De la difficulté de faire des prévisions (quand on a peu de données)

Arthur Charpentier

Professeur d'actuariat, Université du Québec, Montréal

Depuis plusieurs mois, on observe un engouement (probablement légitime) pour le big data. *Si beaucoup peut être fait pour utiliser les volumes énormes de données à la disposition des assureurs, il convient de garder en mémoire que, dans de nombreux cas, les données sont rares et que la technologie ne devrait pas pouvoir y changer grand-chose. Le manque de données (fiables) crée une variabilité importante.*

Loi des grands nombres, approximations et statistique asymptotique

Tout comme les assureurs, les statisticiens aiment les gros volumes de données. Pour les assureurs, les gros portefeuilles sont considérés comme moins incertains, et, pour les mêmes raisons, les statisticiens disposent, avec des grosses bases de données, d'estimateurs moins volatils (on parle d'ailleurs du risque associé à un estimateur).

Supposons que l'on suive un portefeuille de 1 047 assurés pendant un an, et que 159 d'entre eux aient déclaré un sinistre auprès de la société d'assurance. La probabilité qu'un assuré ait un sinistre est

$$p = \frac{159}{1\,047} \approx 15{,}2\ \%$$

Si on admet que l'on dispose d'un nombre suffisant d'observations, on peut utiliser une approximation gaussienne du taux de déclaration et obtenir un intervalle de confiance à 95 %, de telle sorte qu'il y a 95 chances sur 100 que le taux de déclaration appartienne à l'intervalle (1).

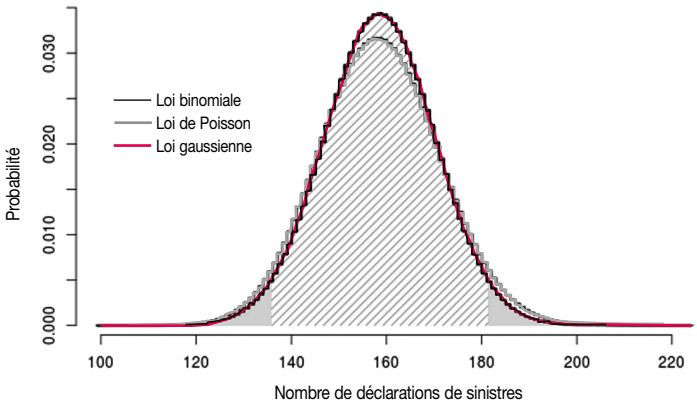
$$\left[p \pm 1{,}96\sqrt{\frac{p(1-p)}{n}}\right] \approx [13\ \%; 17\ \%]$$

On utilise ici l'approximation (2) de la loi binomiale (le nombre de personnes qui déclarent suit une loi binomiale) par une loi gaussienne, comme sur la figure 1 p. 110.

Avec cette approximation, on peut valoriser des contrats de réassurance. On peut ainsi approcher les probabilités d'avoir des dérives de la sinistralité dans notre portefeuille. Dans cet exemple, on notera que la probabilité d'avoir un taux de déclaration excédant 20 % est de l'ordre de 0,0014 %.

Mais tous ces calculs ne sont valides que parce que l'on dispose d'assez d'observations. Malheureusement, dans beaucoup de situations, ce n'est pas le cas.

Figure 1 - Approximation gaussienne du nombre de déclarations de sinistres dans un portefeuille



Source : Arthur Charpentier.

Prédire des « cygnes noirs »

Supposons qu'un assureur décide de vendre à des collectivités locales des contrats d'assurance en cas de marée noire. Dans un modèle actuariel standard, la prime à payer devrait être $p \cdot c \cdot i$, où p est la probabilité qu'une marée noire survienne par kilomètre de côte, c la longueur de la côte de la région assurée et i le montant de l'indemnité versée. Après cinq ans d'expérience, l'assureur souhaite faire le bilan de son produit. Sauf qu'aucune marée noire n'a été observée. Que peut-il dire sur son tarif ?

Cet exemple (à peine) fictif est à rapprocher d'une question posée dans les années 1950 à l'actuaire L. H. Longley-Cook : est-il possible de prédire (ou d'estimer) la probabilité d'observer, une année donnée, une collision en plein vol entre deux avions ? Il n'y avait jamais eu de (grave) collision d'avions commerciaux lorsque la question a été posée. Et, sans aucune expérience passée, les statisticiens ne savaient trop quoi répondre. Pourtant, Longley-Cook avait prédit "*anything from 0 to 4 […] collisions over the next ten years*", comme le rapporte McGrayne [2012]. Deux ans plus tard, 128 personnes perdirent la vie au-dessus du Grand Canyon lors d'une telle collision, et, quatre ans plus tard, ce sont 133 personnes qui périrent au-dessus de New York.

