UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ESTIMATION ROBUSTE DANS DES MODÈLES DE POISSON

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

JOËLLE ROUSSEAU TRÉPANIER

FÉVRIER 2019

REMERCIEMENTS

En premier lieu, j'aimerais remercier mon directeur, Jean-François Coeurjolly, sans qui la rédaction de ce mémoire n'aurait pas été possible. Merci pour ton appui moral et financier, tes idées, corrections et suggestions durant la rédaction de ce mémoire. Merci aussi pour ta présence tout au long de ma maîtrise.

Je remercie aussi les autres professeurs du département qui m'ont aidé à développer mon intérêt et mes habiletés pour les statistiques et de m'avoir offert des démonstrations. Merci à Sorana Froda pour l'accueil incroyable au sein du département.

Je tiens aussi à remercier Florence, Patrick, Renaud, Anthony et Olivier pour leur présence durant cette aventure tant au bureau qu'à l'extérieur. Un petit merci spécial pour Florence, une rencontre formidable, d'avoir été fidèle au poste tous les jours et pour ton soutien. Merci aussi à Olivier pour la fonction *pretty_hist* qui m'a permis de faire de beaux histogrammes rapidement.

Finalement, un merci à mes parents, mes frères et ma soeur qui m'ont encouragée et soutenue pendant cette expérience et durant tout mon parcours scolaire.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Pour une distribution de Poisson de paramètre λ et de médiane notée Me_{N_λ}, il est connu que les meilleures bornes pour $\mathrm{Me}_{N_\lambda} - \lambda$ sont -log(2) et 1/3. Asymptotiquement, il est même prouvé que ces bornes deviennent -2/3 et 1/3. Motivés par l'obtention d'une loi limite pour la médiane empirique d'un échantillon issu d'une loi de Poisson, nous avons étudié la médiane d'une variable aléatoire obtenue comme la somme d'une loi de Poisson de paramètre λ et d'une loi uniforme sur [0,1]. De façon étonnante, nous sommes parvenus à démontrer que la médiane d'une telle variable est proche de $\lambda + 1/3$, et d'autant plus proche que λ est grand. Il en résulte une procédure très simple, efficace et satisfaisant des propriétés asymptotiques standards (convergence, théorème central limite) pour l'estimateur de λ. Par la suite, nous avons comparé par simulation l'estimateur de λ issu de cette procédure avec d'autres techniques existantes estimant de façon robuste le paramètre λ. Notons que l'estimateur est adapté dans un contexte de très grande dimension.

mots clés : Poisson, médiane, jittering, robuste, estimateur

INTRODUCTION

Un jeu de données contient fréquemment certaines observations qui s'éloignent fortement de la masse. Ces observations sont appellées valeurs aberrantes. Les méthodes classiques d'estimation, par exemple l'estimateur des moindres carrés pour un modèle de régression linéaire simple, sont largement affectées par de telles valeurs. La droite ne sera alors pas représentative des observations. Évidemment, la moyenne et la variance sont aussi deux estimateurs largement influencés par les valeurs aberrantes, contrairement à la médiane. Dans ce mémoire, on s'intéresse à estimer le paramètre λ d'une distribution de Poisson de façon robuste, c'est-à-dire obtenir un estimateur qui n'est pas trop influencé par les valeurs aberrantes et qui doit bien représenter la masse des données.

Au chapitre 1, on explore des estimateurs standard et connus du paramètre λ d'une loi de Poisson, tels que l'estimateur de maximum de vraisemblance, les M-estimateurs, les moyennes tronquées et les estimateurs de maximum de vraisemblance pondérée aux sections 1.1, 1.2 1.3 et 1.4 respectivement. En particuler, à la section 1.2, on présente les M-estimateurs basés sur la fonction de Huber et de Tukey ainsi que leurs versions modifiées en les mettant à l'échelle et en ajoutant un terme correctif. On retrouve ces estimateurs dans les articles de Cadigan et Chen (2001) et Elsaied et Fried (2016).

Ensuite, au chapitre 2, on présente un nouvel estimateur du paramètre λ basé sur la médiane empirique. Afin d'y parvenir, on commence par s'intéresser, à la section 2.3.1, à l'écart entre la médiane et la moyenne d'une loi de Poisson, soit $\mathrm{Me}_{N_\lambda} - \lambda$. Choi (1994) a montré que cet écart est borné par $-\log(2)$ et $1/3$, puis

Adell et Jodrá (2005) ont démontré qu'asymptotiquement ces bornes devenaient $-2/3$ et $1/3$. Cependant, notre estimateur se base plutôt sur un estimateur de la médiane théorique d'une variable aléatoire obtenue par la somme d'une loi de Poisson et d'une loi uniforme sur [0,1], on note cette variable $Z_\lambda = N_\lambda + U$ où $U \sim \mathcal{U}(0,1)$. Une telle variable suit une loi de Poisson perturbée. Ce processus, appellé *jittering*, est présenté à la section 2.2. Ce processus de lissage permet de récupérer des propriétés asymptotiques pour l'estimateur de la médiane de Z_λ. Ensuite, par une étude empirique, on a montré que l'écart entre la moyenne et la médiane théorique d'une loi de Poisson perturbée semblait se rapprocher de zéro lorsque λ augmente. Puis, on a formellement trouvé des bornes pour l'écart entre la médiane théorique et la moyenne à la section 2.3.2

Finalement, au chapitre 3, on analyse le comportement asymptotique de l'estimateur issu de la procédure présentée au chapitre 2 et on compare numériquement sa robustesse avec celle des autres estimateurs connus présentés au chapitre 1.

CHAPITRE I

ESTIMATEURS STANDARD ET ROBUSTES CONNUS DU PARAMÈTRE D'UNE LOI DE POISSON

Dans ce chapitre, on fera une revue des estimateurs standard et robustes connus pour estimer le paramètre d'une loi de Poisson. À la Section 1.1, on présente d'abord l'estimateur de maximum de vraisemblance. Ensuite, à la Section 1.2, on présente les M-estimateurs dont celui de Huber et de Tukey ainsi que leur version modifiée respectivement par Cadigan et Chen (2001) et Elsaied *et al.* (2012) afin d'obtenir une meilleure performance dans un cadre poissonien. Puis, à la Section 1.3, on présente la méthode de moyenne tronquée ainsi que sa version modifiée par Elsaied *et al.* (2012). Finalement, à la Section 1.4, on retrouve l'estimateur de vraisemblance pondérée de Markatou *et al.* (1997).

1.1 Estimateur de maximum de vraisemblance

Soit un échantillon aléatoire $\mathbf{Y} = (Y_1, \ldots, Y_n)$ i.i.d. provenant d'une loi de Poisson. On connaît la forme de l'estimateur de maximum de vraisemblance. En effet, on sait

— fonction de vraisemblance :

$$L(\lambda) = \prod_{i=1}^{n} f_{\mathbf{Y}}(y_i; \lambda) = \frac{\mathrm{e}^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!}$$

— estimateur de maximum de vraisemblance $\widehat{\lambda} = \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$

S'exprimant comme une moyenne empirique, il est clair que cet estimateur n'a aucune chance d'être robuste aux valeurs aberrantes.

1.2 Estimateurs robustes basés sur les M-estimateurs

Pour obtenir un estimateur robuste unidimensionnel, on peut se baser sur les M-estimateurs. On se réfère entre autres à Huber et Ronchetti (2009) dans cette section.

Définition 1.2.1. *Soit* $\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ *et* $\lambda \in \mathbb{R}$, *alors un M-estimateur du paramètre unidimensionnel* λ, *noté* $\widehat{\lambda}$, *est l'argument qui minimise une fonction estimante, c'est-à-dire*

$$\arg\min \sum_{i=1}^{n} \rho(y_i,\lambda) \tag{1.1}$$

ou dans le cas où $\rho(y,\cdot)$ *a une dérivée* $\psi(y,\lambda) = \frac{\partial}{\partial\lambda}\rho(y,\lambda)$ *pour tout* $y \in \mathbb{R}$, *il est défini comme la la racine d'une équation estimante , c'est-à-dire qu'il est la solution de*

$$\sum_{i=1}^{n} \psi(y_i,\widehat{\lambda}) = 0. \tag{1.2}$$

Par exemple, si la fonction ρ est la log-vraisemblance, le M-estimateur consiste en l'estimateur de maximum de vraisemblance. En présence de valeurs aberrantes, cet estimateur n'est plus optimal au sens qu'il ne possède plus la plus petite variance asymptotique dans la classe d'estimateurs qui sont des solutions de $\sum_{i=1}^{n} \psi(y_i,\lambda) = 0$.

On s'intéresse particulièrement aux M-estimateurs de position qui sont définis

ainsi par Maronna *et al.* (2006) :

$$\hat{\lambda} = \arg\min \sum_i \rho(y_i - \lambda) \tag{1.3}$$

où $\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ et $\lambda \in \mathbb{R}$. De même, ils peuvent aussi être définis par la solution de

$$\sum_{i=1}^{n} \psi(y_i - \hat{\lambda}) = 0. \tag{1.4}$$

Par exemple, on trouve la médiane empirique pour $\arg\min \sum_{i=1}^{n} \rho(y_i - \lambda)$, lorsque $\rho(y) = |y|$.

En effet,

$$\rho(y) = |y| \Leftrightarrow \psi(y) = \text{sgn}(y) = \mathbb{1}_{\{y>0\}} - \mathbb{1}_{\{y<0\}}$$

et

$$\begin{aligned}
&\sum \psi(y_i - \hat{\lambda}) = 0 \\
\Leftrightarrow &\sum \text{sgn}(y_i - \hat{\lambda}) = 0 \\
\Leftrightarrow &\sum \left(\mathbb{1}_{\{y_i - \hat{\lambda} > 0\}} - \mathbb{1}_{\{y_i - \hat{\lambda} < 0\}} \right) = 0 \\
\Leftrightarrow &\sum \mathbb{1}_{\{y_i - \hat{\lambda} > 0\}} - \sum \mathbb{1}_{\{y_i - \hat{\lambda} < 0\}} = 0 \\
\Leftrightarrow &\text{card}\left(y_i - \hat{\lambda} > 0\right) - \text{card}\left(y_i - \hat{\lambda} < 0\right) = 0 \\
\Leftrightarrow &\text{card}\left(y_i - \hat{\lambda} > 0\right) = \text{card}\left(y_i - \hat{\lambda} < 0\right).
\end{aligned}$$

Ainsi, la solution λ est bien la médiane échantillonale.

Si on prend plutôt $\rho(y) = y^2/2$, l'estimateur correspond à la moyenne puisque

$$\rho(y) = y^2/2 \Leftrightarrow \psi(y) = y$$

et donc, l'Équation 1.4 prend la forme

$$\sum_{i=1}^{n} (y_i - \lambda) = 0;$$

cette équation a bien la moyenne empirique comme solution.

Afin de comparer la performance des différents M-estimateurs, on peut s'appuyer sur leur distribution asymptotique. Pour une distribution donnée F, en supposant la fonction ψ croissante, on définit $\lambda = \lambda(F)$ comme solution de $\mathrm{E}\,\psi(Y-\lambda) = 0$. La distribution asymptotique de l'estimateur $\widehat{\lambda}$ est donnée par

$$\sqrt{n}(\widehat{\lambda} - \lambda) \xrightarrow{d} \mathcal{N}(0,v) \text{ où } v = \frac{\mathrm{E}\,(\psi(Y-\lambda)^2)}{(\mathrm{E}(\psi'(Y-\lambda)))^2}$$

selon Maronna *et al.* (2006). L'efficacité asymptotique relative de $\widehat{\lambda}$ par rapport à l'estimateur de maximum de vraisemblance, noté λ_{MLE}, est donnée par :

$$\mathrm{ARE}(\widehat{\lambda},\lambda_{\mathrm{MLE}}) = \frac{v_o}{v}$$

où v et v_0 sont les variances asymptotiques de $\widehat{\lambda}$ et de l'estimateur de maximum de vraisemblance respectivement.

1.2.1 Estimateur de Huber et de Tukey

La moyenne empirique est l'estimateur le plus efficace pour λ sous l'hypothèse de normalité des observations, cependant elle n'est pas robuste. La médiane, quant à elle, est robuste mais n'est pas efficace sous la même hypothèse. L'estimateur de Huber est défini par :

$$\rho_k(y-\lambda) = \begin{cases} (y-\lambda)^2 & \text{si } |y-\lambda| \leq k \\ 2k|y-\lambda| - k^2 & \text{si } |y-\lambda| > k \end{cases}$$

et

$$\psi_k(y-\lambda) = \begin{cases} y-\lambda & \text{si } |y-\lambda| \leq k \\ \mathrm{sgn}(y-\lambda)k & \text{si } |y-\lambda| > k \end{cases}$$

où k est une constante d'ajustement permettant d'assurer une certaine variance asymptotique. Il s'agit d'un compromis connu entre efficacité et robustesse puisque le M-estimateur de Huber, $\widehat{\lambda}$, tend vers la moyenne quand k tend vers l'infini et

vers la médiane quand k tend vers 0. Il est à noter qu'on peut écrire la fonction ψ comme

$$\psi_k(y-\lambda) = (y-\lambda)\min\left(1, \frac{k}{|y-\lambda|}\right).$$

Pour calculer le M-estimateur, on peut utiliser la méthode itérative de moindres carrées repondérés (IRLS). On utilise les poids :

$$\mathrm{W}_k(y-\lambda) = \frac{\psi(y-\lambda)}{y-\lambda} = \begin{cases} 1 & \text{si } |y-\lambda| \leq k \\ \frac{k}{y-\lambda} & \text{si } |y-\lambda| > k \end{cases}.$$

Ensuite, on peut remplacer la fonction ψ par les poids W dans l'Équation (1.4), donc on résoud :

$$\sum_{i=1}^{n} \mathrm{W}(y_i - \widehat{\lambda})(y_i - \widehat{\lambda}) = 0$$

et, si la fonction W est bornée et décroissante pour $y_i - \lambda$, on trouve la solution qui exprime l'estimateur comme une moyenne pondérée :

$$\widehat{\lambda} = \frac{\sum_{i=1}^{n} \omega_i y_i}{\sum_{i=1}^{n} \omega_i}$$

où $\omega_i = \mathrm{W}(y_i - \lambda)$.

Dans le cas d'un M-estimateur de Huber, la fonction de poids est donnée par :

$$\mathrm{W}_k(y-\lambda) = \frac{\psi_k(y-\lambda)}{y-\lambda} = \begin{cases} 1 & \text{si } |y-\lambda| \leq k \\ \frac{k}{|y-\lambda|} & \text{si } |y-\lambda| > k \end{cases} = \min\left(1, \frac{k}{|y-\lambda|}\right).$$

Lorsque la fonction ψ est continue et strictement croissante, la solution de l'Équation (1.4) est unique et appellée M-estimateur monotone. Dans le cas de la fonction de Huber, on a un M-estimateur monotone. Si la fonction ψ n'est pas monotone, la solution de l'Équation (1.4) n'est pas unique et s'appelle un M-estimateur redescendant (la fonction ψ tend vers zéro à l'infini). Les M-estimateurs redescendants

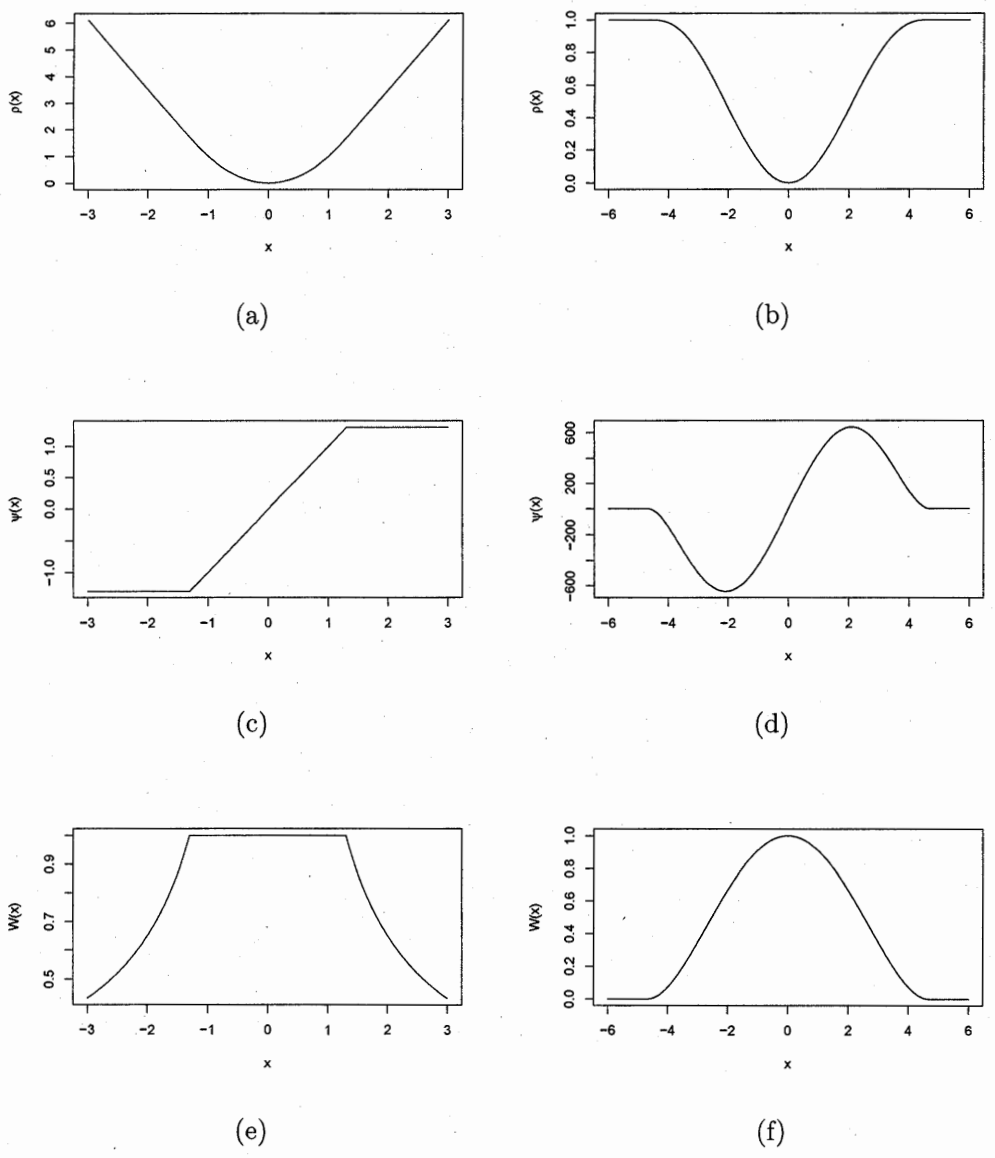


Figure 1.1: Fonctions ρ, ψ et W_k de haut en bas pour Huber (colonne de gauche) et Tukey (colonne de droite) pour k=1.3 et k=4.68 respectivement.

sont plus robustes aux très grandes valeurs aberrantes. L'estimateur bipondéré de Tukey en est un exemple connu. Il est défini par :

$$\rho_k(y-\lambda) = \begin{cases} 1-\left(1-\left(\frac{y-\lambda}{k}\right)^2\right)^3 & \text{si } |y-\lambda| \leq k \\ 1 & \text{si } |y-\lambda| > k \end{cases}$$

$$\psi_k(y-\lambda) = \begin{cases} (y-\lambda)\left(1-\left(\frac{y-\lambda}{k}\right)^2\right)^2 & \text{si } |y-\lambda| \leq k \\ 0 & \text{si } |y-\lambda| > k \end{cases}$$

et

$$\mathrm{W}_k(y-\lambda) = \frac{\psi_k(y-\lambda)}{y-\lambda} = \begin{cases} \left(1-\left(\frac{y-\lambda}{k}\right)^2\right)^2 & \text{si } |y-\lambda| \leq k \\ 0 & \text{si } |y-\lambda| > k. \end{cases}$$

1.2.2 Estimateurs modifiés de Huber et de Tukey

Par la suite, Cadigan et Chen (2001) ont proposé une version modifiée de l'estimateur de Huber en utilisant la nature poissonnienne de la variable aléatoire. En effet, ils utilisent le paramètre de la loi pour centrer et réduire :

$$\psi_{k,a}(y,\lambda) = \left(\frac{y-\lambda}{\sqrt{\lambda}} - a\right) \min\left(1, \frac{k\sqrt{\lambda}}{|y-\lambda-\sqrt{\lambda}|}\right)$$

où $a = a(\lambda,k)$ est un terme correctif pour obtenir un estimateur asymptotiquement non biaisé. Afin d'estimer λ, il faut résoudre les deux équations suivantes pour λ et a :

$$\mathrm{E}(\psi_{k,a}(Y,\lambda)) = 0 \tag{1.5}$$

et

$$\sum_{i=1}^{n} \psi_{k,a}(y_i,\lambda) = 0. \tag{1.6}$$

Dans le même ordre d'idée, Elsaied et Fried (2016) ont, quant à eux, proposé une

version modifiée du M-estimateur de Tukey en utilisant λ pour centrer et réduire et en introduisant un terme correctif a :

$$\psi_{k,a}(y,\lambda) = \left(\frac{y-\lambda}{\sqrt{\lambda}} - a\right)\left(k^2 - \left(\frac{y-\lambda}{\sqrt{\lambda}}\right)^2\right)^2 \mathrm{I}_{[-k,k]}\left(\frac{y-\lambda}{\sqrt{\lambda}} - a\right) \tag{1.7}$$

où $a = a(\lambda,k)$ satisfait la condition (1.5).

1.3 Moyenne tronquée

Une autre façon d'obtenir un estimateur robuste consiste à tronquer les très grandes et très petites valeurs lors de l'estimation du paramètre. Maronna *et al.* (2006) définissent la moyenne tronquée ainsi :

Définition 1.3.1. *Soit $\alpha \in [0,1/2)$ et $m = [(n-1)\alpha]$ où $[\cdot]$ est la fonction partie entière ; alors la moyenne α-tronquée est définie par :*

$$\bar{x}_\alpha = \frac{1}{n-2m}\sum_{i=m+1}^{n-m} x_{(i)}.$$

Dans le même odre d'idées, on peut aussi estimer la moyenne par troncature adaptative des données. On estime l'étendue possible des données, par exemple en ajoutant et en soustrayant un multiple de l'erreur absolue médiane ($MAD = \mathrm{Me}(|X_i - \mathrm{Me}_X|)$) à la médiane empirique dans le cas gaussien, et on tronque toutes les valeurs qui ne sont pas comprises dans cette étendue. Puis, on calcule la moyenne des données dans l'étendue.

Cependant, ces deux approches ne fonctionnent bien que pour des distributions symétriques. Par conséquent, on ne peut utiliser ces estimateurs pour des données poissoniennes. Elsaied et Fried (2016) proposent un estimateur adapté à cette loi. Ils utilisent un estimateur pour l'initialisation, par exemple la médiane, puis ils tronquent les observations supérieures et inférieures aux quantiles $1-\alpha/2$ et $\alpha/2$

respectivement. Ensuite, ils réestiment le paramètre par la moyenne tronquée et itèrent cette procédure jusqu'à convergence.

1.4 Estimateur de vraisemblance pondérée

On peut aussi faire de l'estimation robuste de paramètre par la méthode de vraisemblance pondérée, c'est-à-dire qu'on enlève du poids aux valeurs aberrantes. De plus, cette méthode, introduite par Markatou *et al.* (1997), n'enlève pas automatiquement du poids à une certaine proportion de données, seulement à celles qui ne concordent pas avec le modèle.

Suppsons que $\{X_1,\ldots,X_n\}$ est un échantillon aléatoire et $u(x,\lambda) = \nabla_\lambda \ln(m_\lambda(x))$ la fonction de score de maximum de vraisemblance où ∇_λ est le gradient par rapport à λ et $m_\lambda(x)$ est la fonction de masse définie sur l'espace échantionnal $R_x = \{0,1,\ldots,T\}$ où T peut être infini. L'estimateur de maximum de vraisemblance de λ est solution de $\sum_{i=1}^{n} u(X_i;\lambda) = 0$ sous la condition que la famille de distributions soit régulière. L'idée est d'accorder un poids $w(x; M_\lambda,\widehat{F})$ à chaque point x de l'espace échantillonnal dépendant de x, de la distribution de probabilité supposée M_λ et de sa fonction de répartition empirique $\widehat{F}$. Ainsi, les estimateurs de vraisemblance pondérés sont solutions de $\sum_{i=1}^{n} w(X_i; M_\lambda,\widehat{F})u(X_i;\lambda) = 0$.

