L'Institut qui dompte les chiffres

Comment utiliser correctement la statistique? Un service de consultation est depuis peu proposé par l'Institut de recherche en statistique. Réservé pour l'instant aux chercheurs en sciences économiques et sociales, cet appui devrait être rapidement élargi à d'autres facultés avant d'être étendu à la cité.
Dans l'intervalle, un collaborateur scientifique, dont le poste est financé par le fonds d'innovation pédagogique, est chargé de répondre aux questions sur lesquelles butent les chercheurs.
L'Institut propose par ailleurs des cours spécifiques dans le cadre d'écoles doctorales en fonction de demandes reçues. Proposé depuis 2006, le Master en statistique accueille, quant à lui, chaque année entre 20 et 30 étudiants. La formation met l'accent sur la pratique de l'analyse de données, la résolution de problèmes méthodologiques, l'apprentissage de logiciels de statistique ou la statistique mathématique.
La filière offre des débouchés dans l'industrie (pharma, alimentation, finance, marketing, santé, consulting) et dans le secteur public, sans oublier les instituts de recherche et la carrière académique.
www.stat-center.unige.ch/index.html

La «science des données», nouveau défi statistique

L'énorme quantité de données chiffrées qui circule aujourd'hui dans le monde au travers des technologies de l'information ouvre des opportunités de recherche inédites aux statisticiens. A condition d'y donner du sens

Les technologies de l'information ont fait entrer la science dans l'ère des grands nombres. Chaque jour, des milliards de textes, de photos et de vidéos sont échangés à travers le monde via le courrier électronique et les réseaux sociaux. Chaque recherche ou achat effectué sur le Web, à l'aide d'un téléphone mobile ou d'une carte client dans un magasin laisse des traces numériques. Toutes ces données sont ensuite stockées sur des serveurs, attisant la convoitise des gouvernements et des entreprises. La publicité et les services personnalisés sont déjà une réalité, laissant entrevoir une utilisation possible de cette précieuse information.

La capacité des serveurs à mémoriser et à traiter d'énormes quantités de données représente également une opportunité pour les chercheurs, en particulier en sciences économiques et sociales, ainsi que dans des domaines comme la génétique, les neurosciences, la physique des particules ou encore les sciences humaines avec l'essor des humanités numériques.

Image: iStockphoto

POINT DE VUE INÉDIT

Ce nouveau champ d'investigation, appelé «Big Data» ou, de manière plus académique, «Data Science», la «science des données», constitue un nouveau défi pour les statisticiens. S'il était possible d'ouvrir une fenêtre sur cet univers numérique, un point de vue sur la réalité totalement inédit et forcément chaotique s'offrirait en effet à nos yeux. Un point de vue qui confirmerait l'aphorisme du mathématicien suédois Andrejs Dunkel: «Il est facile de faire mentir les statistiques, mais il est difficile de dire la vérité sans elles.»

Potentiellement, les données numériques représentent un avantage considérable. Parce qu'elles sont écrites dans un langage de base relativement simple, elles sont facilement accessibles et peuvent être dupliquées et combinées entre elles de manière infinie et quasi instantanée.

En recoupant, par exemple, les admissions dans des hôpitaux, les achats de médicaments et les recherches sur Internet, il devient possible de suivre la propagation d'une épidémie en temps réel. Lors de l'épidémie de grippe porcine, en 2009, Google a ainsi tracé la progression de la maladie en suivant les requêtes des usagers.

Les données numériques brutes, surtout lorsqu'elles se présentent en quantité aussi bruyante, n'offrent cependant que peu d'intérêt si elles ne sont pas intégrées à des modèles susceptibles de leur conférer du sens et, à terme, de produire de nouvelles connaissances.

Pour pouvoir les traiter, il est en effet nécessaire de structurer l'information de façon pertinente, en corrigeant notamment les biais par lesquels elle a été collectée, organisée et parfois inférée lorsque les données originelles étaient manquantes. Plus il y a d'information, plus il est facile de la faire mentir, souvent involontairement.

TRAVAIL TITANESQUE

Pour s'atteler à ce travail titanesque, auquel collabore l'Institut de recherche en statistique de l'UNIGE, les statisticiens se doivent de faire appel aux connaissances d'autres disciplines, au sein de cette nouvelle «science des données». Celles des mathématiciens et des informaticiens en premier lieu. Mais aussi celles des psychologues et des sociologues, tant le facteur humain tend à occuper une place importante dans le traitement des grandes bases de données.

lusoire. Les chercheurs sont par conséquent condamnés à extrapoler, en limitant au maximum la marge d'erreur induite par cette démarche, s'ils veulent produire une description valable pour toute une population.

Les statisticiens considèrent généralement qu'une information est significative, lorsqu'il est possible de l'extrapoler avec une marge d'erreur se situant en-dessous de la barre des 5%, sa grandeur précise étant déterminée en fonction du domaine, du problème et des données.

Traqueurs de connaissances, les statisticiens s'efforcent aussi de rendre leurs collègues chercheurs attentifs à des résultats trompeurs. «Une marge d'erreur de 5% est relativement faible, explique Maria-Pia Victoria-Feser. Mais si on la cumule en répétant l'analyse, la possibilité de produire un résultat erroné grimpe rapidement.»

SAUMON MORT-VIVANT

Des chercheurs américains se sont ainsi amusés à fournir la «preuve» d'une activité cérébrale chez un saumon mort depuis plusieurs heures, en employant la méthode statistique standard utilisée en neurosciences. Un résultat pour le moins surprenant qui s'explique par le fait que cette approche standard est mal adaptée à la mesure de l'activité cérébrale. Elle contient forcément une marge d'erreur qui, si elle est cumulée sur une grande quantité de tests, augmente fortement la probabilité d'obtenir un résultat absurde.

Autre exemple d'erreurs fréquentes, celles liées au conditionnement des données: en ne tenant compte – volontairement ou non – que d'un seul paramètre dans une analyse, laissant de côté d'autres aspects signifiants, il est possible de décrire un phénomène de cause à effet là où il n'existe qu'une corrélation.

Un chercheur de l'Université Columbia a ainsi établi un lien entre la consommation de chocolat et les capacités cognitives d'une population mesurées au nombre de ses Prix Nobel: plus on mange de chocolat, plus on produit de Prix Nobel. Cause à effet ou simple corrélation?

Selon Maria-Pia Victoria-Feser, nombre d'erreurs scientifiques involontaires pourraient être évitées en s'assurant les services d'un statisticien. «Les universités, du moins en Europe, sont sous-dotées sur ce plan, dans nombre de domaines qui recourent à la statistique comme la médecine ou la biologie.» Raison pour laquelle l'Institut de recherche en statistique a dégagé des ressources pour offrir à la communauté universitaire un service de consultation (*lire ci-dessus*).