La statistique à l'ère des big data

Pierre-Yves Louis

La statistique à l'ère des « big data »

La période actuelle avec son déferlement de données peut-elle être qualifiée de révolution numérique ? Seul le recul historique le déterminera. De profondes mutations sont cependant déjà en cours. Les statisticiens sont en première ligne.

D'incroyables quantités de nombres jouent dorénavant un rôle décisif dans des domaines techniques, sociétaux et économiques variés. Des mutations culturelles sont engagées. La statistique, science des données, voit sa position de connaissance élémentaire renforcée. L'urgence d'une numératie, à l'instar de l'alphabétisation, est régulièrement soulignée par les spécialistes.

L'utilisation de nouvelles technologies et de l'outil informatique génère énormément de données. La quantité produite est d'un ordre de grandeur sans précédent. La production en 2015 devrait être de l'ordre de 5.600 milliards de GB, le double de 2012. La variété des données, ainsi que leur modification très rapide dans le temps posent de nouvelles questions : acquisition, stockage, traitement, mais également information utile extraite, décisions en conséquence, réglementation et respect de la vie privée et du citoyen, accès libre aux données publiques (open data).

Dès 2014, le gouvernement français a été le premier au monde à nommer un administrateur général des données, comme les entreprises se dotent de directeur des données/ Chief Data Officer (CDO).

La terminologie de big data, données massives ou mégadonnées est sur toutes les lèvres depuis le rapport McKinsey*. Elle porte des espoirs de croissance économique. Comme une matière première, les données seraient le pétrole du XXI[ème] siècle.

« Il ne s'agit pas uniquement de manipuler des nombres...»

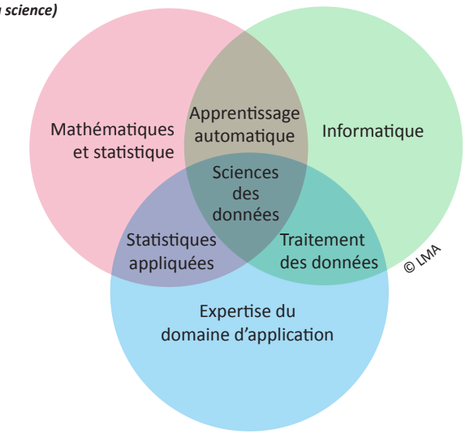
De nouveaux horizons statistiques

La science des données est en fait la statistique du 21[ème] siècle, portée sur de nouveaux champs d'exploration et de défis. Dans ces ordres de grandeur inédits, l'acquisition, le stockage, la sécurisation, la recherche, le partage, l'analyse et la visualisation des données doivent être redéfinis. La statistique s'est développée et structurée autour de méthodes mathématiques utilisées pour raffiner la matière brute que constituent les données. Ainsi, classiquement, quand une grandeur mesurée ne prend que des valeurs numériques, elle est dite quantitative.

Des fonctions mathématiques (appliquées aux données), des « statistiques », peuvent synthétiser l'information observée. Une valeur « centrale » est par exemple obtenue grâce à la moyenne usuelle qui possède des propriétés utiles à de nombreuses généralisations. L'écart-type est un indicateur possible pour quantifier et estimer la manière dont les différentes mesures se dispersent autour de cette valeur moyenne. Une grandeur dite catégorielle qui ne prend que des valeurs qualitatives, comme, par exemple, le sexe d'une personne, ne peut être résumée par une moyenne numérique. Une solution pour représenter l'information consiste à travailler avec les proportions d'hommes et de femmes observées dans les mesures. On est alors ramené à des nombres compris entre 0 et 1 et dont la somme vaut 1.

Il ne s'agit pas uniquement de manipuler des nombres, au-delà des stricts aspects mathématiques, il faut tenir compte du contexte dans lequel les mesures sont obtenues. Et ce contexte implique des choix en termes d'objets mathématiques retenus pour représenter l'information. Ainsi la couleur des yeux peut-elle être modélisée grâce à une variable qualitative par quelques couleurs courantes. Si toutefois, dans le cadre d'une

Qu'est ce que la science des données ? (data science)



expérience à l'aide d'un capteur spécifique, on dispose d'une mesure plus précise, par exemple, quantifiée en code rouge/vert /bleu, il sera plus judicieux, afin d'exploiter la richesse de l'information acquise, de traiter la grandeur comme une variable numérique quantitative. Le besoin de mathématiques demeure afin de choisir une représentation adéquate de l'information afin de la traiter et de la visualiser : graphes, modèles et algorithmes aléatoires, analyse matricielle, optimisation et même topologie et géométrie !

