

La statistique : pour quoi faire ?

Les mathématiques, vous aimez bien, mais ça vous semble trop déconnecté de la « vraie vie ». Et si vous regardiez du côté de la statistique ? Les applications directes de cette discipline sont innombrables, et d'autres encore sont sans aucun doute à imaginer ou à découvrir !

La statistique est une discipline des mathématiques appliquées et s'est imposée comme un élément indispensable de la science des données (ou *data science* en anglais). Nous vivons entourés de données, mais que sont-elles ? Ce sont tout simplement des observations. Elles peuvent être mesurées ; on parle alors de *données quantitatives* : par exemple, la taille, le poids, la température de l'air, la durée d'utilisation d'un téléphone, les revenus d'un salarié... Lorsqu'elles ne peuvent pas être mesurées, on parle de *données qualitatives* : rentrent dans cette catégorie la couleur préférée, les moyens de transport habituels, si on est atteint de diabète ou non...

Extraire de l'information utile

La statistique permet ainsi, à l'aide de méthodes et d'outils mathématiques, d'extraire de l'information utile à par-

tir de données. D'une part, on souhaite décrire, résumer, les informations qui nous sont fournies. C'est ce qu'on appelle la statistique *descriptive* et nous la côtoyons tous les jours. Par exemple, à partir des notes d'un élève, on peut construire leur moyenne et obtenir un bulletin de notes. Dans les jeux vidéo, on croise souvent des « taux de victoires » ; il s'agit là de la proportion de parties gagnées par rapport au nombre de parties jouées. D'autre part, on peut vouloir extrapoler les résultats observés sur un échantillon d'individus à un groupe plus large. En reprenant l'exemple scolaire, on pourrait utiliser les notes des élèves d'une classe de troisième pour tirer des conclusions sur le niveau scolaire de l'ensemble des élèves de troisième en France. C'est la statistique *inférentielle*.

À l'ère des données, voire des données massives (*big data* en anglais), les applications de la statistique sont

nombreuses. Un grand nombre de problèmes consistent à adapter nos actions en fonction des observations que nous faisons du réel. L'approche scientifique elle-même requiert en général de collecter des données expérimentales et d'évaluer si les observations sont dues au hasard ou à un phénomène en particulier. Avec cette idée en tête, explorons à présent dans quels domaines particuliers les statisticiens travaillent.

Sondages et enquêtes

Les *sondages* constituent une application de la statistique à laquelle nous sommes régulièrement confrontés. Lors des campagnes électorales, notamment, des instituts de sondage contactent un échantillon représentatif de la population française et les interrogent sur leurs intentions de vote. Cela permet de prédire les résultats des élections. Les *enquêtes*, menées par exemple pour connaître la fréquentation d'une ligne de train ou la satisfaction des passagers d'une gare, sont des méthodes similaires.

De plus en plus dans le sport

De la Coupe du monde de football aux Jeux olympiques, en passant par les séances d'entraînement au quotidien, le sport bénéficie également des compétences des statisticiens. En effet, le monde du sport a, depuis de nombreuses années, massivement intégré l'utilisation de données dans sa pratique. D'abord par la mise en place de divers capteurs (GPS, accéléromètres, montres connectées...) et autres dispositifs, notamment vidéo, lors des entraînements puis des compétitions. Cela conduit à la collecte de nombreuses données sur les terrains.

Les marges d'erreur

Les résultats d'un sondage ne sont pas exacts. Il s'agit d'estimations de proportions faites à un certain niveau de confiance (usuellement à 95 %). L'incertitude liée à cette estimation est donnée par la formule de l'intervalle de confiance pour l'estimation d'une proportion, comme suit :

