JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

PAUL DEHEUVELS

Statistique des événements rares

Journal de la société française de statistique, tome 140, n° 3 (1999), p. 53-66

<http://www.numdam.org/item?id=JSFS_1999__140_3_53_0>

STATISTIQUE DES ÉVÉNEMENTS RARES

Paul DEHEUVELS *

RÉSUMÉ

Nous exposons quelques-unes des méthodes pouvant être utilisées pour l'analyse statistique des événements rares. Nous montrons en particulier que le modèle de Poisson, particulièrement bien adapté à l'analyse de séries d'observations indépendantes, peut tomber en défaut et mener à des conclusions erronées dans le cas de dépendances sérielles marquées.

1. INTRODUCTION – ÉVÉNEMENTS RARES – LE MODÈLE DE POISSON

Au cours des dernières décennies, on a pu observer dans le monde scientifique un intérêt renouvelé pour la modélisation et l'analyse statistique des événements rares. Nous nous référons par exemple à l'ouvrage d'Aldous (1989) qui décrit une grande variété de phénomènes rentrant dans cette catégorie. D'une manière générale, on appellera événement rare tout événement dont la probabilité p de se produire est «très petite». Nous donnerons plus loin une formulation plus quantitative de cette définition en liant la valeur du paramètre p au nombre n des observations disponibles pour évaluer cette première quantité par des méthodes d'inférence statistique. L'objet de cette note, qui ne prétend aucunement à l'exhaustivité, est de présenter de manière concise quelques uns des développements récents et anciens de cette théorie permettant son usage appliqué. Il deviendra évident par la suite que l'analyse statistique de phénomènes où le nombre d'événements observables est, par essence, réduit du fait de leur faible fréquence, impose une très grande rigueur pour l'adoption d'hypothèses de modélisation, dont l'influence peut être considérable sur les conclusions finales d'une analyse.

Considérons un exemple simple mettant en évidence l'intérêt de l'étude des événements rares. On observe un nombre n élevé (par exemple de plusieurs milliers) de patients recevant un traitement nouveau dont l'efficacité thérapeutique a été reconnue au cours d'expérimentations antérieures. Il peut se faire, et c'est d'ailleurs malheureusement presque toujours le cas, qu'une petite proportion de ces patients développe une réaction de nature *toxique* (ou allergique) vis à vis du composant actif qui leur est administré. Il est alors nécessaire d'évaluer le risque correspondant afin de vérifier qu'il ne dépasse pas

* L.S.T A , Université Paris VI, 7 avenue du Château, 92340 Bourg la Reine,
e mail : pd@ccr jussieu fr

une limite admissible, au delà de laquelle l'apport thérapeutique du traitement sera compensé négativement par des effets indésirables imposant des réserves à son utilisation. Pour cela, on observe la suite $X_1, ..., X_n$ des indicatrices définies pour le $k^{\text{ème}}$ patient par

$$X_k = \begin{cases} 1 & \text{si un événement (de toxicité) est observé;} \\ 0 & \text{si ce n'est pas le cas.} \end{cases}$$

Il paraîtra naturel à toute personne familière avec le calcul des probabilités de supposer que les observations $X_1, ..., X_n$ constituent une suite de variables aléatoires indépendantes de même loi. En notant $\mathbb{P}(E)$ la probabilité d'un événement E, les variables $X_1, ..., X_n$ suivent alors une loi, dite de Bernoulli, définie, indépendamment de $1 \leqslant k \leqslant n$, par

$$\mathbb{P}(X_k = 1) = 1 - \mathbb{P}(X_k = 0) = p.$$

Le nombre $0 < p < 1$, supposé positif et très petit (par exemple $p = 0.001 = 0.1\%$), représente la proportion moyenne des patients pour lesquels une réaction toxique peut être observée. Sa valeur est inconnue et doit être évaluée à partir de $X_1, ..., X_n$. La solution de ce problème est bien connue et se trouve dans tous les ouvrages de statistique élémentaire. Il s'agit d'abord de construire une statistique $\hat{p}$, fonction de $X_1, ..., X_n$ qui estime p à partir des observations. Ensuite, pour toute valeur spécifiée de $0 < \alpha < 1$, on cherche à déterminer des intervalles de confiance $[\tilde{p}_L, \tilde{p}_U]$ où $\tilde{p}_L$ et $\tilde{p}_U$ sont également des statistiques, fonctions des observations $X_1, ..., X_n$, et vérifiant l'égalité approximative

$$\mathbb{P}(p \in [\tilde{p}_L, \tilde{p}_U]) \approx 1 - \alpha.$$

La résolution du premier problème est relativement simple. Celle du second pose des difficultés importantes qui seront évoquées plus loin.