Un travail tout en finesse

Un des aspects des données massives réside dans la quantité très importante de grandeurs mesurées. Une autre de leurs spécificités réside dans le croisement de données de sources diverses, hétérogènes. Elles doivent être considérées conjointement. Un exemple de taille relativement restreinte : les données issues des biopuces. Cette technologie permet de mesurer simultanément un niveau d'expression de nombreux gènes à travers les molécules d'ARN produites. La question du biologiste est de distinguer les gènes sous/surexprimés en fonction de conditions expérimentales différentes dans lesquelles sont placées les cellules. Le statisticien doit relever deux défis :

- travailler avec un grand nombre de variables quantitatives, typiquement ici de l'ordre du millier ;
- ne pas résumer des groupes à des moyennes dont on peut considérer l'écart mais différencier des gènes dont le comportement est déviant par rapport au groupe.

En moyennant trop brusquement, les finesses nécessaires pour répondre au biologiste sont perdues. La nécessité est alors de disposer de modèles aléatoires plus précis, qui prennent en compte une plus grande hétérogénéité dans les données. Une manière abstraite de représenter le résultat d'une telle expérience est de considérer que les valeurs mesurées pour chaque gène correspondent à des coordonnées. Chaque cellule est alors représentée par un point dans un espace mathématique de très grande dimension. Une difficulté s'impose au statisticien : les espaces de grande dimension ont des propriétés mathématiques très différentes de l'espace à deux ou trois dimensions auxquels chacun est habitué. En particulier, la notion de localité, de voisinage, devient inadaptée pour répondre à ce type de questions. Le nombre de cellules qui constitue ce nuage de points est par ailleurs faible au regard de la très grande dimension.

De nouvelles approximations sont nécessaires, où l'on ne fait plus seulement augmenter la taille de l'échantillon pour améliorer la finesse de l'analyse.
L'exemple des biopuces relève d'un cadre expérimental bien défini et pensé au préalable. L'approche big data consiste cependant à croiser des données préexistantes sans plan d'expérience statistique réfléchi et sans avoir défini les questions auxquelles on souhaite répondre. Les informaticiens et spécialistes de la fouille des données voient ainsi également leurs méthodes mises à l'épreuve. Les techniques d'apprentissage automatique qu'ils déploient apportent des réponses. La société américaine Oracle, leader mondial des bases de données, estime que les données non-structurées représentent 80 % des données.

Les données massives présentent de nombreuses particularités qui rendent leur exploitation très limitée avec les outils traditionnels de l'analyse de données. Hors des sentiers traditionnels, les écueils possibles sont nombreux. Le risque est grand de tirer des conclusions et d'établir des liens de causalité à partir de manipulations kabbalistiques fortuites de nombres.

La « science des données » étend les champs d'exploration de la statistique largement au-delà de la discipline mathématique vers les domaines applicatifs et d'autres sciences fondamentales, comme l'informatique. Le rôle d'interface de la statistique est renforcé. Ce nouveau champ d'activités confirme les besoins en termes de représentation de l'information, de quantification, de nouvelles approximations, de modèles qui tiennent compte d'une plus grande hétérogénéité. De belles perspectives s'offrent alors aux futur(e)s mathématicien-s/-nes.

Pierre-Yves LOUIS < LMA
pierre.yves.louis@univ-poitiers.fr

http://rech-math.sp2mi.univ-poitiers.fr

*Le Cabinet McKinsey conseille les directions générales de grandes entreprises françaises et internationales, ainsi que celles d'institutions publiques et d'organisations à but non lucratif. Il a publié en 2011 une étude sur l'impact du numérique : *Big data: the next frontier for innovation, competition, and productivity.*