REVUE DE STATISTIQUE APPLIQUÉE

D. MIZÈRE
C. C. KOKONENDJI
S. DOSSOU-GBÉTÉ

Quelques tests de la loi de Poisson contre des alternatives générales basées sur l'indice de dispersion de Fisher

Revue de statistique appliquée, tome 54, n° 4 (2006), p. 61-84
<http://www.numdam.org/item?id=RSA_2006__54_4_61_0>

Rev. Statistique Appliquée, 2006, **LIV** (4), 61-84

QUELQUES TESTS DE LA LOI DE POISSON CONTRE DES ALTERNATIVES GÉNÉRALES BASÉES SUR L'INDICE DE DISPERSION DE FISHER

D. MIZÈRE[(a)(b)], C.C. KOKONENDJI[(a)], S. DOSSOU-GBÉTÉ[(a)]

[(a)] *Université de Pau et des Pays de l'Adour*
Laboratoire de Mathématiques Appliquées - UMR 5142 CNRS
Avenue de l'Université - 64000 Pau, France
[(b)] *Université Marien Ngouabi - Faculté des Sciences*
Brazzaville, République du Congo

RÉSUMÉ

La principale motivation de ce travail est d'évaluer les performances de quelques procédures de tests destinés à valider l'adéquation de la loi de Poisson avec des données de comptage contre des alternatives générales de surdispersion et de sousdispersion. Ainsi, nous comparons le test du khi-deux de Pearson à des tests construits à partir des statistiques obtenues par l'application d'une transformation de Box-Cox à l'indice de dispersion de Fisher puis à son inverse. Des simulations permettent d'étudier les différentes propriétés de ces tests.

Mots-clés : *Données de comptage, sousdispersion, surdispersion, transformation de Box-Cox.*

ABSTRACT

The main motivation of this paper is to evaluate the performances of some tests procedures devoted to valid the fitness of count data by Poisson distribution against general alternatives of overdispersion and underdispersion. Thus, we compare the chi-square test of Pearson to tests constructed from statistics which are obtained by the Box-Cox transformation of the Fisher dispersion index and of its inverse. Some simulations are done for pointing out various properties of these tests.

Keywords : *Box-Cox transformation, count data, overdispersion, underdispersion*

1. Introduction

Les données de comptage résultent généralement du dénombrement des occurences d'événements contemporains ou voisins dans l'espace.

Bien que le modèle poissonnien soit le cadre probabiliste le plus utilisé pour l'analyse des données de comptage, ce modèle est approprié seulement si le processus sous-jacent à l'occurence des événements considérés vérifient les hypothèses suivantes :

1. la probabilité de l'occurence simultanée de deux événements est nulle;
2. aucune occurence d'événements au commencement de la période d'observation;
3. la probabilité d'une occurence pendant la période d'observation est, d'une part, indépendante des occurences d'événement antérieures à la date de cette occurence dans la période d'observation et, d'autre part, indépendante de la date de cette occurence (hypothèse d'homogénéité).

La famille exponentielle des lois de Poisson est constituée de lois de probabilités

$$p(x;\theta) = \frac{1}{x!}\exp\{x\theta - e^{\theta}\},\ \ x \in \mathbb{N},\ \theta \in \mathbb{R},$$

de support égal à $\mathbb{N}$, associées chacune de manière injective au paramètre canonique $\theta \in \mathbb{R}$, où

$$\exp\,(\theta) = \log \sum_{x=0}^{+\infty} \frac{\exp\,(x\theta)}{x!}$$

est appelé fonction cumulante. Puisqu'on a à la fois

$$\lambda = \exp\,(\theta) = \sum_{x=0}^{+\infty} xp(x;\theta) \ \text{ et } \ \lambda = \sum_{x=0}^{+\infty} (x-\lambda)^2\, p(x;\theta)\,,$$

pour tout $\theta \in \mathbb{R}$, l'égalité entre la moyenne et la variance d'une loi de Poisson exprimée par les expressions ci-dessus caractérise la famille des lois de Poisson parmi les familles exponentielles de lois de probabilités.

Lorsque le modèle poissonnien n'est pas approprié, principalement parce que les résultats issus du traitement des données ne permettent pas de le valider, la question se pose de trouver un modèle probabiliste alternatif pour décrire les variations observées dans les données étudiées. Il est courant de chercher ces familles alternatives de lois de probabilités en se fondant sur la position du rapport variance/moyenne relativement à 1 :

— lorsque ce rapport est significativement supérieur à 1 la situation est qualifiée de surdispersion

— si par contre ce rapport est inférieur à 1 on est en situation de sousdispersion.

DÉFINITION 1. – *Soit $\mathcal{F}$ une famille de lois de probabilités sur $\mathbb{N}$. On dira que la famille $\mathcal{F}$ est surdispersée (resp. sousdispersée) relativement à la famille des lois de Poisson si pour toute distribution de probabilité appartenant à cette famille, de moyenne μ et de variance σ^2, on a $\sigma^2/\mu > 1$ (resp. $\sigma^2/\mu < 1$).*

La surdispersion a retenu beaucoup l'attention et a été étudiée de manière extensive : détection de la surdispersion et modélisation de la surdispersion (*e.g.* Cox, 1983; Gelfand & Dalal, 1990; Giano & Schafer, 1992; Kokonendji *et al.*, 2004, et aussi pour quelques références). Par contre, la sousdispersion demeure peu étudiée

et les travaux convainquants en matière de modélisation de la sousdispersion ne sont pas nombreux (*e.g.* Castillo & Pérez-Casany, 1998). Les résultats du présent travail concernent les différents aspects de l'inadéquation de la loi de Poisson avec des données, la surdispersion comme la sousdispersion. Dans la section 2, nous étudions quelques statistiques en rapport avec l'indice de dispersion de Fisher. La section 3 est consacrée à une comparaison empirique des performances des tests construits à partir des statistiques étudiées dans la section précédente. La section 4 propose une dernière remarque sur les statistiques de données de comptage selon que l'alternative soit de surdispersion ou de sousdispersion.

2. Statistiques des données de comptage

On considère $(x_i)_{i=1,\cdots,n}$ une série de n dénombrements indépendants des occurences d'un certain type d'événement. On note

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Sous l'hypothèse que la série statistique est constituée par les réalisations de variables aléatoire indépendantes distribuées suivant la même loi de Poisson de moyenne λ, $\overline{x}_n$ est l'estimation de λ par la méthode du maximum de vraisemblance. Notons par ailleurs la variance empirique des données par

$$\overline{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{n}{n-1}\, s_n^2.$$

Une question récurrente du traitement statistique des données de comptage est la validation ou non de l'hypothèse que les données sont compatibles avec le modèle poissonnien. On considère dans la suite comme hypothèse nulle, H_0, l'hypothèse que les dénombrements observés sont des réalisations indépendantes d'une même loi de Poisson.

2.1. Indice de dispersion de Fisher

L'indice de dispersion de Fisher est une des statistiques les plus utilisées pour discriminer le modèle poissonnien par rapport à des modèles alternatifs. Il est défini par le quotient $\overline{s}_n^2/\overline{x}_n$. On peut observer que le numérateur $\overline{s}_n^2$ mesure la variabilité observée dans les données et que $\overline{x}_n$ est l'estimation de cette variabilité prévue par le modèle poissonnien. Il parait donc raisonnable de considérer un écart trop important entre ces deux mesures de dispersion comme l'évidence de l'inadéquation du modèle poissonien avec les données.