Pour définir les valeurs qui s'éloignent trop du modèle, on doit introduire une notion de résidu. La définition d'un résidu varie de ce pourquoi il est employé. Par exemple, on connaît bien la définition d'un résidu géométrique dans le cas d'une régression linéaire. Par contre, lorsque la relation entre les observations et les paramètres est probabiliste, il est nécessaire de définir une valeur aberrante comme une observation t qui a une probabilité $m_\lambda(t)$ très petite de se produire sous le modèle supposé.

Définition 1.4.1. *Pour tout $t \in R_X$, on définit le résidu de Pearson δ par*

$$\delta(t) = \frac{d(t)}{m_\lambda(t)} - 1$$

où $d(t)$ est la proportion d'observations de valeur t et de probabilité $m_\lambda(t)$ sous le modèle supposé.

Ainsi, la méthode de vraisemblance pondérée enlève du poids aux données qui ont des résidus de Pearson élevés selon la fonction

$$w(t,M_\lambda,\widehat{F}) = w(\delta(t)) = \frac{A(\delta(t)) + 1}{\delta(t) + 1}$$

où $A : \mathbb{R} \to [-1,\infty)$ est une fonction d'ajustement résiduelle (RAF) supposée deux fois différentiable, strictement croissante et telle que $A(0) = 0$ et $A'(0) = 1$.

Par exemple, on peut utiliser la distance de Hellinger (HD) comme fonction d'ajustement résiduelle.

Définition 1.4.2. *Soit deux distributions $P = (p_1,\ldots,p_n)$ et $Q = (q_1,\ldots,q_n)$ alors la distance de Hellinger entre P et Q est donnée par :*

$$h(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})^2}.$$

La distance de Hellinger peut aussi s'exprimer en fonction du résidu de Pearson :

$$A(\delta) = 2\left((\delta + 1)^{1/2} - 1\right)$$

CHAPITRE II

ESTIMATEUR DE λ BASÉ SUR LA MÉDIANE

Au chapitre 2, on présente notre estimateur pour le paramètre d'une loi de Poisson : la médiane empirique d'une loi de Poisson perturbée auquel on soustrait 1/3. Pour ce faire, on commence d'abord par rappeler les définitions des quantiles et de la médiane théorique et empirique à la Section 2.1. Ensuite, à la Section 2.2, on explique le principe de perturbation d'une loi de Poisson pour lisser artificiellement la loi discrète. Finalement, à la Section 2.3, on explique ce qui a inspiré l'estimateur, soit la différence entre la médiane et la moyenne. On commence, à la Sous-section 2.3.1, par faire une revue des bornes qui ont été trouvées pour l'écart entre la médiane et la moyenne théoriques d'une loi de Poisson standard. Suivant cette idée, à la Sous-section 2.3.2, on s'intéresse ensuite à borner cet écart pour une loi de Poisson perturbée afin de construire l'estimateur basé sur la médiane empirique.

2.1 Quantiles et médiane empirique

Définition 2.1.1. *Pour une variable aléatoire Y, la médiane théorique, notée* Me_Y, *se définit par*

$$\mathrm{P}(Y \leq \mathrm{Me}_Y) \geq \frac{1}{2} \text{ et } \mathrm{P}(Y \geq \mathrm{Me}_Y) \geq \frac{1}{2};$$

cela revient à

$$F_Y(\mathrm{Me}_Y^-) \leq \frac{1}{2} \leq F_Y(\mathrm{Me}_Y)$$

où F_Y *est la fonction de répartition de* Y *et* $F_Y(\mathrm{Me}_Y^-)$ *est la limite à gauche de* F_Y *à* Me_Y.

Ainsi, si la variable Y est continue, la médiane est unique et est la solution de $\mathrm{Me}_Y = F_Y^{-1}(1/2)$ tandis que si Y est discrète, la médiane n'est pas unique.

Définition 2.1.2. *Soit l'échantillon* $\mathbf{Y} = (Y_1, \ldots, Y_n)$ *de* n *variables aléatoires identiquement distribuées de même fonction de répartition* F_Y *; alors la médiane échantillonnale, notée* $\widehat{\mathrm{Me}}(\mathbf{Y})$*, se définit par*

$$\widehat{\mathrm{Me}}(\mathbf{Y}) = \inf\{x \in \mathbb{R} : 1/2 \leq \widehat{F}_Y(x, \mathbf{Y})\}$$

où $\widehat{F}_Y(\cdot, \mathbf{Y})$ *est la fonction de répartition empirique.*

Définition 2.1.3. *Soit* F_Y *une fonction de répartition, alors le quantile d'ordre* $p \in [0,1]$ *se définit par*

$$F_Y^{-1}(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\}$$

tandis que le quantile d'ordre $p \in [0,1]$ *échantillonal est défini par*

$$\widehat{F}^{-1}(p, \mathbf{Y}) = \inf\{x \in \mathbb{R} : p \leq \widehat{F}(x, \mathbf{Y})\}$$

pour l'échantillon $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

Dans Serfling (2009), on montre la normalité asymptotique du quantile échantillonnal si $F_Y(\cdot)$ possède une dérivée à droite ou à gauche à $F_Y^{-1}(p)$. On reproduit ces résultats ici.

Théorème 2.1.4. *Soit $0 < p < 1$ et soit $F_Y(\cdot)$ une fonction de répartition continue au point $F_Y^{-1}(p)$. On a alors les assertions suivantes lorsque $n \to \infty$:*

(i) s'il existe une dérivée à gauche, notée $F'_Y(F_Y^{-1}(p)^-)$, alors, pour $t < 0$,

$$\sqrt{n}\left(\widehat{F}^{-1}(p,\mathbf{Y}) - F_Y^{-1}(p)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{\left[F'_Y(F_Y^{-1}(p)^-)\right]^2}\right)$$

où $\xrightarrow{d}$ désigne la convergence en loi.

(ii) s'il existe une dérivée à droite, notée $F'_Y({F_Y}^{-1}(p)^+)$, alors, pour $t > 0$,

$$\sqrt{n}\left(\widehat{F}^{-1}(p,\mathbf{Y}) - F_Y^{-1}(p)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{\left[F'_Y(F_Y^{-1}(p)^+)\right]^2}\right).$$

(iii) dans tous les cas,

$$\lim_{n\to\infty} \mathrm{P}(n^{1/2}(\widehat{F}^{-1}(p,\mathbf{Y}) - F_Y^{-1}(p)) \leq 0) = \Phi(0) = \frac{1}{2}.$$

On en déduit le résultat général suivant.

Corollaire 2.1.5. *Soit $0 < p < 1$, on a les assertions suivantes lorsque $n \to \infty$*

(i) Si F_Y est différentiable au point $F_Y^{-1}(p)$ et que $F'_Y(F_Y^{-1}(p)) > 0$, alors

$$\sqrt{n}\left(\widehat{F}^{-1}(p,\mathbf{Y}) - F_Y^{-1}(p)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{\left[F'_Y(F_Y^{-1}(p))\right]^2}\right).$$

(ii) Si F_Y possède une densité f_Y dans un voisinage de $F_Y^{-1}(p)$ et que f_Y est strictement positive et continue au point $F_Y^{-1}(p)$, alors

$$\sqrt{n}\left(\widehat{F}^{-1}(p,\mathbf{Y}) - F_Y^{-1}(p)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{{f_Y}^2(F_Y^{-1}(p))}\right).$$

Évidemment, la partie 2 du Corollaire 2.1.5 n'est pas applicable aux lois discrètes puisque la médiane n'est pas unique dans ce cas.

Soit N_λ une variable aléatoire de loi de Poisson de paramètre $\lambda > 0$, on note $N_\lambda \sim \mathcal{P}(\lambda)$. La médiane théorique est notée par $\mathrm{Me}_{N_\lambda} = F^{-1}_{N_\lambda}(1/2)$. En particulier, notant $\mathbf{N}_\lambda$ un échantillon de n variables aléatoires indépendantes identiquement distribuées de même loi que N_λ, on observe que la médiane empirique $\widehat{\mathrm{Me}}(\mathbf{N}_\lambda) = \widehat{F}^{-1}(1/2; \mathbf{N}_\lambda)$ ne peut satisfaire au théorème central limite. Par conséquent, outre le fait qu'on n'a pour le moment pas établi de lien entre Me_{N_λ} et λ, estimer de façon robuste λ par la simple médiane semble être une mauvaise piste.

2.2 Perturbation d'une loi de Poisson

Suivant la section précédente, l'idée est d'appliquer une méthode présentée, entre autres, par Machado et Silva (2005), le *jittering* pour introduire un lissage artificiel d'une variable discrète Y en lui ajoutant une variable aléatoire de loi uniforme sur (0,1), c'est-à-dire $Z = Y + U$. Ainsi, la variable Z est continue. Dans le cadre de la loi de Poisson, on note $Z_\lambda = N_\lambda + U$ où $U \sim \mathcal{U}(0,1)$. On peut démontrer le résultat suivant.

Proposition 2.2.1. *Soit $Z_\lambda = N_\lambda + U$ où $N_\lambda \sim \mathcal{P}(\lambda)$, $U \sim \mathcal{U}(0,1)$ et N_λ et U sont indépendantes, alors la fonction de répartition de Z_λ est donnée par*

$$F_{Z_\lambda}(t) = \mathrm{P}\left(Z_\lambda \leq t\right) = \sum_{k=0}^{\infty} \mathrm{P}\left(N_\lambda = k\right) \min(1, t-k)^+$$

où $\min(a,b)^+ = 0$ si $\min(a,b) < 0$. De plus, la fonction $F_{Z_\lambda}(t)$ est dérivable presque partout et $f_{Z_\lambda}(t) = \mathrm{P}(N_\lambda = \lfloor t \rfloor)$.

Démonstration.

$$\begin{aligned} F_{Z_\lambda}(t) &= \mathrm{P}\left(Z_\lambda \leq t\right) \\ &= \sum_{k=0}^{\infty} \mathrm{P}\left(N_\lambda = k\right) \mathrm{P}\left(U \leq t-k\right) \\ &= \sum_{k=0}^{\infty} \mathrm{P}\left(N_\lambda = k\right) \min(1, t-k)^+. \end{aligned}$$

Soit

$$\delta_\lambda(t,h) = F_{Z_\lambda}(t+h) - F_{Z_\lambda}(t)$$

On montre que pour presque tout t, $\delta_\lambda(t,h)/h$ admet une limite lorsque $h \to 0$. On sépare la preuve en deux cas, selon le signe de h.

1. Soit $h > 0$

$$\delta_\lambda(t,h) = \sum_{k=0}^{\infty} \mathrm{P}(N_\lambda = k) \min(1,t+h-k)^+ - \sum_{k=0}^{\infty} \mathrm{P}(N_\lambda = k) \min(1,t-k)^+.$$

On considère les deux cas suivants :

— Si $0 < t < 1$

$$\begin{aligned}
\delta_\lambda(t,h) &= \sum_{k=0}^{\infty} \mathrm{P}(N_\lambda = k)[\min(1,t+h-k)^+ - \min(1,t-k)^+] \\
&= \mathrm{P}(N_\lambda = 0)[t+h-t] + \sum_{k=1}^{\infty} \mathrm{P}(N_\lambda = k)[0-0] \\
&= h\,\mathrm{P}(N_\lambda = 0) \\
&= h\,\mathrm{P}(N_\lambda = \lfloor t \rfloor).
\end{aligned}$$

— si $t \geq 1$

$$\begin{aligned}
\delta_\lambda(t,h) &= \sum_{k=0}^{\infty} \mathrm{P}(N_\lambda = k)[\min(1,t+h-k)^+ - \min(1,t-k)^+] \\
&= \sum_{k=0}^{\lfloor t \rfloor - 1} \mathrm{P}(N_\lambda = k)[1-1] + \mathrm{P}(N_\lambda = \lfloor t \rfloor)(t+h-\lfloor t \rfloor - (t-\lfloor t \rfloor)) \\
&\quad + \sum_{k=\lfloor t \rfloor + 1}^{\infty} \mathrm{P}(N_\lambda = k)[0-0] \\
&= h\,\mathrm{P}(N_\lambda = \lfloor t \rfloor).
\end{aligned}$$

Ainsi, $\forall t > 0$ et $h > 0$, $\delta_\lambda(t,h) = h\,\mathrm{P}(N_\lambda = \lfloor t \rfloor)$ et par conséquent,

$$\lim_{h \to 0^+} \frac{F_{Z_\lambda}(t+h) - F_{Z_\lambda}(t)}{h} = \mathrm{P}(N_\lambda = \lfloor t \rfloor).$$

2. Soit $h > 0$,

$$\delta_\lambda(t,-h) = \sum_{k=0}^{\infty} \mathrm{P}(N_\lambda = k)\min(1,t-h-k)^+ - \sum_{k=0}^{\infty} \mathrm{P}(N_\lambda = k)\min(1,t-k)^+.$$

On considère les deux cas suivants :

— si $0 < t < 1$ et $t \neq \lfloor t \rfloor$

$$\begin{aligned}\delta_\lambda(t,-h) &= \sum_{k=0}^{\infty} \mathrm{P}(N_\lambda = k)[\min(1,t-h-k)^+ - \min(1,t-k)^+] \\ &= \mathrm{P}(N_\lambda = 0)[t-h-t] + \sum_{k=1}^{\infty} \mathrm{P}(N_\lambda = k)[0-0] \\ &= -h\,\mathrm{P}(N_\lambda = 0) = -h\,\mathrm{P}(N_\lambda = \lfloor t \rfloor).\end{aligned}$$

— si $t \geq 1$ et $t \neq \lfloor t \rfloor$

$$\begin{aligned}\delta_\lambda(t,-h) &= \sum_{k=0}^{\infty} \mathrm{P}(N_\lambda = k)[\min(1,t-h-k)^+ - \min(1,t-k)^+] \\ &= \sum_{k=0}^{\lfloor t \rfloor - 1} \mathrm{P}(N_\lambda = k)[1-1] \\ &\quad + \mathrm{P}(N_\lambda = \lfloor t \rfloor)(t-h-\lfloor t \rfloor - (t-\lfloor t \rfloor)) \\ &\quad + \sum_{k=\lfloor t \rfloor + 1}^{\infty} \mathrm{P}(N_\lambda = k)[0-0] \\ &= -h\,\mathrm{P}(N_\lambda = \lfloor t \rfloor).\end{aligned}$$

Ainsi, $\forall t > 0$, $t \neq \lfloor t \rfloor$ et $h > 0$, $\delta_\lambda(t,-h) = -h\,\mathrm{P}(N_\lambda = \lfloor t \rfloor)$ et par conséquent,

$$\lim_{h \to 0^+} \frac{F_{Z_\lambda}(t-h) - F_{Z_\lambda}(t)}{-h} = \mathrm{P}(N_\lambda = \lfloor t \rfloor).$$

Puisque $F_{Z_\lambda}(\cdot)$ est presque partout dérivable, on peut définir la densité de Z_λ par $f_{Z_\lambda}(t) = \mathrm{P}(N_\lambda = \lfloor t \rfloor)$. On peut aussi vérifer que f_{Z_λ} est Lebesgue intégrable, $\int_{-\infty}^{\infty} f_{Z_\lambda}(t)dt = 1$ et pour $x > 0$

$$\int_{-\infty}^{x} f_{Z_\lambda}(t)dt = \sum_{k=0}^{\lfloor x \rfloor - 1} \mathrm{P}(N_\lambda = k) + (x - \lfloor x \rfloor)\,\mathrm{P}(N_\lambda = \lfloor x \rfloor) = F_{Z_\lambda}(x).$$

□

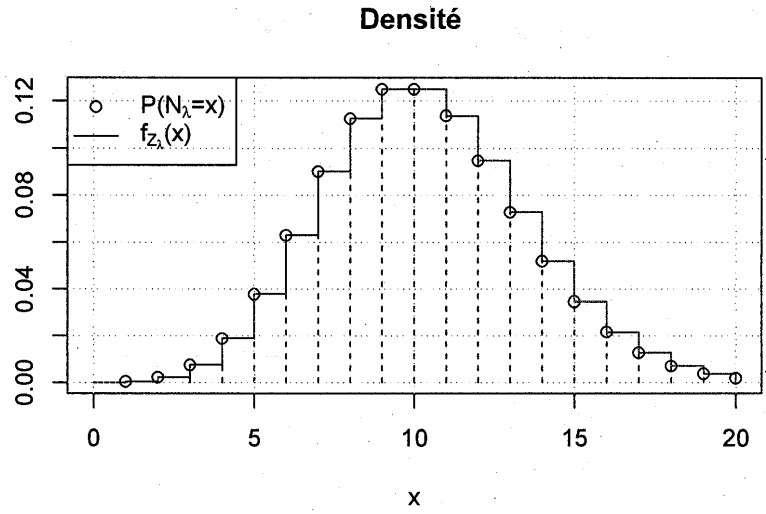


Figure 2.1: Fonction de masse de N_λ et densité de Z_λ pour $\lambda = 10$.

Les lissages des Figures 2.1 et 2.2 illustrent la Proposition 2.2.1 et représentent les fonctions de masse/densité et les fonctions de répartition de N_λ et Z_λ pour $\lambda = 10$.

Pour tout λ,$t > 0$, $f_{Z_\lambda}(t) > 0$. La fonction de répartition de Z_λ est donc strictement monotone et donc la médiane de Z_λ est définie de manière unique.

$$F_{Z_\lambda}(\mathrm{Me}_{Z_\lambda}) = \frac{1}{2} \Leftrightarrow \mathrm{Me}_{Z_\lambda} = F^{-1}_{Z_\lambda}\left(\frac{1}{2}\right) = \arg\min_{\lambda>0} \left| F_{Z_\lambda}(t) - \frac{1}{2} \right|.$$

Proposition 2.2.2. *Lorsque* $n \to \infty$,

$$\sqrt{n}\left(\widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - \mathrm{Me}_{Z_\lambda}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4\,\mathrm{P}(N_\lambda = \lfloor \mathrm{Me}_{Z_\lambda} \rfloor)^2}\right)$$

où $\mathbf{Z}_\lambda = (Z_{1,\lambda}, \ldots, Z_{n,\lambda})$.

La preuve découle du fait qu'on peut appliquer la partie 2 du Corollaire 2.1.5 pour

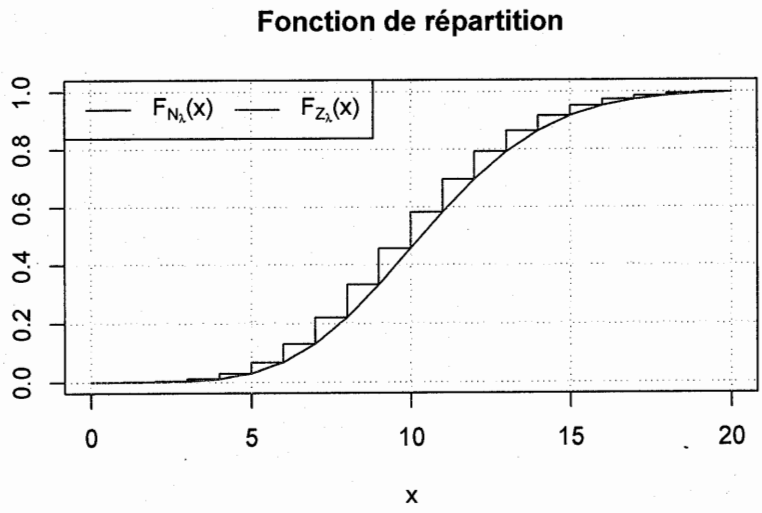


Figure 2.2: Fonction de répartition de N_λ et $Z_\lambda = N_\lambda + U$ pour $\lambda = 10$.

le cas particulier de la médiane

$$\sqrt{n}\left(\widehat{F}^{-1}\left(\frac{1}{2},\mathbf{Z}_\lambda\right) - F_{Z_\lambda}^{-1}\left(\frac{1}{2}\right)\right) \xrightarrow{d} \mathcal{N}\left(0,\frac{1}{4f_{Z_\lambda}^2(F_{Z_\lambda}^{-1}(\frac{1}{2}))}\right).$$

Dans le cas d'une loi de Poisson perturbée, on peut donc déduire

$$\sqrt{n}\left(\widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - \mathrm{Me}_{Z_\lambda}\right) \xrightarrow{d} \mathcal{N}\left(0,\frac{1}{4f_{Z_\lambda}^2(\mathrm{Me}_{Z_\lambda}))}\right).$$

Il est temps maintenant de se concentrer sur la compréhension de Me_{Z_λ} en fonction de λ. Avant cela, on analyse les résultats existants sur les liens entre Me_{N_λ} et λ.

2.3 Médiane d'une loi de Poisson standard et perturbée

Dans cette section, on s'intéresse à la forme de la médiane puisqu'on cherche à borner l'écart entre la moyenne et la médiane d'une loi de Poisson et celui de sa version perturbée. Le but consiste à contrôler le plus possible l'écart entre ces deux valeurs.

2.3.1 Loi de Poisson discrète

On commence par étudier l'écart entre la médiane d'une loi de Poisson et sa moyenne, c'est-à-dire $\mathrm{Me}_{N_\lambda} - \lambda$. À cette fin Chen et Rubin (1986) utilisent la relation Poisson-Gamma qu'on énonce dans le Théorème 2.3.1.

Théorème 2.3.1 (Relation Poisson Gamma). *Soit N_λ une variable aléatoire de loi de Poisson de paramètre λ, $N_\lambda \sim \mathcal{P}(\lambda)$, et X une variable aléatoire suivant une loi Gamma de paramètre α, $X_\alpha \sim \mathcal{G}(\alpha, 1)$, alors*

$$\mathrm{P}(N_\lambda \geq \alpha) = \mathrm{P}(X_\alpha \leq \lambda) \qquad et \qquad \mathrm{P}(N_\lambda \leq \alpha - 1) = \mathrm{P}(X_\alpha \geq \lambda).$$

Chen et Rubin (1986) obtiennent alors le résultat suivant.

Théorème 2.3.2. *Soit N_λ une variable aléatoire de loi de Poisson de paramètre λ, $N_\lambda \sim \mathcal{P}(\lambda)$, alors*

$$-1 < \mathrm{Me}_{N_\lambda} - \lambda < \frac{1}{3}.$$

Chen et Rubin (1986) avaient aussi formulé la conjecture suivante, qui, par la suite, a été démontrée par Choi (1994).

Théorème 2.3.3. *Soit N_λ une variable aléatoire de loi de Poisson de paramètre λ, $N_\lambda \sim \mathcal{P}(\lambda)$; alors on a*

$$-\log(2) < \mathrm{Me}_{N_\lambda} - \lambda < \frac{1}{3}.$$

Il s'agit des meilleures bornes possibles pour l'écart entre la médiane et la moyenne comme l'illustre la Figure 2.3.

Adell et Jodrá (2005) obtiennent des résultats encore plus fins. Pour ce faire, ils définissent d'abord une suite de fonctions $(f_n, n \in \mathbb{N})$

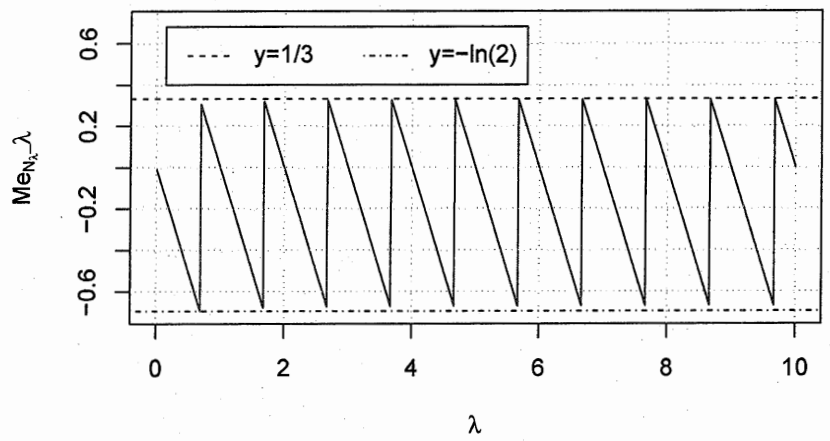


Figure 2.3: Écart entre la médiane et la moyenne d'une loi de Poisson en fonction de λ.

$$f_n(\lambda) := \mathrm{P}\left(N_\lambda \leq n\right) = \sum_{k=0}^{n} \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}, \quad \lambda \geq 0 \tag{2.1}$$

où $N_\lambda \sim \mathcal{P}(\lambda)$.

