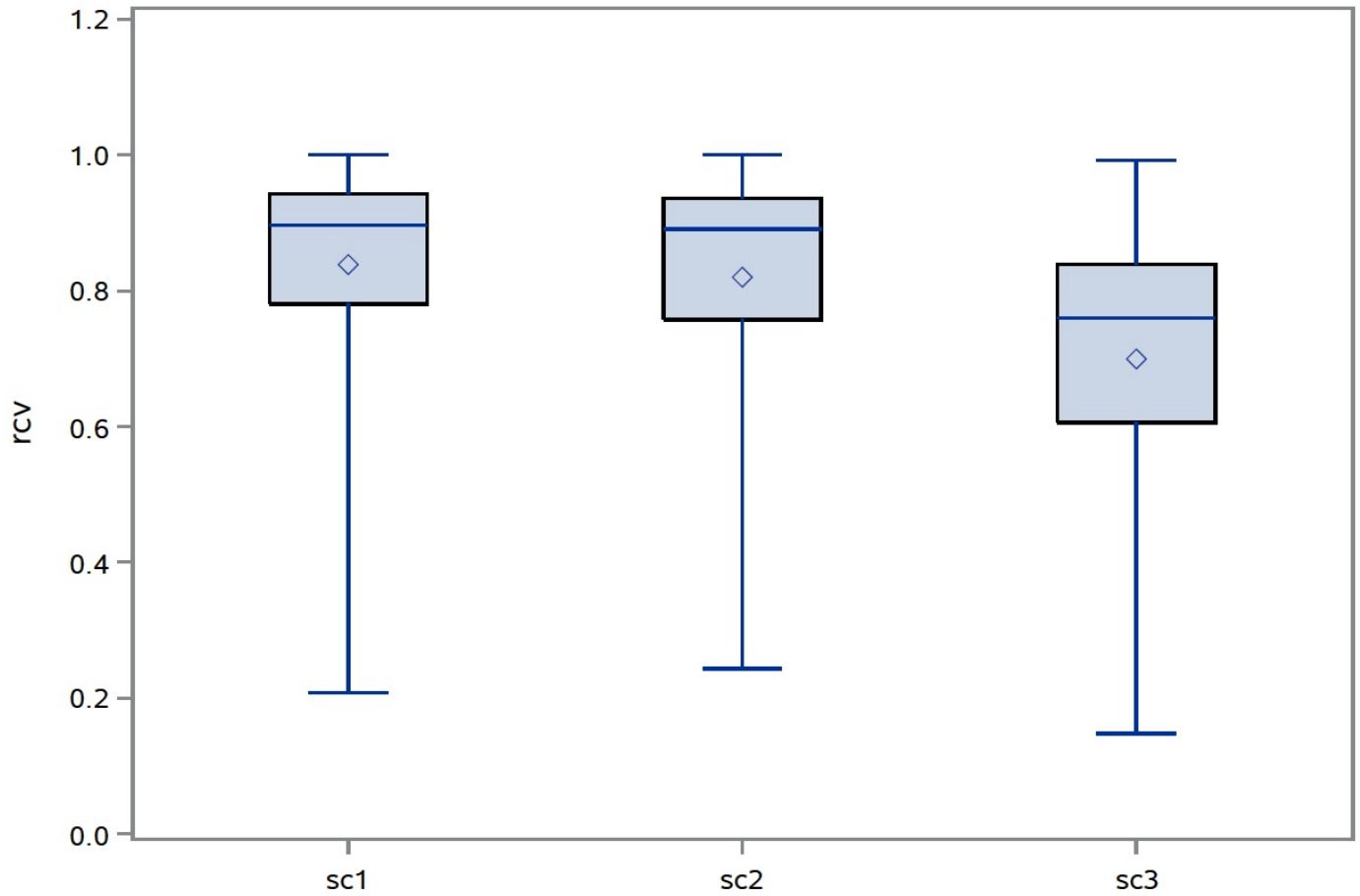
Simulations

50 000 replications du plan de sondage.

Comparaison des estimateurs du total de variables fiscales (disponibles dans la base de sondage) pour les trois scénarios, avec un estimateur sans traitement des valeurs influentes.

Indicateur : rapport entre le CV de l'estimateur étudié et le CV de l'estimateur sans traitement des valeurs influentes.

RCV pour l'estimation du Chiffre d'affaires par activité (200 modalités)



Résultats

- Les trois scénarios conduisent à une amélioration des CV (rapports inférieurs à 1) dans toutes les modalités d'activité (ou quasiment toutes lorsqu'il s'agit d'autres variables que le Chiffre d'affaires, résultats non présentés ici)
- Le scénario 3 parait le meilleur d'un point de vue précision, mais est plus complexe à mettre en place.
- étude à finaliser : rôle de la non-réponse, discussions avec le service en charge de l'enquête...

Les études sur lesquelles se base cette présentation

Adaptation de la winsorisation à un plan de sondage qui n'est pas un SAS stratifié et un paramètre d'intérêt qui est un Ratio

- ECMOSS : Article de Thomas Deroyon et Cyril Favre-Martinoz dans un *Techniques d'enquête* à paraître.