2.1.1. Distribution de probabilités

Soit $(X_i)_{i=1,\cdots,n}$ une suite de variables aléatoires indépendantes identiquement distribuées à valeurs dans $\mathbb{N}$. On pose

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \overline{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{n}{n-1}S_n^2.$$

On ne connaît pas d'expression algébrique de la distribution de la statistique $\overline{S}_n^2/\overline{X}_n$ pour un échantillon de taille n donnée. Sous l'hypothèse que les dénombrements observés sont des réalisations indépendantes de variables de Poisson de même moyenne, la statistique $nS_n^2/\overline{X}_n$ apparaît comme celle de Pearson et elle est approximativement égale à celle du rapport de vraisemblance sous l'hypothèse nulle. Il s'en suit que $T_F = nS_n^2/\overline{X}_n$ est asymptotiquement distribuée comme la loi du khi-deux à $n-1$ degrés de liberté si l'échantillon observée est générée par une loi de Poisson (*e.g.* Hoel, 1943).

Le résultat suivant permet par la suite d'établir les différentes régions critiques des tests asymptotiques basés sur $T_F = nS_n^2/\overline{X}_n$ et son inverse.

PROPOSITION 2. – *Si $(X_i)_{i=1,\cdots,n}$ est une suite de variables aléatoires iid de moyenne μ et de variance σ^2, alors la convergence en loi de $nS_n^2/\overline{X}_n$ vers une variable aléatoire U entraîne celle de $n\overline{X}_n/S_n^2$ vers la variable aléatoire $(\mu/\sigma^2)^2U$.*

Démonstration. – Puisque $(\overline{X}_n/S_n^2)^2$ converge en loi vers la constante $(\mu/\sigma^2)^2$, on en conclut par le théorème classique de Slutsky que $n\overline{X}_n/S_n^2 = (\overline{X}_n/S_n^2)^2(nS_n^2/\overline{X}_n)$ converge en loi vers $(\mu/\sigma^2)^2U$, où U est la variable aléatoire limite de $nS_n^2/\overline{X}_n$ par hypothèse. □

2.1.2. Tests du khi-deux

À l'aide de la Propostion 2, nous ne considérons ici que les tests du khi-deux basés sur l'indice de dispersion pour discriminer la famille des lois de Poisson par rapport à des familles alternatives de lois de probabilité. En effet, si on dispose d'un échantilon de taille suffisante, on peut envisager de tester l'hypothèse nulle que la distribution est une loi de Poisson contre les alternatives de surdispersion ou de sousdispersion et même l'alternative bilatérale. Malheureusement on ne connait pas de manière explicite le comportement asymptotique de $nS_n^2/\overline{X}_n$ et $n\overline{X}_n/S_n^2$ sous des hypothèses alternatives générales pour pouvoir comparer ces procédures de test de point de vue de leurs puissances.

Soit $\chi^2_{n-1,\alpha}$ et $\chi^2_{n-1,1-\alpha}$ respectivement les quantiles d'ordre α et $1-\alpha$ de la loi du khi-deux à $n-1$ degrés de liberté.

– Hypothèse alternative d'une famille surdispersée (la plus fréquente, *e.g.* voir Brown & Zhao, 2002 ; Smyth & Podlich, 2000) : Pour une valeur nominale de la probabilité d'erreur de première espèce fixée à α, on peut envisager l'une des deux procédures

unilatérales suivantes

$$\text{Rejet de } H_0 \text{ si } \frac{nS_n^2}{\overline{X}_n} = T_F > \chi^2_{n-1,1-\alpha}$$

ou

$$\text{Rejet de } H_0 \text{ si } \frac{n\overline{X}_n}{S_n^2} < \chi^2_{n-1,\alpha} \text{ ;}$$

– Hypothèse alternative d'une famille sousdispersée : De la même manière que ci-dessus, on peut envisager de tester l'hypothèse nulle de la distribution de Poisson contre l'alternative de sousdispersion en considérant l'un des deux tests de régions critiques unilatérales suivantes

$$\text{Rejet de } H_0 \text{ si } \frac{nS_n^2}{\overline{X}_n} = T_F < \chi^2_{n-1,\alpha}$$

ou

$$\text{Rejet de } H_0 \text{ si } \frac{n\overline{X}_n}{S_n^2} > \chi^2_{n-1,1-\alpha};$$

– Hypothèse alternative bilatérale : À l'évidence on peut envisager également un test bilatéral de l'hypothèse nulle contre son contraire basé sur l'indice de dispersion de l'une des deux manières suivantes

$$\text{Rejet de } H_0 \text{ si } T_F < \chi^2_{n-1,\alpha/2} \text{ ou } T_F > \chi^2_{n-1,1-\alpha/2}$$

ou

$$\text{Rejet de } H_0 \text{ si } \frac{n\overline{X}_n}{S_n^2} < \chi^2_{n-1,\alpha/2} \text{ ou } \frac{n\overline{X}_n}{S_n^2} > \chi^2_{n-1,1-\alpha/2}.$$

2.2. Autres statistiques en rapport avec l'indice de dispersion de Fisher

Tiago de Oliveira (1965) a suggéré, à partir de l'étude de la variance de la statistique $\overline{S}_n^2 - \overline{X}_n$, l'utilisation de la statistique

$$T_O = \frac{\sqrt{n}(\overline{S}_n^2 - \overline{X}_n)}{\sqrt{1 - 2\overline{X}_n^{1/2} + 3\overline{X}_n}}$$

pour construire un test de l'hypothèse nulle contre l'aternative de surdispersion.

Après avoir observé que l'expression de la variance de la statistique $\overline{S}_n^2 - \overline{X}_n$ obtenue par Tiago de Oliveira sous l'hypothèse nulle était fausse, Böhning (1994) a donné l'expression correcte de la variance de $\overline{S}_n^2 - \overline{X}_n$ et a proposé la statistique

$$T_B = \sqrt{\frac{n-1}{2}}\,\frac{\overline{S}_n^2 - \overline{X}_n}{\overline{X}_n}$$

pour construire un test de l'hypothèse nulle de la loi de Poisson contre l'alternative de la surdispersion.

Böhning (1994) n'a pas explicitement étudié la distribution de la statistique T_B qu'il a proposée ni sous l'hypothèse nulle, ni sous l'hypothèse de surdispersion. Comme la statistique T_B est fondée sur

$$\overline{S}_n^2 - \overline{X}_n = S_n^2 - \overline{X}_n + \frac{S_n^2}{n-1}\,,$$

poursuivant en cela l'idée de Tiago de Oliveira, nous allons étudier le comportement asymptotique de $S_n^2 - \overline{X}_n$ dans la section qui va suivre, et ce, dans le double cadre de surdispersion et sousdispersion.

2.2.1. Comportement asymptotique de la statistique $S_n^2 - \overline{X}_n$

Considérons une suite $(X_i)_{i=1,\cdots,n}$ de n variables aléatoires indépendantes de même loi. On supposera que cette loi admet des moments jusqu'à l'ordre 4. Soit $\mu = \mathbb{E}(X_1)$, $\sigma^2 = \mathbb{E}(X_1 - \mu)^2$ et $\mu_k = \mathbb{E}(X_1 - \mu)^k$, $k \geqslant 3$.

PROPOSITION 3. – *Les statistiques* $\sqrt{n}\left\{\left(S_n^2 - \overline{X}_n\right) - \left(\sigma^2 - \mu\right)\right\}$ *et* $\sqrt{n}\{(\overline{S}_n^2 - \overline{X}_n) - \left(\sigma^2 - \mu\right)\}$ *convergent en loi vers la distribution gaussienne centrée et de variance* $\mu_4 - 2\mu_3 + \sigma^2 - \left(\sigma^2\right)^2$.

