La statistique : un nom galvaudé pour une discipline universelle

Armand Maul et Daniel Vagost

Lors de la conférence donnée par Armand Maul pendant les Journées Nationales de notre association à METZ, ce dernier a réussi à partager avec ses auditeurs sa passion de la statistique, et c'est cette passion pour une science, dont l'importance ne cesse de grandir, que nous voudrions partager avec vous dans cet article écrit à quatre mains.

Armand Maul enseigne les mathématiques à l'IUT de Metz ; Daniel Vagost, jeune retraité, reste un membre actif de l'APMEP.

Quelques citations

En guise d'introduction souriante, voici quelques propos autour de la statistique entendus ça et là :

- *Il y a trois sortes de mensonges : les mensonges, les gros mensonges et les statistiques* (Mark Twain)

- *Les statistiques sont la forme la plus élaborée du mensonge* (Gladstone)

- *Je ne crois aux statistiques que lorsque je les ai moi-même falsifiées* (Winston Churchill)

- *Les statistiques, c'est comme le bikini. Ce qu'elles révèlent est suggestif. Ce qu'elles dissimulent est essentiel* (Aaron Levenstein)

- *La statistique est moins une science qu'un art. Elle est la poésie des nombres. Chacun y trouve ce qu'il y met* (Albert Brie)

Quelques définitions

Plus sérieusement, voici ce que disent de la statistique quelques dictionnaires :

- Étude méthodique des faits sociaux par des procédés numériques (classements, dénombrements, inventaires chiffrés, recensements) destinée à renseigner et à aider les gouvernements (Le Petit Robert)

- Ensemble des méthodes qui ont pour objet la collecte, le traitement et l'interprétation de données d'observation relatives à un groupe d'individus ou d'unités (Larousse)

- Branche des mathématiques ayant pour objet l'analyse (généralement non exhaustive) et l'interprétation de données quantifiables (Le Trésor de la Langue Française).

Ces quelques définitions montrent que la statistique est une branche des mathématiques et qu'elle a, entre autres, vocation à donner des éléments d'aide à la décision.

Pour terminer cette introduction, citons J.P.A. Ioannidis (2005), selon lequel : *Plus de 50 % des conclusions des articles publiés en médecine sont fausses*. Cette déclaration, qui ne manque pas de jeter une certaine suspicion sur une partie de la recherche scientifique, est-elle vraiment justifiée ?

Il faut dire que, trop souvent, la formation de certains scientifiques est insuffisante en statistique : les techniques statistiques sont parfois très sophistiquées et nécessitent, contrairement à ce que l'on croit trop souvent, une grande compétence, ainsi qu'une grande maîtrise des méthodes utilisées. Aussi, parmi les erreurs expéri-

mentales les plus courantes, il convient d'en citer deux :

- la première consiste à « voir » quelque chose, alors que « il n'y a rien » ; cette erreur, appelée erreur de première espèce, se produit, par exemple, lors de la déclaration, à tort, de l'efficacité d'un traitement médicamenteux.

- la seconde, ou erreur *a posteriori*, consiste à vérifier la véracité d'une hypothèse à partir du même échantillon que celui ayant permis d'observer et d'émettre cette hypothèse. Or, la rigueur scientifique requiert l'utilisation d'un nouvel échantillon, indépendant du premier, pour valider ou invalider ladite hypothèse.