L'estimateur du paramètre inconnu p qui possède toutes les « bonnes » propriétés est donné par

$$\hat{p} = n^{-1} S_n \quad \text{où} \quad S_n = \sum_{i=1}^{n} X_i$$

désigne le nombre total d'événements observés. En particulier, $\hat{p}$ résume toute l'information apportée par les observations sur p. Il n'échappera pas au lecteur que la loi exacte de S_n est une loi binômiale, dont l'expression est

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$
$$\text{pour} \quad k = 0, 1, ..., n-1, n.$$

Dans le cas où le nombre d'observations n est tellement grand que, malgré la petitesse de p, on puisse considérer que l'asymptotique $p \to 0$ et $np \to \infty$ est

approximativement satisfaite, on peut utiliser une approximation de la loi de $\hat{p}$ par une loi normale. Plus précisément, si on désigne par

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt = \int_{-\infty}^{z} \varphi(t) dt \quad \text{et} \quad \varphi(t) = \frac{e^{-t^2/2}}{\sqrt{2\pi}} = \Phi'(t)$$

la fonction de répartition et la densité d'une loi normale (ou de Laplace-Gauss) $N(0,1)$ standard, il ressort de travaux anciens d'Uspensky (1937) (voir aussi Raff (1956)) qu'il existe une constante universelle C telle que

$$\sup_{k \in \mathbb{N}} \left| \mathbb{P}(S_n \leqslant k) - \Phi(x_k) - \left\{ \frac{1-2p}{6\sqrt{np(1-p)}} \right\} (1 - x_k^2) \varphi(x_k) \right| \leqslant \frac{C}{np(1-p)},$$

où, pour $k \in \mathbb{N}$,

$$x_k = \frac{k - np}{\sqrt{np(1-p)}}.$$

De manière plus approximative, sous l'hypothèse que $np \to \infty$, on peut obtenir un intervalle de confiance approché pour la valeur inconnue p, en posant

$$\tilde{p}_L = \hat{p} - \nu_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{et} \quad \tilde{p}_U = \hat{p} + \nu_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ce qui revient à écrire que

$$\mathbb{P}\left(p \in \left[\hat{p} - \nu_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + \nu_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right] \right) \approx 1 - \alpha,$$

et

$$\mathbb{P}\left(p \in \left[0, \hat{p} + \nu_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right] \right) \approx 1 - \alpha,$$

où, pour $0 < \delta < 1$, ν_δ désigne la quantile supérieure d'ordre δ de la loi normale $N(0,1)$, solution de l'équation $1 - \Phi(\nu_\delta) = \delta$. Par exemple, $\nu_{\alpha/2} \approx 1.96$ et $\nu_\alpha \approx 1.64$ pour $\alpha = 0.05 = 5\%$. De tels intervalles de confiance peuvent être utilisés, en pratique et avec des erreurs relativement peu importantes sur le niveau de confiance $1 - \alpha$, dès que $np \geqslant 5$ (voir Nayatani et Kurahara (1964)). Prenons un exemple où l'on observe $S_n = 5$ réactions toxiques pour $n = 1000$ patients. Dans ce cas, $\hat{p} = 0.005 = 0.5\%$, et les formules ci-dessus, avec $\alpha = 5\%$ et $\nu_{\alpha/2} \approx 1.96$ fournissent

$$\mathbb{P}\Big(p \in [0.06\%, 0.94\%]\Big) \approx 95\% \quad \text{et} \quad \mathbb{P}\Big(p \in [0\%, 0.87\%]\Big) \approx 95\%.$$

Pour une description détaillée de telles méthodes, on pourra se référer au chapitre 3 du livre de Johnson, Kotz et Kemp (1992).

Nous sommes plus particulièrement intéressés dans la suite de cette note par le cas où p et n se situent dans une asymptotique de la forme

$$p \to 0 \quad \text{et} \quad np \not\to \infty,$$

et pour laquelle la convergence asymptotique de $(S_n - np)/\sqrt{np(1-p)}$ vers la loi normale n'a donc pas lieu. Dans ce cas, la loi de Poisson représente la limite de référence pour la distribution de S_n (Poisson (1837)). Avant de présenter les applications de cette distribution, il convient d'élargir quelque peu le cas de cette étude en revenant sur les hypothèses de base que nous avons postulées au départ.