Démonstration. – En écrivant

$$\begin{aligned}\left(S_n^2 - \overline{X}_n\right) - \left(\sigma^2 - \mu\right) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \left(\overline{X}_n - \mu\right)^2 - \left(\overline{X}_n - \mu\right) - \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{(X_i - \mu)^2 - (X_i - \mu)\right\} - \sigma^2 - \left(\overline{X}_n - \mu\right)^2,\end{aligned}$$

on observe d'une part que la variable aléatoire $(X_i - \mu)^2 - (X_i - \mu)$ est de moyenne égale à σ^2 et de variance égale à $\mu_4 - 2\mu_3 + \sigma^2 - \left(\sigma^2\right)^2$. Il s'ensuit que

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\left\{(X_i - \mu)^2 - (X_i - \mu)\right\} - \sigma^2}{\sqrt{n}\,\sqrt{\mu_4 - 2\mu_3 + \sigma^2 - \left(\sigma^2\right)^2}}$$

converge en loi vers la loi gaussienne centrée et réduite. D'autre part, comme $\sqrt{n}\left(\overline{X}_n - \mu\right)$ converge en loi d'après le théorème limite central et $\overline{X}_n - \mu$ converge presque-sûrement vers 0, alors $\sqrt{n}\left(\overline{X}_n - \mu\right)^2$ converge en probabilité vers 0. On en déduit que $\sqrt{n}\left[\left(S_n^2 - \overline{X}_n\right) - \left(\sigma^2 - \mu\right)\right]$ converge en loi vers la loi gaussienne centrée et de variance $\mu_4 - 2\mu_3 + \sigma^2 - \left(\sigma^2\right)^2$.

Par ailleurs, puisqu'on a

$$\sqrt{n}\left(\overline{S}_n^2 - \overline{X}_n\right) = \sqrt{n}\left(S_n^2 - \overline{X}_n + \frac{S_n^2}{n-1}\right)$$

et $\sqrt{n}S_n^2/(n-1)$ converge presque-sûrement vers 0, il s'ensuit que les statistiques $\sqrt{n}\left\{\left(S_n^2 - \overline{X}_n\right) - \left(\sigma^2 - \mu\right)\right\}$ et $\sqrt{n}\{(\overline{S}_n^2 - \overline{X}_n) - \left(\sigma^2 - \mu\right)\}$ ont la même distribution asymptotique. □

COROLLAIRE 4. – *Si les variables aléatoires X_i, $i = 1, \cdots, n$, sont indépendantes et équidistribuées suivant la loi de Poisson de moyenne λ. Alors les statistiques $\sqrt{n/2}\left(S_n^2 - \overline{X}_n\right)/\lambda$, $\sqrt{n/2}\left(\overline{S}_n^2 - \overline{X}_n\right)/\lambda$ et $\sqrt{(n-1)/2}\left(\overline{S}_n^2 - \overline{X}_n\right)/\lambda$ convergent en loi vers la loi gaussienne centrée et réduite.*

Démonstration. – Ces résultats se déduisent de manière immédiate de la Proposition 3 en vérifiant que $\mu = \sigma^2 = \mu_3 = \lambda$, $\mu_4 = 3\lambda^2 + \lambda$ et donc $\mu_4 - 2\mu_3 + \sigma^2 - \left(\sigma^2\right)^2 = 2\lambda^2$. □

COROLLAIRE 5. – *(i) Les statistiques $\sqrt{n}\left\{\left(S_n^2 - \overline{X}_n\right) - \left(\sigma^2 - \mu\right)\right\}/\overline{X}_n$ et $\sqrt{n}\{(\overline{S}_n^2 - \overline{X}_n) - \left(\sigma^2 - \mu\right)\}/\overline{X}_n$ convergent en loi vers la distribution gaussienne centrée et de variance $\{\mu_4 - 2\mu_3 + \sigma^2 - \left(\sigma^2\right)^2\}/\mu^2$.*

(ii) Les statistiques $\sqrt{n}\left\{\left(S_n^2 - \overline{X}_n\right) - \left(\sigma^2 - \mu\right)\right\}/S_n^2$ et $\sqrt{n}\{(\overline{S}_n^2 - \overline{X}_n) - \left(\sigma^2 - \mu\right)\}/\overline{S}_n^2$ convergent en loi vers la distribution gaussienne centrée et de variance $\{\mu_4 - 2\mu_3 + \sigma^2 - \left(\sigma^2\right)^2\}/\left(\sigma^2\right)^2$.

Démonstration. – On obtient aussi ce corollaire à partir de la Proposition 3 et en tenant compte que $\overline{X}_n$ et S_n^2 convergent presque-sûrement vers μ et σ^2 respectivement. □

COROLLAIRE 6. – *Sous l'hypothèse nulle de la loi de Poisson, les statistiques $\sqrt{n/2}\left(S_n^2/\overline{X}_n - 1\right)$, $\sqrt{n/2}\left(\overline{S}_n^2/\overline{X}_n - 1\right)$, $\sqrt{n/2}\left(1 - \overline{X}_n/S_n^2\right)$ et $\sqrt{n/2}\left(1 - \overline{X}_n/\overline{S}_n^2\right)$ convergent en loi vers la loi gaussienne centrée et réduite, ces résultats restant valables si dans les statistiques précédentes on remplace $\sqrt{n}$ par $\sqrt{n-1}$.*

Démonstration. – Le résultat vient de l'énoncé qui le précéde en tenant compte que sous l'hypothèse poissonnienne, on a $\mu = \sigma^2 = \mu_3 = \lambda$, $\mu_4 = 3\lambda^2 + \lambda$. □

2.2.2. Transformation de Box-Cox

La statistique $(\overline{S}_n^2 - \overline{X}_n)/\overline{X}_n = (\overline{S}_n^2/\overline{X}_n - 1)$ est un cas particulier de la statistique de Box-Cox appliquée à l'indice de dispersion $\overline{S}_n^2/\overline{X}_n$. Ceci suggère de s'intéresser aux statistiques obtenues en appliquant à l'indice de dispersion les transformations de Box-Cox en général.

PROPOSITION 7. – *Considérons $\gamma \in \mathbb{R}$ et les statistiques*

$$T_{n;\gamma} = \begin{cases} \gamma^{-1}\sqrt{\frac{n-1}{2}}\left\{\left(\frac{\overline{S}_n^2}{\overline{X}_n}\right)^{\gamma} - 1\right\} & \text{si} \quad \gamma \neq 0 \\ \sqrt{\frac{n-1}{2}}\log\left(\frac{\overline{S}_n^2}{\overline{X}_n}\right) & \text{si} \quad \gamma = 0. \end{cases}$$

Alors, sous l'hypothèse nulle d'une loi de Poisson, $T_{n;\gamma}$ converge en loi vers la distribution gaussienne centrée et réduite.

Cette proposition résulte immédiatement du lemme suivant et du Corollaire 6.

LEMME 8. – *Soit g une fonction dérivable. Sous l'hypothèse nulle, la statistique $\sqrt{(n-1)/2}(g'(1))^{-1}\left\{g\left(\overline{S}_n^2/\overline{X}_n\right) - g\left(1\right)\right\}$ converge en loi vers la distribution gaussienne centrée et réduite.*

Démonstration. – La fonction g étant dérivable, le lemme resulte de la méthode «delta» (*e.g.* Saporta, 1990, pages 271-273). □

On observe que la statistique de Böhning (1994) n'est autre que

$$T_B = \sqrt{\frac{n-1}{2}}\left(\frac{\overline{S}_n^2}{\overline{X}_n} - 1\right) = T_{n;1}.$$

Par conséquent, sous l'hypothèse nulle que l'échantillon est poissonnien, T_B est asymptotiquement gaussienne centrée et réduite. Par ailleurs, en écrivant

$$T_B = \frac{T_F - (n-1)}{\sqrt{2(n-1)}},$$

où $T_F = nS_n^2/\overline{X}_n$ est asymptotiquement un khi-deux à $n-1$ degrés de liberté, cette expression de T_B en fonction de T_F montre que asymptotiquement T_B correspond à l'approximation de Paul Lévy d'une distribution de khi-deux par une distribution gaussienne. On sait aussi que cette approximation est l'une des moins bonnes. On

pourrait donc envisager les corrections suivantes de la statistique T_B afin d'améliorer la vitesse de convergence :

$$T_B^{'} = \sqrt{2\frac{nS_n^2}{\overline{X}_n}} - \sqrt{2n-3}$$

suggérée par l'approximation de Fisher pour une variable de khi-deux (Tassi, 1989, page 31) et

$$T_B^{''} = \sqrt{\frac{9(n-1)}{2}}\left[\sqrt[3]{\frac{n}{n-1}\frac{S_n^2}{\overline{X}_n}} - \left(1 - \frac{2}{9\,(n-1)}\right)\right]$$

suggérée par l'approximation de Wilson-Hilferty (Tassi, 1989, page 31).

