Extension de la méthode de Kokic et Bell au plan de sondage poissonnien

Thomas Deroyon [1], Cyril Favre-Martinoz[2]

Résumé

Dans le cadre du traitement des valeurs atypiques ou influentes, les méthodes de winsorisation souvent très souvent utilisées. Cette méthode nécessite de déterminer un seuil au delà duquel les unités sont ramenées. En pratique, la méthode de Kokic et Bell (1994) est une méthode très souvent mobilisée. Cette méthode a été développée dans le cas d'un sondage aléatoire simple stratifié et sous une hypothèse modèle particulière. L'objectif est d'étendre dans le cas du plan de sondage poissonnien la méthode de Kokic et Bell en présentant les hypothèses modèles associées. Des résultats par simulation seront présentés pour évaluer les performances de cette nouvelle méthode et sa robustesse à une mauvaise spécification du modèle. Cet estimateur sera également comparé empiriquement à l'estimateur robuste basé sur le biais conditionnel proposé par Beaumont et al.(2013).

Mots-clés. Estimation robuste, estimateur winsorisé, unités influentes, biais conditionnel.

Abstract

In presence of influential units, winsorization is often used to treat this problem. This technique requires the determination of a constant that corresponds to the threshold above which large values are reduced. The Kokic and Bell method is often used to choose this threshold. This method consist in determining the optimal value of the constant by setting up a common mean model in each stratum and minimizing the winsorized estimator ?s mean square error with respect to both relative to the model and the sampling design. This article deals with the extension of the Kokic and Bell method to poisson sampling design and its related model. We provide simulation studies to test the robustness of this estimator in case of model misspecification and to compare its performance with the estimator based on conditionnal bias proposed by Beaumont et al.(2013).

Key Words. Conditional bias, Robust estimation, Winsorized estimator, Influential values.

[1]thomas.deroyon@insee.fr, Insee, Paris, France.

[2]cyril.favre-martinoz@insee.fr, Insee, Saint-Denis de la Réunion, France.

1 Introduction

En statistique d'enquête, une unité de la population est influente si les estimateurs obtenus sur un échantillon tiré dans cette population changent beaucoup suivant que cette unité est échantillonnée ou pas. La notion d'unité influente est donc dépendante de plusieurs facteurs, qui déterminent ce qu'on appelle une configuration :

- un plan de sondage sur une population ;
- une ou plusieurs variables d'intérêt et un paramètre d'intérêt sur la distribution de cette variable
- un estimateur calculé sur l'échantillon pour ce paramètre d'intérêt.

Une unité peut être influente dans une configuration donnée, et pas dans une autre. Par exemple, elle peut avoir un effet important sur l'estimateur du total d'une variable dans un domaine particulier, mais n'avoir qu'une influence négligeable sur l'estimateur du total de cette même variable dans la population totale.

Les estimateurs classiques (estimateur par dilatation, estimateur ajusté pour la non-réponse totale) ne présentent pas (ou asymptotiquement) de biais mais peuvent être très instables en présence de valeurs influentes. Des méthodes d'estimation robuste peuvent alors être mises en œuvre afin de limiter leur impact. Le principe de ces méthodes est de modifier les poids d'estimation ou les valeurs déclarées par les unités influentes, de façon à rendre les estimateurs plus stables, au risque de les biaiser. Plus précisément, les estimateurs auxquels conduisent ces méthodes doivent avoir une erreur quadratique moyenne significativement plus faible que celle des estimateurs par expansion classiques en présence de données influentes, sans perdre trop en efficacité en l'absence de valeurs atypiques dans l'échantillon. Le traitement des valeurs influentes réside donc dans un compromis entre le biais et la variance.

La méthode la plus souvent employée en pratique pour traiter le problème des valeurs influentes est la winsorisation, qui s'applique à l'estimation de totaux de variables d'intérêt. Dans l'application de la winsorisation, le choix des seuils est crucial, un mauvais choix pouvant conduire à des estimateurs winsorisés ayant une erreur quadratique moyenne supérieure à celle des estimateurs classiques. Le choix de ces seuils a fait l'objet de nombreuses études, entre autres par Kokic et Bell (1994), Rivest et Hurtubise (1995) et Favre-Martinoz et al. (2015).