Nous avons supposé jusqu'ici que les indicatrices $X_1, ..., X_n$ étaient de même loi de probabilité. En d'autres termes, ceci suppose que la probabilité d'observer un événement toxique est la même chez tous les patients. Il est parfaitement clair qu'une telle hypothèse est rarement satisfaite dans la pratique. Pour illustrer cette évidence, supposons que la population des patients comprenne des hommes, des femmes, des enfants, ou d'autres catégories spécifiques de personnes présentant par essence des sensibilités différentes à l'action des composants toxiques du produit. Sous cette hypothèse, on devra alors faire appel à un modèle plus général, où

$$\mathbb{P}(X_k = 1) = p_k \quad \text{pour} \quad k = 1, ..., n,$$

dépend de l'indice k. On peut alors rechercher une estimation de la valeur moyenne inconnue

$$\overline{p} = \frac{1}{n} \sum_{i=1}^{n} p_i$$

des probabilités $p_1, ..., p_n$, en utilisant la même statistique $\hat{p} = n^{-1} S_n$ que précédemment. La description des propriétés de $\hat{p}$ dans ce nouveau contexte est alors rendue inextricable du fait que la loi exacte de S_n, caractérisée par les différentes valeurs prises par $\mathbb{P}(S_n = k)$ pour $k = 0, ..., n$, a une expression compliquée en fonction de k et de $p_1, ..., p_n$, dont l'évaluation numérique est délicate même pour de petites valeurs de k. Par exemple, pour $k = 0, 1$, on obtient que

$$\mathbb{P}(S_n = 0) = \prod_{k=1}^{n} (1 - p_k),$$

$$\mathbb{P}(S_n = 1) = \sum_{k=1}^{n} p_k \prod_{\substack{j=1 \\ j \neq k}}^{n} (1 - p_j) - \sum_{1 \leqslant k < \ell \leqslant n} p_k p_\ell \prod_{\substack{j=1 \\ j \neq k, \ell}}^{n} (1 - p_j).$$

Pour pallier cette difficulté, on fait alors usage de la technique, dite d'approximation Poissonienne, qui consiste à remplacer la loi exacte de S_n par celle, plus simple, d'une variable aléatoire T_n suivant une loi de Poisson. Par des méthodes de couplage, il est d'ailleurs possible de construire formellement S_n et T_n simultanément, de sorte qu'on puisse évaluer les probabilités de la forme $\mathbb{P}(S_n = r, T_n = s)$ pour tout couple $r, s \in \mathbb{N}$. On peut ainsi définir une variable aléatoire T_n sur le même espace de probabilités que la suite $X_1, ..., X_n$, suivant une loi de Poisson définie par

$$\mathbb{P}(T_n = k) = \frac{\mu^k}{k!} \exp(-\mu) \quad \text{pour} \quad k \in \mathbb{N},$$

où l'espérance $\mu = \mathbb{E}(T_n)$ de T_n est choisie de telle sorte que

$$\mu = \mathbb{E}(T_n) = n\overline{p} = \sum_{i=1}^{n} p_i.$$

On constate alors que les lois de S_n et de T_n sont généralement très voisines l'une de l'autre. Pour exprimer cette proximité, citons une remarquable inégalité, due à Barbour et Hall (1984). Il est possible de construire T_n et S_n sur le même espace de probabilités, de sorte que

$$\begin{aligned}\mathbb{P}(S_n \neq T_n) &= \sup_{A \subset \mathbb{N}} |\mathbb{P}(S_n \in A) - \mathbb{P}(T_n \in A)| = \frac{1}{2}\sum_{i=1}^{\infty} |\mathbb{P}(S_n = i) - \mathbb{P}(T_n = i)| \\ &\leqslant \frac{\sum_{i=1}^{n} p_i^2}{\sum_{i=1}^{n} p_i} \left\{1 - \exp\left(-\sum_{i=1}^{n} p_i\right)\right\} \\ &\leqslant \min\left\{\sum_{i=1}^{n} p_i^2, \max\{p_1, ..., p_n\}\right\}.\end{aligned}$$

On voit par ces formules que l'erreur induite dans les calculs de probabilités par le remplacement de S_n par T_n n'excède pas $\max\{p_1, ..., p_n\}$, expression qui, par hypothèse, reste très petite dans les applications considérées. De tels résultats permettent donc de remplacer directement la loi (complexe) de S_n par celle (plus simple) de T_n, avec une erreur pouvant être bien maîtrisée.