Le résultat suivant montre une forme de croissance en $\gamma \in \mathbb{R}$ des statistiques $T_{n;\gamma}$ de Box-Cox.

THÉORÈME 9. – *Pour n fixé, les statistiques $T_{n;\gamma}$ sont continues en $\gamma \in \mathbb{R}$ et sont telles que, pour $\gamma_1 < 0 < \gamma_2 < 1 < \gamma_3$,*

$$T_{n;\gamma_1} \leqslant T_{n;0} \leqslant T_{n;\gamma_2} \leqslant T_{n;1} \leqslant T_{n;\gamma_3}.$$

Démonstration. – En posant $z = \overline{S}_n^2/\overline{X}_n$ et

$$g(z;\gamma) = \begin{cases} (z^\gamma - 1)/\gamma & \text{si} \quad \gamma \neq 0 \\ \log z & \text{si} \quad \gamma = 0, \end{cases}$$

il suffit de montrer que, pour tout $z > 0$, la fonction $\gamma \longmapsto g(z;\gamma)$ est continue (triviale) et vérifie : $g(z;\gamma_1) \leqslant g(z;0) \leqslant g(z;\gamma_2) \leqslant g(z;1) \leqslant g(z;\gamma_3)$, pour $\gamma_1 < 0 < \gamma_2 < 1 < \gamma_3$. En effet, on a besoin de deux étapes autour de $\gamma = 1$ et de $\gamma = 0$.

Pour $\gamma > 0$, la fonction $z \longmapsto g(z;\gamma)$ est convexe pour $\gamma = \gamma_3 > 1$, concave pour $\gamma = \gamma_2 \in]0,1[$ et, de plus, la droite $y = g(z;1)$ est la tangente de $g(z;\gamma)$ en $z = 1$; on en déduit les inégalités : $g(z;\gamma_2) \leqslant g(z;1) \leqslant g(z;\gamma_3)$.

Pour $\gamma < 1$, la fonction $z \longmapsto g(z;\gamma) - g(z;0)$ est négative pour $\gamma = \gamma_1 < 0$, nulle pour $\gamma = 0$ et positive pour $\gamma = \gamma_2 \in]0,1[$, et donc on obtient aisément le reste des inégalités : $g(z;\gamma_1) \leqslant g(z;0) \leqslant g(z;\gamma_2)$. □

2.3. Propriétés des tests statistiques

Puisque les distributions exactes des statistiques $T_F = nS_n^2/\overline{X}_n$ et $T_{n;\gamma}$ ne sont pas explicitement connues, on peut s'appuyer sur les distributions asymptotiques pour déterminer les régions critiques pour des tests d'adéquation à la distribution de Poisson. Notons $z_\alpha = -z_{1-\alpha}$ (resp. $\chi^2_{n-1,\alpha}$) le quantile d'ordre α de la loi

gaussienne centrée et réduite (resp. de la loi du khi-deux à $n-1$ degrés de liberté). Nous considérons les tests asymptotiques définis ci-dessous :

1. pour l'alternative de surdispersion

$$(a) \text{ Rejet de } H_0 \text{ si } T_F > \chi^2_{n-1,1-\alpha}; \quad (b) \text{ Rejet de } H_0 \text{ si } T_{n;\gamma} > z_{1-\alpha}$$

2. pour l'alternative de sousdispersion

$$(a) \text{ Rejet de } H_0 \text{ si } T_F < \chi^2_{n-1,\alpha}; \quad (b) \text{ Rejet de } H_0 \text{ si } T_{n;\gamma} < z_{\alpha}.$$

2.3.1. Equivalences asymptotiques des tests

Le résultat suivant montre que, pour n grand, la probabilité de rejeter l'hypothèse nulle à tort avec l'un des tests et de l'accepter avec l'autre est faible. En particulier, on pourrait prendre comme référence le test de Böhning où $\gamma = 1$.

PROPOSITION 10. – *Soit $\alpha \in]0,1[$ et $\gamma \neq \gamma'$ deux réels.*

(i) Si l'hypothèse alternative est la surdispersion (resp. sousdispersion), alors les tests de régions critiques $T_{n;\gamma} > z_{1-\alpha}$ et $T_{n;\gamma'} > z_{1-\alpha}$ (resp. $T_{n;\gamma} < z_{\alpha}$ et $T_{n;\gamma'} < z_{\alpha}$) sont asymptotiquement équivalents.

(ii) Si l'hypothèse alternative est la surdispersion (resp. sousdispersion), alors les tests de régions critiques $T_F > \chi^2_{n-1,1-\alpha}$ et $T_{n;\gamma} > z_{1-\alpha}$ (resp. $T_F < \chi^2_{n-1,\alpha}$ et $T_{n;\gamma} < z_{\alpha}$) sont asymptotiquement équivalents.

Démonstration. – Nous ne montrons les résultats (i) et (ii) que dans le cas de la surdispersion, car le cas de la sousdispersion se démontre de manière analogue. Aussi, sans perte de généralité, on suppose $\gamma < \gamma'$.

(i) On obtient le résultat à travers la Proposition 7 et le Théorème 9 de la façon suivante :

$$\lim_{n\to+\infty} \Pr(T_{n;\gamma} > z_{1-\alpha} \text{ et } T_{n;\gamma'} \leqslant z_{1-\alpha}) = \Pr(z_{1-\alpha} < Z \leqslant z_{1-\alpha}) = 0,$$

où Z est une variable aléatoire gaussienne centrée et réduite.

(ii) Puisque $T_{n;1} = [T_F - (n-1)]/\sqrt{2(n-1)}$, on obtient également le résultat à travers la Proposition 7 et le Théorème 9 comme suit :

$$\begin{aligned}
&\lim_{n\to+\infty} \Pr(T_F > \chi^2_{n-1,1-\alpha} \text{ et } T_{n;\gamma} \leqslant z_{1-\alpha}) \\
&\qquad = \lim_{n\to+\infty} \Pr\left(\frac{T_F - (n-1)}{\sqrt{2(n-1)}} > \frac{\chi^2_{n-1,1-\alpha} - (n-1)}{\sqrt{2(n-1)}} \text{ et } T_{n;\gamma} \leqslant z_{1-\alpha} \right) \\
&\qquad = \lim_{n\to+\infty} \Pr\left(T_{n;1} > z_{1-\alpha} \text{ et } T_{n;\gamma} \leqslant z_{1-\alpha} \right) \\
&\qquad = \Pr(z_{1-\alpha} < Z \leqslant z_{1-\alpha}) \\
&\qquad = 0,
\end{aligned}$$

où Z est une variable aléatoire gaussienne centrée et réduite. □

2.3.2. Probabilités d'erreur de seconde espèce

Nous montrons ci-dessous que le risque de seconde espèce β_γ est une fonction « décroissante » de γ sous une alternative de surdispersion et une fonction « croissante » de γ sous une alternative de sousdispersion au sens du Théorème 9. Ainsi, on peut déduire le comportement de la puissance $1 - \beta_\gamma$ des tests correspondants.

