Sur l'ajustement du nombre de sinistres

ROSE-MARIE KESTEMONT et JOSÉ PARIS, Louvain-la-Neuve

Sur l'ajustement du nombre de sinistres

1 Introduction

Dans [5], Gossiaux et Lemaire comparent plusieurs méthodes d'ajustement du nombre de sinistres sur différents exemples. Ils tirent plusieurs conclusions. La première indique que la loi de Poisson simple ne convient jamais. La deuxième confirme le fait bien connu suivant: la méthode des moments est inadéquate pour estimer les paramètres de la loi de probabilité d'une variable aléatoire discrète quand la distribution n'est pas symétrique. La troisième conclusion indique qu'ils n'ont pas mis en évidence une loi unique capable d'ajuster valablement les exemples analysés. Nous nous proposons d'indiquer ici qu'une telle loi existe et qu'il est possible d'en estimer efficacement les paramètres. Enfin nous proposerons deux mesures simples et efficaces de la qualité de l'ajustement obtenu.

2 Le Modèle

Il est bien connu (Feller [3], Hofmann [6], Thyrion [12]), qu'on peut caractériser une loi de probabilité du nombre $N(t)$ d'événements survenant dans l'intervalle $[0, \tau]$ par

$$\begin{aligned} \Pr[N(t)=0] &= \pi(0,t)=e^{-\theta(t)} \\ \Pr[N(t)=n] &= \pi(n,t)=(-1)^n \frac{t^n}{n!}\,\pi^{(n)}(0,t) \qquad n=1,2,\ldots \end{aligned} \tag{1}$$

dans laquelle

$$\theta(t)=\int_0^t \theta'(u)\,du \tag{2}$$

avec θ' complètement monotone sur R^+ et $\theta(0)=0$. Ainsi construite, la fonction $\pi(0,t)$ est complètement monotone et en vertu du théorème de Bernstein-Widder, la loi de probabilité de $N(t)$ peut être interprétée comme une loi de Poisson composée ou pondérée du type

$$\pi(0,t)=\int_0^\infty e^{-\lambda t}\,U(d\lambda) \tag{3}$$

dans laquelle la fonction de structure U décrit l'hétérogénéité a priori des assurés. La loi de probabilité de $N(t)$ peut également être identifiée à une loi de Poisson généralisée du type

$$N(t) = \xi_1 + \xi_2 + \ldots + \xi_{L(t)} \tag{4}$$

dans laquelle les ξ_i sont des v.a. entières et positives, indépendantes entre elles, et aussi indépendantes de la v.a. $L(t)$ qui suit une loi de Poisson simple.
Ce type de loi est donc très général. Pour le rendre opérationnel, il y a lieu de faire un choix convenable de la fonction θ'. Nous utiliserons la fonction

$$\theta'(t) = \frac{p}{(1+ct)^a} \qquad p>0,\ c>0,\ a\geqslant 0 \tag{5}$$

introduite par Hofmann [6] et utilisée par Thyrion [12].
Elle conduit à

$$\theta(t) = \begin{cases} pt & a=0 \\ \dfrac{p}{c(1-a)}\,[(1+ct)^{1-a}-1] & 0<a<1 \\ \dfrac{p}{c}\ln(1+ct) & a=1 \\ \dfrac{p}{c(1-a)}\,[1-(1+ct)^{1-a}] & 1<a. \end{cases} \tag{6}$$

Ce choix se justifie à beaucoup d'égards. Du point de vue mathématique, la fonction θ' est une transformée de Laplace. Du point de vue statistique, ce choix permet d'englober plusieurs lois de probabilité usuelles. C'est le paramètre a qui permet de différencier les lois introduites. Pour $a=0$, on retrouve la loi de Poisson simple; pour $a=1$, on retrouve la loi binomiale négative; pour $a=2$, on retrouve la loi de Polya-Aeppli; si a tend vers l'infini, on retrouve la loi de Neyman type A.
Les différentes probabilités se trouvent par les relations de récurrence

$$\begin{aligned} &\pi(0,t) = e^{-\theta(t)} \\ &\pi(n+1,t) = \frac{tp}{n+1}\,(1+ct)^{-a} \sum_{k=0}^{n} \frac{\Gamma(a+k)}{\Gamma(a)k!}\left(\frac{ct}{1+ct}\right)^k \pi(n-k;t) \end{aligned} \tag{7}$$