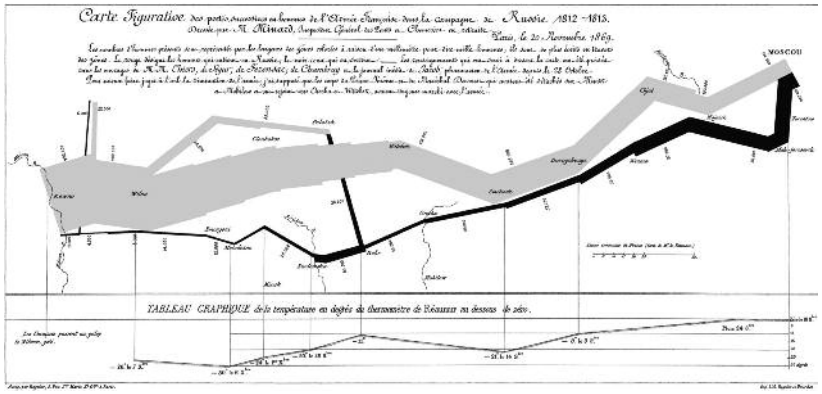
En dehors de ces deux erreurs techniques, on rencontre également des erreurs humaines, liées à une certaine partialité, à l'adhésion à une théorie en vogue, ou encore à une perte d'objectivité en raison d'intérêts financiers, sans compter l'effet amplificateur de médias toujours à la recherche d'un scoop. L'actualité récente nous fournit un bon exemple d'accumulation de ces erreurs : la publication d'une étude de l'effet d'un maïs OGM sur l'apparition de tumeurs chez des rats. Mais attention : dire que cette étude pèche par son manque de rigueur scientifique, qu'elle contient certaines insuffisances méthodologiques, ne signifie nullement qu'il n'y a pas d'effet de l'OGM sur les rats, mais comme l'écrit, et le regrette, Marc Lavielle : *Tout commentaire d'ordre purement scientifique est trop souvent systématiquement assimilé à une prise de position*.

Or, nous venons d'illustrer un paradoxe... car le but essentiel de la statistique, de même que le rôle du statisticien, sont précisément d'éviter le manque d'objectivité ! Le statisticien est assurément le garant de l'objectivité.

Quelques éléments d'histoire

La statistique n'est pas une science récente, son histoire est déjà longue. Le mot statistique vient du latin *statisticum :* ce qui se rapporte à l'État. C'est Gottfried Achenwall qui, en 1746, créa le premier enseignement de statistique à l'Université de Marburg en Allemagne ; mais en fait, l'origine du mot semble plus ancienne car il avait déjà été mentionné dans un texte administratif de Colbert (vers 1666). Plus généralement et depuis l'antiquité, les chefs d'états ont constamment fait appel à la statistique, sans le savoir, que ce soit afin d'évaluer leur puissance, leurs richesses, leur potentiel militaire ou, plus simplement, lorsqu'ils cherchaient à dénombrer leur population. C'est l'idée de recensement qui apparaît ici, recensement que l'on retrouve déjà chez les sumériens (vers 3000 av J.-C.) mais aussi chez les égyptiens qui pratiquaient des recensements systématiques de la population. Ainsi, et de tout temps, la statistique est incontestablement liée à des États forts avec un système administratif fort.

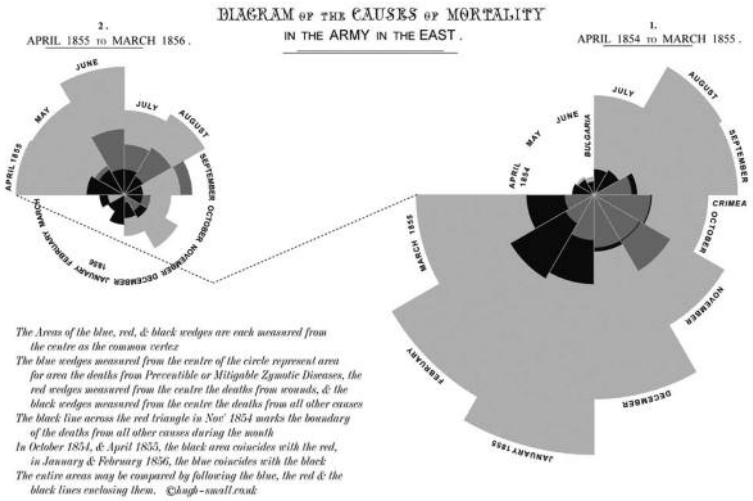
Il faut cependant remarquer que la statistique, du moins à l'origine, n'est qu'une science de l'observation : c'est la **statistique descriptive**, dont le développement a connu un essor particulier au 14ème siècle, avec le début de l'enregistrement des actes civils (naissances, mariages, décès). Les deux figures de la page 4 représentent des fleurons de la statistique descriptive.



La Grande Armée durant la campagne de Russie (1812-1813)

Ce premier diagramme illustre les pertes humaines par une bande, dont la largeur est proportionnelle à l'effectif de la Grande Armée.