L'évaluation précise de bornes optimales raffinant les inégalités ci-dessus fait appel à des méthodes relativement sophistiquées sur le plan mathématique, parmi lesquelles il convient de mentionner les techniques de semi-groupes d'opérateurs (voir Deheuvels et Pfeifer (1986,1988)) et la méthode de Stein-Chen, pour laquelle nous renvoyons au livre de Barbour, Holst et Janson (1992). A titre d'exemple de résultats du genre, citons l'évaluation suivante, due à Deheuvels et Pfeifer (1986), valable lorsque $p_1, ..., p_n$ varient avec n de sorte que $n\overline{p} = \sum_{i=1}^{n} p_i \to \infty$ et $\{\sum_{i=1}^{n} p_i^2\} / \sum_{i=1}^{n} p_i \to 0$. On a alors, sous cette asymptotique,

$$\mathbb{P}(S_n \neq T_n) = \frac{1}{2}\sum_{i=1}^{\infty} |\mathbb{P}(S_n = i) - \mathbb{P}(T_n = i)| = \frac{(1 + o(1))}{\sqrt{2\pi e}} \frac{\sum_{i=1}^{n} p_i^2}{\sum_{i=1}^{n} p_i}.$$

On arrive à ce point à la conclusion que la loi de Poisson semble constituer un modèle universel de description asymptotique pour un nombre cumulé d'événements rares.

Le remplacement approximatif de la loi de S_n par la loi de Poisson de T_n ne suffit pas à résoudre le problème de la construction d'intervalles de confiance pour $\overline{p}$. Il n'existe en effet que des méthodes approximatives, pas toujours très précises, pour cela, et nous renvoyons au chapitre 4 de Johnson, Kotz et Kemp (1992) pour une étude détaillée de cette question. La méthode standard de construction est la suivante. Ayant observé la valeur T d'une loi de Poisson de paramètre inconnu $\mu = \mathbb{E}(T)$, on évalue les quantités μ_L et μ_U, solutions des équations

$$\exp(-\mu_L)\sum_{i=T}^{\infty}\frac{\mu_L^j}{j!} = \frac{\alpha}{2} \quad \text{et} \quad \exp(-\mu_U)\sum_{i=0}^{x}\frac{\mu_U^j}{j!} = \frac{\alpha}{2}.$$

On conclut alors en écrivant qu'approximativement

$$\mathbb{P}(\mu \in [\mu_L, \mu_U]) \approx 1 - \alpha.$$

D'une manière pratique, ces formules sont peu commodes à utiliser, d'une part par le fait qu'elles induisent des calculs numériques nécessitant l'usage de tables ou de logiciels, et d'autre part, parce qu'elles restent approximatives, avec une erreur non quantifiable de manière précise sur le niveau de confiance exact, voisin de mais non égal à $1-\alpha$, finalement obtenu. Revenant à l'exemple décrit initialement, supposons qu'on ait observé $S_n = S = 4$ réactions toxiques pour $n = 1000$ patients. La valeur de p_{max} correspondant à un intervalle de confiance approximatif de la forme

$$\mathbb{P}(\overline{p} \leqslant p_{max}) \approx 0.95 = 95\%$$

est obtenue comme suit. Tout d'abord, on recherche la solution de l'équation

$$\sum_{i=0}^{S}\frac{\lambda^i}{i!}\exp(-\lambda) = \sum_{i=0}^{4}\frac{\lambda^i}{i!}\exp(-\lambda) = 1 - \alpha = 0.95,$$

En faisant, par exemple, appel à la table 9.4 d'Owen (1962), on obtient $\lambda = 9.154$. On trouve alors p_{max} en égalant $np_{\max} = 1000p_{max} = \lambda$. Ainsi, on obtient dans cet exemple que

$$\hat{p} = 0.40\% \quad \text{alors que} \quad p_{max} \approx 0.92\%.$$

A titre d'exemple, et bien que la condition $np \geqslant 5$ mentionnée plus haut ne soit pas satisfaite, il est intéressant de comparer la valeur de p_{max} à celle qui serait obtenue par une approximation normale. Celle-ci fournit une évaluation de cette même borne donnée par

$$p^*_{max} = \hat{p} + 1.64 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.73\%.$$

La comparaison de ces deux bornes, finalement toutes aussi imprécises l'une que l'autre, illustre la difficulté du problème, et l'embarras d'une prise de décision effective en faisant usage de ces méthodes. Supposons, par exemple, que la valeur maximale admissible pour $\overline{p}$ soit 0.8%. Peut-on, avec de tels chiffres, aboutir à une décision rationnelle acceptant ou infirmant cette hypothèse ? Comment également évaluer les risques correspondant à chacune des décisions possibles sous l'une ou l'autre des hypothèses, et plus généralement, en fonction des valeurs possibles de p ?