Dans cet article, nous présentons l'extension de la méthode de Kokic et Bell dans le cas poissonnien. D'autres méthodes ont été proposées pour identifier et traiter les unités influentes en statistique d'enquêtes. L'une d'entre elles, introduite par Beaumont et al. (2013), s'appuie sur la notion de biais conditionnel, une mesure

d'influence proposée par Moreno-Rebollo et al. (1999) et Moreno-Rebollo et al. (2002). Contrairement aux méthodes de winsorisation évoquées *supra*, qui ne sont adaptées qu'à certains plans de sondage et nécessitent une information extérieure à l'échantillon assez riche, la méthode proposée par Beaumont et al. (2013) peut s'appliquer *a priori* à n'importe quel plan de sondage et ne mobilise que les réponses à l'enquête.

L'objectif de cet article est de comparer l'efficacité des méthodes de winsorisation et de biais conditionnel pour le traitement des valeurs influentes dans le cas d'un plan poissonnien. Pour ce faire, nous rappelons dans la section 2 les grands principes de la méthode de winsorisation. Dans la section 3, nous proposons une extension de la méthode de Kokic et Bell dans le cas d'un plan de sondage poissonnien. Enfin, nous présentons dans la section 4 des simulations afin de comparer l'extension dans le cas poissonnien de la méthode de Kokic et Bell avec les méthodes de biais conditionnels.

2 Notation et principe de la winsorisation

Soit une population U de taille N et une variable d'intérêt X observée sur un échantillon S de taille espérée n et dont on cherche à estimer le total $T(X) = \sum_{i \in U} X_i$ sur la population. On suppose de plus que :

- X est une variable positive ou nulle ;
- S est sélectionné suivant un plan de sondage poissonnien P ;
- l'on peut partitionner l'échantillon en parties S_h, $h = 1, .., H$ auxquelles nous allons associer un seuil K_h, $h = 1, .., H$.

On définit ensuite :

- une variable winsorisée $\tilde{X}$ par

$$\tilde{X}_{h,i} = \begin{cases} X_{h,i} & \text{si } d_{h,i} X_{h,i} \leq K_h \\ \frac{X_{h,i}}{d_{h,i}} + \left(1 - \frac{1}{d_{h,i}}\right) \frac{K_h}{d_{h,i}} & \text{si } d_{h,i} X_{h,i} > K_h, \end{cases} \tag{2.1}$$

  où $d_{h,i} = \frac{1}{\pi_i}$ est le poids de l'unité i dans la partie h ;

- l'estimateur winsorisé du total de X par :

$$\hat{T}(\tilde{X}) = \sum_{h=1}^{H} \sum_{i \in S_h} d_{hi} \, \tilde{X}_{hi}. \tag{2.2}$$

Le choix des seuils K_h mobilisés pour la winsorisation est crucial. Dans cet article, nous comparons deux méthodes pour déterminer ces seuils : une extension de la méthode de Kokic et Bell et la méthode de Favre-Martinoz et al.(2015) dont les seuils permettent de retrouver l'estimateur basé sur les biais conditionnels de Beaumont et al.(2013).

3 Extension de la méthode de Kokic et Bell dans le cas poissonnien pour déterminer les seuils

Nous allons supposer qu'il est possible de partitionner la population et l'échantillon en sous-populations U_h et S_h dans lesquelles toutes les valeurs $d_{hi}X_{hi}$ sont des réalisations indépendantes issues d'un même modèle vérifiant :

$$\forall h = 1, .., H, \forall i \in U_h, \, d_{hi}\, X_{hi} \, = \, \mu_h \, + \, \epsilon_{hi}, \tag{3.1}$$

$$\text{avec } \left\{ \begin{array}{ll} E_m(\epsilon_{hi}) & = 0 \\ V_m(\epsilon_{hi}) & = \sigma_h^2 \, < \, +\infty \end{array} \right.$$

en notant E_m et V_m l'espérance et la variance sous le modèle (3.1).
L'hypothèse forte sous-jacente à ce modèle est que les espérances des produits des X_{hi} par les poids d_{hi} sont supposées constantes en espérance dans chaque sous-population. Cela revient à dire que les probabilités d'inclusion au sein de chaque strate sont définies proportionnellement à la variable d'intérêt X. En pratique, ces probabilités d'inclusion sont définies proportionnellement à une variable auxiliaire connue et fortement corrélée à X, ce qui permet d'être proche de l'hypothèse sous jacente au modèle (3.1).
Dans la suite, les variables aléatoires $d_{hi}\, X_{hi}$ étant supposées indépendantes et identiquement distribuées au sein de chaque strate, on ne considère qu'une seule variable aléatoire Z_h suivant le même loi qu'un des $d_{hi}\, X_{hi}$.