Le diagramme suivant, dû à Florence Nightingale, illustre les causes de la mortalité observée dans l'armée britannique au cours de la guerre de Crimée. L'aspect quantitatif est donné par des secteurs dont l'aire est proportionnelle à l'importance de la mortalité correspondant à chaque mois de l'année.



La statistique inférentielle

C'est au cours du 17ème siècle qu'apparaissent, avec les prémisses du calcul probabiliste, les premières tentatives d'estimations et de prévisions, réalisées à partir d'un sous-ensemble de la population statistique étudiée, marquant ainsi les débuts de la **statistique inférentielle**. Le premier exemple d'extrapolation, à partir d'une partie seulement de la population considérée, peut être attribué à William Petty qui, en 1686, a cherché à estimer la population londonienne. À cet effet, il a estimé le nombre de maisons dans Londres, puis le nombre moyens de feux (foyers dirait-on aujourd'hui) par maison et, enfin, le nombre moyen de personnes par feu, pour aboutir à une estimation de 695 000 habitants…, ce qui, au dire des démographes actuels, est tout à fait vraisemblable.

Juger d'après un échantillon n'est cependant pas chose facile, et convaincre l'auditoire qu'il est possible de se faire une idée du tout à partir seulement d'une partie a été un travail de longue haleine. Sur cette longue route citons deux dates :

* 1895 : le Norvégien Anders Kiær présente, lors du congrès de l'*Institut International de Statistique* (IIS) à Berne de 1895, une étude portant sur la distribution des revenus en Norvège. Pour ce faire, il avait sélectionné aléatoirement un certain nombre de communes et, dans ces communes, il avait tiré une lettre au hasard, afin d'interroger les habitants dont le patronyme commençait par la lettre tirée. Les réactions de l'auditoire furent violentes car les scientifiques de l'époque n'étaient pas prêts à accepter la possibilité de juger ainsi à partir d'un échantillon. Entre 1895 et 1925, le problème s'est progressivement transformé : en effet, la question de l'opportunité de pouvoir juger d'après un échantillon s'est très vite déplacée vers le débat sur la représentativité, très présent à l'IIS durant cette période. Comment en effet faut-il constituer un échantillon ? On peut cependant affirmer qu'à partir de 1925 la théorie statistique était devenue parfaitement cohérente. Et c'est là qu'intervient un autre événement important.

∗ le 3 novembre 1936 correspond à la date des élections présidentielles américaines, qui voient le républicain G. Landon affronter le président sortant F. Roosevelt. Le magazine *Literary Digest,* qui ne s'était jamais trompé lors des élections précédentes, prévoit la victoire de Landon, en interrogeant plus de 2 000 000 de personnes issues des lecteurs et listes d'abonnés du périodique. Le statisticien George Gallup, quant à lui, se contente d'interroger 3 000 personnes, issues d'un tirage aléatoire dans la population du pays, ce qui lui permet de prévoir la victoire de Roosevelt, lequel est effectivement réélu à l'issue du scrutin. Ce résultat surprenant pour l'époque a contribué à faire accepter dans l'opinion le principe de la validité d'un sondage par tirage aléatoire.

Puis, l'informatique et l'avènement des ordinateurs, au cours de la seconde moitié du 20ème siècle, ont fait le reste : on n'imagine en effet plus possible de nos jours de faire de la statistique sans l'outil informatique.

Un petit mot à propos de Student, dont on connaît la distribution éponyme et le t-test, popularisés par Fisher. Il s'agit en fait du statisticien William Gosset qui, en 1908, a publié son travail sous le pseudonyme de Student, parce que son contrat avec la brasserie Guinness, où il avait été embauché pour stabiliser le goût de la bière, ne l'autorisait pas à le faire sous son vrai nom !

Il serait inconvenant de terminer ce petit tour d'horizon historique sans citer Ronald Fisher qui a fait de la statistique une science moderne, et à qui l'on doit : la théorie générale de l'estimation avec la méthode du maximum de vraisemblance, l'analyse de la variance (ANOVA), l'analyse discriminante, les plans d'expériences, le test exact de Fisher..., et la liste est loin d'être close ! Si l'on ajoute à cela qu'il a été anobli par la reine d'Angleterre en 1952, on comprend combien la statistique fait partie des valeurs de références de la culture anglo-saxonne. À cet égard, s'il est vrai que le métier de statisticien, dans le monde anglo-saxon, est prestigieux, il n'en va pas de même en France, où nous avons par ailleurs accumulé un énorme retard, en particulier dans la culture probabiliste.