PROPOSITION 11. – *Soit* $\alpha \in]0,1[$ *et notons* $I_1 =]-\infty, 0[$, $I_2 = \{0\}$, $I_3 =]0,1[$, $I_4 = \{1\}$ *et* $I_5 =]1, +\infty[$. *Soit* $i < j$ *dans* $\{1, 2, \cdots, 5\}$. *Lorsque l'hypothèse alternative est la surdispersion (resp. sousdispersion), alors la probabilité d'erreur de seconde espèce du test asymptotique de région critique* $T_{n;\gamma_j} > z_{1-\alpha}$ *(resp.* $T_{n;\gamma_i} < z_\alpha$*) est plus faible que celui du test asymptotique de région critique* $T_{n;\gamma_i} > z_{1-\alpha}$ *(resp.* $T_{n;\gamma_j} < z_\alpha$*), pour tout* $\gamma_i \in I_i$ *et* $\gamma_j \in I_j$.

Démonstration. – Puisque $T_{n;\gamma}$ est une fonction croissante de γ au sens du Théorème 9, on en déduit les résultats à travers les deux propriétés suivantes :

– sous l'hypothèse alternative de la surdispersion, le risque de seconde espèce $\beta_\gamma = \Pr(T_{n;\gamma} \leqslant z_{1-\alpha})$ est une fonction « décroissante » de γ;
– sous l'hypothèse alternative de la sousdispersion, le risque de seconde espèce $\beta_\gamma = \Pr(T_{n;\gamma} \geqslant z_\alpha)$ est une fonction « croissante » de γ. □

3. Analyse empirique des performances

Les quantiles extrêmes de la distribution d'une statistique de test sont souvent celles qui sont intéressantes pour la mise en oeuvre d'un test d'hypothèse basé sur cette statistique. On a donc besoin d'avoir de bonnes approximations pour ces quantiles extrêmes lorsque la distribution exacte de la statistique de test n'est pas connue. L'objectif des simulations dont les résultats sont présentés ci-après est d'évaluer

empiriquement la qualité de l'approximation des quantiles extrêmes des distributions des statistiques $T_F = nS_n^2/\overline{X}_n$, $T_{n;\gamma}$ (avec $\gamma = -1, -0.1, 0, 0.1, 1$), T'_B, T''_B par les quantiles nominaux correspondant aux lois limites en fonction des tailles d'échantillon.

Pour $T_{n;\gamma}$ avec $|\gamma| > 1$, nous avons remarqué sur de nombreuses simulations que la convergence des tests asymptotiques sont moins rapides que la convergence des tests asymptotiques avec $T_{n;\gamma}$ où $|\gamma| \leqslant 1$. En plus, les statistiques $T_{n;-1}$, $T_{n;0}$ et $T_{n;1} = T_B$ nous intéressent particulièrement car elles sont en partie à l'origine de cette étude.

3.1. Probabilités d'erreur de première espèce

Dans la mesure où les régions critiques des tests sont déterminées à partir de distributions asymptotiques des statistiques de test, on peut s'attendre à ce que les probabilités effectives d'erreur de première espèce s'écartent de la valeur nominale correspondant à ces régions critiques. Une probabilité d'erreur de première espèce significativement supérieure à la valeur nominale de 5% indiquerait que la précision de l'approximation des quantiles extrêmes de la distribution de la statistique de test par celles de sa distribution limite n'est pas suffisante. Pour étudier le problème posé ci-dessus nous avons simulé 10000 échantillons pour chaque couple (n, μ) de taille d'échantillon (n) et de moyenne de loi de Poisson (μ). Les tailles n des échantillons examinées sont : 30, 50, 100; et nous considérons trois cas pour la moyenne μ de la loi de Poisson :

- les petites moyennes $\{0.5, 1\}$;
- la moyenne de niveau intermédiaire $\{5\}$;
- les grandes moyennes $\{10, 20\}$.

3.1.1. Tests bilatéraux

On considère un test bilatéral lorsque les informations à priori disponibles sur le phénomène étudié ne permettent pas de choisir entre la sousdispersion et la surdispersion comme situation alternative à l'inadéquation de la loi de Poisson. Les résultats reportés dans les tableaux 1, 2 et 3 suggèrent les observations suivantes pour un test bilatéral :

1. Les probabilités de première espèce sont supérieures à 0.05 (sauf pour $\mu = 0.5$ et $n = 50$, pour lequel on a 0.049) pour la statistique de test $T_{n;-1}$ lorsque l'échantillon est de taille ($n \leqslant 100$); les quantiles extrêmes de la distribution de cette statistique ne sont donc pas approximées avec une précision suffisante par celles de la loi gaussienne centrée et reduite.
2. Pour toutes les autres statistiques de test, et quelle que soit la taille de l'échantillon, les probabilités d'erreur de première espèce sont inférieures à la valeur nominale de 5% ou non significativement différentes de cette valeur nominale.

TABLEAU 1
Test bilatéral : probabilités empiriques d'erreur de première espèce
(on a mis en gras les valeurs supérieures à 0.05 *)*

$n = 30$	$\mu = 0.5$	$\mu = 1$	$\mu = 5$	$\mu = 10$	$\mu = 20$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	0.0357	0.0434	0.0470	0.0467	0.0460
$T_{n;-1}$	**0.0647**	**0.0736**	**0.0791**	**0.0786**	**0.0799**
$T_{n;-0.1}$	0.0347	**0.0521**	**0.0570**	**0.0569**	**0.0554**
$T_{n;0}$	0.0360	0.0487	**0.0539**	**0.0548**	**0.0505**
$T_{n;0.1}$	0.0355	0.0453	**0.0525**	**0.0524**	0.0492
$T_{n;1} = T_B$	0.0393	0.0423	0.0426	0.0439	0.0421
T'_B	0.0324	0.0434	0.0455	0.0442	0.0441
T''_B	0.0357	0.0434	0.0470	0.0466	0.0458

TABLEAU 2
Test bilatéral (suite)

$n = 50$	$\mu = 0.5$	$\mu = 1$	$\mu = 5$	$\mu = 10$	$\mu = 20$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	0.0425	0.0455	**0.0511**	0.0487	0.0458
$T_{n;-1}$	0.0490	**0.0653**	**0.0656**	**0.0683**	**0.0678**
$T_{n;-0.1}$	0.0430	0.0475	**0.0534**	**0.0546**	**0.0527**
$T_{n;0}$	0.0428	0.0464	**0.0517**	**0.0531**	0.0499
$T_{n;0.1}$	0.0415	0.0465	**0.0516**	**0.0519**	0.0490
$T_{n;1} = T_B$	0.0461	0.0472	0.0461	0.0467	0.0437
T'_B	0.0442	0.0455	0.0495	0.0481	0.0446
T''_B	0.0425	0.0455	**0.0507**	0.0486	0.0457

TABLEAU 3
Test bilatéral (fin)

$n = 100$	$\mu = 0.5$	$\mu = 1$	$\mu = 5$	$\mu = 10$	$\mu = 20$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	0.0439	0.0454	0.0477	0.0496	0.0482
$T_{n;-1}$	**0.0521**	**0.0556**	**0.0604**	**0.0602**	**0.0580**
$T_{n;-0.1}$	0.0446	0.0491	**0.0523**	**0.0533**	**0.0516**
$T_{n;0}$	0.0442	0.0486	**0.0506**	**0.0524**	**0.0511**
$T_{n;0.1}$	0.0438	0.0478	**0.0502**	**0.0525**	**0.0504**
$T_{n;1} = T_B$	0.0450	0.0470	0.0480	0.0487	0.0494
T'_B	0.0438	0.0454	0.0492	0.0493	0.0479
T''_B	0.0439	0.0454	0.0477	0.0495	0.0482

3.1.2. Tests unilatéraux de surdispersion

Pour les tests unilatéraux où l'hypothèse alternative est la surdispersion, on observe à partir des résultats de simulation présentés dans les tableaux 4, 5 et 6 que :

1. Les probabilités d'erreur de première espèce sont inférieures à 5% lorsque la statistique de test est de la famille $T_{n;\gamma}$ avec $\gamma \neq 1$, quelle que soit la taille de l'échantillon. On signale que, pour $T_{n;-1}$, ces probabilités sont très faibles.
2. La probabilité d'erreur de première espèce pour le test du khi-deux basé sur la statistique de Pearson T_F est de l'ordre de 5%, de même pour les tests basés sur la statistique de Böhning ($T_{n;1} = T_B$) et ses dérivées (T'_B et T''_B).

