

L

e 15 septembre 2008, la banque américaine Lehman Brothers se déclare en faillite. Cet événement, consécutif de la crise des *subprimes* l'année précédente, ébranle le système financier mondial et affole les bourses de la planète. Les gouvernements et les banques centrales enchaînent les plans de relance, mais rien n'y fait, le monde plonge dans la crise financière la plus grave depuis celle de 1929. Dans quelle mesure aurait-on pu prévoir ce cataclysme ?

De tels événements rares et catastrophiques (aux krachs boursiers, ajoutons les tremblements de terre, les inondations, les accidents nucléaires...) fascinent par leur caractère imprévisible et l'importance des enjeux sociaux et scientifiques qu'ils représentent. Ils sont aussi des défis à la modélisation scientifique. Exceptionnels presque par définition, ou parfois mis de côté comme « aberrations statistiques », les événements rares sont néanmoins accessibles à la théorie statistique... pour peu qu'elle sollicite des outils mieux adaptés que ceux du quotidien.

Dans beaucoup d'applications, les statistiques se résument au calcul de moyennes et d'écarts-types, et la distribution des observations se répartit sagement selon la fameuse « courbe en cloche » ou distribution gaussienne.

LES STATS POUR LES NULS

Le statisticien représente souvent un échantillon d'observations (températures, précipitations, indices boursiers...) comme une suite de tirages indépendants d'une « variable aléatoire » notée X et représentant la grandeur en question. On s'intéresse alors à la moyenne, ou espérance mathématique, de cette variable aléatoire, à son écart-type (la dispersion des valeurs autour de la moyenne) et, plus généralement, à sa distribution de probabilité.

La loi des grands nombres, établie pour la première fois dans sa version la plus simple au XVIII^e siècle par Jacques Bernoulli, assure que, lorsque la taille de l'échantillon est assez grande, la moyenne empirique de l'échantillon tend vers la moyenne théorique de la variable aléatoire qu'il représente. De même, l'écart-type empirique converge vers l'écart-type de

la variable aléatoire. En garantissant qu'un échantillon fournit une bonne estimation d'une variable aléatoire réelle, la loi des grands nombres s'impose comme un des piliers de la statistique.

La loi des grands nombres est une loi asymptotique : la moyenne d'un échantillon ne converge vers l'espérance mathématique que si la taille N de l'échantillon est grande. La différence entre les deux valeurs – l'erreur statistique – est régie par le « théorème de la limite centrale » (voir l'encadré page ci-contre). Ce théorème, second pilier des statistiques, est connu sous des formes diverses depuis le XVIII^e siècle, mais sa forme définitive, due au mathématicien français Paul Lévy, date du début du XX^e siècle. Il affirme que, si la variable aléatoire X a un écart-type fini, alors pour un échantillon de N observations indépendantes de X , la différence entre la moyenne de l'échantillon et la moyenne de X est bien représentée par une distribution gaussienne dont l'écart-type est proportionnel à $1/\sqrt{N}$.

La distribution gaussienne apparaît donc comme une description statistique de toute quantité que l'on peut représenter comme la somme d'un grand nombre de petites contributions indépendantes.

Ces deux principes ont une grande portée. La loi des grands nombres est le fondement des sondages, car elle montre qu'il suffit de sonder

La moyenne et l'écart-type n'ont plus forcément de sens pour les distributions inégales

un échantillon assez grand pour approcher les caractéristiques statistiques de la population réelle. Le théorème de la limite centrale permet, pour sa part, de contrôler l'erreur ainsi commise : l'écart entre la moyenne empirique et la moyenne théorique suit une loi gaussienne.

LOI DES GRANDS NOMBRES ET THÉORÈME DE LIMITE CENTRALE

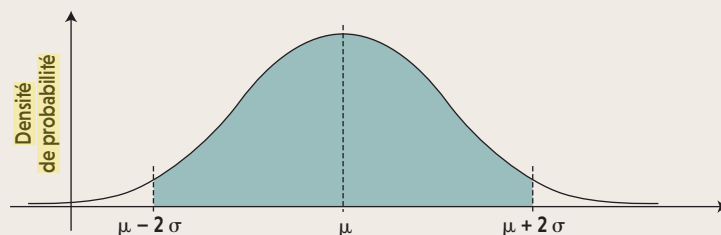
La loi des grands nombres énonce que lorsqu'on échantillonne une variable aléatoire, sous certaines conditions, lorsque la taille de l'échantillon est assez grande, la moyenne empirique de cet échantillon tend vers la moyenne théorique de la variable aléatoire échantillonnée.

