Titre : **Traitement des observations atypiques d'une enquête par winsorization** - Application aux Enquêtes Sectorielles Annuelles

Auteur : Thomas Deroyon (Insee - DMCSI) thomas.deroyon@insee.fr

Thème : Sondages, Traitement des valeurs influentes

Une observation atypique mais non aberrante (ou *representative outlier*) est une unité de l'échantillon d'une enquête dont les réponses sont très différentes de celles des observations qui lui sont proches. Par exemple, dans un sondage stratifié avec sondage aléatoire simple dans chaque strate, une unité atypique aura des réponses très supérieures aux autres unités interrogées de la strate, alors qu'elles ont toutes le même poids de sondage. Ce concept s'applique surtout pour les enquêtes dans lesquelles des variables quantitatives sont mesurées, et notamment aux enquêtes auprès des entreprises dans lesquelles elles doivent renseigner des grandeurs comme leur chiffre d'affaires.

C'est le cas par exemple de l'Enquête Annuelle de Production (EAP) pour les entreprises de l'industrie hors agroalimentaire en métropole et des Enquêtes Sectorielles Annuelles (ESA) sur les autres secteurs et les DOM, qui sont conduites chaque année par l'Insee pour mesurer, sur un échantillon d'entreprises, la ventilation de leur chiffre d'affaires sur leurs différentes activités. Elles complètent les déclarations fiscales et permettent ainsi l'estimation des statistiques structurelles d'entreprise dans le cadre d'Esane.

Les observations atypiques nuisent à la robustesse et à la précision des estimations : suivant qu'elles sont échantillonnées ou pas, les estimateurs issus de l'enquête peuvent prendre des valeurs très différentes.

La winsorization, que nous présentons dans la première partie de l'article, est une technique de traitement des observations atypiques. Elle vise à raboter les valeurs extrêmes sur les variables d'intérêt pour les observations de l'enquête. La distribution de la variable après winsorization est ainsi moins dispersée, si bien que ses caractéristiques - moyenne, total…- peuvent être estimées avec une plus grande précision. L'ajustement qu'entraîne la winsorization doit cependant être calibré de manière que le gain en variance qu'il permet fasse plus que compenser le biais qu'il introduit.
Kokic et Bell ont proposé une méthode pour parvenir à un ajustement quasi optimal. Elle consiste, dans le cas d'un sondage stratifié, à choisir une variable d'intérêt et à définir un seuil dans chaque strate de tirage. Si la valeur déclarée par une unité interrogée dépasse ce seuil, sa réponse est rabotée : seule une fraction de la valeur dépassant le seuil est prise en compte dans le montant de la variable winsorizée, cette fraction étant égale au taux de sondage dans la strate. Kokic et Bell proposent une méthode de calcul des seuils qui s'approche d'un arbitrage optimal entre biais et variance.

Depuis 2009, la winsorization suivant la méthode de Kokic et Bell est appliquée aux données des ESA et de l'EAP. Nous présentons plus en détail ces enquêtes et la winsorization qui leur est appliquée dans la deuxième partie de l'article. Les seuils de winsorization ont en particulier été déterminés à partir des données des « ancêtres » des ESA, les Enquêtes Annuelles d'Entreprise (EAE) de 2007 et n'ont pas été renouvelés depuis, si bien que de plus en plus d'entreprises sont traitées chaque année, pour un impact croissant de la winsorization.

Aussi, nous avons été amenés à mettre en place une procédure d'actualisation des seuils, que nous présentons dans une troisième partie. Mise en place à l'automne 2014, cette procédure a permis de n'appliquer la winsorization qu'à un nombre plus faible d'entreprises, pour un impact total du même ordre de grandeur qu'avec les seuils issus des EAE 2007 mais plus stable d'une année sur l'autre.

Bibliographique :

- P. Brion, F. Guggemos : *Du bon usage de la winsorization…ou comment traiter les entreprises atypiques dans les enquêtes sectorielles annuelles auprès des entreprises*, Lettre du SSE (2010)
- P. Kokic, P. Bell : *Optimal winsorizing cutoffs for a stratified finite population estimator*, Journal, of Official Statistics (1994)