Puis, par la Relation Poisson-Gamma 2.3.1, ils réécrivent la suite

$$f_n(\lambda) = \sum_{k=0}^{n} \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} = \frac{1}{n!}\int_\lambda^\infty \mathrm{e}^{-u}u^n du, \quad \lambda \geq 0.$$

Ils obtiennent notamment le résultat suivant.

Corollaire 2.3.4. *Soit $\lambda \geq 0$ et $k \in \mathbb{N}$, alors, pour $\lambda \geq k$*

$$-\frac{2}{3} - \frac{8(1-c_{k+1})}{81c_{k+1}} < \mathrm{Me}(N_\lambda) - \lambda < \frac{1}{3}$$

où $c_k := \frac{1}{\mathrm{e}}\left(1+\frac{1}{k}\right)^{k-1}$.

On en déduit alors les meilleures bornes asymptotiques pour la différence entre la médiane et la moyenne, soit $-2/3$ et $1/3$, c'est-à-dire

$$-\frac{2}{3} = \liminf_{\lambda\to\infty}(\mathrm{Me}_{N_\lambda} - \lambda) \leq \limsup_{\lambda\to\infty}(\mathrm{Me}_{N_\lambda} - \lambda) = \frac{1}{3}. \tag{2.2}$$

Autrement dit, l'écart entre Me_{N_λ} et λ fluctue dans un intervalle de longueur 1 approximativement, quelle que soit la valeur de l'intensité λ.

2.3.2 Loi de Poisson perturbée

On s'intéresse maintenant à la différence entre la médiane et la moyenne dans le cas d'une loi de Poisson lissée par la méthode de *jittering*. Le but est de contrôler autant que possible cet écart, c'est-à-dire $\mathrm{Me}_{Z_\lambda} - \lambda$. Pour cela, sans utiliser la nature poissonienne de Z_λ, on pourrait s'appuyer sur un résultat général valable pour toute variable aléatoire continue pour laquelle les deux premiers moments sont finis stipulant que $|\mathrm{Me}_Y - \mathrm{E}(Y)| \leq \sqrt{\mathrm{Var}(Y)}$. On peut appliquer cette inégalité à une variable de loi de Poisson perturbée, ce qui implique le résultat suivant.

Proposition 2.3.5. *Soit $N_\lambda \sim \mathcal{P}(\lambda)$ une variable aléatoire suivant une loi de Poisson et Z_λ sa version perturbée, on a*

$$|\mathrm{Me}_{Z_\lambda} - \lambda - 1/2| \leq \sqrt{\lambda + 1/12}.$$

Les bornes données à la Proposition 2.3.5 ne sont clairement pas satisfaisantes ; l'écart entre Me_{Z_λ} et λ semblant être de plus en plus grand lorsque λ augmente. Si l'on cherche à exploiter le résultat 2.3.3 de la section précédente, on pourrait obtenir la proposition suivante :

Proposition 2.3.6. *Soit $N_\lambda \sim \mathcal{P}(\lambda)$ une variable aléatoire suivant une loi de Poisson et Z_λ sa version perturbée, on peut montrer que*

$$-\log(2) \leq \mathrm{Me}_{Z_\lambda} - \lambda \leq \frac{4}{3}.$$

Démonstration. Puisque $0 \leq U \leq 1$, on peut montrer que

$$\mathrm{Me}_{N_\lambda} \leq \mathrm{Me}_{Z_\lambda} \leq 1 + \mathrm{Me}_{N_\lambda}.$$

Puis, il suffit d'utiliser le Théorème 2.3.3 pour conclure. □

Ce résultat, bien qu'intéresant, n'est pas des plus satisfaisants ; l'écart entre les deux bornes étant de l'ordre de $4/3 + \log(2) \approx 2$. On a affiné ce résultat au regard d'une étude empirique. En effet, à la lumière de la Figure 2.4, on peut remarquer que la médiane de Z_λ, calculée numériquement comme l'argument minimum de $|F_{Z_\lambda}(\cdot) - 1/2|$, semble être proche de $\lambda + 1/3$ et d'autant plus proche que λ augmente. Ce fait étonnant encourage à conjecturer que

$$\mathrm{Me}_{Z_\lambda} \approx \lambda + 1/3,$$

voire même que $\mathrm{Me}_{Z_\lambda} - \lambda - 1/3$ semble se rapprocher de 0 lorsque λ augmente. Nous exprimons ce fait dans le résultat suivant qui constitue la majeure contribution de ce mémoire.

Théorème 2.3.7. *Pour tout $\varepsilon > 0$, il existe $n_0 \in \mathbb{N}$ tel que pour tout $n \geq n_0$ et $\lambda \geq n$,*

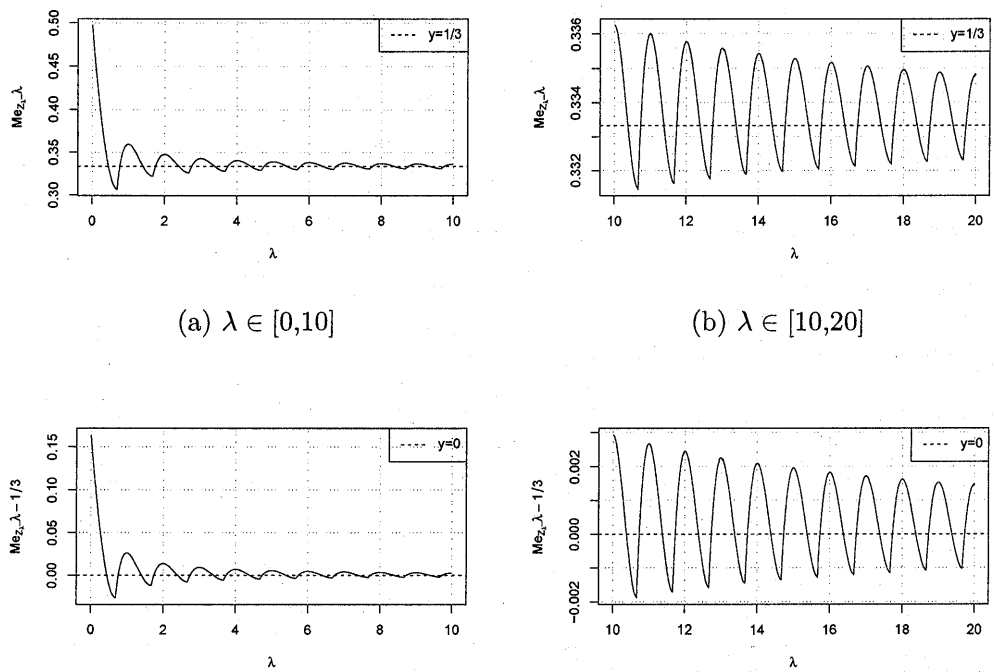
$$\mathrm{Me}_{Z_\lambda} = \lambda + \frac{1}{3} + \frac{\mathcal{H}(\lambda - \lfloor \lambda \rfloor)}{\lambda} + o\left(\frac{1}{\lambda}\right)$$

où la fonction $\mathcal{H} : [0,1] \to \mathbb{R}$ est définie par

$$\mathcal{H}(x) = \begin{cases} \frac{x^2(x-1)}{3} + \frac{4}{135} & , \ si\ x \in [0,2/3] \\ \frac{x}{3}(x^2 - 4x + 5) - \frac{86}{135} & , \ si\ x \in [2/3,1]. \end{cases}$$

La démonstration du Théorème 2.3.7, sectionnée en plusieurs lemmes, est présentée ci-après. Comparant ce résultat avec l'Équation (2.2), notre résultat implique entre autres que

$$\limsup_{\lambda \to \infty} (\mathrm{Me}_{Z_\lambda} - \lambda - 1/3) = \liminf_{\lambda \to \infty} (\mathrm{Me}_{Z_\lambda} - \lambda - 1/3) = 0.$$



(a) $\lambda \in [0,10]$ (b) $\lambda \in [10,20]$

(c) $\lambda \in [0,10]$ (d) $\lambda \in [10,20]$

Figure 2.4: Écarts entre la médiane et la moyenne d'une loi de Poisson perturbée en fonction de λ.

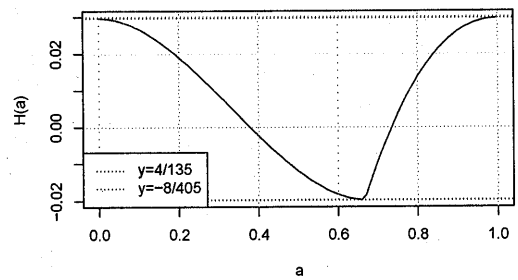


Figure 2.5: $\mathcal{H}(a)$ en fonction de a.

D'ailleurs, il implique de façon plus forte que

$$\liminf_{\lambda\to\infty} (\mathrm{Me}_{Z_\lambda} - \lambda - 1/3)\,\lambda = -\frac{8}{405} \quad \text{et} \quad \limsup_{\lambda\to\infty} (\mathrm{Me}_{Z_\lambda} - \lambda - 1/3)\,\lambda = \frac{4}{135}$$

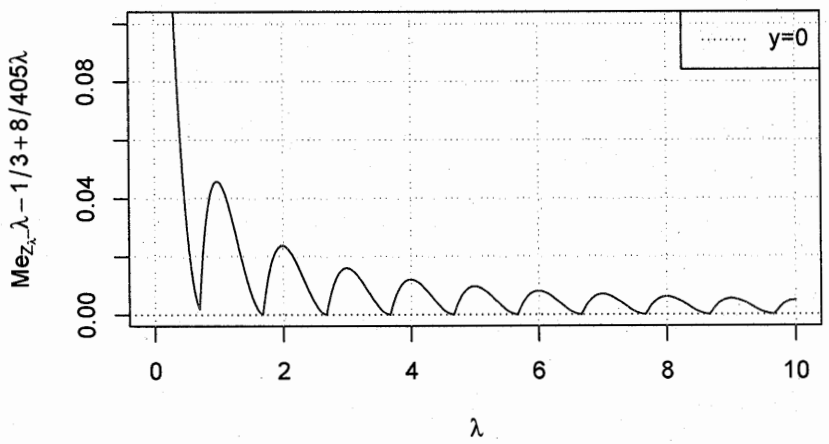
puisque la fonction $\mathcal{H}(a)$ est bornée sur [0,1) tel qu'illustré à la Figure 2.5

$$\frac{-8}{405} \leq \mathcal{H}(a) \leq \frac{4}{135}.$$

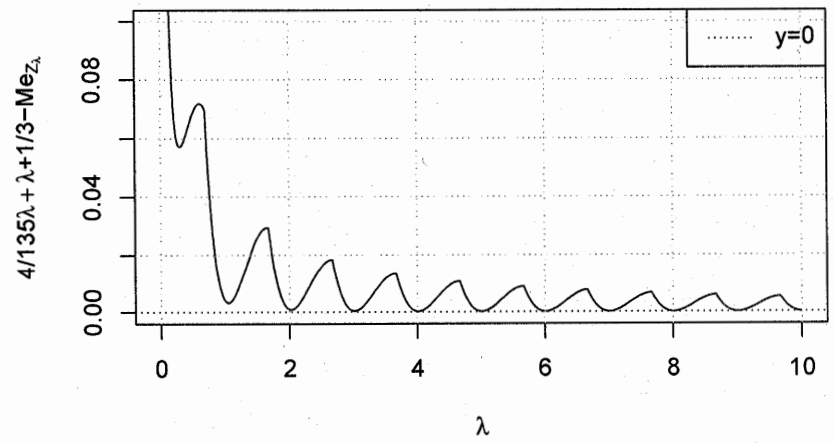
Enfin, on ajoute que l'on conjecture le résultat vrai pour $\varepsilon = 0$ et $\forall\lambda > 0$, soit

$$\frac{-8}{405\lambda} \leq \mathrm{Me}_{Z_\lambda} - \lambda - 1/3 \leq \frac{4}{135\lambda}. \tag{2.3}$$

Ce résultat plus fin, illustré à la Figure 2.6, n'a pas été démontré.

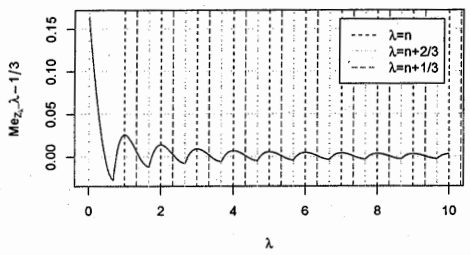


(a) $\mathrm{Me}_{Z_\lambda} - \lambda - 1/3 + 8/405\lambda$ en fonction de λ

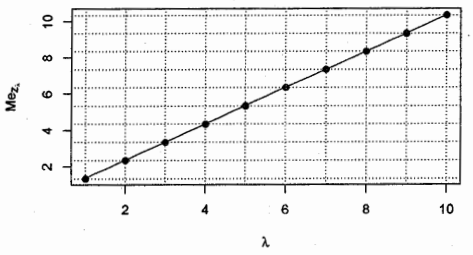


(b) $4/135\lambda + \lambda + 1/3 - \mathrm{Me}_{Z_\lambda}$ en fonction de λ

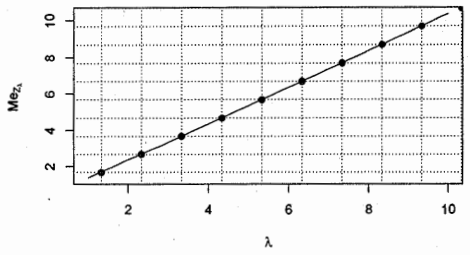
Figure 2.6: Écarts entre la médiane d'une loi de Poisson perturbée et sa borne inférieure et supérieure respectivement en fonction de λ.



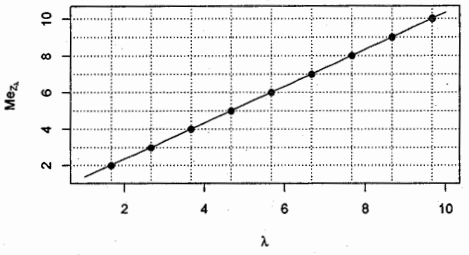
(a) Motif observé



(b) Observations aux $\lambda = n$



(c) Observations aux $\lambda = n + 1/3$



(d) Observations aux $\lambda = n + 2/3$

Figure 2.7: Médiane en fonction de λ où $n = 1,2,\ldots,10$.

En observant la Figure 2.7a, on s'aperçoit d'un certain motif. En effet, on remarque que la fonction $\text{Me}_{Z_\lambda} - \lambda - 1/3$ atteint un maximum aux entiers, croise l'axe aux entiers additionnés d'un tiers et un minimum aux entiers additionnés de deux tiers. Puis, le motif se répète. Afin de bien saisir ce qui se passe à ces points précis, on a tracé les graphique de la Figure 2.7. Par la Figure 2.7b, on voit bien qu'aux entiers, la médiane est très proche de l'entier additionné d'un tiers. Par la Figure 2.7c, on voit bien qu'aux entiers additionnés d'un tiers, la médiane est très proche de l'entier auquel on soustrait un tiers. Finalement, par la Figure 2.7d on voit bien qu'aux entiers additionnés de deux tiers, la médiane est très proche de l'entier supérieur.

Par la suite, l'idée est de trouver une suite $r_n(a)$ qui fait en sorte que la probabilité $\mathrm{P}\left(Z_{n+a} \leq n+a+\frac{1}{3}+r_n(a)\right)$ converge vers $1/2$ par le théorème central limite. On commence d'abord par bien définir cette suite et ensuite regarder l'écart entre deux de ses termes successifs afin de distinguer quand la suite est croissante ou plutôt décroissante. Ainsi, on pourra trouver une borne inférieure pour l'écart entre la médiane et la moyenne quand la suite est croissante vers $1/2$ et une borne supérieure lorsqu'elle est plutôt décroissante vers $1/2$. Cette stratégie est exploitée dans le résultat suivant qui implique le Théorème 2.3.7.

Théorème 2.3.8. *Soit n un entier et $a \in [0,1)$, alors pour tout $\varepsilon > 0$, il existe $n_0 \in \mathbb{N}$ tel que pour tout $n \geq n_0$, la médiane d'une loi de Poisson perturbée est bornée ainsi*

$$\left|\mathrm{Me}_{Z_{n+a}} - (n+a) - \frac{1}{3} - \frac{\mathcal{H}(a)}{n+a}\right| \leq \frac{\varepsilon}{n+a}$$

où

$$\mathcal{H}(a) = \begin{cases} \frac{a^2(a-1)}{3} + \frac{4}{135}, & si\ a \in [0,2/3] \\ \frac{a}{3}(a^2-4a+5) - \frac{86}{135}, & si\ a \in [2/3,1]. \end{cases}$$

Démonstration.

Soit la suite $w_n(a) = \mathrm{P}\left(Z_{n+a} \leq n+a+\frac{1}{3}+r_n(a)\right)$ où $r_n(a) = o(1)$. On peut démontrer que $w_n(a) \to \frac{1}{2}$ lorsque $n \to \infty$, $\forall a \in [0,1)$. En effet, par la définition de Z_{n+a},

$$Z_{n+a} - \left(n+a+\frac{1}{3}+r_n(a)\right) \overset{d}{=} \sum_{i=1}^{n}(Y_i-1) + (N_a + U - a - 1/3 - r_n(a))$$

où $\overset{d}{=}$ désigne l'égalité en distribution et où $Y_1, \ldots, Y_n$ sont des variables aléatoires i.i.d. de loi $\mathcal{P}(1)$ et $N_a \sim \mathcal{P}(a)$.

Ensuite, par le théorème central limite, on a que

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i-1) \xrightarrow{d} \mathcal{N}(0,1).$$

En outre,

$$\frac{N_a + U - a - 1/3 - r_n(a)}{\sqrt{n}} \xrightarrow{P} 0$$

où $\xrightarrow{P}$ désigne la convergence en probabilité. Ainsi, d'après le théorème de Slutsky

$$\frac{Z_{n+a} - (n + a + 1/3 + r_n(a))}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

lorsque $n \to \infty$ et donc

$$\lim_{n\to\infty} w_n(a) = \lim_{n\to\infty} \mathrm{P}(Z \leq 0) = \frac{1}{2}$$

où $Z \sim \mathcal{N}(0,1)$

L'idée est d'ajuster $r_n(a)$ pour montrer que $(w_n(a))_{n\geq 1}$ est croissante (respectivement décroissante) à partir d'un certain rang n_0 afin d'obtenir :

$$w_n(a) \leq w_{n+1}(a) \leq \frac{1}{2} \Leftrightarrow \mathrm{Me}_{Z_{n+a}} \geq n + a + \frac{1}{3} + r_n(a) \tag{2.4}$$

et, respectivement,

$$w_n(a) \geq w_{n+1}(a) \geq \frac{1}{2} \Leftrightarrow \mathrm{Me}_{Z_{n+a}} \leq n + a + \frac{1}{3} + r_n(a). \tag{2.5}$$

On commence par partitionner l'intervalle [0,1]. Pour tout $\varepsilon > 0$, il existe $n_0 \in \mathbb{N}$ tel que pour tout $n \geq 0$

$$a + r_n(a) \in [-1/3,2/3) \qquad \text{ou} \qquad a + r_n(a) \in [2/3,5/3)$$

en fonction du signe de k où $r_n(a) = \frac{k(1+\varepsilon)}{n+a}$. Par la suite, on note n_0 ce rang et on considère les cas

— $a + r_{n_0}(a) \in [-1/3,2/3)$

— $a + r_{n_0}(a) \in [2/3,5/3)$.

Ainsi, quel que soit le signe de k, on couvre tous les cas possibles de $a \in [0,1)$.

D'après le Lemme 2.3.12, $\text{sign}(w_{n+1}(a)-w_n(a)) = \text{sign}(\Delta_n(a))$ où $\Delta_n(a)$ est donné par

$$\Delta_n(a) = \frac{(n+1)!}{g_n(n+1+a)}\left(w_{n+1}(a) - w_n(a)\right)$$

où $g_n(u) = u^n e^{-u}$. Selon le cas, $\Delta_n(a)$ prend les formes suivantes

— si $a + r_{n_0}(a) \in [-1/3,2/3)$ alors

$$\Delta_n(a) = \frac{3}{2(n+1+a)^2}\left(\frac{a^2(a-1)}{3} + \frac{4}{135} - k\right) + o\left(\frac{1}{n^2}\right).$$

— si $a + r_{n_0}(a) \in [2/3,5/3)$ alors

$$\Delta_n(a) = \frac{3}{2(n+1+a)^2}\left(\frac{a}{3}(a^2 - 4a + 5) - \frac{86}{135} - k\right) + o\left(\frac{1}{n^2}\right).$$

Alors, on peut réécrire $\Delta_n(a)$ ainsi

$$\Delta_n(a) = \frac{3}{2(n+1+a)^2}\left(\mathcal{H}(a) - k\right) + o\left(\frac{1}{n^2}\right).$$

Les Lemmes 2.3.9 à 2.3.14 visent spécifiquement à étudier la monotonie de la suite $(w_n(a))_{n\geq 1}$. On obtient les cas suivants

(i) Soit $a \in [0,2/3)$, dans ce cas, à partir d'un certain rang n_0, $a + r_{n_0}(a) \in [-1/3,2/3)$ $\forall k$. Ainsi, en choisissant

  (a) $k = \mathcal{H}(a) - \varepsilon$ alors $\text{sign}(\Delta_n(a)) > 0$ à partir d'un certain rang

  (b) $k = \mathcal{H}(a) + \varepsilon$ alors $\text{sign}(\Delta_n(a)) < 0$ à partir d'un certain rang.

(ii) Soit $a \in (2/3,1)$, dans ce cas, à partir d'un certain rang n_0, $a + r_{n_0}(a) \in [2/3,5/3)$ $\forall k$. Ainsi, à nouveau en choisissant

(a) $k = \mathcal{H}(a) - \varepsilon$ alors $\text{sign}(\Delta_n(a)) > 0$ à partir d'un certain rang

(b) $k = \mathcal{H}(a) + \varepsilon$ alors $\text{sign}(\Delta_n(a)) < 0$ à partir d'un certain rang.

(iii) Finalement, si $a = 2/3$

(a) $k = \frac{-8(1+\varepsilon)}{405}$, $\text{sign}(\Delta_n(2/3)) > 0$

(b) $k = \frac{4(1+\varepsilon)}{135}$, $\text{sign}(\Delta_n(2/3)) < 0$.

Ainsi, à partir d'un certain rang, $\forall a \in [0,1)$ et en choisissant

- $r_n(a) = \frac{\mathcal{H}(a)-\varepsilon}{n+a}$ alors la suite $(w_n(a))_{n\geq 1}$ est croissante
- $r_n(a) = \frac{\mathcal{H}(a)+\varepsilon}{n+a}$ alors la suite $(w_n(a))_{n\geq 1}$ est décroissante,

et on conclut, d'après (2.4) et (2.5)

$$-\frac{8}{405}\frac{1+\varepsilon}{n+a} \leq \frac{\mathcal{H}(a)-\varepsilon}{n+a} \leq \text{Me}(Z_{n+a}) - (n+a) - \frac{1}{3} \leq \frac{\mathcal{H}(a)+\varepsilon}{n+a} \leq \frac{4}{135}\frac{1+\varepsilon}{n+a}.$$

Ces bornes sont illustrées à la Figure 2.8. □

Les lemmes suivants sont utilisés dans la démonstration du théorème précédent. Ils reposent essentiellement sur de fastidieux développements limités que le lecteur peut éventuellement s'abstenir de lire.