Nous nous plaçons de plus dans le même cadre asymptotique que Kokic et Bell(1994) :

- $N_\nu, \, n_\nu \underset{\nu \to +\infty}{\to} +\infty$, où ν désigne ici l'indice des populations imbriquées dont la taille tend vers l'infinie, cette indice sera omis par la suite pour faciliter la lecture ;
- le nombre de strate H est fixe.

et adaptant l'hypothèse portant sur les probabilités d'inclusion :

$$\forall \, h = 1..H, \forall i \in U_h \, \min(\pi_i) > \lambda_{1h} > 0 \text{ et } \max(\pi_i) < \lambda_{2h} < 1 \tag{3.2}$$

Les seuils K_h sont déterminés de manière à minimiser l'erreur quadratique moyenne de l'estimateur winsorisé $\hat{T}(\tilde{X})$ sous le modèle de la variable X et sous le plan

de sondage P, *i.e.* en moyenne sur l'ensemble des populations possibles compte-tenu du modèle de super-population posé sur X et en moyenne sur l'ensemble des échantillons tirés dans ces populations compte-tenu du plan de sondage P :

$$(K^\star)_{h=1..H} \in Argmin_{(K_h)_{h=1..H}} E_m E_P [(\hat{T}(\tilde{X} - T(X)))^2]$$

Il est possible de montrer qu'à l'optimum et asymptotiquement, en notant $J_h = \mathbb{I}(Z_h > K_h)$:

$$\forall\, h = 1..H,\ K_h \sim -\frac{A_h}{C_h + D_h} B \tag{3.3}$$

$$\text{avec} \begin{cases} A_h &= \sum_{i \in U_h} \frac{1}{d_{hi}} \left(1 - \frac{1}{d_{hi}}\right) \\ C_h &= \sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)^2 \left(1 - \frac{1}{d_{hi}}\right)^2 \\ D_h &= \sum_{i \in U_h} \frac{1}{d_{hi}} \left(1 - \frac{1}{d_{hi}}\right)^3 \end{cases}$$

$$\text{et } B = \sum_{h=1}^{H} A_h \left[K_h E_m(J_h) - E_m(J_h Z_h)\right] \tag{3.4}$$

B est le biais de l'estimateur winsorisé optimal $\hat{T}(\tilde{X})$. A l'optimum, le seuil K_h est donc égal à un terme positif près, à l'opposé du biais multiplié par le terme $\frac{A_h}{C_h + D_h}$. Quand le taux de sondage moyen est faible dans la sous-population U_h, et les poids de sondage d_{hi} élevés, le terme $\frac{A_h}{C_h + D_h}$ tend vers 0. Ainsi, si très peu d'unités sont tirées, il peut être utile d'en winsoriser beaucoup, car les valeurs élevées des $d_{hi} X_{hi}$, ont une grande incidence sur les estimateurs.
Quand le taux de sondage est à l'inverse proche de 1, le terme $\frac{A_h}{C_h + D_h}$ tend vers l'infini : quand le sondage est proche d'être exhaustif, seules les valeurs très atypiques des $d_{hi} X_{hi}$ méritent d'être winsorisées, car elles seules sont susceptibles d'avoir un effet important sur $\hat{T}(X) = \sum_{h=1}^{H} \sum_{i \in S_h} \frac{X_{hi}}{d_{hi}}$ suivant qu'elles sont échantillonnées ou pas.

Si nous notons $L = -B$ et $\frac{C_h + D_h}{A_h} Z_h = X_h^\star$, alors asymptotiquement $J_h = J_h^\star = \mathbb{I}(X_h^\star > L)$ en utilisant la relation 3.3.