2. UNE APPROCHE BAYÉSIENNE DE LA PRISE DE DÉCISION

La difficulté de prendre une décision à partir des techniques classiques de l'analyse statistique standard peut être surmontée de manière particulièrement élégante à l'aide de techniques bayésiennes (*cf.* Lindley (1972), Lecoutre (1984)). Ces méthodes supposent, d'une manière générale, que le paramètre d'intérêt (ici p) est partiellement connu au départ par le biais d'une loi de probabilité dite *a priori*. Il paraîtra conforme à l'intuition que, en l'absence de toute information sur $p \in [0,1]$, on choisisse comme candidat logique pour une telle loi *a priori* la loi uniforme sur $[0,1]$. Ayant observé $X_1, ..., X_n$, on évalue ensuite une nouvelle loi de probabilité, dite *a posteriori*, et incorporant toute l'information contenue par la loi *a priori* et les observations. On se sert ensuite de cette nouvelle distribution pour évaluer les probabilités correspondant à toute décision qui pourra être prise en fonction de la valeur du paramètre.

L'utilisation de méthodes bayésiennes a toujours fait l'objet de controverses, par le caractère subjectif lié au choix d'une loi *a priori*. Dans le cas particulier de notre étude, cette objection n'a pas de raison d'être dans la mesure où la plupart des choix logiques possibles de lois *a priori* pour p n'influent pas significativement de manière pratique sur les décisions qui pourront être déduites de la loi *a posteriori* correspondante.

Pour illustrer ce phénomène, revenons au cas où $X_1, ..., X_n$ sont des indicatrices indépendantes de même loi de probabilité de paramètre p. Dans ce cas, on modélise l'incertitude sur p par une loi, dite *a priori* non informative (*cf.* §3.4 dans Robert (1992)). Ceci revient à supposer (voir également Villegas (1977) pour la généralisation à des modèles *a priori* impropres) que p suit une loi *a priori* $\beta(1,1)$ ou $\beta(\frac{1}{2},\frac{1}{2})$ sur $[0,1]$. D'une manière générale, Z est dite suivre une loi Béta de paramètres $r > 0$ et $s > 0$, notée $\beta(r,s)$, si, pour tout borélien $A \subseteq \mathbb{R}$

$$\mathbb{P}(Z \in A) = \int_A f(z)dz, \quad \text{avec} \quad f(z) = \frac{z^{r-1}(1-z)^{s-1}}{\beta(r,s)} \quad \text{pour} \quad 0 < z < 1,$$

où

$$\beta(r,s) = \int_0^1 t^{r-1}(1-t)^{s-1}dt = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} \quad \text{et} \quad \Gamma(r) = \int_0^\infty t^{r-1}e^{-t}dt$$

sont les fonctions β et Γ d'Euler, définies pour $r > 0$ et $s > 0$ comme ci-dessus. En particulier, on observera que la loi $\beta(1,1)$ n'est autre que la loi uniforme sur $[0,1]$, mentionnée plus haut comme «candidat logique» pour la loi a priori de p. D'une manière générale, un choix de loi a priori $\beta(r,s)$ pour p, combiné à l'observation de $S_n = S$ suivant une loi binômiale de paramètres n et p, implique que p suit une loi *a posteriori* $\beta(r+S, s+n-S)$ (*cf.* Exemple 1.19, p.33 dans Robert (1992)). Une telle distribution permet d'évaluer par

une intégrale simple la probabilité qu'une hypothèse portant sur ce paramètre soit vérifiée ou non. On constate également que, pour les grandes valeurs de n, le résultat ainsi obtenu est peu sensible au choix de la loi a priori ($\beta(1,1)$ ou $\beta(\frac{1}{2},\frac{1}{2})$).

A titre d'exemple, je donnerai les résultats d'une expérience de bio-équivalence développée par G. Derzko (Sanofi-Recherche). Elle comprend deux lots de patients d'un effectif de 200 chacun. Le nombre d'événements observés pour le premier lot est de 11. Pour le second lot, il y a 3 événements. Il s'agit ici de vérifier une hypothèse, dite de bio-équivalence relative, caractérisée par le fait que les probabilités p_1 et p_2 des événements associés aux lots 1 et 2 respectivement vérifient les inégalités

$$0.8 \times p_1 \leqslant p_2 \quad \text{et} \quad 0.8 \times p_2 \leqslant p_1.$$

On obtient une probabilité *a posteriori* de cette condition égale à 3.87% qui justifie amplement le rejet de l'hypothèse de bio-équivalence.