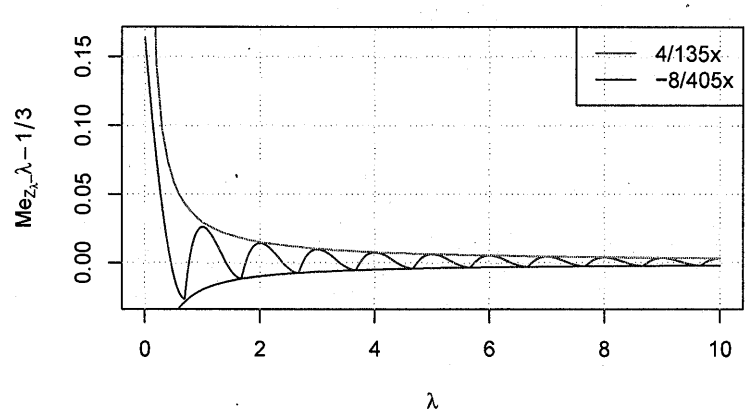
Lemme 2.3.9. *Soit $a \in [0,1)$, $r_n(a) = \frac{k}{n+a}$ pour une certaine constante k et la suite $w_n(a) := \text{P}\left(Z_{n+a} \leq n + a + \frac{1}{3} + r_n(a)\right)$,*

(i) Si $a + r_n(a) \in [-1/3,2/3)$, alors

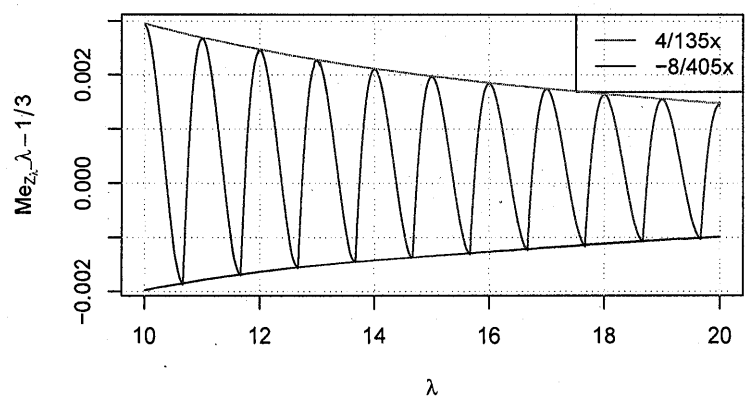
$$w_n(a) = \text{P}\left(N_{n+a} \leq n\right) + \left(a - \frac{2}{3} + r_n(a)\right)\text{P}\left(N_{n+a} = n\right).$$

(ii) Si $a + r_n(a) \in [2/3,5/3)$, alors

$$w_n(a) = \text{P}\left(N_{n+a} \leq n\right) + \left(a - \frac{2}{3} + r_n(a)\right)\text{P}(N_{n+a} = n+1).$$



(a) $\lambda \in [0,10]$



(b) $\lambda \in [10,20]$

Figure 2.8: Bornes pour l'écart entre la médiane et la moyenne.

Démonstration.

(i) Supposons d'abord que $a + r_n(a) \in [-1/3,2/3)$,

$$\begin{aligned} w_n(a) &= \mathrm{P}\left(Z_{n+a} \leq n + a + \frac{1}{3} + r_n(a)\right) \\ &= \mathrm{P}\left(N_{n+a} \leq \left\lfloor n + a + \frac{1}{3} + r_n(a)\right\rfloor - 1\right) \\ &\quad + \left(n + a + \frac{1}{3} + r_n(a) - \left\lfloor n + a + \frac{1}{3} + r_n(a)\right\rfloor\right) \\ &\quad \mathrm{P}\left(N_{n+a} = \left\lfloor n + a + \frac{1}{3} + r_n(a)\right\rfloor\right) \end{aligned}$$

par définition de la fonction de répartition de Z_{n+a}. Ainsi,

$$\begin{aligned} w_n(a) &= \mathrm{P}\left(N_{n+a} \leq n - 1\right) + \left(a + \frac{1}{3} + r_n(a)\right) \mathrm{P}\left(N_{n+a} = n\right) \\ &= \mathrm{P}\left(N_{n+a} \leq n\right) + \left(a - \frac{2}{3} + r_n(a)\right) \mathrm{P}\left(N_{n+a} = n\right). \end{aligned}$$

(ii) Supposons maintenant que $a + r_n(a) \in [2/3,5/3)$,

$$\begin{aligned} w_n(a) &= \mathrm{P}\left(Z_{n+a} \leq n + a + \frac{1}{3} + r_n(a)\right) \\ &= \mathrm{P}\left(N_{n+a} \leq \left\lfloor n + a + \frac{1}{3} + r_n(a)\right\rfloor - 1\right) \\ &\quad + \left(n + a + \frac{1}{3} + r_n(a) - \left\lfloor n + a + \frac{1}{3} + r_n(a)\right\rfloor\right) \\ &\quad \mathrm{P}\left(N_{n+a} = \left\lfloor n + a + \frac{1}{3} + r_n a\right\rfloor\right) \end{aligned}$$

par définition de la fonction de répartition de Z_{n+a}. Ainsi,

$$\begin{aligned} w_n(a) &= \mathrm{P}\left(N_{n+a} \leq n + 1 - 1\right) + \left(n + a + \frac{1}{3} + r_n(a) - (n+1)\right) \\ &\quad \times \mathrm{P}(N_{n+a} = n + 1) \\ &= \mathrm{P}\left(N_{n+a} \leq n\right) + \left(a - \frac{2}{3} + r_n(a)\right) \mathrm{P}(N_{n+a} = n + 1). \end{aligned}$$

□

Lemme 2.3.10. *Soit* $a \in [0,1)$, $r_n(a) = \frac{k}{n+a}$ *pour une certaine constante* k, $g_n(u) = u^n e^{-u}$ *et la suite* $w_n(a) := \mathrm{P}\left(Z_{n+a} \leq n + a + \frac{1}{3} + r_n(a)\right)$, *l'écart entre deux termes successifs de la suite est donné par*

(i) Si $a + r_n(a) \in [-1/3, 2/3)$, *alors*

$$\begin{aligned} w_{n+1}(a) - w_n(a) = \frac{g_{n+1}(n+1+a)}{(n+1)!} \Bigg[& \int_0^1 (1 - c_n(v))dv + (c_n(0) - 1) \\ & + \left(a - \frac{2}{3}\right)\left(1 - c_n(0)\frac{n+1}{n+a}\right) \\ & + r_{n+1}(a) - \frac{n+1}{n+a} c_n(0) r_n(a) \Bigg]. \end{aligned}$$

(ii) Si $a + r_n(a) \in [2/3, 5/3)$, *alors*

$$\begin{aligned} w_{n+1}(a) - w_n(a) = \frac{g_{n+1}(n+1+a)}{(n+1)!} \Bigg[& \int_0^1 (1 - c_n(v))dv + (c_n(0) - 1) \\ & + \left(a - \frac{2}{3} + r_{n+1}(a)\right)\left(\frac{n+1+a}{n+2}\right) \\ & - \left(a - \frac{2}{3} + r_n(a)\right) c_n(0) \Bigg]. \end{aligned}$$

Démonstration.

(i)

On suppose $a + r_n(a) \in [-1/3,2/3)$, alors en utilisant le Lemme 2.3.9(i), on a

$$\begin{aligned}
w_{n+1}(a) - w_n(a) &= \mathrm{P}\left(N_{n+1+a} \leq n+1\right) - \mathrm{P}\left(N_{n+a} \leq n\right) \\
&\quad + \left(a - \frac{2}{3} + r_{n+1}(a)\right) \mathrm{P}\left(N_{n+1+a} = n+1\right) \\
&\quad - \left(a - \frac{2}{3} + r_n(a)\right) \mathrm{P}\left(N_{n+a} = n\right) \\
&= \mathrm{P}\left(N_{n+1+a} \leq n+1\right) - \left(\mathrm{P}\left(N_{n+a} \leq n+1\right) - \mathrm{P}\left(N_{n+a} = n+1\right)\right) \\
&\quad + \left(a - \frac{2}{3} + r_{n+1}(a)\right) \mathrm{P}\left(N_{n+1+a} = n+1\right) \\
&\quad - \left(a - \frac{2}{3} + r_n(a)\right) \mathrm{P}\left(N_{n+a} = n\right) \\
&= \frac{1}{(n+1)!} \int_{n+1+a}^{\infty} \mathrm{e}^{-u} u^{n+1} du - \frac{1}{(n+1)!} \int_{n+a}^{\infty} \mathrm{e}^{-u} u^{n+1} du \\
&\quad + \frac{e^{-(n+a)}(n+a)^{n+1}}{(n+1)!} + \left(a - \frac{2}{3} + r_{n+1}(a)\right) \mathrm{P}\left(N_{n+1+a} = n+1\right) \\
&\quad - \left(a - \frac{2}{3} + r_n(a)\right) \mathrm{P}\left(N_{n+a} = n\right) \\
&= \frac{1}{(n+1)!} \int_{n+a}^{n+1+a} -g_{n+1}(u) du + \frac{g_{n+1}(n+a)}{(n+1)!} \\
&\quad + \left(a - \frac{2}{3} + r_{n+1}(a)\right) \mathrm{P}\left(N_{n+1+a} = n+1\right) \\
&\quad - \left(a - \frac{2}{3} + r_n(a)\right) \mathrm{P}\left(N_{n+a} = n\right).
\end{aligned}$$

On sait que

$$\begin{aligned}
\mathrm{P}(N_{n+a} = n) &= \frac{(n+a)^n \mathrm{e}^{-(n+a)}}{n!} \\
&= \frac{(n+1)(n+a)^n \mathrm{e}^{-(n+a)}}{(n+1)!} = \frac{(n+1) g_n(n+a)}{(n+1)!}
\end{aligned}$$

alors

$$w_{n+1}(a) - w_n(a) = \frac{g_{n+1}(n+1+a)}{(n+1)!}\left[\int_{n+a}^{n+1+a} \frac{-g_{n+1}(u)}{g_{n+1}(n+1+a)}du \right.$$
$$+ \frac{g_{n+1}(n+a)}{g_{n+1}(n+1+a)} + \left(a - \frac{2}{3} + r_{n+1}(a)\right)$$
$$\left. - \left(a - \frac{2}{3} + r_n(a)\right)\frac{(n+1)g_n(n+a)}{g_{n+1}(n+1+a)}\right].$$

Posons

$$c_n(v) = \frac{g_{n+1}(n+a+v)}{g_{n+1}(n+1+a)} = \left(\frac{n+a+v}{n+1+a}\right)^{n+1} \mathrm{e}^{1-v}$$

$$w_{n+1}(a) - w_n(a) = \frac{g_{n+1}(n+1+a)}{(n+1)!}\left[\int_0^1 -c_n(v)dv + c_n(0)\right.$$
$$+ \left(a - \frac{2}{3} + r_{n+1}(a)\right)$$
$$\left. - \left(a - \frac{2}{3} + r_n(a)\right)\frac{(n+1)c_n(0)}{n+a}\right]$$
$$= \frac{g_{n+1}(n+1+a)}{(n+1)!}\left[\int_0^1 (1 - c_n(v))dv + (c_n(0) - 1)\right.$$
$$+ \left(a - \frac{2}{3}\right)\left(1 - c_n(0)\frac{n+1}{n+a}\right)$$
$$\left. + r_{n+1}(a) - \frac{n+1}{n+a}c_n(0)r_n(a)\right].$$

(ii)

On suppose maintenant que $a + r_n(a) \in [2/3,5/3)$, alors en utilisant le Lemme 2.3.9(ii), on a

$$
\begin{aligned}
w_{n+1}(a) - w_n(a) &= \mathrm{P}\left(N_{n+1+a} \leq n+1\right) - \left(\mathrm{P}\left(N_{n+a} \leq n+1\right) - \mathrm{P}\left(N_{n+a} = n+1\right)\right) \\
&\quad + \left(a - \frac{2}{3} + r_{n+1}(a)\right) \mathrm{P}\left(N_{n+1+a} = n+2\right) \\
&\quad - \left(a - \frac{2}{3} + r_n(a)\right) \mathrm{P}\left(N_{n+a} = n+1\right) \\
&= \frac{1}{(n+1)!} \int_{n+1+a}^{\infty} \mathrm{e}^{-u} u^{n+1} du - \frac{1}{(n+1)!} \int_{n+a}^{\infty} \mathrm{e}^{-u} u^{n+1} du \\
&\quad + \frac{e^{-(n+a)}(n+a)^{n+1}}{(n+1)!} + \left(a - \frac{2}{3} + r_{n+1}(a)\right) \mathrm{P}\left(N_{n+1+a} = n+2\right) \\
&\quad - \left(a - \frac{2}{3} + r_n(a)\right) \mathrm{P}\left(N_{n+a} = n+1\right).
\end{aligned}
$$

On sait que

$$
\begin{aligned}
\mathrm{P}(N_{n+1+a} = n+2) &= \frac{(n+1+a)^{n+2}\mathrm{e}^{-(n+1+a)}}{(n+2)!} \\
&= \frac{1}{(n+2)(n+1)!}(n+1+a)^{n+1}\mathrm{e}^{-(n+1+a)}(n+1+a) \\
&= \frac{1}{(n+1)!} g_{n+1}(n+1+a)\frac{n+1+a}{(n+2)}
\end{aligned}
$$

et

$$
\begin{aligned}
\mathrm{P}(N_{n+1+a} = n+1) &= \frac{(n+a)^{n+1}\mathrm{e}^{-(n+a)}}{(n+1)!} \\
&= \frac{(n+1+a)^{n+1}}{(n+1)!}(n+a)^{n+1}\mathrm{e}^{-(n+1+a)}\mathrm{e}\frac{1}{(n+1+a)^{n+1}} \\
&= \frac{(n+1+a)^{n+1}}{(n+1)!}\mathrm{e}^{-(n+1+a)}\left(\frac{n+a}{n+1+a}\right)^{n+1}\mathrm{e} \\
&= \frac{1}{(n+1)!} g_{n+1}(n+1+a)c_n(0).
\end{aligned}
$$

Alors, on obtient

$$
\begin{aligned}
w_{n+1}(a) - w_n(a) &= \frac{1}{(n+1)!}\int_{n+a}^{n+1+a} -g_{n+1}(u)du + \frac{g_{n+1}(n+a)}{(n+1)!} \\
&\quad + \left(a - \frac{2}{3} + r_{n+1}(a)\right) \mathrm{P}\left(N_{n+1+a} = n+2\right) \\
&\quad - \left(a - \frac{2}{3} + r_n(a)\right) \mathrm{P}\left(N_{n+a} = n+1\right) \\
&= \frac{g_{n+1}(n+1+a)}{(n+1)!}\left[\int_{n+a}^{n+1+a} \frac{-g_{n+1}(u)}{g_{n+1}(n+1+a)}du + \frac{g_{n+1}(n+a)}{g_{n+1}(n+1+a)}\right. \\
&\quad \left. + \left(a - \frac{2}{3} + r_{n+1}(a)\right)\left(\frac{n+1+a}{n+2}\right) - \left(a - \frac{2}{3} + r_n(a)\right) c_n(0)\right] \\
&= \frac{g_{n+1}(n+1+a)}{(n+1)!}\left[\int_0^1 -c_n(v)dv + c_n(0)\right. \\
&\quad \left. + \left(a - \frac{2}{3} + r_{n+1}(a)\right)\left(\frac{n+1+a}{n+2}\right) - \left(a - \frac{2}{3} + r_n(a)\right) c_n(0)\right] \\
&= \frac{g_{n+1}(n+1+a)}{(n+1)!}\left[\int_0^1 (1 - c_n(v))dv + (c_n(0) - 1)\right. \\
&\quad \left. + \left(a - \frac{2}{3} + r_{n+1}(a)\right)\left(\frac{n+1+a}{n+2}\right) - \left(a - \frac{2}{3} + r_n(a)\right) c_n(0)\right].
\end{aligned}
$$

□

Lemme 2.3.11. *Pour tout $v \in [0,1]$ et $a \in [0,1)$, on a que*

(i)

$$c_n(v) = 1 + \frac{1}{n+1+a}\left(a(1-v) - \frac{(1-v)^2}{2}\right) + \frac{1}{(n+1+a)^2}\left(\frac{(1-v)^2}{2}(a^2+a) - (1-v)^3\left(\frac{a}{2}+\frac{1}{3}\right) + \frac{(1-v)^4}{8}\right) + o\left(\frac{1}{n^2}\right) \tag{2.6}$$

et

$$c_n(0) = 1 + \frac{1}{n+1+a}\left(a - \frac{1}{2}\right) + \frac{1}{(n+1+a)^2}\left(\frac{a^2}{2} - \frac{5}{24}\right) + o\left(\frac{1}{n^2}\right). \tag{2.7}$$

(ii)

$$\int_0^1 c_n(v)dv = 1 + \frac{1}{n+1+a}\left(\frac{a}{2} - \frac{1}{6}\right) + \frac{1}{(n+1+a)^2}\left(\frac{a^2}{6} + \frac{a}{24} - \frac{7}{120}\right) + o\left(\frac{1}{n^2}\right).$$

(iii) Soit $\rho'_n = \frac{n+1}{n+a}$, alors

$$\rho'_n = 1 + \frac{1-a}{n+1+a} + \frac{1-a}{(n+1+a)^2} + o\left(\frac{1}{n^2}\right).$$

(iv)

$$1 - c_n(0)\rho'_n = \frac{-1/2}{n+1+a} + \frac{1}{(n+1+a)^2}\left[\frac{a^2}{2} - \frac{a}{2} - \frac{7}{24}\right] + o\left(\frac{1}{n^2}\right).$$

Démonstration.

(i) Par des développements limités, on a

$$
\begin{aligned}
c_n(v) &= \left(\frac{n+a+v}{n+1+a}\right)^{n+1} \mathrm{e}^{1-v} \\
&= \left(\frac{n+1+a+v-1}{n+1+a}\right)^{n+1} \mathrm{e}^{1-v} \\
&= \left(1+\frac{1-v}{n+1+a}\right)^{n+1} \mathrm{e}^{1-v} \\
&= \exp\left\{(n+1)\log\left(\frac{1-v}{n+1+a}\right)\right\}\exp(1-v) \\
&= \exp\left\{(n+1)\left[-\frac{(1-v)}{n+1+a}-\frac{(1-v)^2}{2(n+1+a)^2}-\frac{(1-v)^3}{3(n+1+a)^3}+o\left(\frac{1}{n^3}\right)\right]\right\} \\
&\qquad \times \exp(1-v) \\
&= \exp\left\{(1-v)-\frac{(n+1)(1-v)}{n+1+a}\right\} \\
&\qquad \times \exp\left\{-\frac{(1-v)^2}{2(n+1+a)^2}-\frac{(1-v)^3}{3(n+1+a)^3}\right\}\left(1+o\left(\frac{1}{n^2}\right)\right) \\
&= \exp\left\{(1-v)\left[1-\frac{(n+1)}{n+1+a}\right]\right\}\exp\left\{-\frac{(1-v)^2}{2(n+1+a)^2}\right\} \\
&\qquad \times \exp\left\{-\frac{(1-v)^3}{3(n+1+a)^3}\right\}\left(1+o\left(\frac{1}{n^2}\right)\right) \\
&= \left(1+\frac{(1-v)a}{n+1+a}+\frac{(1-v)^2a^2}{2(n+1+a)^2}\right) \\
&\qquad \times \left(1-\frac{(1-v)^2(n+1)}{2(n+1+a)^2}+\frac{(1-v)^4(n+1)^2}{8(n+1+a)^4}\right) \\
&\qquad \times \left(1-\frac{(1-v)^3(n+1)}{3(n+1+a)^3}\right)\left(1+o\left(\frac{1}{n^2}\right)\right) \\
&= 1+\frac{1}{n+1+a}\left((1-v)a-\frac{(1-v)^2}{2}\frac{n+1}{n+1+a}\right) \\
&\qquad +\frac{1}{(n+1+a)^2}\left[-\frac{a(1-v)^3}{2}\frac{n+1}{n+1+a}+\frac{a^2(1-v)^2}{2}\right. \\
&\qquad \left.+\frac{(1-v)^3}{3}\frac{n+1}{n+1+a}+\frac{(1-v)^4}{8}\left(\frac{n+1}{n+1+a}\right)^2\right]+o\left(\frac{1}{n^2}\right).
\end{aligned}
$$

Posons $\rho_n = \frac{n+1}{n+1+a} = 1-\frac{a}{n+1+a}$, alors

$$
\begin{aligned}
c_n(v) &= 1 + \frac{1}{n+1+a}\left((1-v)a - \frac{(1-v)^2}{2}\rho_n\right) \\
&\quad + \frac{1}{(n+1+a)^2}\left[-\frac{a(1-v)^3}{2}\rho_n + \frac{a^2(1-v)^2}{2} + \frac{(1-v)^3}{3}\rho_n\right. \\
&\quad \left. + \frac{(1-v)^4}{8}\rho_n^2\right] + o\left(\frac{1}{n^2}\right) \\
&= 1 + \frac{1}{n+1+a}\left(a(1-v) - \frac{(1-v)^2}{2} + \frac{a(1-v)^2}{2(n+1+a)}\right) \\
&\quad + \frac{1}{(n+1+a)^2}\left(-\frac{a(1-v)^3}{2} + \frac{a^2(1-v)^2}{2} - \frac{(1-v)^3}{3}\right. \\
&\quad \left. + \frac{(1-v)^4}{8}\right) + o\left(\frac{1}{n^2}\right) \\
&= 1 + \frac{1}{n+1+a}\left(a(1-v) - \frac{(1-v)^2}{2}\right) \\
&\quad + \frac{1}{(n+1+a)^2}\left(\frac{(1-v)^2}{2}(a^2+a) - (1-v)^3\left(\frac{a}{2}+\frac{1}{3}\right)\right. \\
&\quad \left. + \frac{(1-v)^4}{8}\right) + o\left(\frac{1}{n^2}\right).
\end{aligned}
$$

On peut en déduire $c_n(0)$.