En injectant la relation d'équivalence 3.3 dans la formule 3.4 définissant B, nous obtenons qu'à l'optimum et asymptotiquement, B est l'opposé du point d'annulation de la fonction F définie par

$$F(L) = L \left(1 + \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} E_m(J_h^\star)\right) - \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} E_m(J_h^\star\ X_h^\star) \tag{3.5}$$

Les espérances $E_m(J_h^\star)$ et $E_m(J_h^\star\ X_h^\star)$ sont inconnues, mais elles peuvent être estimées sous réserve de disposer par exemple de données historiques non issues

de l'échantillon mobilisé pour ce traitement. Pour chaque sous-population h, nous notons $\breve{X}_{hi}$ les p_h réalisations tirées selon la loi de X_h et indépendantes de l'échantillon S. Avec ces observations, nous pouvons estimer F par

$$\hat{F}(L) = L\left(1 + \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} \frac{\sum_{i=1}^{p_h} \mathbb{I}(\breve{X}_{hi}^{\star} > L)}{p_h}\right) - \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} \frac{\sum_{i=1}^{p_h} \breve{X}_{hi}^{\star} \mathbb{I}(\breve{X}_{hi}^{\star} > L)}{p_h} \tag{3.6}$$

et estimer B par $\hat{B}$, l'opposé du point d'annulation de $\hat{F}$.

4 Quelques résultats par simulation

Nous avons effectué une étude par simulations afin d'étudier les propriétés l'estimateur issu de l'extension de Kokic et Bell dans le cas poissonnien et l'estimateur de Beaumont, Haziza et Ruiz-Gazen (2013). Nous avons réalisé quatre scénarios pour comparer l'efficacité des deux estimateurs, mais aussi pour étudier dans le cas de l'estimateur de Kokic et Bell, la robustesse à une mauvaise spécification du modèle, *i.e.* à une modification entre le modèle d'apprentissage et le modèle ayant généré les données de l'échantillon.

La simulation se déroule de la façon suivante :

- nous considérons $L = 1000$ réalisations d'un certain modèle qui permet de générer notre base d'apprentissage de $N = 5000$ unités ;
- pour chacune de ces réalisations, nous calculons le seuil optimal K_l selon la méthode proposée à la section précédente ;
- puis nous créons $M = 10000$ bases de sondages de test générées selon un (autre) modèle sur lesquelles nous sélectionnons un échantillon suivant un tirage poissonnien et calculons l'estimateur robuste $\hat{\theta}_{(m)}$ avec le seuil K_l calculé. En guise de comparaison, nous calculons également l'estimateur robuste issu de la méthode basée sur le biais conditionnel.

Les probabilités d'inclusion, ainsi que les valeurs de la variable X ont été générées selon le modèle suivant :

$$U_i \sim \mathcal{L}\text{og-}\mathcal{N}(1, 1.1)$$

$$\pi_i = n \times \frac{U_i}{\sum_{i=1}^{N} U_i}$$

$$X_i = 2000 \times \pi_i + \pi_i \epsilon_i + \delta_i V_i,$$

$$\epsilon_i \sim \mathcal{N}(0, 100), \delta_i \sim \mathcal{B}(\omega)$$

où ω est le paramètre de la Bernoulli, reflétant la proportion de valeurs influentes et V_i est une variable aléatoire dont la distribution est décrite dans le Tableau 1. La notation $\mathcal{L}$og-$\mathcal{N}$ désigne une distribution log-normale.

	Modèle d'apprentissage		Modèle de test	
Scénario	ω	Loi de V_i	ω	Loi de V_i
2	0.01	$\mathcal{L}$og-$\mathcal{N}(log(500), 1.2)$	0.01	$\mathcal{L}$og-$\mathcal{N}(log(500), 1.2)$
3	0.01	$\mathcal{L}$og-$\mathcal{N}(log(500), 1.2)$	0.1	$\mathcal{L}$og-$\mathcal{N}(log(500), 1.2)$
4	0.1	$\mathcal{L}$og-$\mathcal{N}(log(500), 1.2)$	0.01	$\mathcal{L}$og-$\mathcal{N}(log(500), 1.2)$

TABLE 1: Modèles et paramètres utilisés afin de générer les populations

Comme mesure du biais d'un estimateur $\hat{\theta}$, nous avons calculé le biais relatif Monte Carlo (en %) :

$$BR_{MC}(\hat{\theta}) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - t\right)}{t} \times 100,$$

où $\hat{\theta}_{(m)}$ désigne l'estimateur $\hat{\theta}$ dans l'échantillon m, $m = 1, \ldots, M$.

Nous avons également calculé l'efficacité relative des estimateurs robustes relativement à l'estimateur par dilatation, $\hat{t}$:

$$RE_{MC}(\hat{\theta}) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - t\right)^2}{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{t}_{(m)} - t\right)^2} \times 100.$$

Les tableaux 2 et 3 représentent les statistiques descriptives associées au $L = 1000$ valeurs Monte-Carlo calculées selon la population d'apprentissage considérée.

