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

Arnaud Fizzala [1] & Thomas Deroyon [2] & Cyril Favre-Martinoz [3]

[1] *Insee, Direction Générale - 88 avenue Verdier CS 70058 92541 Montrouge Cedex, France - arnaud.fizzala@insee.fr*
[2] *Insee, Direction Générale - 88 avenue Verdier CS 70058 92541 Montrouge Cedex, France - thomas.deroyon@insee.fr*
[3] *Insee, Direction Régionale de la Réunion-Mayotte - 10 rue Demarne - 97743 Saint Denis, France - cyril.favre-martinoz@insee.fr*

Résumé

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. Elles sont de plus en plus souvent traitées à l'Insee par winsorisation. Le but de la winsorisation est de limiter l'impact de ces unités en réduisant leur poids de sondage ou la valeur associée à la (ou les) variable pour laquelle l'unité est influente. Ce faisant, les estimateurs gagnent en stabilité mais un biais est introduit.

En pratique, la winsorisation, appliquée à un échantillon sélectionné par sondage aléatoire simple stratifié, se base sur des seuils définis dans chaque strate de tirage : les valeurs qui dépassent le seuil sont « rognées », et les valeurs se situant sous le seuil sont conservées à l'identique. Le choix de ces seuils est déterminant pour la qualité de la winsorisation. Kokic et Bell ont proposé une méthode pour calculer des seuils optimaux, c'est-à-dire minimisant l'erreur quadratique moyenne anticpée sous un certain modèle de l'estimateur winsorisé. Cependant, ces seuils sont optimaux sous certaines hypothèses parmi lesquelles le fait que le plan de sondage de l'enquête soit aléatoire simple stratifié, ce qui est le cas de la majorité des enquêtes auprès d'entreprises du système statistique public français.

Deux demandes concernant le traitement des unités atypiques représentatives dans des cas de plans de sondage « plus complexes » ont récemment été formulées à la Division Sondages.

L'article décrira plus en détails les simulations réalisées dans le but de répondre à ces demandes, et qui montrent dans les deux cas que les seuils de Kokic et Bell peuvent donner de bons résultats même lorsque le plan de sondage n'est pas aléatoire simple stratifié.

Mots-clés. Estimation robuste, Winsorisation, Enquêtes économiques et dans les entreprises.

Abstract

Representative outliers are population units with a strong impact on the results whether if they are selected or not. The french national institute of statistics (INSEE) apply more and more winsorization to deal with representative outliers in business surveys. Winsorization aim is to limit the impact of the representative outliers by reducing the expansion weight or the value corresponding to the variable for which the unit is influential. Doing so, estimators become more stable but potentially biased.

Winsorization is based on thresholds defined in each drawing stratum. Values above the thresholds are truncated. The choice of the thresholds is important for the quality of the winsorization. Kokic and Bell have proposed optimum thresholds, minimizing the anticipated mean square error of the winsorized estimator. This optimum thresholds are calculated under the assumption that the sampling design is a stratified random sampling.

Most of the french business survey are based on a stratified random sampling, but recently, The INSEE Survey methodology unit has been asked about representative outliers treatments for two surveys not based on stratified random sampling.

The paper will describe the simulations realised in this context. They tend to show that Kokic and Bell thresholds perfom well even if the sampling design is not a stratified random sampling.

Keywords. Robust estimation, winsorization, Business surveys.

1 Résumé long

Dans les enquêtes auprès des entreprises, la présence de points atypiques est quasiment inévitable. Un point atypique est une unité de la population qui, suivant qu'elle appartient ou pas à l'échantillon, change beaucoup le niveau des estimateurs calculés et diffusés. Il s'agit le plus souvent d'entreprises dont le chiffre d'affaires, l'effectif salarié ou tout autre variable dont la distribution est positive et asymétrique, parait élevé par rapport à celles des entreprises avec qui elles devraient être comparables. Le plan de sondage usuellement appliqué pour une enquête auprès des entreprises étant le sondage aléatoire simple stratifié, les points atypiques correspondent ainsi à des entreprises ayant un chiffre

d'affaires ou un effectif très éloigné des valeurs moyennes de leur strate [1]. Face à un tel cas, deux situations menant à des traitements différents peuvent se présenter :

- Il s'agit d'une erreur de saisie : la valeur collectée doit alors être remplacée par la « bonne » valeur s'il est possible de l'obtenir ou à défaut être mise à blanc et traitée comme une non-réponse partielle. On parle alors d'unité atypique non-représentative, et des procédures spécifiques existent pour les repérer et hiérarchiser leur importance (Gros (2012) ; Di Zio et Guarnera (2013)).