C'est sur cette loi que reposent notamment les sondages. Considérons un échantillon d'observations $x_1, x_2, \dots, x_n$ d'une variable aléatoire X , d'espérance μ et d'écart-type σ finis. La loi forte des grands nombres affirme que quand n tend vers l'infini, la moyenne empirique $M_n = (x_1 + x_2 + \dots + x_n)/n$ converge presque sûrement vers μ , c'est-à-dire que la probabilité que la moyenne empirique converge vers la moyenne théorique est égale à 1.

$$\mathbb{P} \left(\lim_{n \rightarrow \infty} M_n = \mu \right) = 1$$

En d'autres termes, la moyenne d'un grand nombre d'observations aléatoires indépendantes n'est plus aléatoire ! La loi des grands nombres assure ainsi que la moyenne empirique est un estimateur convergent (ou « consistant ») de l'espérance mathématique. Cela reste vrai tant que la contribution de chaque terme x_i à la moyenne reste négligeable : aucune observation n'est assez grande pour dominer la somme. Cette situation est qualifiée de hasard « sage ».

Le théorème de la limite centrale est l'autre pilier de l'échantillonnage. Considérons des variables aléatoires X_i indépendantes et



de même distribution (d'espérance μ et d'écart-type σ). Alors la somme centrée $(X_1 + X_2 + \dots + X_n - n\mu)/\sqrt{n}$ tend vers une variable aléatoire normale (ou gaussienne) d'espérance nulle et d'écart-type σ quand n tend vers l'infini. La densité de probabilité d'une variable gaussienne de moyenne m et écart-type σ est définie par la fonction

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

et sa représentation graphique est la célèbre courbe en cloche.

Le théorème de la limite centrale reste valable même si les distributions individuelles sont différentes, pour autant que l'écart-type de chacun des termes soit négligeable vis-à-vis de l'écart-type de la somme.

Ce théorème implique que la différence entre la moyenne empirique M_n de l'échantillon et l'espérance μ de la variable aléatoire X est représentée par une distribution gaussienne dont l'écart-type est proportionnel à $1/\sqrt{n}$. Cela permet de contrôler l'erreur faite en approximant la moyenne théorique μ par la moyenne empirique M_n . La probabilité que M_n se trouve dans l'intervalle $[\mu - t\sigma, \mu + t\sigma]$ est représentée, pour n assez grand, par l'aire sous la courbe en cloche comprise entre les abscisses $\mu - t\sigma$ et $\mu + t\sigma$. Il y a par exemple 95 % de chances que l'écart entre la moyenne empirique M_n et l'espérance μ soit inférieur à 2σ .

Le théorème de la limite centrale permet de quantifier la probabilité que la moyenne empirique approche la moyenne réelle avec une précision donnée.

Par exemple, si un sondage électoral porte sur 10 000 personnes et que 46 % d'entre elles se déclarent favorables au candidat A, le théorème de la limite centrale indique qu'il y a 95 % de chance que lors du vote, il recueille entre 45 et 47 % des voix. Concrètement, il n'est pas nécessaire de sonder un très grand échantillon de la population pour avoir une estimation fiable.

Le théorème de la limite centrale est invoqué pour justifier la représentation par une distribution gaussienne dans des domaines allant des sciences sociales à la physique, en passant par l'économie, la biologie ou la finance. La courbe en cloche est mise à toutes les sauces. On l'utilise même pour « normaliser » la distribution des notes dans les concours...

Cependant, l'omniprésence de la distribution gaussienne en statistique (d'où son qualificatif de « loi normale ») est peut-être due à la trop grande confiance des statisticiens plutôt

qu'à la réalité : tout phénomène ne peut pas être ramené à une somme de petites contributions indépendantes. De fait, en sciences économiques et sociales, la loi normale paraît plutôt être l'exception que la règle.

Dès le début du xx^e siècle, l'économiste Vilfredo Pareto s'est par exemple intéressé à la richesse nationale italienne, et a constaté que les 20 % les plus riches de la population détenaient 80 % de la richesse totale du pays. Cette situation d'inégalité n'est manifestement pas bien représentée par une distribution gaussienne des richesses, où la richesse individuelle serait concentrée autour de la richesse moyenne et où les individus très riches ou très pauvres seraient très rares.

Pareto a proposé une distribution plus réaliste qui porte aujourd'hui son nom : la part d'individus ayant une richesse supérieure à un niveau u est proportionnelle à $1/u^\alpha$. Plus l'exposant α de cette loi de puissance est petit, plus >