La réponse bayésienne

L'idée de Longley-Cook était d'utiliser des techniques bayésiennes en actuariat. Comme le notait Liu [Liu *et al.*, 1996] : "*Statistical methods with a Bayesian flavor […] have long been used in the insurance industry.*" Dans Charpentier [2007], nous avions vu l'apport des techniques de crédibilité et les liens avec la statistique bayésienne. Et, comme le raconte McGrayne [2012], Arthur Bailey a joué un rôle essentiel, précisément en proposant des techniques dans le cas d'événements (très) rares : "*[Arthur] Bailey spent his first year in New York [in 1918] trying to prove to himself that 'all of the fancy actuarial [Bayesian] procedures of the casualty business were mathematically unsound.' After a year of intense mental struggle, however, he realized to his consternation that actuarial sledgehammering worked. He even preferred it to the elegance of frequentism. He positively liked formulae that described 'actual data': 'I realized that the hard-shelled underwriters were recognizing certain facts of life neglected by the statistical theorists.' He wanted to give more weight to a large volume of data than to the frequentists small sample; doing so felt surprisingly 'logical and reasonable'. He concluded that only a 'suicidal' actuary would use Fishers method of maximum likelihood, which assigned a zero probability to non-events. Since many businesses file no insurance claims at all, Fishers method would produce premiums too low to cover future losses.*" (3)

Pour rappel, le théorème de Bayes permet d'écrire

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[A]}{\Pr[B]} \cdot \Pr[B|A] \propto \Pr[A] \cdot \Pr[B|A]$$

si l'on suppose disposer d'un échantillon d'observations $\{x_1,..., x_n\}$ de loi F_θ, où, classiquement, θ est vu comme un paramètre inconnu que l'on cherche à estimer. Ici, on va supposer qu'il s'agit d'une variable aléatoire dont on va pouvoir obtenir une distribution (dite « a posteriori ») compte tenu des observations dont on dispose. On écrit alors

$$\Pr[\Theta=\theta|x_1,\cdots,x_n]\propto\Pr[\Theta=\theta]\cdot\Pr[x_1,\cdots,x_n|\Theta=\theta]$$

où le premier terme est une loi a priori que l'on se donne sur le paramètre Θ, et le second est la vraisemblance de l'échantillon.

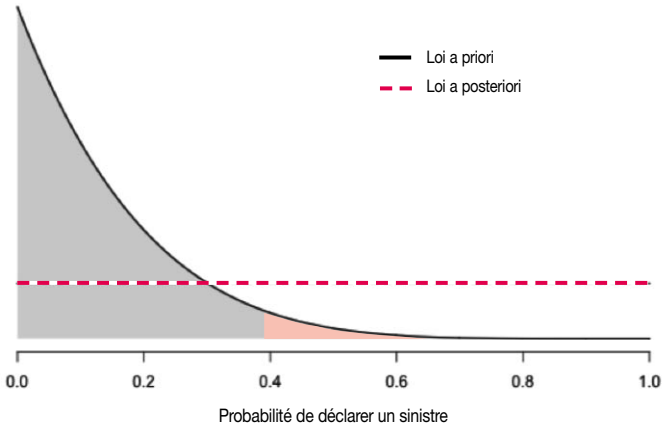
Afin d'illustrer un peu cela, reprenons notre exemple initial, où l'on dispose de cinq assurés, dont aucun n'a déclaré de sinistre. Que peut-on dire sur la probabilité qu'un assuré ait un sinistre ? Avec les approches classiques, la probabilité estimée serait nulle. Ce qui est gênant pour calculer ensuite une prime convenable. La réponse bayésienne consiste à supposer que la probabilité de déclarer un sinistre Θ est une variable aléatoire, dont on peut calculer la loi a posteriori, compte tenu du fait que l'on a procédé – pour l'instant – à cinq observations nulles. En utilisant la relation précédente, la densité a posteriori est ici

$$\pi(\theta|x_1=0,\cdots,x_5=0)\propto\pi(\theta)\cdot\Pr[x_1=0,\cdots,x_5=0|\Theta=\theta]$$

où le second terme est la probabilité d'avoir cinq fois zéro avec une loi de Bernoulli, ce qui vaut $(1-\theta)^5$, et où le premier terme est une loi que l'on se donne, a priori pour Θ.

