# Estimation robuste pour des populations asymétriques

Cyril Favre Martinoz[1], Jean-François Beaumont[2] et David Haziza [3]

14 février 2014

[1] *Laboratoire de Statistique d'Enquête, Crest/Ensai, Campus de Ker Lann, 35170 Bruz, France,* cyril.favremartinoz@ensai.fr

[2] *Division de la recherche et de l'innovation en statistique, Statistique Canada,Tunney's Pasture, R.H. Coats Building, 16th floor, K1A 0T6, Ottawa, Canada.* jean-francois.beaumont@statcan.gc.ca

[3] *Département de mathématiques et de statistique, Université de Montréal, Montréal, Canada, H3C 3J7* David.Haziza@umontreal.ca

**Résumé**

L'estimation de la moyenne dans le cas d'une population asymétrique est un problème très important en pratique. En effet, il est très courant d'observer des variables dont la distribution est asymétrique, c'est le cas par exemple du chiffre d'affaire des entreprises ou le revenu des ménages. En pratique, l'échantillon d'observation possèdent des unités qui sont très influentes sur la moyenne empirique, qui est l'estimateur souvent privilégié.

Rivest (1994) propose un estimateur non paramétrique pour la moyenne d'une population asymétrique en winsorisant la plus grande ou les deux plus grandes observations de l'échantillon, il montre que cet estimateur possède de bonnes propriétés en terme d'erreur quadratique moyenne. Sa stratégie consiste à réduire voire supprimer l'influence des plus grandes valeurs de l'échantillon.

Notre démarche consiste à quantifier l'influence des unités de l'échantillon et de construire un estimateur robuste en réduisant l'impact des unités influentes. Pour cela, nous allons utiliser le biais conditionnel comme mesure d'influence. Nous donnerons les propriétés de cet estimateur en terme d'erreur quadratique moyenne et nous développerons une approximation de cette erreur quadratique moyenne suivant les différents domaines d'attraction possibles pour la loi considérée et nous effectuerons une étude par simulation pour comparer les performances de cet estimateur avec celui proposé par Rivest (1994).

*Mots clés :* Théorie des valeurs extrêmes, domaines d'attraction, statistiques d'ordre, distribution de Pareto, biais conditionnel

The purpose of this work is to estimate in a robust way the mean of a skewed distribution. To construct the robust estimator, we derive the conditional bias "under the model" of the empirical mean, and then we propose an estimator to achieve the compromise between bias and variance. We derive some approximations to the mean square using the theory of extreme values. For each max-domain of attraction (Frechet, Weibull, Gumbel), we derive an approximation of the mean square error of the robust estimator. We have shown for the special case of normal distribution, that the robust estimator is as efficient as the empirical mean. We propose a simulation study to compare the efficiency of the robust estimator versus the winsorized mean proposed by Rivest (1994) and the maximum likelihood estimator for different distributions in order to test the efficiency of these estimators in the case of model misspecification.

*Key words:* Extreme value theory; Max-domain of attraction; Orders statistics; Pareto distribution, Conditionnal bias

On observe un échantillon $(X_1, ...X_n)$ de $n$ variables aléatoires indépendantes et identiquement distribuées selon une loi continue de fonction de répartition $F$ définie sur $\mathbb{R}$. On suppose que la loi $F$ possède une moyenne finie $\mu$ et une variance finie $\sigma^2$. On note $(X_{(1)}, .., X_{(n)})$ l'échantillon des statistiques d'ordre et $\mu_i$ et $\mu_{ij}$ désignent les moments et les moments croisés de la statistique d'ordre $i$ : $\mu_i = \mathbb{E}(X_{(i)})$ et $\mu_{ij} = \mathbb{E}(X_{(i)}X_{(j)})$. On désigne par $\bar{X}$ la moyenne de l'échantillon.

Nous allons dans un premier temps quantifier l'influence de chacune des unités de l'échantillon à l'aide du biais conditionnel introduit dans le cas d'une population finie par Moreno-Rebello et al.[3] (1999).