La fonction génératrice des cumulants factoriels de $N(t)$, à savoir

$$\xi(s) = \ln E[(1+s)^{N(t)}] = \ln P[0, -ts] \tag{8}$$

fournit

$$\begin{aligned} K_{[1]} &= pt \\ K_{[2]} &= pcat^2 \\ K_{[3]} &= pc^2a(a+1)t^3 \\ K_{[4]} &= pc^3a(a+1)\,(a+2)t^4. \end{aligned} \tag{9}$$

On en déduit facilement l'espérance et la variance de $N(t)$, soit

$$\begin{aligned} EN(t) &= pt \\ \operatorname{var} N(t) &= p(1+ac)t^2. \end{aligned} \tag{10}$$

La formule (8) montre encore que la somme de deux variables aléatoires indépendantes $N_1(t)$ et $N_2(t)$ qui suivent une loi de probabilité du type indiqué plus haut, de paramètre respectif p_1, p_2 et de mêmes paramètres a et c, obéit encore à une loi du même type de paramètres p_1+p_2, a, c.
Les données analysées étant annuelles, nous serons conduits à poser $t=1$. Dans ce cas, on notera que ξ est également la fonction génératrice des cumulants de la variable aléatoire Λ, dont U est la fonction de répartition. Elle fournit

$$\begin{aligned} \chi_1 &= p \\ \chi_k &= p(ac)^{k-1}\left(1+\frac{1}{a}\right)\left(1+\frac{2}{a}\right)\cdots\left(1+\frac{k-1}{a}\right) \quad k\geqslant 2. \end{aligned} \tag{11}$$

On en déduit

$$\begin{aligned} E\Lambda &= p \\ \operatorname{var} \Lambda &= pac \end{aligned} \tag{12}$$

et le coefficient de variation de Λ, à savoir $\sqrt{\frac{ac}{p}}$, est une mesure de l'hétérogénéité relative a priori du portefeuille.
Il peut également s'avérer judicieux d'examiner les coefficients d'asymétrie (skewness) et d'aplatissement (leptokurtosis) exprimés à l'aide des cumulants factoriels (Anscombe [1])

$$\begin{aligned} \frac{K_{[3]}}{p(ac)^2} &= 1+\frac{1}{a} \\ \frac{K_{[4]}}{p(ac)^3} &= \left(1+\frac{1}{a}\right)\left(1+\frac{2}{a}\right). \end{aligned} \tag{13}$$

Il s'agit de deux fonctions décroissantes de a qui pourront également aider au

choix de la loi. Les relations (13) sont également vraies si on remplace les cumulants factoriels de $N(t)$ par les cumulants du même ordre de U.
Le réalisme en assurance conduit à imposer la restriction

$$\lim_{t \to +\infty} \pi(0, t) = 0. \tag{14}$$

Elle implique immédiatement que $a \leqslant 1$. Ainsi la loi de Poisson simple et la loi binomiale négative apparaissent comme des cas limites des lois proposées.
Remarque: Si nous considérons la loi de Poisson simple

$$\begin{aligned} P(0, t) &= e^{-pt} \\ P(n, t) &= e^{-pt} \frac{(pt)^n}{n!} \end{aligned} \tag{15}$$

de même espérance que les lois de Poisson composées introduites, la comparaison peut être amplifiée. Nous avons:

$$\begin{gathered} P(0, t) \leqslant \pi(0, t) \\ \frac{P(1, t)}{P(0, t)} = pt \geqslant \frac{\pi(1, t)}{\pi(0, t)} = t\theta'(t). \end{gathered} \tag{16}$$

Ces relations sont bien connues (Feller [3]). Pour les lois introduites, la première traduit le fait que, pour t fixé, θ' est une fonction décroissante de a et dès lors, π est une fonction croissante de a. De ce fait, elle constitue l'une des raisons qui conduisent à écarter les lois de Poisson simples.
La deuxième relation montre l'importance de θ'. Par contre, la relation

$$\pi(1, t) \leqslant P(1, t) \tag{17}$$

n'est pas vraie en général.