Le retard français et ses conséquences

Pourquoi un tel retard en France ? Les raisons sont nombreuses, mais nous ne citerons que les plus importantes :

∗ la statistique a toujours été un parent pauvre des probabilités et des mathématiques ; la forte imprégnation cartésienne et le peu de goût pour l'aléatoire en sont probablement la cause principale,

∗ l'« incertain » est considéré anti-pédagogique car trop déstabilisant ; il existe une nette préférence pour l'enseignement de certitudes et puis : *un chiffre ne peut être qu'exact, exempt de toute indétermination*. Ainsi, personne ne trouverait opportun de contester la justesse, par exemple, d'un taux de cholestérol ; car chercher à relativiser un tel nombre, en l'accompagnant d'une information sur la précision ou la reproductibilité de la mesure, pourrait paraître suspect et relèverait d'un manque de confiance,

∗ l'approche pluridisciplinaire est encore mal acceptée ; peut-être faut-il y voir un effet préjudiciable du cloisonnement de la classification des sciences établie par Auguste Comte ?

La principale conséquence de ce retard est que le bon sens statistique est insuffisamment développé dans notre pays, ce qui permet toutes les manipulations possibles. Citons deux exemples :

∗ Lorsqu'une étude établit le lien étroit entre tabagisme et cancer du poumon, nombreuses sont les personnes qui, en dépit de toutes les études similaires aux résultats convergents, sont tentées de les mettre en doute en évoquant l'exemple d'un grand-père qui, bien que gros fumeur depuis son plus jeune âge, est mort de sa belle mort à plus de 90 ans ! C'est oublier que le statisticien ne s'intéresse guère à l'individu mais uniquement aux tendances générales. Citons alors Léon Schwartzenberg : « *Les statistiques sont vraies quant à la maladie et fausses quant au malade ; elles sont vraies quant aux populations et fausses quant à l'individu*. »

∗ Dans un autre domaine, il faut savoir que la multiplication des expériences génère l'« *insolite* ». Les téléspectateurs admirateurs d'Uri Geller, dans les années 1980, n'étaient manifestement pas conscients de ce phénomène. En effet, cet animateur prétendait pouvoir agir directement sur la matière par l'esprit ; en particulier, lorsqu'il prétendait détenir le pouvoir de faire claquer des ampoules électriques dans les foyers des téléspectateurs, simplement par la pensée. À chacune de ces interventions, ils étaient des centaines à lui donner raison ! De façon similaire, c'est oublier un peu vite que si rencontrer une personne de plus de 2 mètres dans un groupe de 10 personnes est peu fréquent (environ 1 chance sur 100), cette possibilité passe à 10 % dans un groupe de 100 personnes, à 65 % dans un groupe de 1 000 personnes et la probabilité, cette fois, de ne pas trouver une personne de plus de 2 mètres dans un groupe de 10 000 personnes est de l'ordre du cent-millième !

Ces quelques exemples montrent que, d'une manière générale, la présentation de données statistiques est un exercice « difficile » comme l'illustre également une étude épidémiologique des années 60, dont l'objectif était d'étudier le lien éventuel entre le poids des nouveau-nés et certaines caractéristiques physiques et sociologiques de la mère. L'étude, portant sur près de 10 000 femmes enceintes, a permis d'identifier trois facteurs prépondérants sur le poids des nouveau-nés : la consommation de tabac, la taille, ainsi que le niveau d'instruction (i.e. le nombre d'années d'études) de la mère. Quelles ne furent pas les critiques faites à l'encontre de cette étude ! Comment, en effet, le nombre d'années d'études pouvait-il avoir un lien avec le poids du bébé à la naissance ? C'était tout simplement oublier que, derrière ce nombre, se cachaient en réalité d'autres facteurs comme, par exemple, la capacité à respecter certaines règles d'hygiène de vie, la pénibilité de l'activité professionnelle, qui est peut-être moindre chez des femmes ayant un niveau d'éducation élevé... De manière très générale, et ainsi que l'a souligné Gérald Bronner dans un livre remarquable sur les coïncidences (2007) : *Le hasard est un hôte indésirable de la pensée humaine,* et ce ne sont pas les rapports d'experts, parfois contradictoires, qui nous rassurent. Il y a un véritable antagonisme entre, d'une part, la volonté d'être simple, clair et précis et, d'autre part, la présentation d'une réalité qui, très souvent, est beaucoup plus complexe.