On définit le biais conditionnel de l'unité $i$ associé à l'estimateur $\hat{\mu}$ du paramètre $\mu$ par :

$$B_i(\hat{\mu}) = \mathbb{E}(\hat{\mu}|X_i) - \mu$$

On notera $b_i$, la réalisation de $B_i$ conditionnellement à $X_i = x_i$. La réalisation du biais conditionnel $b_i = \frac{1}{n}(x_i - \mu)$ est inconnu, il est nécessaire de l'estimer.

Un estimateur conditionnellement sans biais du biais conditionnel $b_i$ est donné par :

$$\hat{B}_i(\bar{X}) = \frac{1}{n}\left(X_i - \bar{X}_{(i)}\right)$$

où

$$\bar{X}_{(i)} = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} X_j.$$

En utilisant le biais conditionnel, on construit un estimateur robuste $\bar{X}^R$, en reprenant la démarche de Beaumont et al.[1] (2013) :

$$\bar{X}^R = \bar{X} - \sum_{j=1}^{n} B_j(\bar{X}) + \sum_{j=1}^{n} \psi\left(B_j(\bar{X}), c\right)$$

où $\psi$ désigne la fonction de Huber.

On va maintenant déterminer le biais conditionnel associé à l'estimateur robuste $\bar{X}^R$ :

$$\begin{aligned}
b_i\left(\bar{X}^R, c\right) =& \mathbb{E}(\bar{X}^R|X_i = x_i) - \mu \\
=& \mathbb{E}(\bar{X} + \sum_{j=1}^{n}\left[\psi\left(B_j(\bar{X}), c\right) - B_j(\bar{X})\right]|X_i = x_i) - \mu \\
=& \mathbb{E}(\bar{X} + \sum_{j=1}^{n}\left[\psi\left(B_j(\bar{X}), c\right) - B_j(\bar{X})\right]|X_i = x_i) - \mu \\
=& \mathbb{E}(\bar{X}|X_i = x_i) + \mathbb{E}\left(\sum_{j=1}^{n}\left[\psi\left(B_j(\bar{X}), c\right) - B_j(\bar{X})\right]|X_i = x_i\right) - \mu \\
=& b_i(\bar{X}) + \psi\left(b_i(\bar{X}), c\right) - b_i(\bar{X}) + \mathbb{E}\left(\sum_{j=1, j\neq i}^{n}\left[\psi\left(B_j(\bar{X}), c\right) - B_j(\bar{X})\right]\right) \\
=& \psi\left(b_i(\bar{X}), c\right) + \mathbb{E}\left(\sum_{j=1, j\neq i}^{n}\left[\psi\left(B_j(\bar{X}), c\right) - B_j(\bar{X})\right]\right).
\end{aligned}$$

Ce biais conditionnel est inconnu, on peut l'estimer par :

$$\hat{B}_i(\bar{X}^R, c) = \psi\left(\hat{B}_i(\bar{X}), c\right) + \sum_{j=1, j\neq i}^{n}\left[\psi\left(\hat{B}_j(\bar{X}), c\right) - \hat{B}_j(\bar{X})\right] = \hat{B}_i(\bar{X}) + \sum_{j=1}^{n}\left[\psi\left(\hat{B}_j(\bar{X}), c\right) - \hat{B}_j(\bar{X})\right].$$

Ce biais conditionnel estimé associé à l'estimateur robuste peut se réécrire :

$$\hat{B}_i(\bar{X}^R, c) = \hat{B}_i(\bar{X}) + n\bar{\Delta}(c)$$

où

$$\bar{\Delta}(c) = \frac{1}{n}\sum_{j=1}^{n}\left[\psi\left(\hat{B}_j(\bar{X}), c\right) - \hat{B}_j(\bar{X})\right].$$