3 Estimation

En sciences actuarielles, on accepte volontiers que la loi de probabilité qui caractérise le nombre de sinistres est une loi de Poisson composée. Quand on veut aborder le problème de l'estimation, deux voies se présentent. La première consiste à essayer d'estimer la loi de structure U. Dans [11], Teicher a montré que celle-ci était identifiable et Simar [10] a donné un algorithme qui permet d'estimer U par la méthode du maximum de vraisemblance. Cette direction de recherche a suscité de nombreux travaux au cours des dernières années. Il est cependant important de noter que puisque les valeurs observées de N pendant une année sont limitées aux entiers $\leqslant 7$, le nombre de classes homogènes identifiables par ce procédé est également $\leqslant 7$ (Lindsay [9]). Ainsi, ce procédé met nécessairement en évidence une fonction de structure discrète. Conceptuellement, cela présente l'inconvénient de n'identifier que peu de classes homogènes d'assurés. Dès lors, cette méthode non paramétrique peut avantageusement être remplacée par une méthode paramétrique. En effet, les paramètres qui figurent dans U se retrouvent dans la loi de probabilité de $N(t)$. Dès lors, on peut choisir une classe de lois de probabilité de $N(t)$ suffisamment larges et en estimer les paramètres. La forme de la loi proposée au paragraphe 2 présente cet avantage. On pourrait penser en estimer les paramètres par la méthode du maximum de vraisemblance. Mais d'une part, l'expression (7) des probabilités ne laisse aucun espoir de trouver la forme analytique des estimateurs. D'autre part, même les méthodes numériques ne s'avèrent pas simples. Dans [12], Thyrion propose d'estimer les paramètres en utilisant la moyenne et la variance de l'échantillon ainsi que la proportion des assurés n'ayant provoqué aucun sinistre. Le choix de la moyenne expérimentale comme estimateur de p est raisonnable. Selon toute vraisemblance, elle constitue un estimateur exhaustif pour p. Cette affirmation, nullement prouvée, semble cependant réaliste car elle est vraie pour $a=0$ et pour $a=1$.

Comme déjà rappelé plus haut, utiliser la variance expérimentale comme estimateur est contre-indiqué. Dès lors, pour estimer a et c, nous utiliserons le résultat de Lambert et Tierney [8] qui montre que pour les lois de Poisson composées, estimer les probabilités par identification aux proportions observées est asymptotiquement aussi efficace que la méthode du maximum de vraisemblance. En résumé, nous proposons de prendre pour estimation de p, a, c la solution du système

$$\begin{cases} \bar{N}=p \\[2mm] \dfrac{N_0}{N}=e^{-\theta(1)} \\[2mm] \dfrac{N_1}{N_0}=\theta'(1)=\dfrac{p}{(1+c)^a}. \end{cases}$$

La résolution du système des deux dernières équations peut nécessiter un assez grand nombre d'itérations.

4 Analyse des Résultats

Le premier tableau en appendice reprend, pour les exemples analysés dans [5] et dans l'ordre indiqué, l'estimation des paramètres p, a, c. Celle-ci conduit aux effectifs attendus des tableaux suivants. Ils confirment la qualité de la procédure attestée par les valeurs uniformément faibles de la statistique chi-carré usuelle d'ajustement. La loi proposée est toujours facilement acceptée.
Comme mesure de la qualité de l'ajustement, nous proposons plutôt les deux statistiques

$$T = S^2 - p(1+ac)$$

$$V = k_3 - (pc^2a(a+1) + 3pca + p)$$

dans lesquelles S^2 est la variance estimée, k_3 la k_3-statistique de Fisher. Ces deux statistiques ont une espérance nulle. Leur variance peut être facilement calculée. L'estimation de ces deux statistiques, reprise dans le tableau 1, confirme la qualité des ajustements proposés.
L'estimation H de l'hétérogénéité relative a priori qui complète ce tableau indique que la méthode proposée fonctionne bien même pour des exemples assez diversifiés.
Appliquée à d'autres exemples relevés dans la littérature, elle a fourni également d'excellents résultats.