Une loi qui pourrait sembler naturelle serait la loi uniforme, pouvant être perçue comme non informative. Dans ce cas, la densité de Θ, a posteriori, est proportionnelle à $(1-\theta)^5$. Comme le montre la figure ci-contre (voir figure 2), on a ici « beaucoup » de chances que Θ prenne une valeur proche de zéro (ce qui a du sens, car on n'a jamais observé de déclaration, jusqu'à présent). A priori, il y avait 95 chances sur 100 que la probabilité soit inférieure à 95 % (c'est l'idée de la loi uniforme comme loi a priori). Après cinq expériences sans sinistre, a posteriori, il y a 95 chances sur 100 que la probabilité de déclarer soit inférieure à 40 %. Si nous avions observé dix fois la valeur nulle, il y aurait, a posteriori, 95 chances sur 100 que la probabilité de ne pas déclarer soit inférieure à 25 %.

Figure 2 - Lois a priori et a posteriori pour la probabilité de déclarer un sinistre si aucun sinistre n'a encore été observé



Source : Arthur Charpentier.

Les outils statistiques pour le *small data*

De même que le *big data* pose des problèmes aux techniques statistiques classiques et permet de découvrir les techniques d'apprentissage, le *small data* pose des problèmes auxquels la statistique classique n'apporte aucune réponse. Les actuaires redécouvrent les techniques bayésiennes, mais aussi d'autres outils qui permettent malgré tout d'apporter des réponses quantitatives pertinentes. Pour quantifier l'incertitude lors du calcul des provisions pour sinistres à payer, les actuaires utilisent des techniques de rééchantillonnage (*bootstrap*), depuis une vingtaine d'années [Charpentier *et al.*, 2010], pour tenir compte du peu de données dans les triangles de liquidation.

Si le *big data* ouvre des perspectives fascinantes à explorer pour les actuaires (en particulier en souscription),

il convient de garder en mémoire que, dans beaucoup de cas, les actuaires doivent quantifier des risques avec très peu de données. L'avenir n'est aucunement au *big data* ! L'avenir est au contraire dans la connaissance de techniques permettant à la fois de traiter des gros volumes de données, mais aussi de faire face à une insuffisance de données.

Notes

1. On retrouve ici les intervalles de confiance mentionnés désormais dans les sondages.

2. Dans le cas des événements (très) rares, l'approximation par une loi de Poisson sera nettement meilleure, mais, ici, la probabilité est de l'ordre de 15 %, ce qui est trop élevé.

3. « [Arthur] Bailey passa sa première année à New York [c'était en 1918] à tenter de se prouver que "toutes les méthodes actuarielles fantaisistes [bayésiennes] de l'assurance dommages étaient infondées". Toutefois, après une année de lutte dans son esprit, il se rendit compte avec consternation que le "bourrinage" actuariel fonctionnait. Il le préférait même à l'élégance de l'approche fréquentiste. Il aimait indubitablement les formules qui décrivaient les "vraies données" : "Je me rendis compte que les briscards de la souscription savaient reconnaître des caractéristiques distinctives de la réalité que les statisticiens théoriciens ignoraient." Il voulait donner plus de poids à un grand ensemble de données qu'au petit échantillon dont disposaient les fréquentistes. Il lui parut étonnamment "logique et raisonnable" de procéder ainsi. Il en conclut que seul un actuaire "suicidaire" utiliserait la méthode de Fisher (maximum de vraisemblance), qui confère une probabilité nulle aux non-événements. Comme beaucoup d'entreprises ne déclarent jamais aucun sinistre, la méthode de Fisher aboutirait à des primes trop basses pour couvrir les pertes futures. » (traduction de l'auteur).

Bibliographie

CHARPENTIER A., « La crédibilité : un pasteur et un philosophe pour soutenir les actuaires », *Risques*, n° 71, 2007, pp. 122-126.

CHARPENTIER A. ; DEVINEAU L. ; NESSI J.-M., « Mesurer le risque lors du calcul des provisions pour sinistres à payer », *Risques*, n° 83, 2010, pp. 93-100.

LIU Y.-H. ; MAKOV U. E. ; SMITH A. F. M., "Bayesian Methods in Actuarial Science", *The Statistician*, n° 45, 1996, pp. 503-515.

MC GRAYNE S. B., *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines and Emerged Triumphant from Two Centuries of Controversy*, Yale University Press, 2012.