3. LA LOI DES SÉRIES ET LA MISE EN DÉFAUT DU MODÈLE DE POISSON

Nous avons vu au §1 que la loi de Poisson représentait le modèle asymptotique le plus naturel pour décrire la distribution de $S_n = X_1 + ... + X_n$ lorsque $X_1, ..., X_n$ sont des indicatrices d'événements rares indépendantes, mais pas nécessairement de même loi. Une telle représentation peut devenir toutefois inexacte lorsque les variables aléatoires $X_1, ..., X_n$ sont dépendantes. En fait, ce n'est que très récemment qu'on a pu obtenir une caractérisation des familles de distribution de probabilités pouvant être utilisées pour modéliser la loi de S_n dans ce dernier cas.

Pour comprendre ce qui va suivre, il est utile de présenter un deuxième exemple de phénomène générant des événements rares. Considérons désormais une chronique $Z_1, ..., Z_n$ représentant, par exemple, les précipitations annuelles observées dans une zone donnée. On s'intéresse aux événements catastrophiques correspondant à des dépassements de ces relevés au delà d'un niveau exceptionnel u. On génère alors une suite d'indicatrices de la forme

$$X_k = \begin{cases} 1 & \text{si } Z_k \geqslant u, \\ 0 & \text{autrement.} \end{cases}$$

Si les observations $Z_1, ..., Z_n$ constituaient une suite de variables aléatoires indépendantes de même loi (voir par exemple Leadbetter, Lindgren et Rootzén (1983) pour des conditions de stationnarité et d'indépendance asymptotique impliquant le même résultat), on serait ramené au cas traité au §1, avec $p = \mathbb{P}(Z_i \geqslant u)$ pour $i = 1, ..., n$. On conclurait ainsi que la loi de S_n tende à devenir arbitrairement voisine d'une loi de Poisson d'espérance np lorsque $n \to \infty$ et $p \to 0$. Pour cette raison, le modèle de Poisson a été abondamment utilisé dans le cadre de la théorie des valeurs extrêmes depuis son origine (*cf.* Gumbel (1958)).

En fait, si l'on suppose seulement la stationnarité dans le temps des observations sérielles $X_1, ..., X_n$, un autre phénomène se produit. Pour le décrire, désignons par $t_1 < ... < t_{S_n}$ les valeurs de l'entier k pour lesquelles $X_k = 1$. Considérons le processus ponctuel aléatoire sur $[0, 1]$ dont les temps d'arrivée sont $\tau_1 = n^{-1}t_1, ..., \tau_{S_n} = n^{-1}t_{S_n}$. Il est possible de vérifier que, sous les hypothèses du §1, lorsque $np \to \lambda$, $p \to 0$ et $n \to \infty$, $\{\tau_k : 1 \leqslant k \leqslant S_n\}$ converge vers un processus de Poisson (voir Kingman (1993) pour un exposé général sur ces modèles). En particulier, il n'est pas permis d'observer des grappes de points confondus, puisque, pour le processus limite, la probabilité d'observer plus de deux points τ dans un intervalle de longueur h est donnée par la formule

$$P(h) = 1 - \{1 + \lambda h\}\exp(-\lambda h) = O(h^2) \quad \text{lorsque} \quad h \to 0.$$

Le nombre moyen de sites où l'on observerait des points τ confondus est alors obtenu comme limite de $h^{-1} \times O(h^2) = O(h) \to 0$ lorsque $h \to 0$, et est ainsi nul.

Inversement, si la condition d'indépendance n'est pas satisfaite, on peut obtenir une nouvelle famille (englobant le processus de Poisson comme cas particulier) de processus ponctuels limites faisant partie de la classe des processus de Poisson par grappes. Ces processus sont construits avec des temps d'arrivée τ générés par un processus de Poisson homogène, et en chaque point τ apparaît une grappe, composée d'un nombre aléatoire π_τ d'événements simultanés. De ce fait, le nombre total S_n d'événements observés ne se comporte plus comme une variable de Poisson, mais comme une somme composée, de la forme

$$S_n = \sum_{i=1}^{T_n} \pi_i,$$

où $\pi_1, \pi_2, ...$ sont des variables aléatoires indépendantes de même loi, représentant les tailles respectives des grappes d'événements, et T_n suit une loi de Poisson indépendante de $\pi_1, \pi_2, ...$.