- Il ne s'agit pas d'une erreur mais bien de la « bonne » valeur : l'entreprise a pu par exemple connaître une forte croissance pendant les une à deux années séparant la période de référence des informations présentes dans la base de sondage (où l'entreprise était considérée « petite » au sens économique) et la période de référence pour laquelle les réponses à l'enquête sont demandées (où l'entreprise est devenue « importante » au sens économique). L'information de la base de sondage sur la base de laquelle les strates ont été constituées peut également être erronée. On parle alors d'unités atypiques représentatives et ce sont ces unités auxquelles nous nous intéressons dans cet article.

Les unités atypiques représentatives ont longtemps été traitées à l'Insee en ramenant leur poids de sondage à 1 de façon à limiter leur influence, mais depuis environ 10 ans des méthodes plus optimales de winsorisation sont progressivement mises en place (Brion et Guggemos, 2010). Le but de la winsorisation est de limiter l'impact des unités atypiques représentatives en réduisant leur poids de sondage (mais pas forcément jusqu'à 1) ou la valeur associée à la (ou les) variable pour laquelle l'unité est influente, bien que cette valeur soit correcte. Ce faisant, les estimateurs gagnent en stabilité mais un biais est introduit.

En pratique, la winsorisation appliquée à un échantillon sélectionné par sondage aléatoire simple stratifié et pour l'estimation du total d'une variable se base sur des seuils définis dans chaque strate de tirage : les valeurs qui dépassent le seuil sont « rognées » [2] de différentes façons suivant le type de winsorisation appliqué. Plus précisément, la valeur d'une variable Y après winsorisation de type 2, qui est le type de winsorisation le plus employé à l'Insee, est :

$$Y_i^w = \begin{cases} Y_i & \text{si} \quad Y_i < K_h \\ \frac{n_h}{N_h} Y_i + (1 - \frac{n_h}{N_h}) K_h & \text{sinon} \end{cases} \tag{1}$$

1. Le traitement de la non-réponse totale par repondération, lorsque cette dernière est réalisée à un niveau différent des strates de tirage, peut provoquer des écarts de poids entre entreprises d'une même strate de tirage et accentuer alors le caractère atypique de certaines entreprises.

2. Réduites à la valeur du seuil lorsque l'on applique une winsorisation dite de type 1, ou à une valeur comprise entre le seuil et la valeur initiale lorsque l'on applique une winsorisation de type 2.

Avec :
- K_h : le seuil de winsorisation dans la strate h ;
- n_h : le nombre d'unités sélectionnées dans l'échantillon dans la strate h ;
- N_h : le nombre d'unités présentes dans la base de sondage dans la strate h.

L'estimateur winsorisé correspond ensuite simplement à l'estimateur par expansion de la variable winsorisée :

$$\hat{t}_y^w = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} Y_i^w \tag{2}$$

Contrairement à l'estimateur d'Horvitz-Thompson qui est sans biais, l'estimateur winsorisé comporte un biais négatif[3] dans l'estimation du total de la variable d'intérêt. En revanche, sa variance dépend de la variance empirique de la variable Y^w dans chaque strate, qui est par construction plus petite que celle de la variable Y. Ainsi, la winsorisation repose sur un compromis entre biais et variance, qui sera efficace si la perte de variance est plus « avantageuse » que le biais introduit, ce qui plus formellement correspond à une diminution de l'erreur quadratique moyenne de l'estimateur.

On se doute, d'après la formule utilisée pour winsoriser une variable, que le choix des seuils K_h est déterminant pour que la winsorisation soit de bonne qualité. Le choix de ces seuils a été étudié par Kokic et Bell en 1994 (Kokic et Bell, 1994), qui ont proposé une méthode pour calculer des seuils optimaux, c'est-à-dire minimisant l'erreur quadratique moyenne anticpée sous un certain modèle, lorsque l'estimateur à winsoriser est le total d'une variable.