Adaptation de la winsorisation à un partage de poids

- ESA en EP : étude non finalisée, *nous contacter*.

Adaptation de la winsorisation à un tirage stratifié de grappes

- ESA en UL : Papier d'Arnaud Fizzala présenté aux *JMS 2018*

Extension de la méthode de Kokic et Bell au cas d'un plan de sondage poissonien

- Présentation à 10h30 dans la session Estimation de variance et Robustesse

Merci pour votre attention

Contact

thomas.deroyon@insee.fr
cyril.favre-martinoz@insee.fr
arnaud.fizzala@insee.fr

Bibliographie I

- Beaumont J-F, Haziza D, Ruiz-Gazen A, *A unified approach to robust estimation in finite population sampling*. Biometrika 100 (3), 555-569
- Brion, P. et Guggemos, F. (2010). *Du bon usage de la winsorisation... ou comment traiter les entreprises atypiques dans les enquêtes sectorielles annuelles*. Lettre du SSE n. 65.
- Gros, E. (2012) : *Assessment and improvement of the selective editing process in Esane*. Work Session on Statistical Data Editing.
- Guggemos, F. et Sautory, O. (2012), *La coordination d'échantillons d'enquêtes auprès des entreprises mise en place à l'Insee*, 11e Journées de méthodologie statistique de l'Insee.

Bibliographie II

- Kokic P.N., Bell P.A. (1994), *Optimal winsorizing cut-offs for a stratified finite population estimator.* Journal of Official Statistics, vol. 10, n. 4 : 419-435.
- Deroyon T. (2015). *Traitement des valeurs atypiques d'une enquête par winsorization - application aux enquêtes sectorielles annuelles.* Acte des Journées de Méthodologie Statistique.
- Deroyon T. Favre-Martinoz (à paraître). *?* Techniques d'enquête.
- Favre-Martinoz C., Haziza D. et Beaumont J-F. (2016) *Robust Inference in Two-phase Sampling Designs with Application to Unit Non-response.* Scandinavian journal of statistics vol. 43 :1019-1034.

Bibliographie III

- Fizzala A. (2018) *Comment redresser un échantillon d'unités légales tirées via leurs entreprises ?* Acte des Journées de Méthodologie Statistique.
- Di Zio M., Guarnera U. (2013), *A contamination model for selective editing*, Journal of Official Statistics, Vol. 29, n. 4, pp. 539–555.

Mise en place de Kokic et Bell I

$\Rightarrow$ Calcul des seuils qui minimisent l'erreur quadratique moyenne de l'estimateur winsorizé sous

- l'aléa résultant du plan de sondage
- la distribution de X dans chaque strate

RÉSULTATS :

- à l'optimum
- et asymptotiquement (quand $N_h \to +\infty$ et $n_h \to +\infty$)
- les seuils K_h^* et le biais de l'estimateur winsorizé B^* vérifient :

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

Mise en place de Kokic et Bell II

- $K_h \to +\infty$ quand $\frac{n_h}{N_h} \to 1$
- K_h est proche de μ_h quand le taux de sondage est très faible

RÉSULTATS :

- le biais de l'estimateur winsorizé B^* est le point où la fonction F s'annule avec :

$$F(B) = -B\,[1 + \sum_{h=1}^{H} n_h\, E_h(J_h^*)] - \sum_{h=1}^{H} n_h\, E_h(X_h^* J_h^*) \text{ avec}$$

  - E_h espérance sous la loi de X dans la strate h
  - $X_h^* = (\frac{N_h}{n_h} - 1)(X_h - \mu_h)$
  - J_h^* indicatrice que l'unité est winsorizée $(X > K_h)$, aussi égale à l'indicatrice que X_h^* dépasse $-B$

Mise en place de Kokic et Bell III

- $E_h(J_h^*)$ probabilité qu'une unité de la strate soit winsorizée
- $E_h(J_h^* X_h^*)$ moyenne de la variable égale à X_h^* sur les unités winsorizées et 0 ailleurs

Mise en place Kokic et Bell, en pratique

(Nouvelle) Hypothèse : nous disposons d'observations $\tilde{X}$ de X dans chaque strate **indépendantes de l'échantillon winsorizé**

- estimation de μ_h dans chaque strate par la moyenne empirique des $\tilde{X}$ dans la strate
- calcul des $\tilde{X}_h^*$
- pour chaque valeur possible du biais B, estimation de $E_h(J_h^*)$ par la part des valeurs de $\tilde{X}$ supérieure à $-B$ dans la strate
- pour chaque valeur possible du biais B, estimation de $E_h(X_h^* J_h^*)$ par la moyenne de la variable égale à $\tilde{X}_h^*$ si $\tilde{X}_h^* > -B$ et 0 sinon
- estimation du zéro de la fonction $F \Rightarrow$ estimation du biais optimal B^*
- estimation des seuils optimaux K_h^* par $K_h^* = \hat{\mu}_h - \frac{-\hat{B}^*}{\frac{N_h}{n_h} - 1}$