On retrouve ici un phénomène bien connu de manière empirique et appellé «loi des séries». En termes heuristiques celle-ci correspond à l'idée qu'une catastrophe n'est, en général pas isolée dans le temps, car le temps source τ où elle se produit est associé à une grappe π de catastrophes concomitantes approximativement simultanées. Il est très remarquable que la description systématique de cette propriété n'ait été achevée qu'en 1988 par Hsing, Hüsler et Leadbetter (1988) et Hsing (1989). On se réfèrera, par exemple, à Nandagopalan (1994) pour des versions multivariées de ce résultat.

Il ressort de tout ceci que les dépendances mutuelles des observations peuvent aboutir à sortir du cadre de la loi de Poisson pour aboutir à des lois de Poisson composées. L'inférence statistique pour de tels phénomènes est alors bien plus complexe que ce que nous avons décrit et fait appel à la notion nouvelle d'index extrémal (voir Nandagopalan (1994)). En particulier, la connaissance seule du nombre S_n total d'événements $X_k = 1$ observés lorsque l'indice k varie entre 1 et n devient insuffisante pour évaluer convenablement la loi de probabilité

sous-jacente. De ce fait, elle ne permet plus de construire des intervalles de confiance fiables pour p. Ceci mène à réviser systématiquement toutes les procédures de contrôle basées exclusivement sur l'usage des lois binômiales et de Poisson.

4. CONCLUSION

L'analyse d'événements rares est un domaine de recherche loin encore d'avoir atteint sa maturité scientifique, et où bon nombre de résultats importants sont relativement récents. La loi de Poisson reste la règle standard, bien que son utilisation sans discernement aboutisse nécessairement à l'obtention de conclusions inexactes si les conditions qui justifient son emploi ne sont pas satisfaites. L'utilisation de méthodes bayésiennes semble bien adaptée ici à une approche décisionnelle.

RÉFÉRENCES BIBLIOGRAPHIQUES

ALDOUS D. (1989). *Probability Approximations via the Poisson Clumping Heuristic.* Springer-Verlag, New York.

BARBOUR A.D. et HALL P. (1984). On the rate of Poisson convergence. *Mathematical Proceedings of the Cambridge Philosophical Society.* **95** 473–480.

BARBOUR A.D., HOLST L. et JANSON S. (1992). *Poisson Approximation.* Clarendon Press, Oxford.

DEHEUVELS P. et PFEIFER D. (1986). A semigroup approach to Poisson approximation. *Annals of Probability.* **14** 663–676.

DEHEUVELS P. et PFEIFER D. (1988). On a relationship between Uspensky's theorem and Poisson approximation. *Annals of the Institute of Statistical Mathematics.* **40** 671–681.

GUMBEL E.J. (1958). *Statistics of Extremes.* Columbia University Press, New York.

HSING T. (1989). Extreme value theory for multivariate stationary sequences. *J. Multivariate Analysis.* **29** 274–291.

HSING T., HÜSLER J. et LEADBETTER M.R. (1988). On the exceedance point process for a stationary sequence. *Probab. Theor. Related Fileds.* **78** 97–112.

JOHNSON N.L., KOTZ S. et KEMP A.W. (1992). *Univariate Discrete Distributions.* Wiley, New York.

KINGMAN J.F.C. (1993). *Poisson Processes.* Oxford Studies in Probability **3**, Clarendon Press, Oxford.

LEADBETTER M.R., LINDGREN G. et ROOTZÉN H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer-verlag, New York.

LECOUTRE B. (1984). *L'Analyse Bayésienne des Comparaisons.* Presses Universitaires de Lille, Lille.

LINDLEY D.V. (1972). *Bayesian Statistics, A Review.* CBMS-NSF Regional Conference Series in Applied Mathematics **2**, SIAM, Philadelphia.

NANDAGOPALAN S. (1994). On the multivariate extremal index. *Journal of Research of the National Institute of Standards and Technology.* **99** 543–550.

NAYATANI Y. et KURAHARA B. (1964). A condition for using the approximation by the normal and the Poisson distribution to compute the confidence intervals. Reports of Statistical Application Research, JUSE, **11** 99–105.

OWEN D.B. (1962). *Handbook of Statistical Tables.* Addison-Wesley, Reading.

POISSON S.D. (1837). *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile, Précédée des Règles Générales du Calcul des Probabilités.* Bachelier, Paris.

RAFF M.S. (1956). On approximating the point binomial. *Journal of the American Statistical Association.* **51** 293–303.

ROBERT C. (1992). *L'Analyse Statistique Bayesienne.* Economica, Paris.

USPENSKY J.V. (1937). *Introduction to Mathematical Probability.* McGraw-Hill, New York.