Bibliographie

[1] F. J. Anscombe: Sampling theory of the negative binomial and logarithmic series distribution. Biometrika, 37 (1950), pp. 358–381.
[2] D. Böhning: Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. The Annals of statistics, vol. 10, No 3 (1982), 1006–1008.
[3] W. Feller: On a general class of "contagious" distribution. Annals of Mathematical Statistics, 14 (1943), pp. 389–400.
[4] W. Feller: An introduction to probability theory and its application. Volume 2 (1971), Wiley.
[5] A.-M. Gossiaux et J. Lemaire: Méthodes d'ajustement de distributions de sinistres. Bulletin des actuaires suisses, I (1981), pp. 87–95.
[6] M. Hofmann: Über zusammengesetzte Poisson-Prozesse und ihre Anwendungen in der Unfallversicherung. Bulletin des actuaires suisses, vol. 55–3 (1955).
[7] N. Laird: Nonparametric maximum likelihood estimation of a mixing distribution. J. Amer. Statist. Assoc. 73 (1978), 805–811.
[8] D. Lambert and L. Tierney: Asymptotic properties of maximum likelihood estimates in the mixed Poisson model. The Annals of Statistics, vol. 12, No 4 (1984), 1388–1399.
[9] B. Lindsay: The geometry of mixture likelihoods: a general theory. The Annals of statistics, vol. 11, No 1 (1983), 86–94.

[10] L. Simar: Maximum likelihood estimation of a compound Poisson Process. Ann. Statist. 4 (1976), 1200–1209.
[11] H. Teicher: Identifiability of mixtures. Ann. Math. Statist. 32 (1961), 244–248.
[12] P. Thyrion: Contribution à l'étude du bonus pour non-sinistre en assurance auto. Astin I (1960), pp. 143–162.

Rose-Marie Kestemont
José Paris
Université Catholique de Louvain
2, Chemin du Cyclotron
B-1348 Louvain-la-Neuve

Résumé

Cet article indique une famille de lois de probabilité suffisamment large et une méthode efficace d'estimation des paramètres qui permet d'ajuster avec succès les distributions de sinistres recontrées dans la littérature et d'en expliquer les similitudes et les différences.

Zusammenfassung

Im vorliegenden Artikel wird eine ausreichend grosse Klasse von Wahrscheinlichkeitsverteilungen definiert sowie eine effiziente Methode zur Schätzung ihrer Parameter angegeben. Dies erlaubt eine einfache und wirksame Anpassung an numerische Daten.

Summary

This article defines a sufficiently large class of probability distributions as well as an efficient method of estimating its parameters, which allows a better fit to numerical data compared to traditional approaches.

Appendice

Tableau 1

	$\hat{p}$	$\hat{a}$	$\hat{c}$	T	V	H
1	0,10108	0,57415	0,10843	0,006372	0,000617	0,784789
2	0,08650	0,68892	0,60714	−0,000132	−0,003911	2,198975
3	0,21435	0,40766	0,81835	0,003067	0,029563	1,247542
4	0,15514	0,44060	0,35457	−0,000061	−0,000757	1,003479
5	0,14422	0,32082	0,42468	0,000001	0,000127	0,971957
6	0,13174	0,27648	0,18638	−0,000004	−0,000108	0,625415

Tableau 2

Exemple 1
$N=106974$
$\bar{X}=0{,}10108 \quad S^2=0{,}10745$

k	observés	attendus
0	96978	96978
1	9240	9240
2	704	699,67
3	43	52,02
4	9	3,97
⩾5	0	0,34

Exemple 2
$N=4000$
$\bar{X}=0{,}0865 \quad S^2=0{,}12255$

k	observés	attendus
0	3719	3719
1	232	232
2	38	37,43
3	7	8,45
4	3	2,21
5	1	0,63
⩾6	0	0,27

Exemple 3
$N=9461$
$\bar{X}=0{,}21435 \quad S^2=0{,}28893$

k	observés	attendus
0	7840	7840
1	1317	1317
2	239	231,43
3	42	52,00
4	14	14,09
5	4	4,29
6	4	1,41
7	1	0,77
⩾8	0	0,00

Exemple 4
$N=119853$
$\bar{X}=0{,}15514 \quad S^2=0{,}17932$

k	observés	attendus
0	103704	103704
1	14075	14075
2	1766	1766,78
3	255	255,39
4	45	42,26
5	6	7,69
6	2	1,49
⩾7	0	0,39

Exemple 5
$N=23589$
$\bar{X}=0{,}14422 \quad S^2=0{,}16387$

k	observés	attendus
0	20592	20592
1	2651	2651
2	297	297,40
3	41	40,28
4	7	6,70
5	0	1,28
6	1	0,27
⩾7	0	0,07

Exemple 6
$N=421240$
$\bar{X}=0{,}13174 \quad S^2=0{,}13852$

k	observés	attendus
0	370412	370412
1	46545	46545
2	3935	3935,16
3	317	317,07
4	28	27,74
5	3	2,70
⩾6	0	0,33