Miracles et coïncidences : évènements rares et théorie des valeurs extrêmes.

Bouchra Nasri et Bruno Rémillard

May 4, 2017

Dans son article Le miracle : une supercherie statistique?, paru dans *Convergence* en 2016, Pierre Lavallée réfère à une définition de miracle provenant du Larousse comme « *un fait, un résultat étonnant, extraordinaire, ou encore un hasard merveilleux, une chance exceptionnelle* ». Du point de vue statistique, cela revient à un évènement rare (ou extrême) auquel on attribue une probabilité d'apparition très faible.

Les évènements rares (tremblements de terre, inondations, crises financières, chocs pétroliers ou autres) captent l'attention par leurs caractères inattendus et récurrents. Souvent négligés par certains non-statisticiens et traités comme des observations aberrantes, les évènements rares et extrêmes depuis les deux dernières décennies ont fini par être sous le feu des projecteurs notamment pour l'importance de leurs impacts sociaux et économiques. Mais comment peut-on estimer le caractère rare d'une variable aléatoire?

Pour plusieurs domaines scientifiques, les statistiques se réduisaient souvent à l'estimation de la moyenne et son écart type. On poussait quelques fois plus loin en allant jusqu'à étudier la distribution de probabilités des observations. La statistique "classique" ou "fréquentiste", basée sur la loi des grands nombres résumant les travaux de Pierre-Simon Laplace et Paul Lévy aux 19e et 20e siècles [6], se focalise habituellement sur l'analyse du comportement moyen des variables aléatoires. Essentiellement, la loi des grands nombres indique que pour un tirage aléatoire de grande taille,

plus la taille de l'échantillon augmente, plus les caractéristiques du tirage (i.e. la moyenne, l'écart type) se rapprochent des caractéristiques de la population. En outre, le fameux théorème central limite (TCL) énonce que, la moyenne d'un grand nombre de variables indépendantes et identiquement distribuées (iid) suit approximativement une distribution normale même si celles-ci suivent individuellement une autre loi de probabilité.

Bien que ce résultat impressionnant demeure un des piliers de la statistique, il s'avère souvent impuissant pour décrire des phénomènes au-delà de leurs moyennes. En réalité, dans plusieurs études de fiabilité, ce qui importe : c'est de déterminer la probabilité de dysfonctionnement d'un système en s'appuyant sur la détermination de la probabilité de son élément qui se rompt brusquement et complètement lorsque la limite élastique est atteinte. Par exemple, en génie civil, pour évaluer la dimension d'un barrage, il est indispensable de déterminer la pression maximale qu'il peut supporter dans des conditions extrêmes. La pression moyenne reliée à des conditions normales ne sera pas aussi déterminante dans ce choix. En finance, on s'intéresse plutôt aux pertes extrêmes et à l'évaluation de certaines mesures de risque liées à ces pertes, comme les quantiles d'ordre 0,1%, donnant un niveau de perte ne survenant en moyenne qu'une fois en 1000 ans.

Pour cela, il faut bien caractériser les valeurs rares de notre échantillon. Une des voies empruntées pour remédier au problème des valeurs rares est la théorie des valeurs extrêmes (TVE). Celle-ci est basée sur l'approximation asymptotique des maxima (ou minima) d'un grand nombre de variables supposées iid. Un des résultats fondamentaux de la TVE est le théorème établi par Ronald Fisher et Leonard Tippett en 1928 [4], qui énonce que, sous certaines conditions, la distribution des maxima (minima) d'un grand nombre de variables iid converge vers une des trois distributions suivantes : distribution de Gumbel, distribution de Fréchet ou distribution de Weibull. Étant donné qu'il est difficile de travailler avec trois distributions à la fois, von Mises en 1945 et Jenkinson en 1955 [5, 8] ont proposé une famille paramétrique de distribution, appelée la distribution généralisée des valeurs extrêmes (Generalized Extreme Value; GEV), qui permet d'unifier les trois types de distributions extrêmes et par ailleurs facilite l'estimation des valeurs extrêmes.

Quoique très utilisé en pratique, ce théorème a été fortement critiqué dans la littérature puisqu'il ne tient compte que d'une seule observation, la plus grande (ou la plus petite). On a donc le sentiment de perdre de l'information et notamment toute celle contenue dans les autres grandes valeurs de l'échantillon. Par conséquent, un théorème alternatif a été proposé par De Hann et Balkema en 1974 et Pickands en 1975 [1, 7]. Il repose sur l'étude de la distribution asymptotique des excès au-delà d'un seuil fixé d'un grand nombre de variables aléatoires iid. En effet, au lieu de considérer

le maximum (ou le minimum), on considère toutes les observations qui dépassent un seuil déterministe. Les excès représentent donc la différence entre ces observations et le seuil. Le théorème énonce que, sous certaines conditions, la distribution de ces excès est approximativement la distribution généralisée de Pareto (Generalized Pareto Distribution; GPD).

La TVE a connu un grand développement dans les deux dernières décennies, comme en témoignent les nombreux ouvrages récents : Embrechts et al. [3] et Coles [2], parmi d'autres. Certains de nos membres tels Christian Genest et Louis-Paul Rivest ont aussi déjà publié d'excellents travaux sur ce sujet.

Les modèles des valeurs extrêmes sont appliqués à une grande variété de problèmes provenant de plusieurs domaines, tels que l'estimation des fortes crues ou l'estimation des sévères étiages en environnement, l'évaluation du risque de grands sinistres ou l'évaluation du risque de pertes sévères en finance et assurance, l'évaluation du degré de résistance des matériaux face aux pressions extrêmes en fiabilité, ou encore la mortalité aux grands âges et la durée extrême de la vie humaine en démographie.

Pour conclure, nous vous rappelons la blague suivante : un statisticien, c'est quelqu'un qui met la tête dans le réfrigérateur et les pieds dans un four chaud et qui dit : « en moyenne, je me sens bien ». Donc, sa moyenne n'indique nullement sa vraie condition!

References

[1] Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804.

[2] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics.

[3] Embrechts, P., Kluppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events:for Insurance and Finance*. Springer-Verlag Berlin Heidelberg.

[4] Fisher, R. and Tippett, L. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190.

[5] Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) of meteorological elements. *Q. J. R. Meteorol. Soc*, 81:158–171.

[6] Laplace, P., Truscott, F., and Emory, F. (1902). *A Philosophical Essay on Probabilities*. A Philosophical Essay on Probabilities. Wiley.

[7] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131.

[8] von Mises, R. (1954). La distribution de la plus grande de n valeurs. *Reprinted in Selected Papers Volumen II. American Mathematical Society, Providence, R. I.*, pages 271–294.