(ii) En partant de (i), on trouve

$$
\begin{aligned}
\int_0^1 c_n(v)dv &= \int_0^1 \left[1 + \frac{1}{n+1+a}\left(a(1-v) - \frac{(1-v)^2}{2}\right)\right. \\
&\quad + \frac{1}{(n+1+a)^2}\left(\frac{(1-v)^2}{2}(a^2+a) - (1-v)^3\left(\frac{a}{2}+\frac{1}{3}\right)\right. \\
&\quad \left.\left. + \frac{(1-v)^4}{8}\right)\right] dv + o\left(\frac{1}{n^2}\right) \\
&= 1 + \frac{1}{n+1+a}\left(\frac{a}{2} - \frac{1}{6}\right) \\
&\quad + \frac{1}{(n+1+a)^2}\left(\frac{a^2}{6} + \frac{a}{6} - \frac{a}{8} - \frac{1}{12} + \frac{1}{40}\right) + o\left(\frac{1}{n^2}\right) \\
&= 1 + \frac{1}{n+1+a}\left(\frac{a}{2} - \frac{1}{6}\right)) \\
&\quad + \frac{1}{(n+1+a)^2}\left(\frac{a^2}{6} + \frac{a}{24} - \frac{7}{120}\right) + o\left(\frac{1}{n^2}\right)
\end{aligned}
$$

car

$$\int_0^1 (1-v)^k dv = \int_0^1 u^k du = \frac{1}{k+1}.$$

(iii) Par des développements limités, on obtient

$$\begin{aligned}
\rho'_n &= \frac{n+1}{n+a} \\
&= 1 + \frac{1-a}{n+a} \\
&= 1 + \frac{1-a}{n+1+a}\frac{n+1+a}{n+a} \\
&= 1 + \frac{1-a}{n+1+a}\left(\frac{1}{\frac{n+1+a-1}{n+1+a}}\right) \\
&= 1 + \frac{1-a}{n+1+a}\left(\frac{1}{1-\frac{1}{n+1+a}}\right) \\
&= 1 + \frac{1-a}{n+1+a}\left(1 + \frac{1}{n+1+a}\right) + o\left(\frac{1}{n^2}\right) \\
&= 1 + \frac{1-a}{n+1+a} + \frac{1-a}{(n+1+a)^2} + o\left(\frac{1}{n^2}\right).
\end{aligned}$$

(iv) On développe $c_n(0)\rho'_n$ par (i) et (iii)

$$\begin{aligned}
c_n(0)\rho'_n &= \left(1 + \left(\frac{1}{n+1+a}\right)\left(a - \frac{1}{2}\right) + \frac{1}{(n+1+a)^2}\left(\frac{a^2}{2} - \frac{5}{24}\right)\right) \\
&\quad \times \left(1 + \frac{1-a}{n+1+a} + \frac{1-a}{(n+1+a)^2}\right) + o\left(\frac{1}{n^2}\right) \\
&= 1 + \frac{1}{n+1+a}\left[a - \frac{1}{2} + 1 - a\right] + \frac{1}{(n+1+a)^2} \\
&\quad \times \left[\frac{a^2}{2} - \frac{5}{24} + (1-a) + \left(a - \frac{1}{2}\right)(1-a)\right] + o\left(\frac{1}{n^2}\right) \\
&= 1 + \frac{1/2}{n+1+a} + \frac{1}{(n+1+a)^2}\left[\frac{a^2}{2} - \frac{5}{24} + 1 - a - a^2 - \frac{1}{2} + \frac{a}{2}\right] \\
&\quad + o\left(\frac{1}{n^2}\right) \\
&= 1 + \frac{1/2}{n+1+a} + \frac{1}{(n+1+a)^2}\left[-\frac{a^2}{2} + \frac{a}{2} + \frac{7}{24}\right] + o\left(\frac{1}{n^2}\right).
\end{aligned}$$

Par conséquent,

$$1 - c_n(0)\rho'_n = \frac{-1/2}{n+1+a} + \frac{1}{(n+1+a)^2}\left[\frac{a^2}{2} - \frac{a}{2} - \frac{7}{24}\right] + o\left(\frac{1}{n^2}\right)$$

□

Lemme 2.3.12. *Soit* $a \in [0,1)$, $r_n(a) = \frac{k}{n+a}$ *pour une certaine constante* k, $g_n(u) = u^n e^{-u}$, *la suite* $w_n(a) := \mathrm{P}\left(Z_{n+a} \leq n + a + \frac{1}{3} + r_n(a)\right)$ *et* $\Delta_n(a) = \frac{(n+1)!}{g_n(n+1+a)}\left(w_{n+1}(a) - w_n(a)\right)$ *alors à partir d'un certain rang* n_0

(i) Supposons $a + r_n(a) \in [-1/3, 2/3)$, *alors*

$$\Delta_n(a) = \frac{3}{2(n+1+a)^2}\left(\frac{a^2(a-1)}{3} + \frac{4}{135} - k\right) + o\left(\frac{1}{n^2}\right).$$

(ii) Supposons $a + r_n(a) \in [2/3, 5/3)$, *alors*

$$\Delta_n(a) = \frac{3}{2(n+1+a)^2}\left(\frac{a^3}{3} - \frac{4a^2}{3} + \frac{5a}{3} - \frac{86}{135} - k\right) + o\left(\frac{1}{n^2}\right).$$

Démonstration.

(i) Supposons d'abord que $a + r_n(a) \in [-1/3, 2/3)$.

On commence par développer $\Delta_n(a)$ par le résultat 2.3.10(i) en utilisant la notation $\rho'_n = \frac{n+1}{n+a}$.

$$\begin{aligned}
\Delta_n(a) &= \int_0^1 (1 - c_n(v))dv + (c_n(0) - 1) \\
&\quad + \left(a - \frac{2}{3}\right)(1 - c_n(0)\rho'_n) + r_{n+1}(a) - \rho'_n c_n(0) r_n(a) \\
&= \frac{1}{n+1+a}\left(-\frac{a}{2} + \frac{1}{6}\right) + \frac{1}{(n+1+a)^2}\left(-\frac{a^2}{6} - \frac{a}{24} + \frac{7}{120}\right) \\
&\quad + \frac{1}{n+1+a}\left(a - \frac{1}{2}\right) + \frac{1}{(n+1+a)^2}\left(\frac{a^2}{2} - \frac{5}{24}\right) \\
&\quad + \left(a - \frac{2}{3}\right)\left[\frac{-1/2}{n+1+a} + \frac{1}{(n+1+a)^2}\left(\frac{a^2}{2} - \frac{a}{2} - \frac{7}{24}\right)\right] \\
&\quad + r_{n+1}(a) - \rho'_n c_n(0) r_n(a) + o\left(\frac{1}{n^2}\right)
\end{aligned}$$

par le Lemme 2.3.11(ii), l'Équation (2.7) de 2.3.11(i) et 2.3.11(iv).

$$\begin{aligned}\Delta_n(a) &= \frac{1}{n+1+a}\left[-\frac{a}{2}+\frac{1}{6}-a-\frac{1}{2}-\frac{a}{2}+\frac{1}{3}\right] \\ &\quad + \frac{1}{(n+1+a)^2}\left[-\frac{a^2}{6}-\frac{a}{24}+\frac{7}{120}+\frac{a^2}{2}-\frac{5}{24}+\left(a-\frac{2}{3}\right)\right. \\ &\quad \left.\left(\frac{a^2}{2}-\frac{a}{2}-\frac{7}{24}\right)\right] + r_{n+1}(a) - c_n(0)\rho_n' r_n(a) + o\left(\frac{1}{n^2}\right) \\ &= \frac{1}{(n+1+a)^2}\left[\frac{a^3}{2}-\frac{a^2}{2}-\frac{7a}{24}-\frac{a^2}{3}+\frac{a}{3}-\frac{14}{72}-\frac{a^2}{6}-\frac{a}{24}+\frac{7}{120}\right. \\ &\quad \left.+\frac{a^2}{2}-\frac{5}{24}\right] + r_{n+1}(a) - c_n(0)\rho_n' r_n(a) + o\left(\frac{1}{n^2}\right) \\ &= \frac{1}{(n+1+a)^2}\left[\frac{a^3}{2}-\frac{a^2}{2}+\frac{2}{45}\right] + r_{n+1}(a) - c_n(0)\rho_n' r_n(a) + o\left(\frac{1}{n^2}\right).\end{aligned}$$

On développe $\Delta_n(a)$ en utilisant le Lemme 2.3.14(i)

$$\begin{aligned}\Delta_n(a) &= \frac{1}{(n+1+a)^2}\left(\frac{a^3}{2}-\frac{a^2}{2}+\frac{2}{45}-\frac{3k}{2}\right)+o\left(\frac{1}{n^2}\right) \\ &= \frac{3}{2(n+1+a)^2}\left(\frac{a^3}{3}-\frac{a^2}{3}+\frac{4}{135}-k\right)+o\left(\frac{1}{n^2}\right) \\ &= \frac{3}{2(n+1+a)^2}\left(\frac{a^2(a-1)}{3}+\frac{4}{135}-k\right)+o\left(\frac{1}{n^2}\right) \\ &== \frac{3}{2(n+1+a)^2}\left(\mathcal{H}(a)-k\right)+o\left(\frac{1}{n^2}\right).\end{aligned}$$

(ii) On suppose maintenant que $a + r_n(a) \in [2/3{,}5/3]$, en utilisant le Lemme 2.3.10(ii)

$$\begin{aligned}
\Delta_n(a) &= \left(\frac{(n+1)!}{g_n(n+1+a)}\right)(w_{n+1}(a) - w_n(a) \\
&= \int_0^1 (1 - c_n(v))dv + (c_n(0) - 1) + \left(a - \frac{2}{3} + r_{n+1}(a)\right)\left(\frac{n+1+a}{n+2}\right) \\
&- \left(a - \frac{2}{3} + r_n(a)\right) c_n(0) \\
&= \frac{1}{n+1+a}\left(-\frac{a}{2} + \frac{1}{6}\right) + \frac{1}{(n+1+a)^2}\left(-\frac{a^2}{6} - \frac{a}{24} + \frac{7}{120}\right) \\
&\quad + \frac{1}{n+1+a}\left(a - \frac{1}{2}\right) + \frac{1}{(n+1+a)^2}\left(\frac{a^2}{2} - \frac{5}{24}\right) \\
&\quad + \left(a - \frac{2}{3} + r_{n+1}(a)\right)\left(\frac{n+1+a}{n+2}\right) - \left(a - \frac{2}{3} + r_n(a)\right) c_n(0).
\end{aligned}$$

par le Lemme 2.3.11(ii) et l'Équation (2.7) du Lemme 2.3.11(i).

$$\begin{aligned}
\Delta_n(a) &= \frac{1}{n+1+a}\left(-\frac{a}{2} + \frac{1}{6}\right) + \frac{1}{(n+1+a)^2}\left(-\frac{a^2}{6} - \frac{a}{24} + \frac{7}{120}\right) \\
&\quad + \frac{1}{n+1+a}\left(a - \frac{1}{2}\right) + \frac{1}{(n+1+a)^2}\left(\frac{a^2}{2} - \frac{5}{24}\right) \\
&\quad + \left(a - \frac{2}{3}\right)\left(\frac{n+1+a}{n+2}\right) - \left(a - \frac{2}{3}\right) c_n(0) \\
&\quad + r_{n+1}(a)\frac{n+1+a}{n+2} - r_n(a)c_n(0).
\end{aligned}$$

En utilisant les Lemmes 2.3.14(ii) et 2.3.13,

$$\begin{aligned}
\Delta_n(a) &= \frac{1}{(n+1+a)^2}\left(\frac{a^3}{2} - 2a^2 + \frac{5a}{2} - \frac{43}{45} - \frac{3k}{2}\right) + o\left(\frac{1}{n^2}\right) \\
&= \frac{3}{2(n+1+a)^2}\left(\frac{a^3}{3} - \frac{4a^2}{3} + \frac{5a}{3} - \frac{86}{135} - k\right) + o\left(\frac{1}{n^2}\right) \\
&= \frac{3}{2(n+1+a)^2}\left(\mathcal{H}(a) - k\right) + o\left(\frac{1}{n^2}\right)
\end{aligned}$$

□

Lemme 2.3.13. *Soit $a \in [0,1)$ et $c_n(0)$ tel que défini à l'Équation (2.7) du Lemme 2.3.11(i), on a que*

$$\left(a-\frac{2}{3}\right)\left(\frac{n+1+a}{n+2}\right)-\left(a-\frac{2}{3}\right)c_n(0) = \frac{1}{n+1+a}\left(-\frac{a}{2}+\frac{1}{3}\right)+\frac{1}{(n+1+a)^2}\left(\frac{a^3}{2}-\frac{7a^2}{3}+\frac{61a}{24}-\frac{29}{36}\right) + o\left(\frac{1}{n^2}\right).$$

Démonstration.

On utilise le développement limité de $\frac{n+1+a}{n+2}$, soit $1+\frac{a-1}{n+1+a}+\frac{(a-1)^2}{(n+1+a)^2}$

$$\begin{aligned}
&\left(a-\frac{2}{3}\right)\left(\frac{n+1+a}{n+2}\right)-\left(a-\frac{2}{3}\right)c_n(0)\\
&\quad=\left(a-\frac{2}{3}\right)\left(\frac{n+1+a}{n+2}-c_n(0)\right)\\
&\quad=\left(a-\frac{2}{3}\right)\left(1+\frac{a-1}{n+1+a}+\frac{(a-1)^2}{(n+1+a)^2}-1-\frac{a-1/2}{n+1+a}\right.\\
&\qquad\left.-\frac{a^2/2-5/24}{(n+1+a)^2}\right)+o\left(\frac{1}{n^2}\right)\\
&\quad=\left(a-\frac{2}{3}\right)\left(\frac{-1/2}{n+1+a}-\frac{(a-1)^2-a^2/2+5/24}{(n+1+a)^2}\right)+o\left(\frac{1}{n^2}\right)\\
&\quad=\frac{1}{n+1+a}\left(-\frac{a}{2}+\frac{1}{3}\right)+\frac{1}{(n+1+a)^2}\left(a^2-2a+1-\frac{a^2}{2}+\frac{5}{24}\right)\\
&\qquad\left(a-\frac{2}{3}\right)+o\left(\frac{1}{n^2}\right)\\
&\quad=\frac{1}{n+1+a}\left(-\frac{a}{2}+\frac{1}{3}\right)+\frac{1}{(n+1+a)^2}\left(\frac{a^2}{2}-2a+\frac{29}{24}\right)\left(a-\frac{2}{3}\right)\\
&\qquad+o\left(\frac{1}{n^2}\right)\\
&\quad=\frac{1}{n+1+a}\left(-\frac{a}{2}+\frac{1}{3}\right)+\frac{1}{(n+1+a)^2}\left(\frac{a^3}{2}-\frac{a^2}{3}-2a^2+\frac{4a}{3}\right.\\
&\qquad\left.+\frac{29a}{24}-\frac{58}{72}\right)+o\left(\frac{1}{n^2}\right)\\
&\quad=\frac{1}{n+1+a}\left(-\frac{a}{2}+\frac{1}{3}\right)+\frac{1}{(n+1+a)^2}\left(\frac{a^3}{2}-\frac{7a^2}{3}+\frac{61a}{24}-\frac{29}{36}\right)\\
&\qquad+o\left(\frac{1}{n^2}\right).
\end{aligned}$$

□

Lemme 2.3.14. *Soit $a \in [0,1)$ et $r_n(a) = \frac{k}{n+a}$ pour une certaine constante k, $\rho'_n = \frac{n+1}{n+a}$ et $c_n(0)$ tel que défini à l'Équation (2.7) du Lemme 2.3.11(i) alors*

(i) Supposons $a + r_n(a) \in [-1/3,2/3)$, alors

$$r_{n+1}(a) - c_n(0)\rho'_n r_n(a) = \frac{-3k}{2}\frac{1}{(n+1+a)^2} + o\left(\frac{1}{n^2}\right).$$

(ii) Supposons $a + r_n(a) \in [2/3,5/3)$, alors

$$r_{n+1}(a)\frac{n+1+a}{n+2} - r_n(a)c_n(0) + o\left(\frac{1}{n^2}\right) = \frac{-3k}{2}\frac{1}{(n+1+a)^2} + o\left(\frac{1}{n^2}\right).$$

Démonstration.

(i) Supposons $a + r_n(a) \in [-1/3,2/3)$, alors

$$\begin{aligned}
r_{n+1}(a) - c_n(0)\rho'_n r_n(a) &= r_{n+1}(a) - r_n(a) + r_n(a)(1 - c_n(0)\rho'_n) \\
&= \frac{k}{n+1+a} - \frac{k}{n+a} + \frac{k}{n+a}\left(\frac{-1/2}{n+1+a}\right) + o\left(\frac{1}{n^2}\right) \\
&= \frac{k(n+a) - k(n+1+a)}{(n+a)(n+1+a)} + \frac{k}{n+a}\left(\frac{-1/2}{n+1+a}\right) + o\left(\frac{1}{n^2}\right) \\
&= \frac{-k - k/2}{(n+a)(n+1+a)} + o\left(\frac{1}{n^2}\right) \\
&= \frac{-3k/2}{(n+a)(n+1+a)} + o\left(\frac{1}{n^2}\right) \\
&= \frac{-3k/2}{(n+1+a)^2}\frac{n+1+a}{n+a} + o\left(\frac{1}{n^2}\right) \\
&= \frac{-3k}{2}\frac{1}{(n+1+a)^2} + o\left(\frac{1}{n^2}\right).
\end{aligned}$$

(ii) On suppose maintenant $a + r_n(a) \in [2/3,5/3)$. D'abord, on pose $\rho''_n = \frac{n+1+a}{n+2}$,

et par un développement limité, on obtient

$$\begin{aligned}\frac{n+1+a}{n+2} &= \frac{1}{\frac{n+2}{n+1+a}} = \frac{1}{\frac{(n+1+a)+(1-a)}{n+1+a}} \\ &= \frac{1}{1+\frac{1-a}{n+1+a}} = \frac{1}{1-\frac{a-1}{n+1+a}} \\ &= 1+\frac{a-1}{n+1+a}+o\left(\frac{1}{n^2}\right).\end{aligned}$$

Ensuite, on obtient que

$$\begin{aligned}r_{n+1}(a)\rho_n'' - r_n(a)c_n(0) &= \frac{k}{n+1+a}\frac{n+1+a}{n+2} - \frac{k}{n+a}c_n(0) \\ &= \frac{k}{n+1+a}\left(1+\frac{a-1}{n+1+a}\right) \\ &\quad - \frac{k}{n+a}\left(1+\frac{a-1/2}{n+1+a}\right)+o\left(\frac{1}{n^2}\right) \\ &= \frac{k}{n+1+a}+\frac{k(a-1)}{n+1+a}-\frac{k}{n+a} \\ &\quad - \frac{k(a-1/2)}{n+a}+o\left(\frac{1}{n^2}\right) \\ &= \frac{-k}{(n+a)(n+1+a)} \\ &\quad + \frac{k(a-1)(n+a)-k(a-1/2)(n+1+a)}{(n+a)(n+1+a)^2)}+o\left(\frac{1}{n^2}\right) \\ &= \frac{-k(n+1+a)+k(a-1)(n+a)}{(n+a)(n+1+a)^2} \\ &\quad \frac{-k(a-1/2)(n+1+a)}{(n+a)(n+1+a)^2}+o\left(\frac{1}{n^2}\right) \\ &= \frac{(n+a)\left[-k\frac{n+1+a}{n+a}+k(a-1)-k(a-1/2)\frac{n+1+a}{n+a}\right]}{(n+a)(n+1+a)^2)}+o\left(\frac{1}{n^2}\right) \\ &= \frac{-k\frac{n+1+a}{n+a}+k(a-1)-k(a-1/2)\frac{n+1+a}{n+a}}{(n+1+a)^2)}+o\left(\frac{1}{n^2}\right) \\ &= \frac{-k+k(a-1)-k(a-1/2)}{(n+1+a)^2)}+o\left(\frac{1}{n^2}\right) \\ &= \frac{-3k}{2}\frac{1}{(n+1+a)^2}+o\left(\frac{1}{n^2}\right).\end{aligned}$$

□

2.4 Synthèse du chapitre

Suite au Théorème 2.3.7, on peut conclure que $\mathrm{Me}_{Z_\lambda} \approx \lambda + 1/3$. Ainsi, on peut estimer le paramètre λ par $\widehat{\lambda}_J = \widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - 1/3$ qui est en principe asymptotiquement sans biais et satisfait le théorème central limite. Dans le chapitre suivant, on étudie cet estimateur.

CHAPITRE III

ÉTUDE EN SIMULATION

D'abord, on rappelle la forme de notre estimateur. Il consiste en la médiane empirique de réalisations d'une loi de Poisson perturbée à laquelle on soustrait un tiers : $\widehat{\lambda}_J = \widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - 1/3$. Dans ce chapitre, on s'intéresse aux propriétés de notre estimateur : son comportement asymptotique, sa loi limite, le rapport de variance avec l'estimateur de maximum de vraisemlance et son intervalle de confiance. Ensuite, on a réalisé une étude en simulation dans laquelle on compare les biais et les racines des erreurs quadratiques moyennes (RMSE) de plusieurs estimateurs connus et robustes de λ et de $\widehat{\lambda}_J$.

3.1 Propriétés de l'estimateur $\widehat{\lambda}_J$

3.1.1 Comportement asymptotique

On s'est intéressé au comportement aymptotique, mentionné au résultat 2.2.2 du Chapitre 1, de notre estimateur En effet, on se rappelle qu'on peut utiliser la théorie standard pour connaître la loi limite de $\widehat{\mathrm{Me}}(\mathbf{Z}_\lambda)$ puisque Z_λ est la version lissée de la distribution de Poisson. Lorsque $n \to \infty$,

$$\sqrt{n}\left(\widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - \mathrm{Me}_{Z_\lambda}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4\,\mathrm{P}(N_\lambda = \lfloor \mathrm{Me}_{Z_\lambda} \rfloor)^2}\right). \tag{3.1}$$

À la Figure 3.1, on remarque que les simulations, pour $n = 1000, 5000, 10000$ et $\lambda = 10$, vont dans le même sens que les résultats théoriques.

3.1.2 Loi limite pour λ grand

Suivant l'idée de la section précédente, on s'est aussi intéressé à loi limite de la différence entre notre estimateur $\widehat{\lambda}_J$ et le paramètre λ de la loi de Poisson. Soit $\widehat{\lambda}_J = \widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - 1/3$, alors on sait que

$$\sqrt{n}(\widehat{\lambda}_J - \lambda) \sim \mathcal{N}\left(0, \frac{1}{4\,\mathrm{P}(N_\lambda = \lfloor \mathrm{Me}_{Z_\lambda} \rfloor)^2}\right). \tag{3.2}$$

En effet, de l'équation (3.1), on peut déduire l'équation (3.2) puisqu'on a montré que $\mathrm{Me}_{Z_\lambda} \approx \lambda + 1/3$. Si λ est grand, comme présenté à la Figure 3.2, on peut même déduire que

$$\sqrt{n}(\widehat{\lambda}_J - \lambda) \approx \mathcal{N}\left(0, \frac{\pi}{2}\lambda\right).$$

En effet,

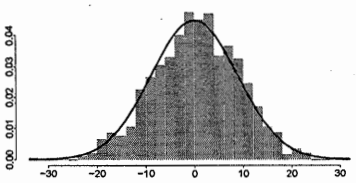
$$\begin{aligned}
\mathrm{P}(N_\lambda = \lfloor \mathrm{Me}_{Z_\lambda} \rfloor) &\approx \mathrm{P}(N_\lambda = \lfloor \lambda + 1/3 \rfloor) \\
&\approx \mathrm{P}(N_\lambda = \lambda) \\
&= \frac{\mathrm{e}^{-\lambda}\lambda^\lambda}{\lambda!} \\
&\sim \frac{\mathrm{e}^{-\lambda}\lambda^\lambda}{\sqrt{2\pi\lambda}(\lambda/\mathrm{e})^\lambda}, \qquad \text{quand } \lambda \to \infty
\end{aligned}$$

par l'approximation de Stirling. Ainsi, on a que

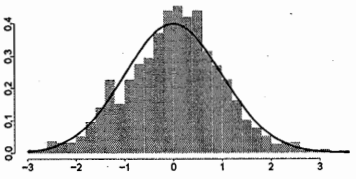
$$\mathrm{P}(N_\lambda = \lfloor \mathrm{Me}_{Z_\lambda} \rfloor)^2 \approx \frac{1}{2\pi\lambda}.$$

3.1.3 Rapport de variances

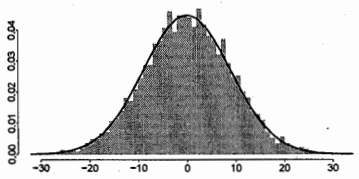
Par la suite, on a aussi voulu regarder le rapport des variances de notre estimateur $\widehat{\lambda}_J$ et de l'estimateur de maximum de vraisemblance obtenu au Chapitre 1, soit



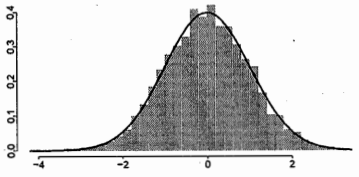
(a) $\sqrt{n}\left(\widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - \mathrm{Me}_{Z_\lambda}\right)$ pour $n = 1000$



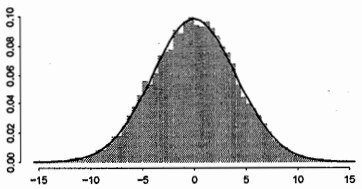
(b) $\sqrt{n}\dfrac{\left(\widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - \mathrm{Me}_{Z_\lambda}\right)}{2\,\mathrm{P}\left(N_\lambda = \lfloor \mathrm{Me}_{Z_\lambda} \rfloor\right)}$ pour $n = 1000$



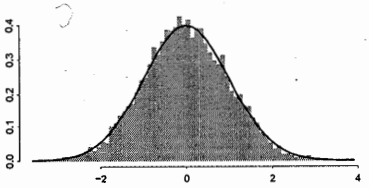
(c) $\sqrt{n}\left(\widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - \mathrm{Me}_{Z_\lambda}\right)$ pour $n = 5000$



(d) $\sqrt{n}\dfrac{\left(\widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - \mathrm{Me}_{Z_\lambda}\right)}{2\,\mathrm{P}\left(N_\lambda = \lfloor \mathrm{Me}_{Z_\lambda} \rfloor\right)}$ pour $n = 5000$

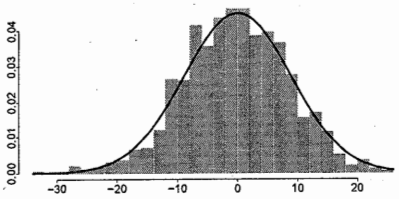


(e) $\sqrt{n}\left(\widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - \mathrm{Me}_{Z_\lambda}\right)$ pour $n = 10000$

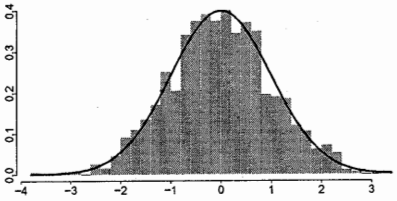


(f) $\sqrt{n}\dfrac{\left(\widehat{\mathrm{Me}}(\mathbf{Z}_\lambda) - \mathrm{Me}_{Z_\lambda}\right)}{2\,\mathrm{P}\left(N_\lambda = \lfloor \mathrm{Me}_{Z_\lambda} \rfloor\right)}$ pour $n = 10000$

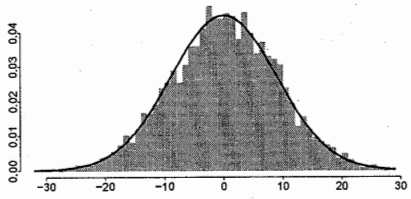
Figure 3.1: Comportement asymptotique de $\widehat{\mathrm{Me}}(\mathbf{Z}_\lambda)$ pour $\lambda = 10$.