COMPTES RENDUS DE LECTURE

L'aléatoire et le vivant

Yves et Maurice GIRAULT

1 vol., 190 pages, Diderot Multimedia, 1999, 129 F, ISBN : 2-84352-335-4

Les lecteurs du Journal de la Société Française de Statistique (et particulièrement du présent volume) savent combien le développement de la statistique est intimement lié à celui des sciences du vivant. L'existence même de revues aux noms évocateurs, comme Biometrika ou Biometrics, le leur rappelle en permanence. Ils peuvent en oublier que ces développements sont récents et que, si le grand public attache une importance majeure aux progrès spectaculaires de la biologie, il a souvent du mal à appréhender correctement l'approche non déterministe. D'ailleurs, les scientifiques eux-mêmes n'ont-ils pas longtemps considéré les modèles déterministes comme universels ? Un ouvrage à vocation pédagogique visant à expliquer le pourquoi et le comment des modèles stochastiques dans la réflexion sur le vivant est donc a priori bienvenu, surtout s'il insiste sur les aspects historiques et épistémologiques plus que sur les aspects techniques, c'est-à-dire s'il cherche plus à convaincre de l'utilité d'une démarche et à montrer sa nature profonde, qu'à donner des recettes. Et c'est bien là l'objectif du livre d'Yves et Maurice Girault ; selon les auteurs eux-mêmes, *il apparaît que* [dans les phénomènes de la vie] *des modèles trop rigides ne sauraient convenir ; il est donc nécessaire de recourir aux modèles aléatoires (...) ; pour effectuer efficacement cette démarche, il convient de connaître les modèles de base (...) mais il est encore plus important d'acquérir un état d'esprit...*

L'ouvrage commence par « un peu d'histoire », selon le titre du chapitre 1. Mais, en fait, le point de vue historique est présent tout au long, les auteurs insistant beaucoup, à bon escient, sur la genèse des divers développements. Pour comprendre ceux-ci, il faut d'abord s'attacher à l'environnement expérimental (chapitre 2), puis à la question vitale de la variabilité (et malgré tout permanence) des phénomènes biologiques (chapitre 3). Expérimentation et variabilité amènent naturellement aux modèles aléatoires et à la statistique. Sont donc abordés successivement la lecture et le décryptage des données (chapitre 4) c'est-à-dire les principes de la statistique descriptive, puis l'introduction des modèles probabilistes et de l'inférence statistique (chapitre 5). Les auteurs sont alors en mesure de préciser le rôle de la statistique en biologie (chapitre 6), comment elle intervient dans la révélation d'un phénomène inconnu (depuis la fièvre puerpérale au milieu du XIX^{ème} siècle jusqu'au SIDA récemment), comment on évalue une population animale ou la biodiversité. Mais il faut parfois rappeler que la manipulation des outils statistiques est

pleine de pièges : le chapitre 7 sur le bon usage du plus banal d'entre eux, la moyenne, est à ce titre fort utile. Puisqu'un des enjeux de la connaissance est la prise de décision, le chapitre 8 introduit à la prise de décision en avenir aléatoire. Enfin, le chapitre 9 souligne et explique l'importance de l'aléatoire dans la compréhension de l'évolution.

Tout cela est présenté à grand renfort d'exemples précis, extrêmement variés, bien choisis, bien documentés (de sorte qu'en citer ici un nombre nécessairement limité serait trompeur...). Les références précises à l'évolution de la pensée scientifique sont constantes (de Claude Bernard à Jacques Monod). Par contre, comme nous l'avons dit, les aspects historiques et épistémologiques sont prioritaires : il ne s'agit pas d'un manuel de statistique à l'usage des biologiste (mais il y en a déjà beaucoup !). Cela conduit évidemment à quelques difficultés que les puristes pourront regretter (par exemple l'utilisation de lois de probabilité qui ne sont pas complètement définies au préalable) ; mais cela ne nous parait pas susceptible d'égarer le lecteur. Au contraire, cela permet une lecture plus facile par le non mathématicien auquel ce livre est plus particulièrement destiné. Quant au statisticien de profession, il y trouvera certainement de nombreux exemples qu'il ne connaissait pas (ou du moins qu'il ne connaissait pas de façon aussi documentée) et une synthèse très intéressante de l'évolution simultanée de sa discipline et des sciences du vivant. A l'instar des auteurs, les signataires, l'un d'origine mathématique et statistique, l'autre de formation bio-médicale, se sont mis à deux pour apprécier ce livre pluridisciplinaire : d'un commun accord, ils le recommandent à un public très varié.

Emmanuel et Henri Caussinus