3.1.3. Tests unilatéraux de sousdispersion

Les résultats des simulations (tableaux 7, 8 et 9) suggèrent que pour les tests unilatéraux où l'hypothèse alternative est la sousdispersion :

1. La probabilité d'erreur de première espèce est majorée par la valeur nominale de 5% pour le test du khi-deux de Pearson ainsi que pour les tests basés sur la statistique de Böhning et les corrections associées.
2. La probabilité d'erreur de première espèce est significativement supérieure à la valeur nominale de 5% lorsque la statistique de test est de la forme $T_{n;\gamma}$ pour $\gamma \leqslant 0.1$ et $\mu \geqslant 1$. Pour $\gamma = 1$, correspondant à T_B, la probabilité empirique de première espèce est nettement inférieure à 5%.

TABLEAU 4
Test unilatéral de surdispersion : probabilités empiriques d'erreur de première espèce (on a mis en gras les valeurs supérieures à 0.05*)*

$n = 30$	$\mu = 0.5$	$\mu = 1$	$\mu = 5$	$\mu = 10$	$\mu = 20$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	0.0482	**0.0512**	0.0487	0.0473	0.0482
$T_{n;-1}$	0.0093	0.0087	0.0050	0.0074	0.0064
$T_{n;-0.1}$	0.0269	0.0302	0.0271	0.0274	0.0264
$T_{n;0}$	0.0309	0.0327	0.0295	0.0296	0.0289
$T_{n;0.1}$	0.0321	0.0355	0.0322	0.0332	0.0316
$T_{n;1} = T_B$	**0.0529**	**0.0598**	**0.0608**	**0.0588**	**0.0622**
T'_B	0.0482	**0.0541**	**0.0517**	0.0498	**0.0522**
T''_B	0.0482	**0.0512**	0.0488	0.0476	0.0482

TABLEAU 5
Test unilatéral de surdispersion (suite)

$n = 50$	$\mu = 0.5$	$\mu = 1$	$\mu = 5$	$\mu = 10$	$\mu = 20$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	**0.0538**	**0.0529**	**0.0506**	0.0475	0.0493
$T_{n;-1}$	0.0195	0.0155	0.0151	0.0145	0.0137
$T_{n;-0.1}$	0.0405	0.0369	0.0328	0.0322	0.0318
$T_{n;0}$	0.0420	0.0392	0.0350	0.0339	0.0349
$T_{n;0.1}$	0.0429	0.0406	0.0380	0.0366	0.0366
$T_{n;1} = T_B$	**0.0633**	**0.0599**	**0.0598**	**0.0570**	**0.0603**
T'_B	**0.0597**	**0.0544**	**0.0536**	0.0499	**0.0520**
T''_B	**0.0538**	**0.0529**	**0.0506**	0.0475	0.0493

TABLEAU 6
Test unilatéral de surdispersion (fin)

$n = 100$	$\mu = 0.5$	$\mu = 1$	$\mu = 5$	$\mu = 10$	$\mu = 20$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	0.0493	**0.0503**	**0.0504**	**0.0523**	**0.0521**
$T_{n;-1}$	0.0253	0.0248	0.0236	0.0230	0.0238
$T_{n,-0.1}$	0.0370	0.0380	0.0384	0.0386	0.0396
$T_{n;0}$	0.0377	0.0394	0.0402	0.0411	0.0417
$T_{n;0.1}$	0.0413	0.0418	0.0413	0.0435	0.0429
$T_{n;1} = T_B$	**0.0568**	**0.0584**	**0.0568**	**0.0581**	**0.0589**
T'_B	0.0493	**0.0524**	**0.0522**	**0.0541**	**0.0541**
T''_B	0.0493	**0.0503**	**0.0504**	**0.0524**	**0.0521**

TABLEAU 7
Test unilatéral de sousdispersion : probabilités empiriques d'erreur de première espèces (on a mis en gras les valeurs supérieures à 0.05*)*

$n = 30$	$\mu = 0.5$	$\mu = 1$	$\mu = 5$	$\mu = 10$	$\mu = 20$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	0.0228	0.0410	0.0483	0.0474	0.0477
$T_{n;-1}$	**0.0839**	**0.0959**	**0.1115**	**0.1126**	**0.1110**
$T_{n;-0.1}$	0.0411	**0.0641**	**0.0757**	**0.0748**	**0.0764**
$T_{n;0}$	0.0400	**0.0585**	**0.0713**	**0.0706**	**0.0720**
$T_{n;0.1}$	0.0369	**0.0553**	**0.0673**	**0.0669**	**0.0679**
$T_{n;1} = T_B$	0.0102	0.0214	0.0297	0.0273	0.0275
T'_B	0.0201	0.0365	0.0444	0.0424	0.0417
T''_B	0.0228	0.0410	0.0483	0.0474	0.0477

TABLEAU 8
Test unilatéral de sousdispersion (suite)

$n = 50$	$\mu = 0.5$	$\mu = 1$	$\mu = 5$	$\mu = 10$	$\mu = 20$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	0.0293	0.0431	0.0466	0.0474	0.0481
$T_{n;-1}$	**0.0720**	**0.0931**	**0.0909**	**0.0960**	**0.0963**
$T_{n;-0.1}$	0.0471	**0.0663**	**0.0648**	**0.0688**	**0.0691**
$T_{n;0}$	0.0460	**0.0636**	**0.0627**	**0.0655**	**0.0659**
$T_{n;0.1}$	0.0396	**0.0576**	**0.0595**	**0.0620**	**0.0625**
$T_{n;1} = T_B$	0.0199	0.0274	0.0330	0.0334	0.0325
T'_B	0.0265	0.0398	0.0429	0.0443	0.0442
T''_B	0.0293	0.0431	0.0465	0.0474	0.0481

TABLEAU 9
Test unilatéral de sousdispersion (fin)

$n = 100$	$\mu = 0.5$	$\mu = 1$	$\mu = 5$	$\mu = 10$	$\mu = 20$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	0.0379	0.0415	0.0500	0.0480	0.0496
$T_{n;-1}$	**0.0732**	**0.0742**	**0.0857**	**0.0791**	**0.0837**
$T_{n;-0.1}$	**0.0545**	**0.0549**	**0.0663**	**0.0626**	**0.0648**
$T_{n;0}$	**0.0506**	**0.0530**	**0.0643**	**0.0602**	**0.0624**
$T_{n;0.1}$	0.0496	**0.0510**	**0.0617**	**0.0585**	**0.0606**
$T_{n;1} = T_B$	0.0308	0.0344	0.0396	0.0393	0.0392
T'_B	0.0364	0.0393	0.0470	0.0455	0.0471
T''_B	0.0379	0.0415	0.0500	0.0480	0.0496

3.2. Probabilités d'erreur de seconde espèce

Toujours sous l'hypothèse nulle du modèle de Poisson, nous avons considérés deux modèles pour l'étude empirique de la probabilité d'erreur de seconde espèce, donc de la puissance :

— le modèle de Poisson tronqué en zéro qui est un cas de sousdispersion

— le modèle binomial négatif pour la surdispersion.