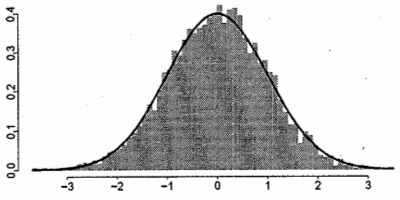
(a) $\sqrt{n}(\widehat{\lambda}_J - \lambda)$ pour $n = 1000$



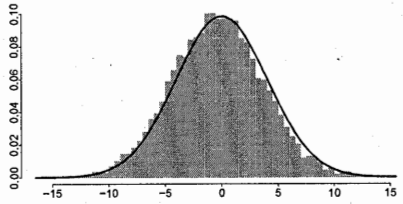
(b) $\frac{\sqrt{n}(\widehat{\lambda}_J - \lambda)}{\sqrt{\frac{\pi\lambda}{2}}}$ pour $n = 1000$



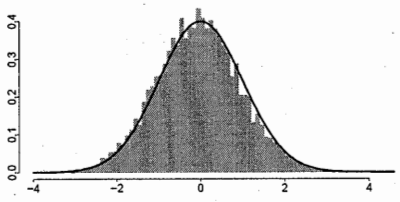
(c) $\sqrt{n}(\widehat{\lambda}_J - \lambda)$ pour $n = 5000$



(d) $\frac{\sqrt{n}(\widehat{\lambda}_J - \lambda)}{\sqrt{\frac{\pi\lambda}{2}}}$pour $n = 5000$



(e) $\sqrt{n}(\widehat{\lambda}_J - \lambda)$ pour $n = 10000$



(f) $\frac{\sqrt{n}(\widehat{\lambda}_J - \lambda)}{\sqrt{\frac{\pi\lambda}{2}}}$ pour $n = 10000$

Figure 3.2: Comportement asymptotique de $\widehat{\lambda}_J$ pour $\lambda = 10$.

la moyenne. Ce résultat est illustré à la Figure 3.3. À nouveau, on peut s'appuyer sur l'approximation de Stirling

$$\begin{aligned}\rho(\lambda) &= \left\{ \lim_{\lambda\to\infty} \frac{\mathrm{Var}(\widehat{\lambda}_J)}{\mathrm{Var}(\widehat{\lambda}_{mle})} \right\}^{1/2} \\ &= \frac{1}{2\sqrt{\lambda}\,\mathrm{P}(N_\lambda = \lfloor \mathrm{Me}_{Z_\lambda} \rfloor)} \sim \sqrt{\frac{\pi}{2}} \approx 1.253.\end{aligned}$$

Autrement dit, $\widehat{\lambda}_J$ a un écart-type de l'ordre 25% plus grand que celui de l'estimateur de maximum de vraisemblance.

3.1.4 Intervalle de confiance

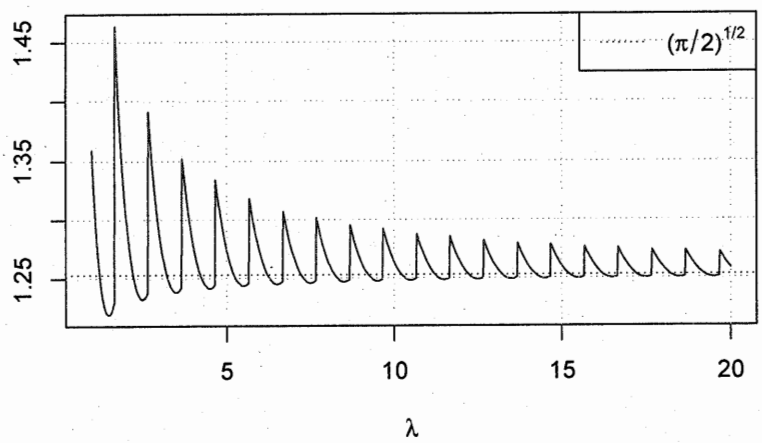
Ainsi, ayant une loi limite pour la différence entre notre estimateur et le paramètre de la loi, on peut bâtir un intervalle de confiance asymptotique pour λ basé sur l'estimateur $\widehat{\lambda}_J$. En effet,

$$\begin{aligned}&\mathrm{P}\left(-z_{\alpha/2} < \frac{\sqrt{n}(\widehat{\lambda}_J - \lambda)}{\sqrt{\frac{\pi\lambda}{2}}} < z_{\alpha/2} \right) = 1 - \alpha \\ \Leftrightarrow &\mathrm{P}\left(\widehat{\lambda}_J - z_{\alpha/2}\sqrt{\frac{\pi\lambda}{2n}} < \lambda < \widehat{\lambda}_J + z_{\alpha/2}\sqrt{\frac{\pi\lambda}{2n}} \right) = 1 - \alpha\end{aligned}$$

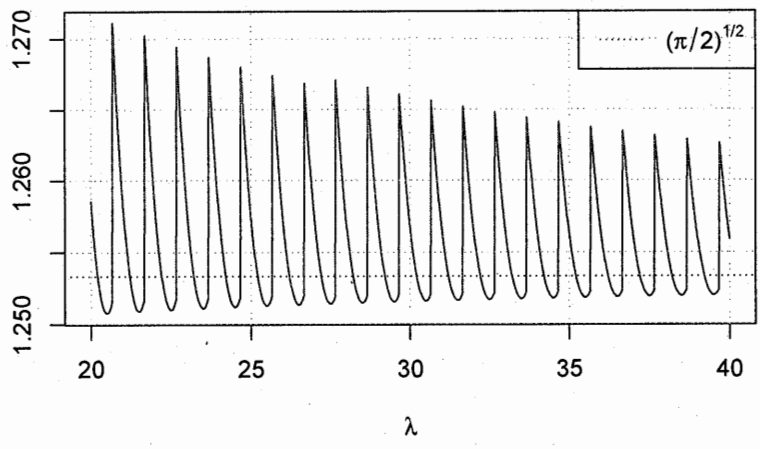
donc, l'intervalle de confiance pour λ grand est donné par :

$$IC_\alpha(\lambda) = \left[\widehat{\lambda}_J - z_{\alpha/2}\sqrt{\frac{\pi\widehat{\lambda}_J}{2n}}, \widehat{\lambda}_J + z_{\alpha/2}\sqrt{\frac{\pi\widehat{\lambda}_J}{2n}} \right].$$

On a réalisé des simulations pour vérifier le taux de couverture de l'intervalle de confiance. La taille de l'échantillon a été fixé à $n = 100$ et le niveau de confiance de l'intervalle à 95%. On peut voir au Tableau 3.1 que l'intervalle de confiance atteint le taux de couverture attendu.



(a) $\lambda \in [1,20]$



(b) $\lambda \in [20,40]$

Figure 3.3: Rapports de variances entre la médiane d'une loi de Poisson perturbée et l'estimateur de maximum de vraisemblance.

Tableau 3.1: Taux de couverture(%) de l'intervalle de confiance de λ basé sur $\widehat{\lambda}_J$ selon différentes valeurs de λ pour m répétitions de $n = 100$ valeurs simulées d'une loi de Poisson perturbée.

	m					
λ	100	1000	2500	5000	10000	100000
2	91	93.0	94.5	93.8	94.1	94.4
5	93	94.9	95.0	95.4	95.2	94.9
10	94 .0	94.5	94.7	94.4	95.1	94.9
50	98.0	95.7	94.8	95.4	94.7	95.3

3.2 Étude en simulation

On procède maintenant à une étude en simulation pour comparer la robustesse des différents estimateurs présentés. Les estimateurs suivants ont été comparés.

1. l'estimateur de maximum de vraisemblance : EMV
2. la médiane empirique : Med
3. l'estimateur $\widehat{\lambda}_J$ basé sur la médiane empirique de la loi de Poisson perturbée : MedPert
4. la moyenne tronquée adaptative : TrimMean
5. l'estimateur de Tukey modifié de Elsaied et Fried (2016) : TukeyPois
6. l'estimateur de Tukey modifié initialisé avec la médiane perturbée : TukeyPoisMedPert
7. l'estimateur de Huber implémenté : glm
8. l'estimateur de vraisemblance pondéré : wle.

On a utilisé le code de Elsaied *et al.* (2012) pour l'implémentation de la moyenne tronquée adaptative et de l'estimateur de Tukey modifié. Ces deux estimateurs

sont initialisés par la fonction *poismall* qui prend des données et calcule un estimateur initial basé sur la fréquence relative de petites valeurs, telles que zéro et un, ou de la médiane empirique. En effet, si la médiane empirique n'est pas petite, on l'utilise comme valeur initiale puisqu'elle correspond approximativement à la moyenne. On a aussi modifié cette initialisation en utilisant l'estimateur $\widehat{\lambda}_J$ plutôt que la médiane ; cette procédure est appellée *TukeyPoisMedPert*. Sinon, on initialise par $-\ln \hat{f}_0$ où $\hat{f}_0$ est la fréquence relative soit de zéros ou de zéros et de uns. Pour l'estimateur de Tukey modifié, on doit aussi calculer un terme correctif $a = a(\lambda,k)$ qui dépend de la moyenne λ et d'une certaine constante à ajuster, k. Pour y parvenir, on choisit un certain éventail de valeurs pour a et on égale à la valeur attendue du coté gauche de l'égalité à l'équation (1.7) à 0 pour plusieurs valeurs de k. Ensuite, on trouve l'estimateur de Tukey modifié par la fonction *TukeyPois* qui prend des données, une valeur k, l'initialisation par *poismall* et une tolérance à fixer. Cette fonction itère entre l'estimation de de λ selon une valeur de a et l'évaluation de a selon la valeur mise à jour de λ. Dans notre simulation, on a choisi $k = 6$ comme suggéré par Elsaied et Fried (2016) et une tolérance de 0.0001, c'est-à-dire que la différence entre deux itérations de l'estimation de a par l'algorithme ne dépasse pas 0.0001. Quant à la fonction *trimmeanfit*, qui calcule la moyenne tronquée adaptative, elle dépend des données, de l'initialisation par *poismall* et d'une proportion α à tronquer à chaque itération qui a été fixée pour être égale à la proportion de valeurs aberrantes.

L'estimateur de Huber est implémenté dans le paquetage *robustbase* par la fonction *glmrob* dans lequel on peut choisir la formule, la famille, les données, les poids et la méthode. Ici, on a donné comme formule $y \sim 1$ comme modèle linéaire généralisé à ajuster afin d'estimer la moyenne et on choisit la famille de Poisson.

On utilise la fonction *wle.poisson* du paquetage *wle* pour obtenir l'estimateur de maximum de vraisemblance pondérée. Cette fonction dépend des données de la

fonction d'ajustement résiduelle (RAF). Dans notre simulation, on a utilisé la distance de Hellinger.

Quant à notre estimateur, son implémentation est très simple, il suffit d'ajouter une variable uniforme à nos données de Poisson, d'en prendre la médiane empirique et d'y soustraire 1/3.

3.2.1 Modèle de données aberrantes

Pour procéder à la simulation, on a ajouté des valeurs aberrantes aux données de la façon suivante

$$(Z_\lambda)_i = (N_\lambda)_i + (1 - \varepsilon_i)\sqrt{h}$$

où $P(\varepsilon_i = 1) = 1 - \pi$ où π est la proportion de contamination et h est une constante.

Définition 3.2.1. *Le rapport signal-bruit est défini par*

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{bruit}}$$

et est donné en décibels par

$$SNR_{dB} = 10 \log_{10}(SNR)$$

Dans notre simulation, le rapport signal-bruit est donc donné par :

$$SNR_{dB} = 10 \log_{10}\left(\frac{\lambda}{h\pi(1-\pi)}\right).$$

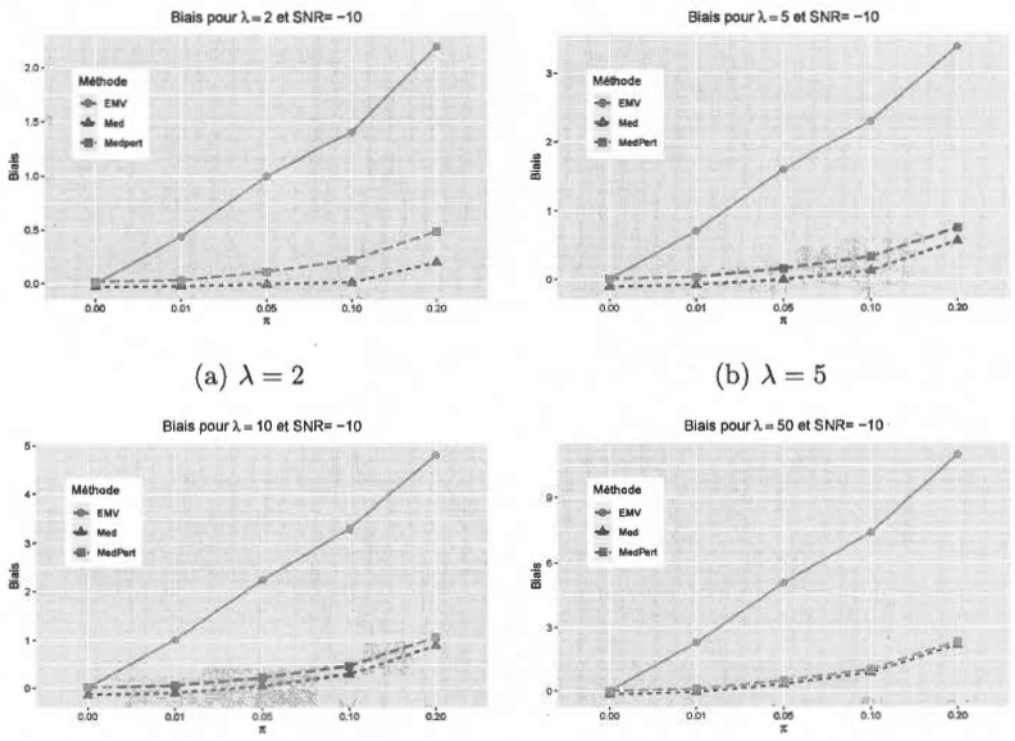
Ainsi, on peut contrôler la constante h à ajouter aux données pour la contamination :

$$h = \frac{\lambda 10^{-SNR_{dB}/10}}{\pi(1-\pi)}$$

où SNR_{dB} est fixé et la proportion de valeurs aberrantes π varie. Ensuite, on utilise le plancher de la constante h afin d'avoir uniquement des valeurs entières puisqu'on veut simuler des valeurs d'une loi discrète. On peut aussi plutôt fixer la proportion de valeurs aberrantes π et faire varier le rapport signal-bruit SNR_{dB}. Dans l'étude en simulation, on a commencé par fixer SNR_{dB} à -10 (on notera SNR_{dB} par seulement SNR par la suite) et on a regardé le biais et la racine de l'erreur quadratique moyenne pour de $\lambda = 2, 5, 10, 50$ puis pour $\lambda = 2.5, 5.5, 10.5, 50.5$, la médiane étant égale au paramètre λ lorsque λ est entier. Ensuite, on refait le même processus, mais pour une proportion de valeurs aberrantes fixée à $\pi = 10\%$. Les résultats suivants ont été établis pour une taille d'échantillon $n = 100$ et un nombre d'échantillons $m = 10000$. Par la suite, nous allons nous intéresser à l'évaluation des biais empiriques et des critères RMSE (root mean square errors).

3.2.2 Comparaison avec la médiane d'une loi de Poisson standard et l'estimateur de maximum de vraisemblance

3.2.2.1 Biais et RMSE à SNR fixé en fonction de π exprimé en pourcentage



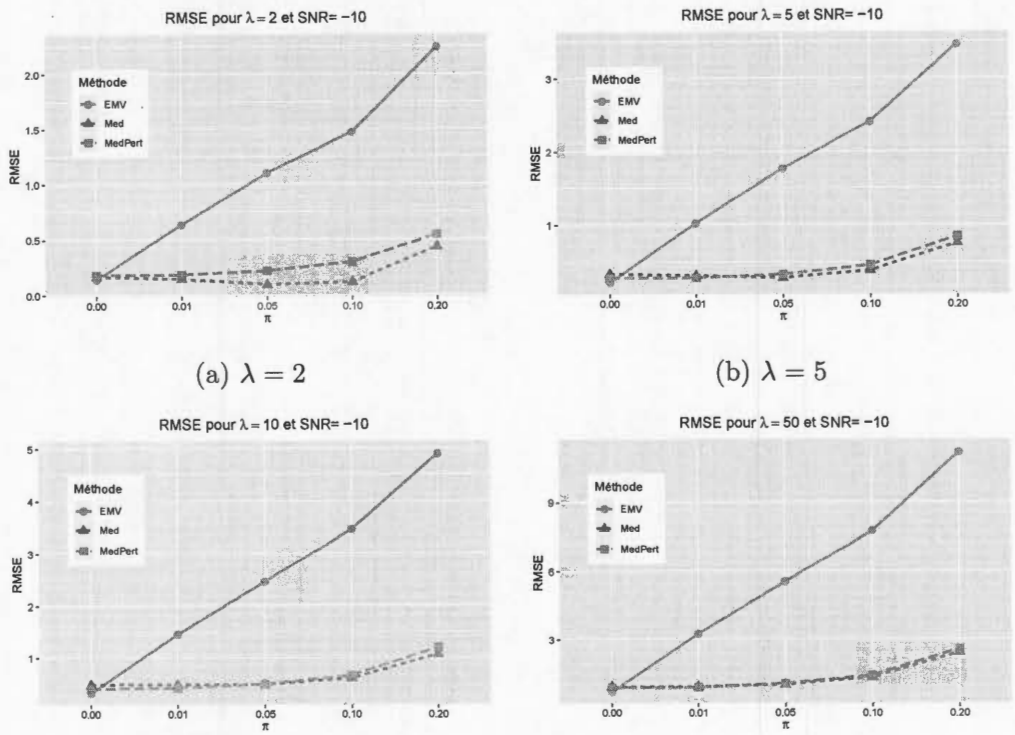
(a) $\lambda = 2$ (b) $\lambda = 5$

(c) $\lambda = 10$ (d) $\lambda = 50$

Figure 3.4: Comparaison entre les biais de la médiane, la médiane de loi de Poisson perturbée et l'estimateur de maximum de vraisemblance pour un paramètre λ entier et un rapport signal-bruit $SNR = -10$.

On a commencé par comparer trois estimateurs simples, soit l'estimateur de maximum de vraisemblance, la médiane et la médiane de la loi de Poisson perturbée. On observe sur la Figure 3.4 que la médiane et la médiane de la loi de Poisson perturbée sont comparables en terme de biais pour un paramètre λ entier et un

rapport-signal bruit fixé. On remarque aussi que les estimateurs ne sont pratiquement pas biaisés lorsque $\pi = 0$. L'estimateur de maximum de vraisemblance n'est pas biaisé lorsqu'il n'y pas de valeurs aberrantes, cependant on peut voir qu'il devient rapidement très biaisé par rapport aux autres estimateurs lorsque π augmente .



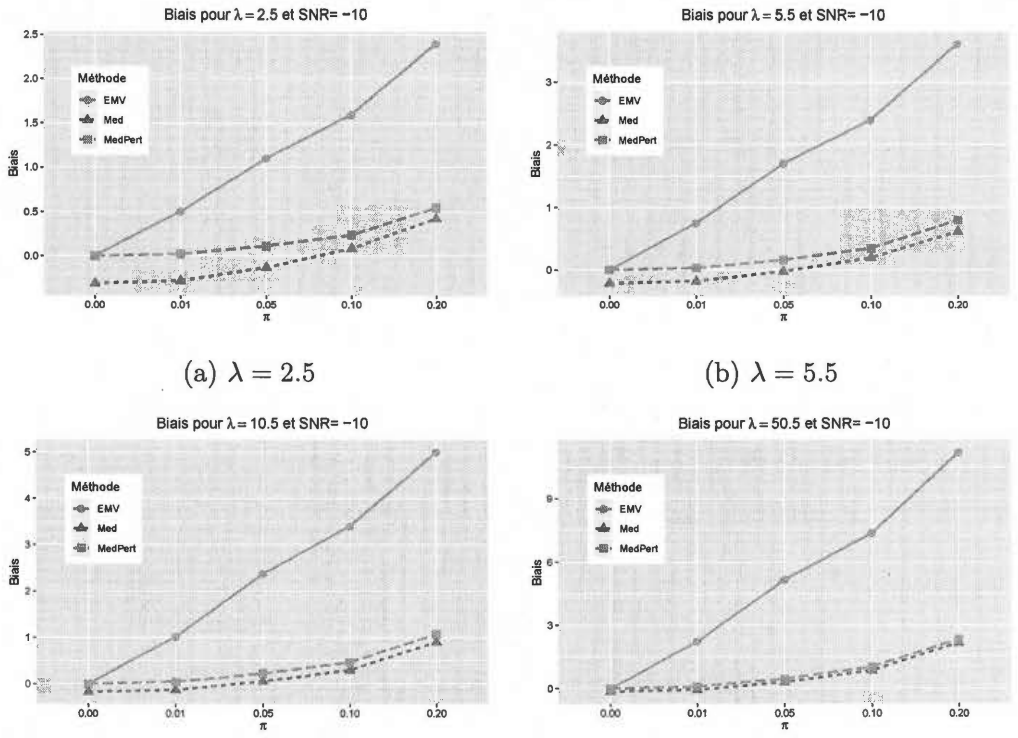
(a) $\lambda = 2$ (b) $\lambda = 5$

(c) $\lambda = 10$ (d) $\lambda = 50$

Figure 3.5: Comparaison entre les racines des erreurs quadratiques moyennes de la médiane, la médiane de loi de Poisson perturbée et l'estimateur de maximum de vraisemblance pour un paramètre λ entier et un rapport signal-bruit $SNR = -10$.

On observe aussi sur la Figure 3.5 que la médiane et la médiane de la loi de Poisson perturbée sont comparables en terme de racine d'erreur quadratique moyenne pour un paramètre λ entier et un rapport-signal bruit fixé. Comme attendu, l'estimateur

de maximum de vraisemblance a des meilleures performances lorsque $\pi = 0$. En revanche, lorsque $\pi > 0$, ses performances deviennent moins bonnes en terme de RMSE comparativement aux deux autres estimateurs.