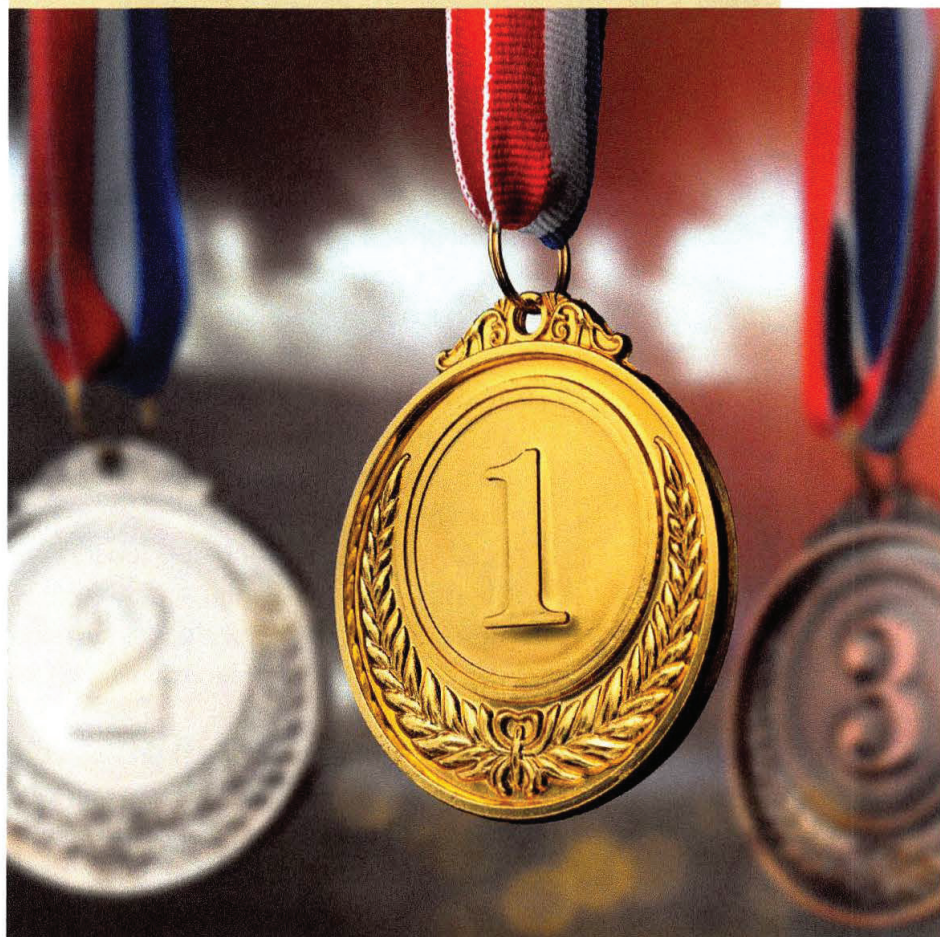
$$\left[p - 1,96 \times \sqrt{\frac{p(1-p)}{n}}; p + 1,96 \sqrt{\frac{p(1-p)}{n}} \right]$$

où p est la proportion estimée et n est la taille de l'échantillon considéré.

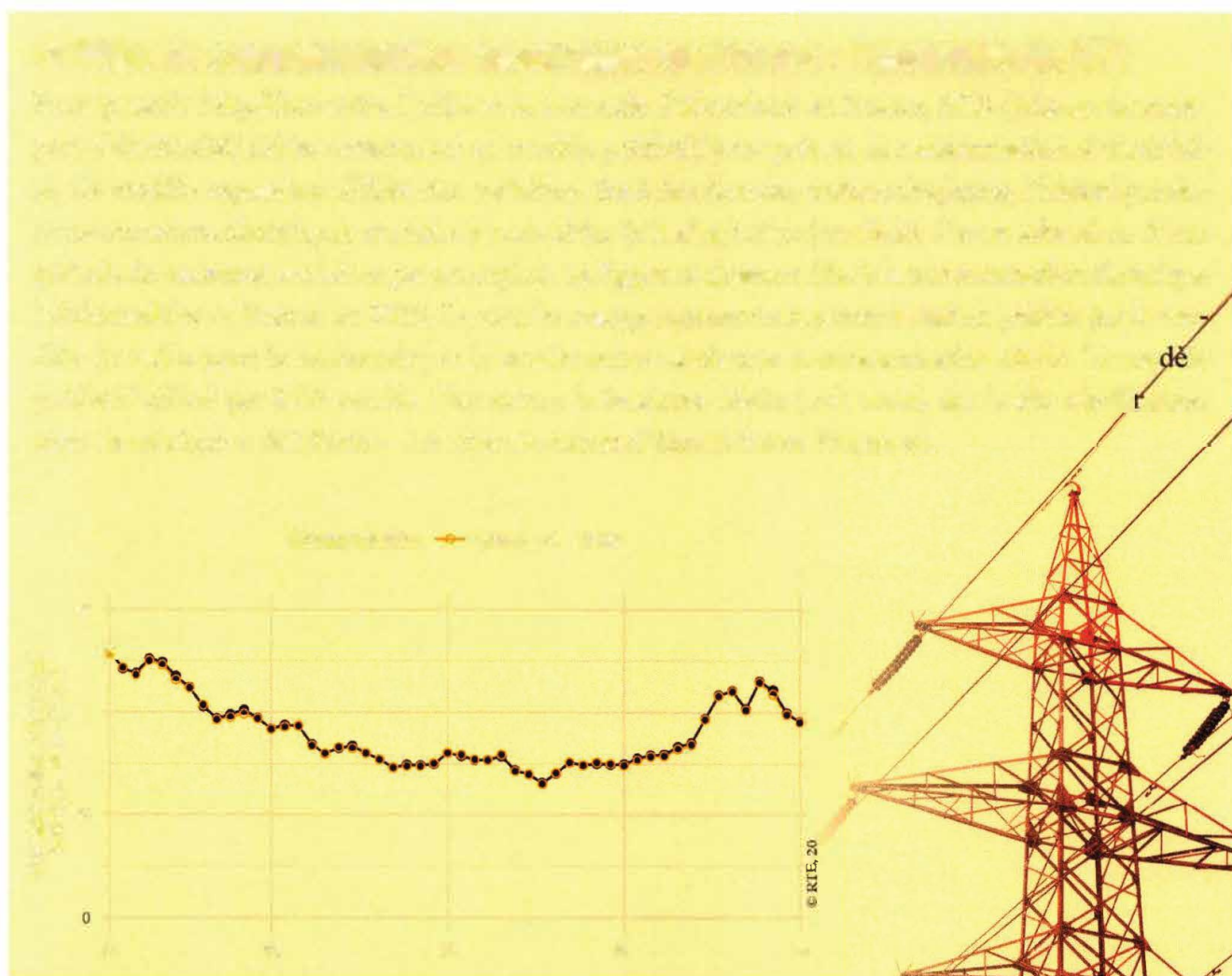
Dans le cas d'un sondage réalisé sur $n = 1\,000$ personnes où la proportion estimée est $p = 0,5$, la marge d'erreur s'écrit :