3.2.1. Poisson tronqué en zéro en modèle alternatif

Le modèle de Poisson tronqué en zéro est défini par la famille des lois de probabilités

$$p\left(x;\lambda\right)=\frac{\lambda^{x}}{x!(e^{\lambda}-1)}\ ,\ x\in\mathbb{N}^{*},\ \lambda>0,$$

où $\theta=\log\lambda$ est le paramètre canonique. Sa moyenne et sa variance sont données respectivement par :

$$\mu\left(\lambda\right)=\sum_{x=1}^{+\infty}xp\left(x;\lambda\right)=\frac{\lambda e^{\lambda}}{e^{\lambda}-1},\ \ \forall\lambda>0\qquad\text{et}$$

$$\sigma^{2}\left(\lambda\right)=\sum_{x=1}^{+\infty}\left(x-\mu(\lambda)\right)^{2}p\left(x;\lambda\right)=\mu\left(\lambda\right)\left[1-e^{-\lambda}\mu\left(\lambda\right)\right],\ \ \forall\lambda>0.$$

Puisqu'on a $\sigma^{2}\left(\lambda\right)<\mu\left(\lambda\right),\forall\lambda>0$, le modèle de Poisson tronqué en zéro de moyenne μ est sousdispersé par rapport au modèle poissonnien de même moyenne. De plus, on a : $\lim_{\lambda\to+\infty}\mu(\lambda)/\lambda=1$ et $\sigma^{2}(\lambda)\sim\mu(\lambda)$ pour $\lambda\to+\infty$; ces résultats signifient que, pour les grandes valeurs de λ, la loi de Poisson tronquée en zéro converge vers la loi de Poisson de même moyenne.

Les simulations dont les résultats sont présentés dans les tableaux 10 et 11 ont été réalisées en considérant les lois de moyenne $\mu=1,2,5,10$ pour les tailles d'échantillon $n=30,50,100$. Les 10000 répliques ont été effectuées pour chaque combinaison (n,μ).

Ces résultats suggèrent que pour une même valeur nominale de la probabilité d'erreur de première espèce de 5%, le test de sousdispersion ayant la probabilité d'erreur de seconde espèce la plus faible est celui construit avec la statistique $T_{n;-1}$. Le test du khi-deux (T_F) et le test de Böhning ($T_{n;1}=T_B$) ont les probabilités d'erreur de seconde espèce les plus élevées. Aucun des tests étudiés dans ce travail ne permet de discriminer entre le modèle de Poisson et le modèle de Poisson tronqué en zéro si la moyenne théorique est élevée ($\mu\geqslant 5$), mais ceci est relativement logique puisque le rapport variance/moyenne est proche de 1 (supérieur à 0.966 pour $\mu\geqslant 5$, et à 0.999 pour $\mu\geqslant 10$) d'après le tableau 12.

TABLEAU 10
Loi de Poisson tronquée en zéro :
probabilités empiriques d'erreur de seconde espèce

	$\mu = 1$			$\mu = 2$		
	$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	0.1322	0.0271	0.0004	0.7324	0.5708	0.2691
$T_{n;-1}$	0.0475	0.0087	0.0002	0.4949	0.3685	0.1608
$T_{n;-0.1}$	0.0820	0.0184	0.0002	0.6227	0.4695	0.2164
$T_{n;0}$	0.0877	0.0200	0.0002	0.6447	0.4801	0.2243
$T_{n;0.1}$	0.0967	0.0208	0.0003	0.6618	0.4947	0.2313
$T_{n;1} = T_B$	0.2445	0.0500	0.0006	0.8530	0.6884	0.3338
T'_B	0.1542	0.0323	0.0004	0.7708	0.5922	0.2821
T''_B	0.1322	0.0271	0.0004	0.7324	0.5708	0.2691

TABLEAU 11
Loi de Poisson tronquée en zéro :
probabilités empiriques d'erreur de seconde espèce (fin)

	$\mu = 5$			$\mu = 10$		
	$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	0.9602	0.9516	0.9494	0.9469	0.9547	0.9515
$T_{n;-1}$	0.9086	0.9110	0.9195	0.9185	0.9318	0.9399
$T_{n;-0.1}$	0.9406	0.9381	0.9370	0.9403	0.9459	0.9479
$T_{n;0}$	0.9457	0.9416	0.9391	0.9431	0.9483	0.9487
$T_{n;0.1}$	0.9472	0.9449	0.9419	0.9456	0.9508	0.9495
$T_{n;1} = T_B$	0.9687	0.9625	0.9589	0.9537	0.9565	0.9527
T'_B	0.9626	0.9549	0.9515	0.9493	0.9546	0.9508
T''_B	0.9602	0.9517	0.9494	0.9472	0.9548	0.9515

TABLEAU 12
Loi de Poisson tronquée en zéro :
variance et indice de dispersion en fonction de la moyenne

Moyennes	Variances	Indices de dispersion
1	0.632	0.632
2	1.457	0.729
5	4.831	0.966
10	9.995	0.999

3.2.2. Binomial négatif en modèle alternatif

On a souvent recours à la loi binomiale négative comme alternative à la loi de Poisson lorsqu'il s'avère que la variabilité observée dans les données est significativement supérieure à celle que prévoit le modèle poissonnien. Rappelons que le modèle binomial négatif est défini par les probabilités individuelles

$$p\,(x;\mu,\phi) = \frac{\Gamma\,(x+\phi)}{x!\Gamma\,(\phi)}\left(\frac{\phi}{\phi+\mu}\right)^{\phi}\left(\frac{\mu}{\phi+\mu}\right)^{x}, \quad x \in \mathbb{N},\ \mu > 0, \phi > 0,$$

où $\theta = \log\{(\mu/\phi)/(1+\mu/\phi)\}$ est le paramètre canonique et ϕ est le paramètre de dispersion; sa moyenne est μ et sa variance est $\mu + \mu^2/\phi$.

Les simulations dont les résultats sont présentés au tableau 13 ont été réalisées en considérant les lois binomiales négatives de moyenne $\mu = 0.5, 5, 10$, avec le même paramètre de dispersion $\phi = 0.1$, pour les tailles d'échantillon $n = 30, 50, 100$. Les 10 000 répliques ont été effectuées pour chaque combinaison (n, μ).

Aucun des tests examinés dans ce travail ne permet de discriminer entre la loi de Poisson et la loi binomiale négative lorsque la moyenne est faible et l'indice de dispersion théorique est proche de 1 (voir le tableau 14). Par ailleurs, les résultats du tableau 13 montrent à l'évidence qu'avec l'alternative de la loi binomiale négative le test construit avec la statistique $T_{n;-1}$ est celui dont la probabilité d'erreur de seconde espèce est la plus élevée, sauf pour $\mu = 0.5$ et $n = 30$. Pour les échantillons de tailles $n \leqslant 100$, le test construit avec la statistique de Böhning $T_{n;1} = T_B$ est celui pour lequel la probabilité d'erreur de seconde espèce est la plus faible pour une même valeur nominale de la probabilité d'erreur de première espèce.

TABLEAU 13
Lois binomiales négatives : probabilités d'erreur de seconde espèce

	$\mu = 0.5$			$\mu = 5$			$\mu = 10$		
	$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$
$\frac{nS_n^2}{\overline{X}_n} = T_F$	0.9463	0.9471	0.9287	0.6222	0.4343	0.1815	0.2309	0.0785	0.0032
$T_{n;-1}$	0.9512	0.9568	0.9546	0.9051	0.7103	0.3230	0.5881	0.2175	0.0106
$T_{n;-0.1}$	0.9647	0.9557	0.9432	0.7215	0.5187	0.2216	0.3135	0.1083	0.0045
$T_{n;0}$	0.9602	0.9559	0.9413	0.7026	0.5022	0.2144	0.2985	0.1024	0.0042
$T_{n;0.1}$	0.9582	0.9550	0.9393	0.6854	0.4889	0.2068	0.2834	0.0969	0.0040
$T_{n;1} = T_B$	0.9339	0.9379	0.9184	0.5655	0.3889	0.1537	0.1941	0.0635	0.0024
T'_B	0.9464	0.9454	0.9268	0.6072	0.4222	0.1730	0.2212	0.0739	0.0030
T''_B	0.9463	0.9471	0.9287	0.6222	0.4343	0.1816	0.2309	0.0785	0.0032

TABLEAU 14
Loi binomiale négative avec $\phi = 0.1$:
indice de dispersion en fonction de la moyenne

Moyennes	Variances	Indices de dispersion
0.5	0.525	1.05
1	1.1	1.1
5	7.5	1.5
10	20	2