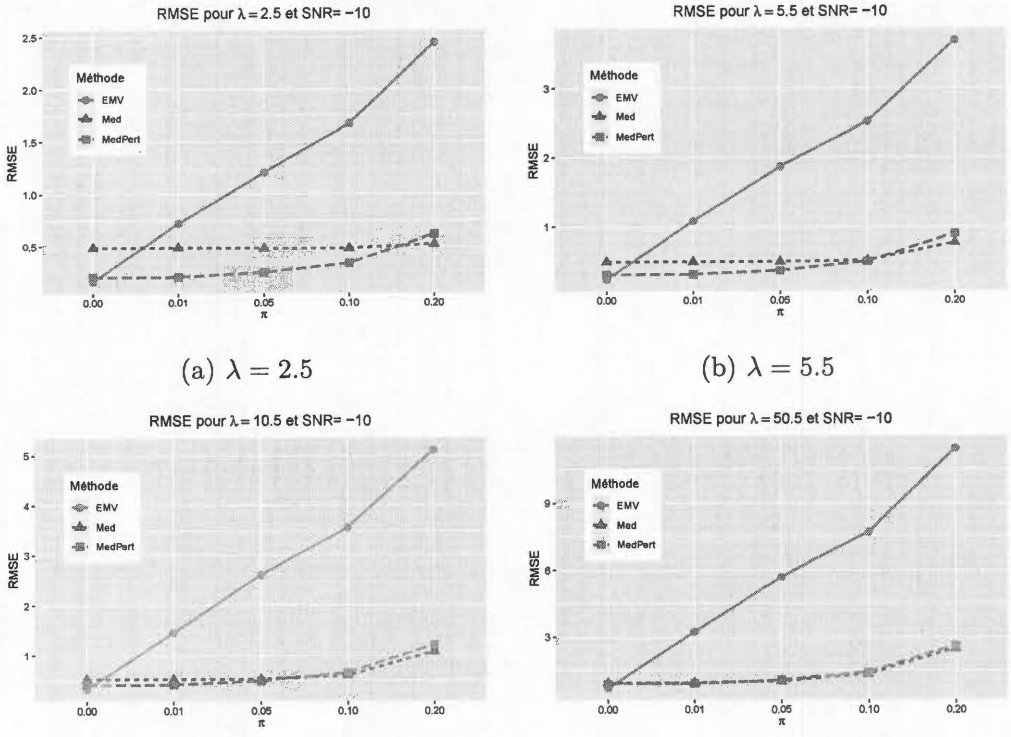
(a) $\lambda = 2.5$ (b) $\lambda = 5.5$

(c) $\lambda = 10.5$ (d) $\lambda = 50.5$

Figure 3.6: Comparaison entre les biais de la médiane, la médiane de loi de Poisson perturbée et l'estimateur de maximum de vraisemblance pour un paramètre λ et un rapport signal-bruit $SNR = -10$.

On peut aussi observer sur la Figure 3.6 que la médiane et la médiane de la loi de Poisson perturbée sont comparables en terme de biais pour un paramètre λ non-entier et un rapport signal-bruit fixé. Par contre, s'il n'y pas présence de valeur aberrantes, la médiane de la loi de Poisson perturbée n'est pratiquement pas biaisée alors que la médiane l'est. En effet, Adell et Jodrá (2005) ont démontré

que $\mathrm{Me}_{N_\lambda} = \lambda$ lorsque λ est entier. Ainsi, choisir un λ non-entier permet de retirer cet effet en faveur de la médiane.



(a) $\lambda = 2.5$

(b) $\lambda = 5.5$

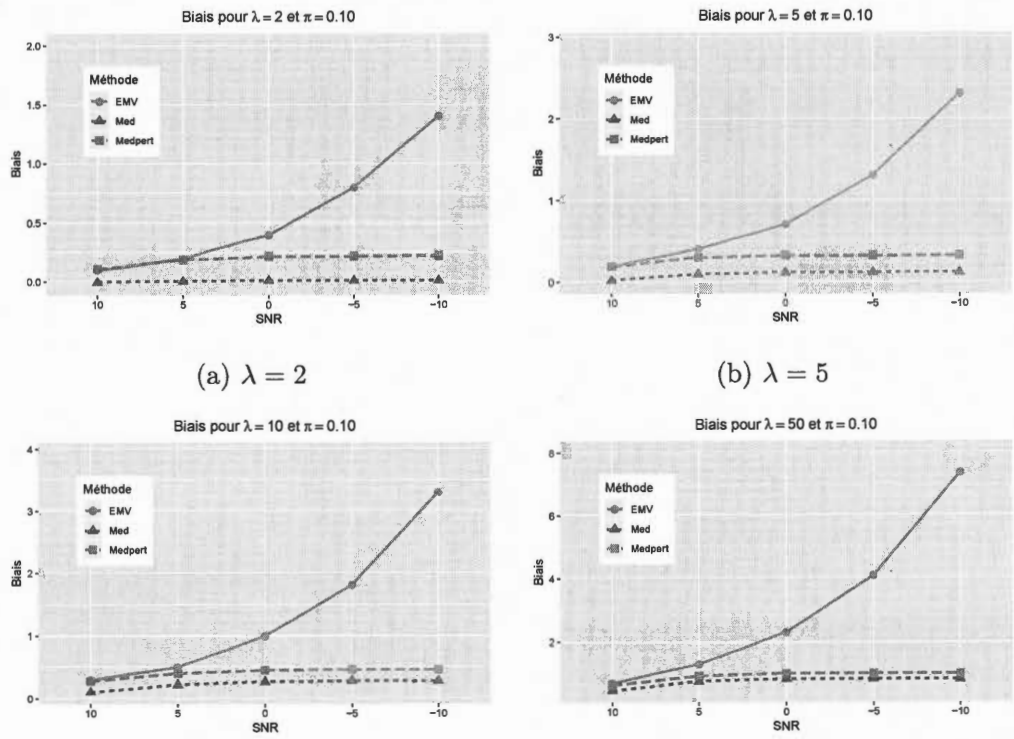
(c) $\lambda = 10.5$

(d) $\lambda = 50.5$

Figure 3.7: Comparaison entre les racines des erreurs quadratiques moyennes de la médiane, la médiane de loi de Poisson perturbée et l'estimateur de maximum de vraisemblance pour un paramètre λ et un rapport signal-bruit $SNR = -10$.

On peut aussi observer sur la Figure 3.7 que la médiane et la médiane de la loi de Poisson perturbée sont comparables en terme de racine d'erreur quadratiques moyennes pour un paramètre λ non-entier et un rapport signal-bruit fixé. Par contre, pour de petites valeurs de λ, on remarque que la médiane de la loi de Poisson perturbée semble mieux performer.

3.2.2.2 Biais et RMSE à π fixé en fonction de SNR

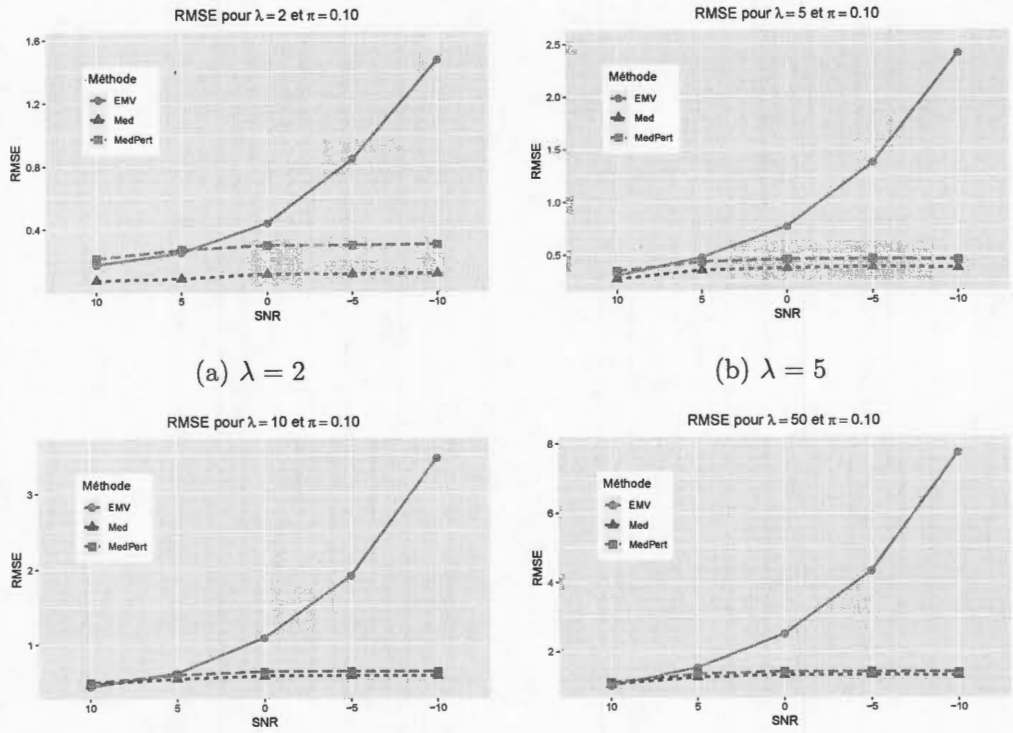


(a) $\lambda = 2$ (b) $\lambda = 5$

(c) $\lambda = 10$ (d) $\lambda = 50$

Figure 3.8: Comparaison entre les biais de la médiane, la médiane de loi de Poisson perturbée et l'estimateur de maximum de vraisemblance pour un paramètre λ entier et une proportion de valeurs aberrantes $\pi = 0.10$.

Maintenant, on peut aussi observer sur la Figure 3.8 que la médiane semble avoir un biais inférieur à la médiane de la loi de Poisson perturbée pour un paramètre λ entier et une proportion de valeurs aberrantes fixée.

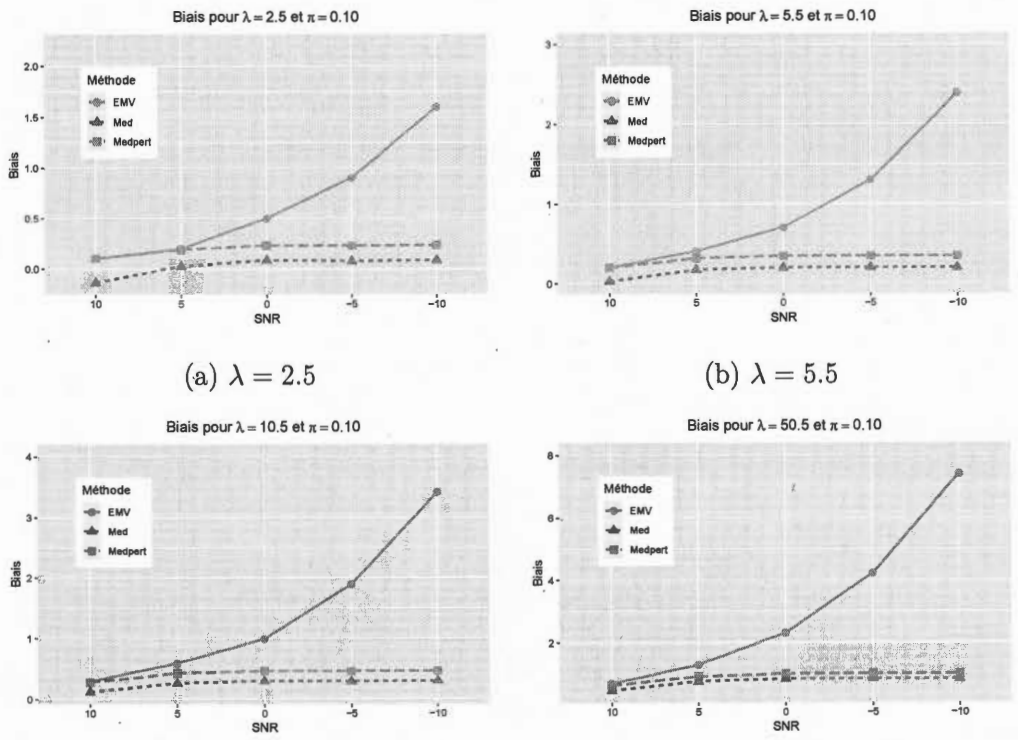


(a) $\lambda = 2$ (b) $\lambda = 5$

(c) $\lambda = 10$ (d) $\lambda = 50$

Figure 3.9: Comparaison entre les racines des erreurs quadratiques moyennes de la médiane, la médiane de loi de Poisson perturbée et l'estimateur de maximum de vraisemblance pour un paramètre λ entier et une proportion de valeurs aberrantes $\pi = 0.10$.

Par contre, en terme de racine d'erreur quadratique moyenne, les deux estimateurs semblent offrir des performances comparables pour un paramètre λ entier et une proportion de valeurs aberrantes fixée comme on peut l'observe à la Figure 3.9.

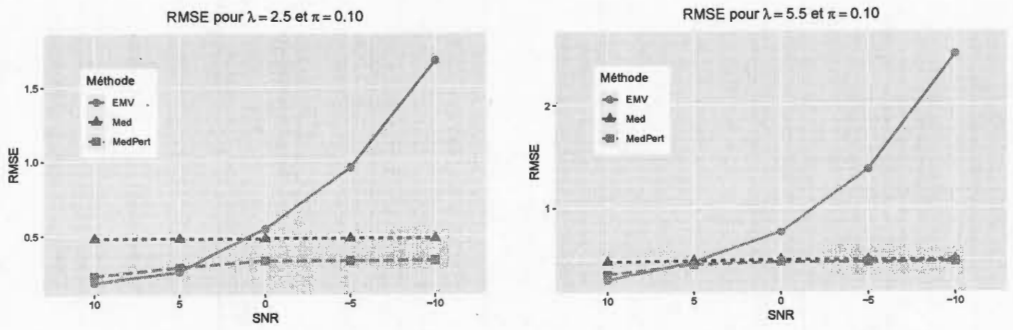


(a) $\lambda = 2.5$ (b) $\lambda = 5.5$

(c) $\lambda = 10.5$ (d) $\lambda = 50.5$

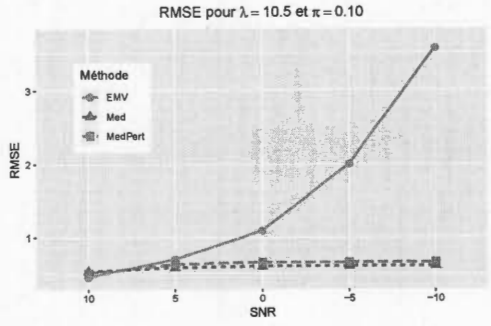
Figure 3.10: Comparaison entre les biais de la médiane, la médiane de loi de Poisson perturbée et l'estimateur de maximum de vraisemblance pour un paramètre λ et une proportion de valeurs aberrantes $\pi = 0.10$.

Maintenant, on peut aussi observer sur la Figure 3.10 que la médiane semble toujours avoir un biais inférieur à la médiane de la loi de Poisson perturbée pour un paramètre λ non-entier et une proportion de valeurs aberrantes fixée.

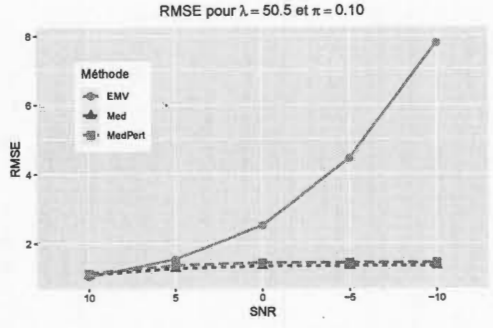


(a) $\lambda = 2.5$

(b) $\lambda = 5.5$



(c) $\lambda = 10.5$



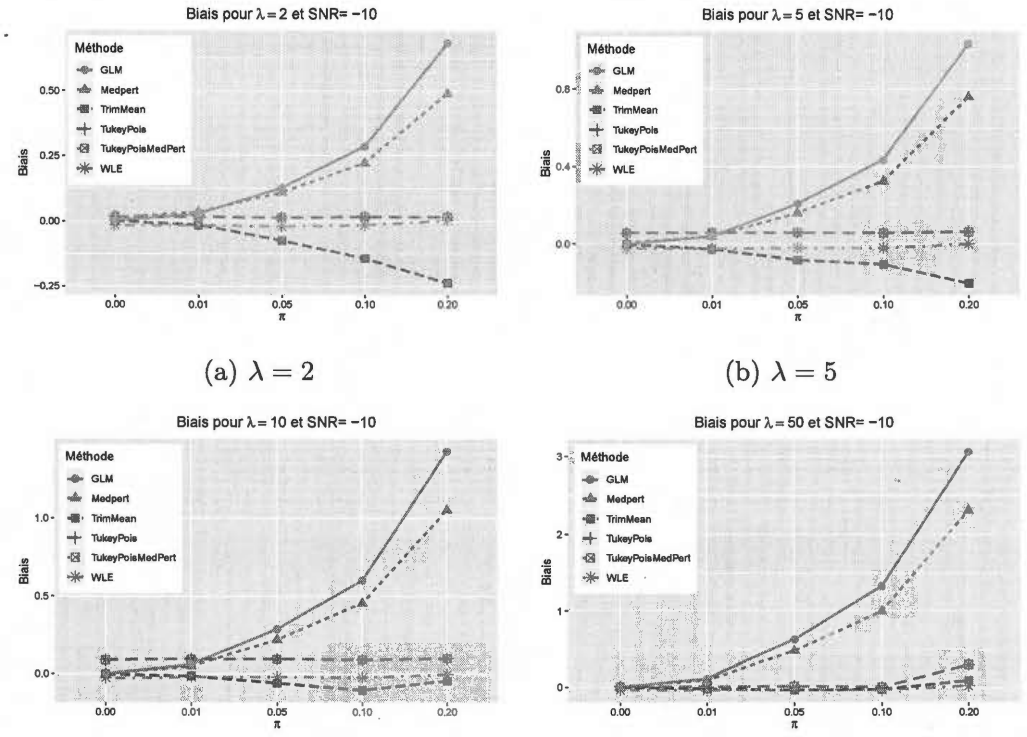
(d) $\lambda = 50.5$

Figure 3.11: Comparaison entre les racines des erreurs quadratiques moyennes de la médiane, la médiane de loi de Poisson perturbée et l'estimateur de maximum de vraisemblance pour un paramètre λ et une proportion de valeurs aberrantes $\pi = 0.10$.

Cependant, en terme de racine d'erreur quadratique moyenne, les deux estimateurs semblent toujours offrir des performances comparables pour un paramètre λ non-entier et une proportion de valeurs aberrantes fixée comme on peut l'observer à la Figure 3.11.

3.2.3 Comparaisons avec d'autres estimateurs connus et robustes

3.2.3.1 Biais et RMSE à SNR fixé en fonction de π



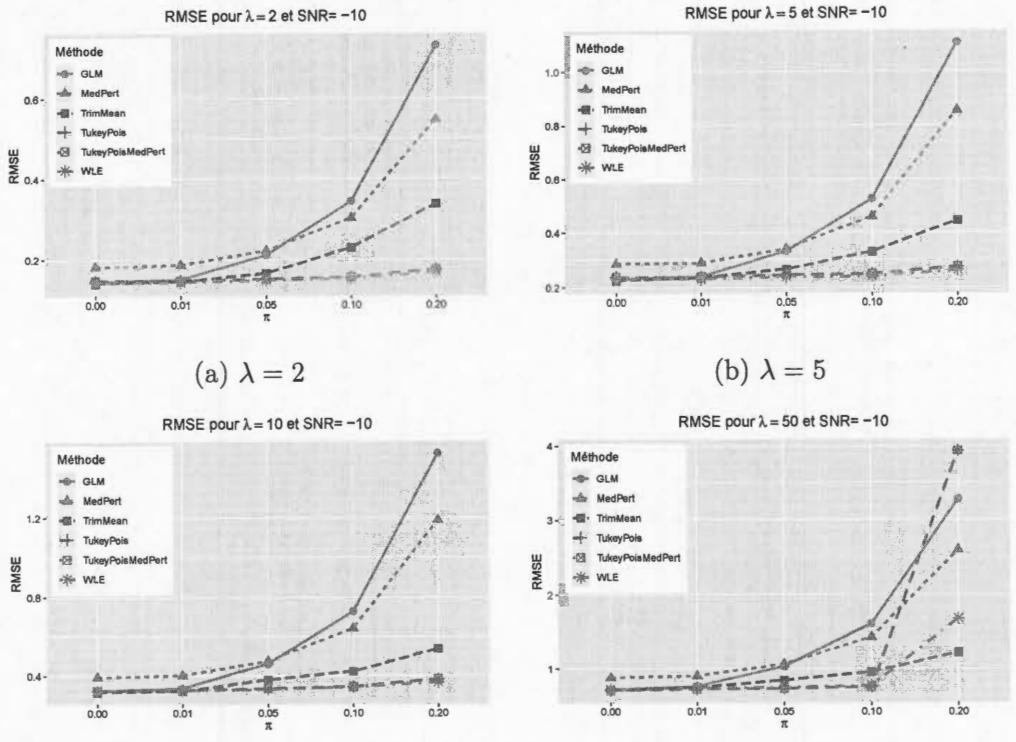
(a) $\lambda = 2$ (b) $\lambda = 5$

(c) $\lambda = 10$ (d) $\lambda = 50$

Figure 3.12: Comparaison des biais des estimateurs pour un paramètre λ entier et un rapport signal-bruit $SNR = -10$.

On a maintenant procédé à une plus importante comparaison d'estimateurs. On a repris nos trois estimateurs simples et on les a comparés avec la moyenne tronquée et l'estimateur de Tukey modifié de Elsaied et Fried (2016) initialisée avec la médiane comme ils l'avaient suggérée et notre estimateur dans le second cas, l'estimateur de Huber du paquetage *robustbase* et l'estimateur de vraisemblance pondérée. Lorsque la proportion de valeurs aberrantes π varie et que le rapport

signal-bruit SNR est fixé, on peut remarquer que l'estimateur de Tukey modifié performe bien en terme de biais et ce peu importe son initialisation. Cependant, le meilleur estimateur semble être l'estimateur de vraisemblance pondérée comme on peut l'observer à la Figure 3.12.



(a) $\lambda = 2$

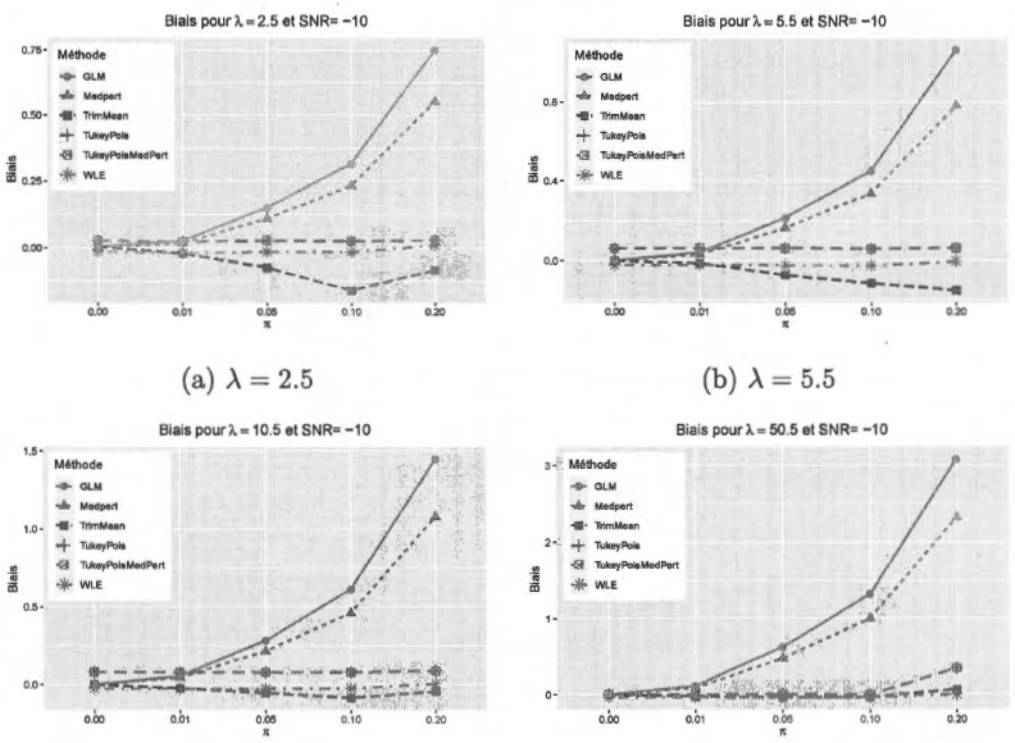
(b) $\lambda = 5$

(c) $\lambda = 10$

(d) $\lambda = 50$

Figure 3.13: Comparaison des racines des erreurs quadratiques moyennes des estimateurs pour un paramètre λ entier et un rapport signal-bruit $SNR = -10$.