On choisit la constante $c$, qui permet de réduire le maximum des influences calculées sur l'estimateur robuste. Autrement dit on choisit la constante $c_{minmax}$ qui minimise le maximum des biais conditionnels calculés sur l'estimateur robuste $\bar{X}^R$ :

$$\begin{aligned} c_{minmax} &= \arg\min_{c\in\mathbb{R}}\left(max\left[|\hat{B}_i(\bar{X}^R, c)|, i \in [1,n]\right]\right) \\ &= \arg\min_{c\in\mathbb{R}}\left(max\left[|\hat{B}_i(\bar{X}) + n\bar{\Delta}(c)|, i \in [1,n]\right]\right). \end{aligned}$$

En résolvant le problème de minimisation, on obtient :

$$n\bar{\Delta}(c_{minmax}) = \frac{1}{2}\left[\min\left(\hat{B}_i(\bar{X})\right) + \max\left(\hat{B}_i(\bar{X})\right)\right].$$

L'estimateur robuste pris en $c = c_{minmax}$ est égale à :

$$\bar{X}^R = \bar{X} - \frac{1}{2}\left[\min\left(\hat{B}_i(\bar{X})\right) + \max\left(\hat{B}_i(\bar{X})\right)\right].$$

En remplaçant $\hat{B}_i(\bar{X})$ par son expression, on obtient :

$$\bar{X}^R = \frac{n}{n-1}\bar{X} - \frac{1}{2(n-1)}\left[X_{(1)} + X_{(n)}\right].$$

Il s'agit maintenant de déterminer les propriétés de cet estimateur.
Le biais de l'estimateur robuste est donnée par :

$$Biais\left(\bar{X}^R\right) = \frac{\mu}{n-1} - \frac{1}{2(n-1)}\left(\mu_1 + \mu_n\right).$$

Remarque : Pour une loi symétrique, on a $\frac{\mu_1+\mu_n}{2} = \mu$, donc $Biais\left(\bar{X}^R\right) = 0$.
On peut montrer que l'estimateur robuste est consistant pour le paramètre $\mu$ à l'aide d'une inégalité proposée par David et al.[2](1981) sur l'espérance du maximum d'un $n$-échantillon issu d'une loi possédant au moins un moment d'ordre deux.
On peut montrer que l'erreur quadratique moyenne de l'estimateur robuste est donnée par :

$$\begin{aligned} MSE\left(\bar{X}^R\right) &= \frac{1}{(n-1)^2}\left(\mu^2 + n\sigma^2\right) - \frac{1}{(n-1)^2}\left[\mu_{n,n} - \mu_{n-1,n} + \mu\mu_{n-1}\right] \\ &+ \frac{1}{4(n-1)^2}\left(\mu_{1,1} + \mu_{n,n} + 2\mu_{1,n}\right) - \frac{1}{(n-1)^2}\left[\mu_{1,1} - \mu_{2,1} + \mu\mu_2\right]. \end{aligned}$$

Les moments des statistiques d'ordre sont le plus souvent inconnus, mais tabulés, voir par exemple Sarhan et Greenberg[5](1962) dans le cas de la loi normale.

Afin de palier ce manque, on peut donner une approximation de cette erreur quadratique moyenne pour les différents domaines d'attractions possibles pour le maximum d'un échantillon et comparer l'efficacité de l'estimateur robuste avec celui proposé par Rivest[4] (1994).

# Références


[1] J-F Beaumont, David Haziza, and Anne Ruiz-Gazen. A unified approach to robust estimation in finite population sampling. *Biometrika*, 2013.

[2] Herbert Aron David and Haikady Navada Nagaraja. *Order statistics*. Wiley Online Library, 1970.

[3] J.L. Moreno-Rebollo, A. Muñoz-Reyes, and J. Muñoz-Pichardo. Miscellanea. influence diagnostic in survey sampling : conditional bias. *Biometrika*, 86(4) :923–928, 1999.

[4] Louis-Paul Rivest. Statistical properties of winsorized means for skewed distributions. *Biometrika*, 81(2) :373–383, 1994.

[5] Ahmed E Sarhan and Bernard G Greenberg. *Contributions to order statistics*. Wiley New York, 1962.