4. Remarque finale

À l'aide aussi des simulations pour des tests unilatéraux, nous avons remarqué que la propriété classifiant les statistiques $T_{n;\gamma}$ (Théorème 9) a des répercutions sur les risques de seconde espèce (Proposition 11). Ceci implique entre autre que le choix de $\gamma \in \mathbb{R}$ de $T_{n;\gamma}$ doit être dicté par celui de l'hypothèse alternative (surdispersion ou sousdispersion). Pour mieux décrire en fait l'influence du signe de γ dans le choix de l'alternative, nous introduisons (au prix d'une redondance) :

$$U_{n;\gamma} = \begin{cases} \gamma^{-1}\sqrt{\frac{n-1}{2}}\left\{\left(\frac{\overline{X}_n}{\overline{S}_n^2}\right)^{\gamma} - 1\right\} & \text{si} \quad \gamma \neq 0 \\ \sqrt{\frac{n-1}{2}}\log\left(\frac{\overline{X}_n}{\overline{S}_n^2}\right) & \text{si} \quad \gamma = 0. \end{cases}$$

Les statistiques $U_{n;\gamma}$ sont aussi de Box-Cox mais obtenues à partir de l'inverse de l'indice de dispersion de Fisher $\overline{X}_n/\overline{S}_n^2$. De manière analogue à $T_{n;\gamma}$, nous avons d'abord la normalité asymptotique des $U_{n;\gamma}$ (Proposition 7) ainsi que la continuité et la croissance de $U_{n;\gamma}$ en $\gamma \in \mathbb{R}$ (Théorème 9). Les tests asymptotiques associés à $U_{n;\gamma}$ sont définis par :

– pour l'alternative de surdispersion

$$\text{Rejet de } H_0 \text{ si } U_{n;\gamma} < z_\alpha$$

– pour l'alternative de sousdispersion

$$\text{Rejet de } H_0 \text{ si } U_{n;\gamma} > z_{1-\alpha}.$$

Puisque la relation entre $U_{n;\gamma}$ et $T_{n;\gamma}$ est donnée par

$$U_{n;\gamma} = -T_{n;-\gamma},$$

nous avons ensuite de façon similaire la Proposition 10 (i) et la Proposition 11 avec $U_{n;\gamma}$. Enfin, pour compléter la Proposition 11, nous avons le principal résultat comparatif entre les tests unilatéraux basés sur $T_{n;\gamma}$ et $U_{n;\gamma}$ (ou simplement l'effet du signe de γ) :

PROPOSITION 12. – *Soit $\alpha \in]0,1[$ et $\gamma' < 0 < \gamma$ deux réels non nuls.*

(i) Lorsque l'hypothèse alternative est la surdispersion, le test asymptotique de région critique $T_{n;\gamma} > z_{1-\alpha}$ (resp. $T_{n;\gamma'} > z_{1-\alpha}$) possède un risque de seconde espèce moins élevé (resp. plus élevé) que le test asymptotique de région critique $U_{n;\gamma} < z_\alpha$ (resp. $U_{n;\gamma'} < z_\alpha$).

(ii) Lorsque l'hypothèse alternative est la sousdispersion, le test asymptotique de région critique $U_{n;\gamma} > z_{1-\alpha}$ (resp. $U_{n;\gamma'} > z_{1-\alpha}$) possède un risque de seconde espèce moins élevé (resp. plus élevé) que le test asymptotique de région critique $T_{n;\gamma} < z_\alpha$ (resp. $T_{n;\gamma'} < z_\alpha$).

Démonstration. – Nous montrons seulement le cas (ii) de la sousdispersion, car (i) se démontre de manière analogue.

On suppose $\gamma > 0$. Pour cela, il suffit de montrer que le risque de seconde espèce $\beta_{U;\gamma}$ du test asymptotique de région critique $U_{n;\gamma} > z_{1-\alpha}$ est plus faible que le risque de seconde espèce $\beta_{T;\gamma}$ du test asymptotique de région critique $T_{n;\gamma} < z_\alpha$. En effet, on a successivement :

$$\begin{aligned}
\beta_{U;\gamma} &= \Pr(U_{n;\gamma} \leqslant z_{1-\alpha}) \\
&= \Pr(T_{n;-\gamma} \geqslant z_\alpha) \\
&< \Pr(T_{n;\gamma} \geqslant z_\alpha) = \beta_{T;\gamma},
\end{aligned}$$

où l'inégalité stricte est obtenue par le Théorème 9 avec $\gamma > 0$. Cette inégalité change de sens si on remplace γ (> 0) par γ' (< 0) et cela démontre la seconde partie du résultat. □

Pour illustrer cette Proposition 12 à travers les simulations déja étudiées dans la section précédente, on peut considérer le cas $\gamma = 1$ correspondant à $T_{n;1} = T_B$ de Böhning (1994) et $U_{n;1} = -T_{n;-1}$. Cela nous conduit en outre à l'observation suivante, laquelle peut être suggérée par le Corollaire 6 : afin d'utiliser un test plus approprié, la «bonne normalisation» de la différence entre $\overline{X}_n$ et $\overline{S}_n^2$ pour tester la surdispersion (resp. sousdispersion) s'obtient par $\overline{X}_n$ (resp. $\overline{S}_n^2$), lesquelles $\overline{X}_n$ et $\overline{S}_n^2$ sont les deux estimateurs sans biais du paramètre de la loi de Poisson. De manière générale, il s'agit de l'effet de l'indice de dispersion de Fisher $\overline{S}_n^2/\overline{X}_n$ et de son inverse $\overline{X}_n/\overline{S}_n^2$ pour une efficacité des tests unilatéraux selon l'alternative de surdispersion et de sousdispersion.

Remerciements

Les auteurs remercient chaleureusement leur collègue Eliette Albert pour ses remarques pertinentes durant la préparation de cet article. Ils remercient également le rédacteur en chef de la revue ainsi que son comité de rédaction pour les commentaires et suggestions.

Références

BÖHNING D. (1994), A note on a test for Poisson overdispersion, *Biometrika* 81, 418-419.

BROWN L.D., ZHAO L.H. (2002), A test for the Poisson distribution, *Sankhyā* A 64 (3), 611-625.

CASTILLO J., PÉREZ-CASANY M. (1998), Weighted Poisson distribution for overdispersion and underdispersion situations, *Ann. Inst. Statist. Math.* 50, 567-585.

COX D.R. (1983), Some remarks on overdispersion, *Biometrika* 70, 269-274.

GELFAND A.E., DALAL S.R.A. (1990), A note on overdispersed exponential families, *Biometrika* 77, 55-64.

GIANO L.M., SCHAFER D.W. (1992), Diagnostics for overdispersion, *J. Amer. Statist. Assoc.* 87, 795-804.

HOEL P.G. (1943), On indices of dispersion, *Ann. Math. Statist.* 14, 155-162.

JOHNSON N.L., KOTZ S., KEMP A.W. (1992), *Univariate Discrete Distributions,* Second Edition, John Wiley & Sons, New York.

KOKONENDJI C.C., DEMÉTRIO C.B.G., DOSSOU-GBÉTÉ S. (2004), Some discrete exponential dispersion models : Poisson-Tweedie and Hinde-Demétrio classes, *SORT* 28 (2), 201-214.

SAPORTA G. (1990), *Probabilités, Analyse des Données et Statistique*, Technip, Paris.

SMYTH G.K., PODLICH H.M. (2000), Score tests for Poisson variation against general alternatives. *Computing Sciences and Statistics* 32, 97-103.

TASSI Ph. (1989), *Méthodes Statistiques*, Economica, Paris.

TIAGO DE OLIVEIRA J. (1965), Some elementary tests of mixtures of discrete distributions. In *Classical and Contagious Discrete Distributions,* Ed. G. P. Patil, pp. 379-384. New York : Pergamon.