En ce qui concerne la racine de l'erreur quadratique moyenne, on peut voir à la Figure 3.13 que l'estimateur de Tukey modifié performe moins bien pour un λ élevé tel qu'indiqué dans l'article de Elsaied et Fried (2016). D'un autre côté, l'estimateur de vraisemblance pondérée est toujours aussi performant.

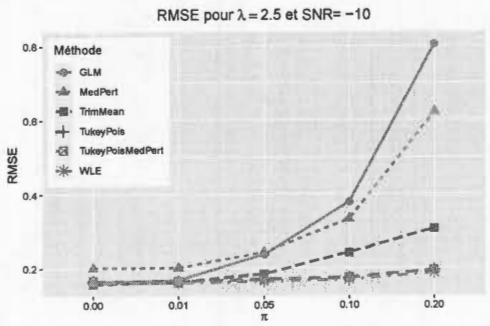


(a) $\lambda = 2.5$ (b) $\lambda = 5.5$

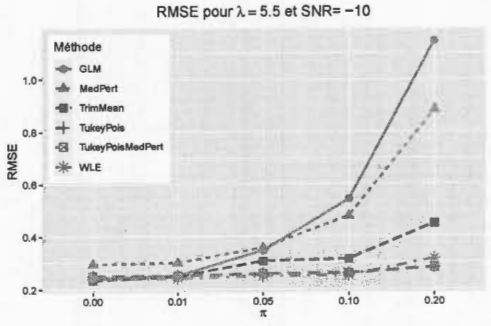
(c) $\lambda = 10.5$ (d) $\lambda = 50.5$

Figure 3.14: Comparaison des biais des estimateurs pour un paramètre λ et un rapport signal-bruit $SNR = -10$.

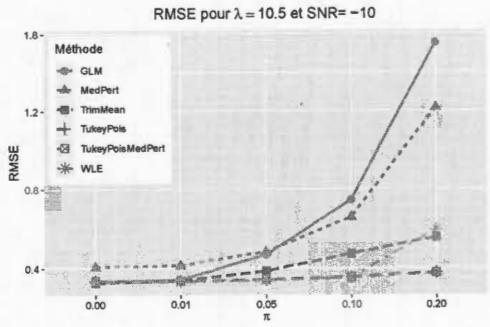
On peut voir à la Figure 3.14 que l'estimateur de Tukey modifié et l'estimateur de vraisemblance pondérée gardent la même performance pour un paramètre λ non-entier. On tire donc les mêmes conclusions.



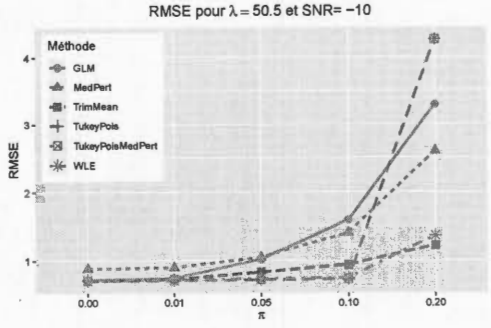
(a) $\lambda = 2.5$



(b) $\lambda = 5.5$



(c) $\lambda = 10.5$

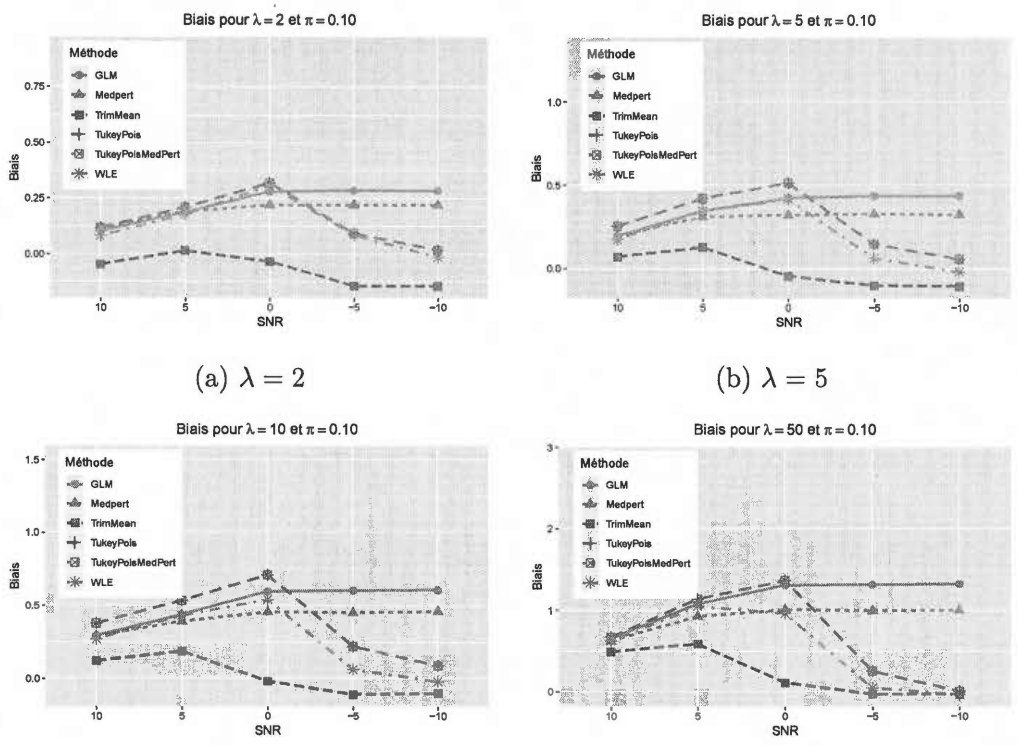


(d) $\lambda = 50.5$

Figure 3.15: Comparaison des racines des erreurs quadratiques moyennes des estimateurs pour un paramètre λ et un rapport signal-bruit $SNR = -10$.

De même, on tire les mêmes conclusions pour la racine de l'erreur quadratique moyenne comme on peut le remarquer à la Figure 3.15.

3.2.3.2 Biais et RMSE à π (h) fixé en fonction de SNR

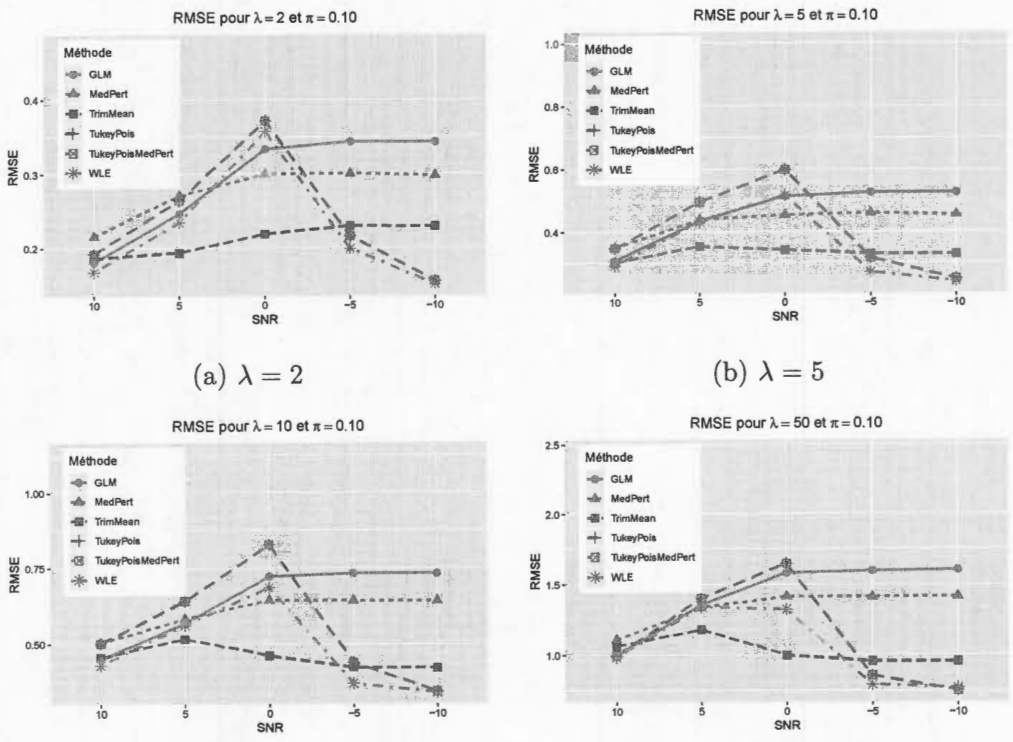


(a) $\lambda = 2$ (b) $\lambda = 5$

(c) $\lambda = 10$ (d) $\lambda = 50$

Figure 3.16: Comparaison des biais des estimateurs pour un paramètre λ entier et une proportion de valeurs aberrantes de $\pi = 0.10$.

Maintenant, si on fixe plutôt la proportion de valeurs aberrantes π à 10% et que l'on fait varier le rapport signal-bruit SNR, on remarque, à la Figure 3.16, que pour un SNR positif, l'estimateur de Tukey modifié, l'estimateur de vraisemblance pondérée et la moyenne tronquée adaptative semblent être biaisés positivement alors que pour un SNR négatif, ces estimateurs ont un biais très faible. On observe aussi un pic lorsque le rapport signal-bruit est nul, c'est-à-dire lorsqu'on ne peut distinguer le bruit du signal.

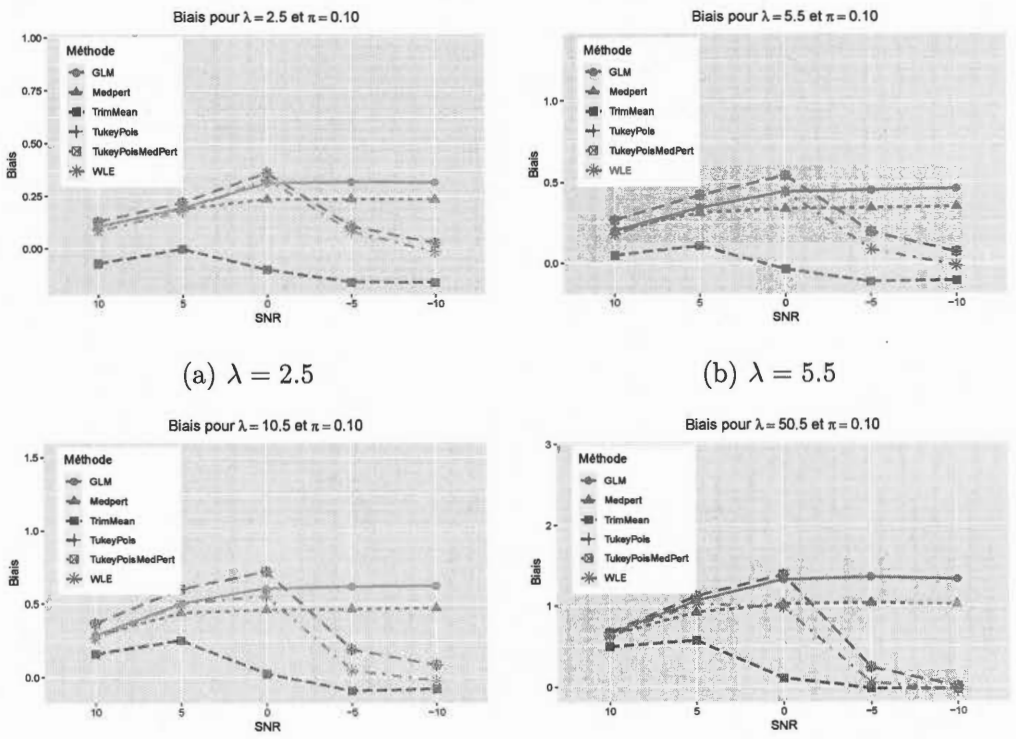


(a) $\lambda = 2$ (b) $\lambda = 5$

(c) $\lambda = 10$ (d) $\lambda = 50$

Figure 3.17: Comparaison des racines des erreurs quadratiques moyennes des estimateurs pour un paramètre λ entier et une proportion de valeurs aberrantes de $\pi = 0.10$.

On observe un phénomène similaire à la Figure 3.17 pour la racine de l'erreur quadratique moyenne. La médiane semble être un meilleur estimateur pour des petites valeurs de λ, mais on se rappelle qu'un λ entier implique que $\text{Me}_{N_\lambda} = \lambda$.

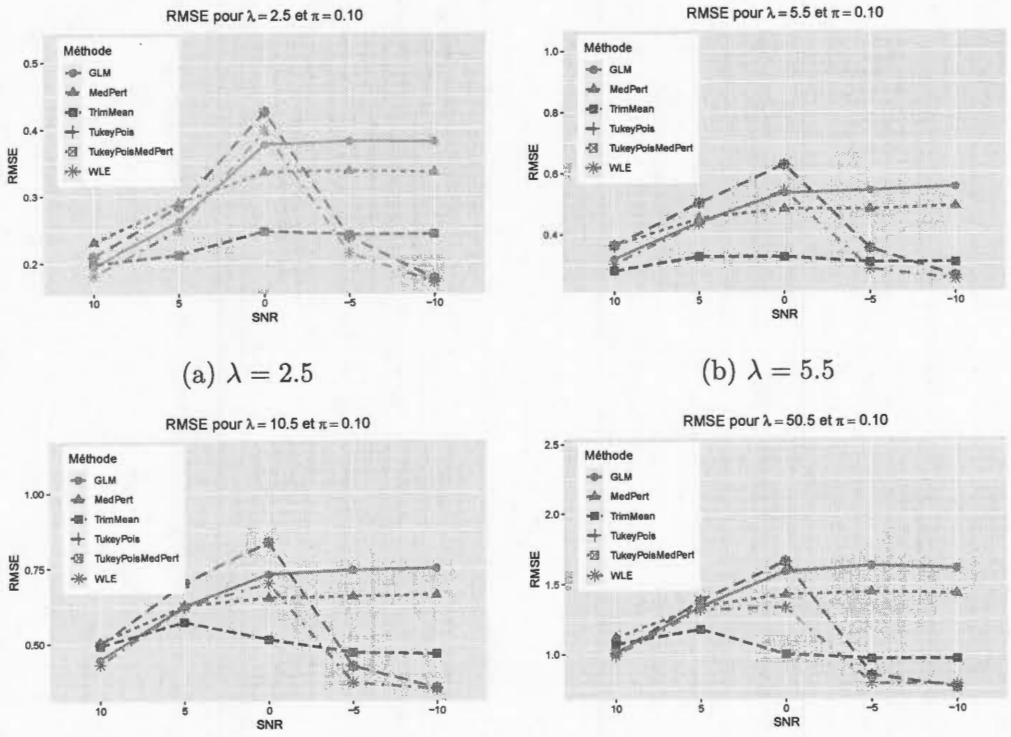


(a) $\lambda = 2.5$ (b) $\lambda = 5.5$

(c) $\lambda = 10.5$ (d) $\lambda = 50.5$

Figure 3.18: Comparaison des biais des estimateurs pour un paramètre λ et une proportion de valeurs aberrantes de $\pi = 0.10$.

À la Figure 3.18, on remarque encore que la moyenne tronquée adaptative, l'estimateur de vraisemblance pondérée et l'estimateur de Tukey modifié semblent être les moins biaisés dans le cas d'un λ non-entier et d'une porportion de valeurs aberrantes fixée.



(a) $\lambda = 2.5$

(b) $\lambda = 5.5$

(c) $\lambda = 10.5$

(d) $\lambda = 50.5$

Figure 3.19: Comparaison des racines des erreurs quadratiques moyennes des estimateurs pour un paramètre λ et une proportion de valeurs aberrantes de $\pi = 0.10$.

Finalement, comme on peut le voir à la Figure 3.19, ils semblent aussi être les estimateurs possédant la plus petite racine d'erreur quadratique moyenne. Cependant pour des petites valeurs de λ, notre estimateur offre une performance comparable.

Par conséquent, nos simulations indiquent que la médiane de loi de Poisson perturbée est un meilleur estimateur que la médiane empirique pour un paramètre λ non-entier par rapport à la racine de l'erreur quadratique moyenne. De plus, dans cette situation, notre estimateur est asymptotiquement sans biais s'il n'y pas présence de valeurs aberrantes contrairement à la médiane. Cependant, on a pu

remarquer que les estimateurs proposés par Elsaied et Fried (2016) ainsi que l'estimateur de vraisemblance pondérée semblent offrir des meilleures performances en terme de biais et de racine d'erreur quadratique moyenne.

3.3 Aspect temps de calcul

Il serait aussi intéressant de comparer les temps de calculs des différents estimateurs. En effet, pour des petits jeu de données, les estimateurs de Tukey modifié et celui de vraisemblance pondérée offrent une meilleure performance en termes de biais et de RMSE. Cependant, le temps de calcul associé à ces deux méthodes est nettement supérieur à celui de notre estimateur comme on peut le voir au Tableau 3.2.

Tableau 3.2: Comparaison des temps de calcul de n simulations d'une loi de Poisson.

n	10^3	10^4	10^5	10^6	10^7	10^8
EMV	0	0	0	0.001	0.016	0.155
Med	0.001	0	0.003	0.026	0.262	2.172
MedPert	0.001	0.001	0.006	0,051	0.428	5.275
TrimMean	0.066	0.002	0.013	0.128	1.921	...
TukeyPois	0.086	0.021	0.503	5.023	24.154	...
glm	0.015	0.093	1.57	13.015	...	...
wle	0.193	0.441	4.364	45.711	...	...

Les entrées représentées par ... du Tableau 3.2 sont manquantes car l'ordinateur de processeur IntelCore i3-2310M n'a pas été en mesure de les calculer sur R 3.4.4.

En conclusion, on sait que notre estimateur est pratiquement sans biais lorsqu'il n'y a pas présence de valeurs aberrantes tout comme l'estimateur de maximum de

vraisemblance. Cependant, contrairement à celui-ci, il performe mieux en terme de biais et de RMSE, de façon similaire à la médiane, lorsque la proportion de valeurs aberrantes augmente. De plus, sachant son comportement asymptotique, on peut produire un intervalle de confiance pour λ basé sur notre estimateur. Par contre, on a pu voir que l'estimateur de Tukey modifié et l'estimateur de vraisemblance pondérée sont plus performants en terme de biais et de RMSE. Mais, en terme de temps de calcul, ces estimateurs sont moins intéressants que le nôtre et deviennent incalculables pour de très grands échantillon.

CONCLUSION

Le but de ce mémoire était de présenter un nouvel estimateur robuste pour le paramètre d'une loi de Poisson. Avant d'y parvenir, on a procédé, au chapitre 1, à une revue des estimateurs connus et standard pour le paramètre λ d'une loi de Poisson. L'estimateur de Tukey modifié, la moyenne tronquée adaptative et l'estimateur de maximum de vraisemblance pondérée y ont été présentés.

Ensuite, au chapitre 2, on présente notre estimateur $\widehat{\lambda}_J = \widehat{\text{Me}}(\mathbf{Z}_\lambda) - 1/3$ où $\mathbf{Z}_\lambda = (Z_{1,\lambda}, \ldots, Z_{n,\lambda})$ et $Z_{i,\lambda}$ sont indépendantes et identiquement distribuées de même loi que $N_\lambda + U$ avec $N_\lambda \sim \mathcal{P}(\lambda)$ et $U \sim \mathcal{U}(0,1)$. L'ajout de U permet de récupérer un théorème central limite pour $\widehat{\lambda}_J$. Ainsi, il est devenu pertinent d'étudier Me_{Z_λ}. La principale contribution de ce mémoire a été de démontrer que $\text{Me}_{Z_\lambda} \approx \lambda + 1/3$, un résultat tout à fait inatendu et d'autant plus précis que λ est grand. La définition de notre estimateur $\widehat{\lambda}_J$ est devenue naturelle.

Finalement, au chapitre 3, on a procédé à une étude en simulation. D'abord, on s'est intéressé au comportement asymptotique de l'estimateur. Suivant cette section, on a construit un intevalle de confiance pour λ basé sur $\widehat{\lambda}_J$. Puis, on l'a comparé à d'autres estimateurs connus et robustes d'une loi de Poisson. De même que l'estimateur de maximum de vraisemblance, lorsqu'un jeu de données ne contient pas de valeurs aberrantes, $\widehat{\lambda}_J$ est asymptotiquement sans biais. Cependant, lorsque la proportion de valeurs aberrantes augmente, $\widehat{\lambda}_J$ est 8nettement moins biaisé. Sa performance est similaire à la médiane pour un paramètre λ entier et supérieure à la médiane pour un paramètre non-entier en terme de RMSE. Lorsqu'on procède à une comparaison plus importante, on peut voir que les M-estimateurs de

Tukey modifié et les estimateurs de maximum de vraisemblance pondérée sont moins biaisés et ont une plus petite RMSE que $\widehat{\lambda}_J$. Cependant, celui-ci a l'avantage d'avoir un temps de calcul nettement inférieur et d'être contenu dans un intervalle de confiance.

Il serait intéressant de savoir si notre méthode peut s'appliquer aux distributions de Poisson à inflation de zéros. Aussi, on pourrait sans doute appliquer la méthode à l'estimation de l'intensité d'un processus ponctuel de Poisson présentant des valeurs aberrantes. Par ailleurs, on a conjecturé le résultat 2.3 à la section 2.3.2, qui est un résultat encore plus fin que celui que nous avons montré et il serait intéressant de savoir si on peut parvenir à le démontrer.

RÉFÉRENCES

Achard, S., Coeurjolly, J.-F. *et al.* (2010). Discrete variations of the fractional brownian motion in the presence of outliers and an additive noise. *Statistics Surveys*, *4*, 117–147.

Adell, J. A. et Jodrá, P. (2005). The median of the Poisson distribution. *Metrika*, *61*(3), 337–346.

Assunção, R. et Guttorp, P. (1999). Robustness for inhomogeneous Poisson point processes. *Annals of the Institute of Statistical Mathematics*, *51*(4), 657–678.

Cadigan, N. et Chen, J. (2001). Properties of robust M-estimators for Poisson and negative binomial data. *Journal of Statistical Computation and Simulation*, *70*(3), 273–288.

Chen, J. et Rubin, H. (1986). Bounds for the difference between median and mean of Gamma and Poisson distributions. *Statistics & probability letters*, *4*(6), 281–283.

Choi, K. P. (1994). On the medians of Gamma distributions and an equation of Ramanujan. *Proceedings of the American Mathematical Society*, *121*(1), 245–251.

Coeurjolly, J.-F. (2017). Median-based estimation of the intensity of a spatial point process. *Annals of the Institute of Statistical Mathematics*, *69*(2), 303–331.

Elsaied, H. et Fried, R. (2016). Tukey's M-estimator of the Poisson parameter with a special focus on small means. *Statistical Methods & Applications*, *25*(2), 191–209.

Elsaied, H., Fried, R., Kuhnt, S. *et al.* (2012). Robust modelling of count data.

Huber, P. J. et Ronchetti, E. M. (2009). Robust Statistics. Hoboken. *NJ : Wiley. doi*, *10*(1002), 9780470434697.

Machado, J. A. F. et Silva, J. S. (2005). Quantiles for counts. *Journal of the American Statistical Association*, *100*(472), 1226–1237.

Markatou, M., Basu, A. et Lindsay, B. (1997). Weighted likelihood estimating equations : The discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference*, *57*(2), 215–232.

Maronna, R., Martin, R. D. et Yohai, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester. ISBN.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Simpson, D. G., Carroll, R. J. et Ruppert, D. (1987). M-estimation for discrete data : asymptotic distribution theory and implications. *The Annals of Statistics*, 657–669.

Wilcox, R. R. et Clark, F. (2013). Robust regression estimators when there are tied values. *Journal of Modern Applied Statistical Methods*, *12*(2), 3.