Cette méthode est par exemple appliquée pour l'enquête sectorielle annuelle (Deroyon, 2015), l'une des enquêtes auprès des entreprises les plus importantes du système statistique public français ; elle fait partie des méthodes de référence recommandées par le département des méthodes statistiques de l'Insee pour le traitement des enquêtes, notamment auprès des entreprises, de la statistique publique.

Cependant, les seuils de Kokic et Bell sont optimaux sous certaines hypothèses ; parmi celles-ci, le plan de sondage de l'enquête doit être aléatoire simple stratifié et le paramètre dont les seuils visent à minimiser l'erreur quadratique moyenne d'estimation doit être le total d'une variable. Bien que la majorité des enquêtes du système statistique public français auprès d'entreprises repose sur un tel plan de sondage[4] et visent à estimer des

3. Dans le sens où l'on sous-estime en moyenne le « vrai » total.

4. La coordination générale des échantillons des enquêtes auprès des entreprises mise en place à l'Insee (Guggemos et Sautory, 2012) nécessite que les échantillons des enquêtes concernées soient tirés selon un plan de sondage aléatoire simple stratifié.

totaux, ce n'est pas le cas pour toutes, et l'application à ces enquêtes de la méthode de winsorisation de Kokic et de Bell pose alors problème.

Dans ces 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.

Ce cas s'est présenté récemment pour deux enquêtes conduites par l'Insee.

La première est l'enquête sur le coût de la main d'oeuvre et la structure des salaires. L'Ecmoss est conduite chaque année par l'Insee et sert notamment à comparer les coûts du travail entre les différents pays européens via l'estimation de salaires horaires moyens dans un ensemble de domaines d'intérêt (secteurs d'activité, régions,...). Le plan de sondage de cette enquête comprend deux degrés : un échantillon d'établissements est sélectionné par un sondage aléatoire simple stratifié ; puis chacun de ces établissements est interrogé sur un échantillon de salariés sélectionné par sondage aléatoire simple stratifié par catégorie sociale dans chaque établissement. Par ailleurs, le paramètre d'intérêt est obtenu comme le ratio des estimateurs de deux totaux (le salaire total divisé par le nombre total d'heures travaillées).

La seconde est l'enquête sectorielle annuelle, dont le plan de sondage a évolué en 2016. Pour permettre la diffusion d'indicateurs sur les entreprises, définies comme des regroupements d'unités légales autonomes économiquement, un échantillon d'entreprises est sélectionné par sondage aléatoire simple stratifié, puis toutes les unités légales de ces entreprises sont interrogées. L'échantillon disponible pour calculer des estimateurs sur la population des unités légales est donc sélectionné par un sondage par grappes.

Pour ces deux enquêtes, nous avons été amenés à proposer des adaptations simples de la méthode de Kokic et Bell : nous avons calculé dans les deux cas les seuils de Kokic et Bell « comme si » l'échantillon de salariés ou d'unités légales avait été sélectionné par un sondage aléatoire simple stratifié dans des strates définies en fonction des domaines d'intérêt de l'enquête. Pour appliquer la méthode à l'Ecmoss, où le paramètre d'intérêt est un ratio, nous avons été par ailleurs amenés à travailler sur la variable linéarisée du ratio.

Ces adaptations reposent sur des simplifications fortes, qui conduisent notamment à négliger les différences de poids de sondage existant dans les « strates » utilisées pour calculer les seuils de winsorisation. Des exercices de simulation nous ont cependant permis de vérifier que les seuils obtenus permettaient des gains sensibles de précision, même si la méthode était appliquée en dehors de ses conditions de validité. De même, les si-

mulations nous ont permis de constater que la méthode de Kokic et Bell conservait des performances équivalentes ou meilleures aux méthodes de biais conditionnel en termes de réduction de l'erreur quadratique moyenne, même si ces dernières sont plus à même de tenir compte du plan de sondage par lequel est sélectionné l'échantillon. Cependant, dans toutes les simulations réalisées, la base d'apprentissage correspond à la base de sondage ce qui favorise « artificiellement » la méthode de winsorisation.

La présentation décrira plus en détails les simulations réalisées et les résultats obtenus.

Bibliographie

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.

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.

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.

Di Zio M., Guarnera U. (2013), *A contamination model for selective editing*, Journal of Official Statistics, Vol. 29, n. 4, pp. 539–555.