$$1,96 \times \sqrt{\frac{p(1-p)}{n}} = 1,96 \sqrt{\frac{0,5 \times (1-0,5)}{1000}},$$

soit environ 3 %. C'est pourquoi les résultats des sondages réalisés sur mille personnes donnent des résultats à ± 3 %.



© Igor Link - stock.adobe.com



La statistique aide à la prévention des blessures, où il est nécessaire d'identifier des variables clés en fonction du sport pour comprendre les facteurs de risques et prédire au mieux les situations à risque. Par exemple, le nombre et l'intensité des collisions au rugby sont aujourd'hui suivis au cours des matchs à l'aide d'accéléromètres présents sous le maillot des joueurs, et fournissent au personnel des équipes des indications préventives pour décider d'éventuels remplacements à effectuer.

La statistique aide aussi à l'amélioration de la performance en tant que telle. La professionnalisation du sport a poussé les athlètes à chercher à optimiser les *micro-gains*, des détails pouvant

faire la différence entre les meilleurs et qui requièrent souvent une approche scientifique pour être optimisés. Le cyclisme britannique a connu des succès fulgurants dans les années 2010 en proposant une approche ultra-quantifiée de la performance ; sans surprise, la systématisation des analyses statistiques a également eu une place de choix dans ces évolutions.

Biostatistique, épidémiologie, environnement

Le corps humain est régulièrement l'objet d'étude de la *biostatistique*, une discipline qui applique la statistique aux sciences du vivant. De la conception des études ou des essais cliniques (planification expérimentale) à l'analyse des données, les rôles du biostatisticien sont multiples.

Les dernières années, marquées par la pandémie de la Covid-19, ont également mis en lumière le rôle de la statistique pour la santé publique, notamment en épidémiologie. Ainsi, à partir du suivi du nombre de personnes malades, hospitalisées, décédées, des modèles ont été construits pour tenter de prédire l'évolution de ces indicateurs. Ces rapports permettent en particulier d'éclairer et de motiver des décisions politiques appropriées.

La statistique n'est pas en marge des enjeux écologiques, comme en témoignent les rapports du Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) des Nations unies. La complexité du système climatique dans son ensemble est telle qu'il est nécessaire d'adopter une approche statistique pour créer des modèles de climat pertinents et efficaces. Ainsi, plutôt que suivre l'évolution de

tous les nuages du monde à chaque instant, les scientifiques établissent des simulations numériques issues des principaux phénomènes physiques connus (évaporation, précipitations, mouvements de l'air...) et tirent parti d'outils statistiques dédiés pour prédire l'évolution climatique à grande échelle et sur le long terme. Les modifications environnementales ont également un impact sur les végétaux. Les données de température, d'humidité, d'ensoleillement permettent d'étudier leur influence sur la croissance des végétaux ou d'évaluer leur interaction avec certains engrais, par exemple.