Statistique	Scénario							
	1				2			
descriptive	K&B		BHR		K&B		BHR	
	BR	RE	BR	RE	BR	RE	BR	RE
min	−0.2	100	−0.43	100	−9.0	1	−4.3	26
$Q1$	−0.1	100	−0.32	100	−2.9	35	−1.9	51
Médiane	0.0	100	−0.27	100	−1.8	50	−1.5	62
Moyenne	0.0	100	−0.27	100	−2.0	50	−1.6	62
$Q3$	0.0	100	−0.23	100	−1.0	64	−1.3	73
max	0.0	100	−0.14	100	−0.1	109	−0.6	91

TABLE 2: Statistiques descriptives pour les scénarios 1 et 2 sur les 1000 simulations pour $n = 500$

Le scénario 1 correspond à une situation où aucune ou très peu d'unités influentes sont présentes dans la population : la performance des estimateurs robustes est alors identique à celle de l'estimateur d'Horvitz-Thompson usuel, avec un biais relatif très proche de 0. Le scénario 2 correspond au cadre pour lequel a été développé l'extension de la méthode de Kokic et Bell au cas poissonnien, avec introduction d'unités influentes. Les deux estimateurs robustes sont plus efficaces que l'estimateur usuel, mais la performance de l'estimateur de Kokic et Bell en termes de gain d'erreur quadratique moyenne est supérieure, avec une efficacité relative médiane sur les 1000 simulations de 50 % contre 62 % pour la méthode du biais conditionnel. Ce résultat est normal, compte-tenu du fait que le seuil de la méthode Kokic et Bell est déterminé explicitement de manière à obtenir l'estimateur ayant la plus faible erreur quadratique moyenne.

Statistique	Scénario							
	3				4			
descriptive	K&B		BHR		K&B		BHR	
	BR	RE	BR	RE	BR	RE	BR	RE
min	−32.2	2	−7.8	27	−4.5	1	−4.3	26
$Q1$	−18.9	50	−5.1	59	−1.8	48	−1.9	51
Médiane	−13.9	82	−4.6	66	−1.5	70	−1.5	62
Moyenne	−14.2	89	−4.7	65	−1.5	68	−1.6	62
$Q3$	−9.3	138	−4.2	72	−1.2	91	−1.3	73
max	−0.01	537	−2.7	89	−0.6	100	−0.6	91

TABLE 3: Statistiques descriptives pour les scénarios 3 et 4 sur les 1000 simulations pour $n = 500$

Les performances des deux méthodes dans le scénario 3 sont plus contrastées. Alors que sur l'ensemble des simulations, la méthode du biais conditionnel parvient à réduire l'erreur quadratique moyenne des estimateurs, avec un gain minimal de 27 % d'erreur quadratique moyenne, la méthode de Kokic et Bell détériore la précision dans plus du quart des cas. La population sur laquelle a été calculé le seuil contient, dans ce scénario, trop peu d'unités influentes par rapport à l'échantillon pour que le seuil calculé soit efficace.
Dans le scénario 4, où la population d'apprentissage contient plus d'unités influentes que l'échantillon, les performances des deux méthodes sont du même ordre de grandeur.

Références

[1] J.F. Beaumont, D. Haziza, A. Ruiz-Gazen, *A unified approach to robust estimation in finite population sampling*, Biometrika, vol. 100, p. 555 - 569, 2013

[2] C. Favre-Martinoz, D. Haziza, J.F. Beaumont, *A method for determining the cut-off points for winsorized estimators with application to domain estimation*, Techniques d'Enquête, vol.41, p.51 - 77, 2015

[3] P.N. Kokic, P.A. Bell, *Optimal winsorizing cut-offs for a stratified finite population estimation*, Journal of Official Statistics, vol.10-4, p.419 - 435, 1994

[4] J.L. Moreno-Rebollo, A.M. Muñoz-Reyez, J.M. Muñoz-Pichardo, *Influence diagnostics in survey sampling : conditional bias*, Biometrika, vol.86, p.923 - 968, 1999

[5] J.L. Moreno-Rebollo, A.M. Muñoz-Reyez, J.L. Jimenez-Gamero, J.M. Muñoz-Pichardo, *Influence diagnostics in survey sampling : estimating the conditional bias*, Metrika, vol.55, p.209 - 214, 2002

[6] L.P. Rivest, D. Hurtubise, *On Searl's winsorized mean for skewed populations*, Techniques d'Enquête, vol.21-2, p.107-116, 1995