La statistique est aussi utilisée en zoologie, pour l'étude des animaux, ainsi qu'en éthologie, qui étudie le comportement des espèces animales dans leur milieu naturel. Le recueil de données spatiales et de comptage permet par exemple d'étudier l'impact des changements environnementaux sur le déplacement de certaines espèces.

Statistiques industrielles, contrôle qualité

La statistique est également un atout pour le secteur industriel. Elle y intervient à plusieurs niveaux. En ce qui concerne la production, elle permet de vérifier si le produit fabriqué est conforme aux normes de qualité. Par exemple, à partir de quelques pièces issues d'un même lot, on peut se demander si toutes les pièces du lot ont bien été fabriquées aux dimensions requises. Dans l'industrie agro-alimentaire, on peut se demander si un biscuit produit ne contient pas trop de sucre pour le consommateur. En cosmétologie, on peut s'intéresser au ressenti des consommateurs lors de la conception d'une crème : la texture est-elle douce ?

collante ? l'odeur est-elle florale ? trop intense ? trop légère ? À chaque question son étude statistique associée, pour tenter d'anticiper au mieux les attentes et préférences des consommateurs.

Une autre application industrielle récurrente concerne la gestion du *risque*, qui est communément défini comme la possibilité d'un évènement indésirable. En gestion des risques, un danger ou une défaillance sont associés à une probabilité d'occurrence, au dommage (quantifiable) en cas d'occurrence, sa gravité et son acceptabilité. Ainsi, on considère notamment le risque sanitaire (la Covid-19, la grippe A...), le risque environnemental (les canicules, les tempêtes, les sécheresses...), le risque géopolitique (les embargos, les guerres ou le terrorisme), le risque industriel (les accidents d'avion, les explosions...) et le risque financier (les taux d'intérêt, les risques de crédit, les crises financières).

Dans le cas particulier de la banque, de la finance et de l'assurance, la statistique côtoie l'actuariat. L'actuaire évalue, quantifie, modélise les risques financiers (voir *Les Mathématiques des assurances*, Bibliothèque Tangente 57, 2016). Concrètement, l'actuaire peut par exemple concevoir et tarifier des produits d'assurances, analyser les risques de marché, réaliser des projections financières selon différents scénarios...

Parfois, les outils statistiques peuvent nous aider à gérer des contraintes du monde physique. Par exemple, l'électricité n'est pas stockable facilement ; il est donc important de pouvoir prévoir la consommation électrique afin de piloter la production électrique, en

particulier de certaines centrales qui sont peu réactives (voir encadrés). Mais comment savoir si l'on peut faire confiance à ces prédictions ? Quels risques prend-on alors ? De nombreuses méthodes statistiques existent pour effectuer cette tâche essentielle que l'on nomme *quantifier l'incertitude*. C'est aussi pour cette raison que les statistiques sont intimement liées aux calculs de probabilités. Même en présence d'aléatoire, on peut formuler des prédictions. Ne pas être certain d'un résultat à 100 % ne veut pas dire que l'on ne sait rien... bien au contraire. Dans notre monde complexe, la certitude est rare, et il est précieux de pouvoir déterminer quels évènements sont plus ou moins probables.

La psychologie et l'archéologie aussi !

La statistique trouve également des applications dans les domaines des sciences humaines et sociales. Les psychologues sont notamment encouragés à considérer des approches fondées par la preuve. Le confinement de la population en 2020 a-t-il conduit à plus de cas de dépression et d'anxiété ? Former les futurs parents à l'accouchement réduit-il le stress qui y est lié ? Quel est l'effet d'un deuil sur le sommeil ? Telle nouvelle approche thérapeutique est-elle efficace dans l'accompagnement des enfants atteints d'un trouble de l'attention ? Autant de questions qui requièrent également une validation par les outils statistiques.

Les vestiges archéologiques sont également une grande source de données : le comptage des pierres, des os, des tessons de poterie, des graines ou la mesure de leur longueur, leur largeur,

leur forme, leur couleur... Ainsi, parmi les travaux ayant recours aux statistiques, on peut citer l'analyse spatiale des occupations anciennes sur un territoire, la reconstitution d'un paysage par la quantification et la qualification des pollens et des graines, ou encore le traçage de l'origine géographique des matières premières.

À travers ces quelques exemples se révèle la diversité des applications de la statistique. Cette discipline mathématique est utile, même nécessaire, dans beaucoup de domaines très variés. Si vous ne savez pas quelle orientation choisir, plongez-vous dans le monde riche et passionnant de la statistique !

M. C.

Références

- *Les statistiques*. Bibliothèque Tangente 34, 2008.
- *Mathématiques et biologie*. Bibliothèque Tangente 42, 2011.