Enfin, afin de mesurer l'ampleur du désintérêt français pour la statistique, il est possible de citer quelques chiffres. L'*Institut International de Statistique* (IIS) a organisé, depuis 1853, 58 congrès internationaux, dont le dernier en date s'est déroulé en 2011, à Dublin. Comme à l'accoutumée, cette manifestation de grande envergure a rassemblé près de 2 250 participants venant de 105 pays, et a donné lieu à plus de 1 500 présentations orales... Or, la délégation française comportait moins de 20 personnes : on aurait pu s'attendre à plus !

Ce constat est d'autant plus regrettable que la statistique est la seule discipline qui revendique le droit à l'erreur et qui, de surcroît, propose de quantifier le risque d'aboutir à une conclusion erronée.

En effet, la statistique nous dit comment :
- effectuer les mesures,
- extraire l'information des données,
- appréhender l'incertitude,
- quantifier le risque d'erreur,

dans le but de :
- trouver et décrire une relation (par exemple entre le risque cardio-vasculaire et la consommation de tabac),
- prendre une décision (par exemple l'efficacité d'un médicament avant sa mise sur le marché),
- prévoir et planifier (par exemple le budget prévisionnel d'une commune, du gouvernement…).

On peut conclure en citant Léon Brunschvicg : *Connaître, c'est mesurer.*

La méthode expérimentale

En 1865, Claude Bernard expose les principes de la méthode expérimentale dans son ouvrage *Introduction à l'étude de la médecine expérimentale*. Les principales étapes de cette approche, qui est aussi celle utilisée par le statisticien, sont les suivantes :

* observation d'un phénomène : par exemple, il semblerait que la criminalité (ou le nombre de naissances) augmente les nuits de pleine lune,

* formulation d'une hypothèse (un modèle) : par exemple, le nombre de naissances est plus important les nuits de pleine lune ,

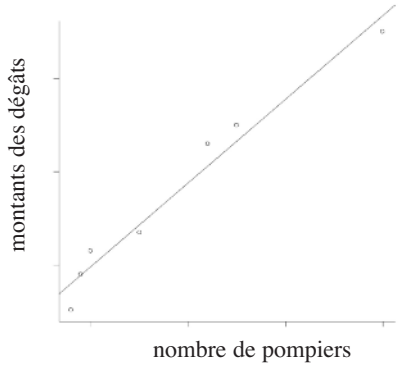
* vérification de cette hypothèse par l'expérimentation, c'est-à-dire à partir d'un échantillon de la population, de tous les résultats possibles.

Cette troisième étape, qui est parfois négligée voire omise, est absolument indispensable. Car sans elle, il ne s'agit pas de science !

Il s'ensuit que le statisticien, qui joue un rôle majeur dans l'étape de vérification, fait quelquefois figure de rabat-joie : par exemple, lorsque les études expérimentales rigoureuses montrent que les nuits de pleine lune ne sont propices, ni à une augmentation des naissances, ni à celle de la criminalité ! Le statisticien sait aussi que les modèles (les hypothèses) validés ne sont que provisoires ; ils sont destinés à être améliorés. En effet, la réalité est souvent complexe, et tout scientifique digne de ce nom acceptera de faire sienne la citation remarquable de George Box : *Tous les modèles sont faux ; certains sont utiles*.

Dans un autre registre, le statisticien sait aussi parfaitement qu'une corrélation entre deux caractères ne signifie pas nécessairement l'existence d'une relation de cause à effet. Nombreux sont ceux, en effet, qui tombent dans ce piège, qu'il est aisé d'illustrer par l'exemple ci-dessous. La figure laisse apparaître le lien existant

entre le nombre de pompiers dépêchés sur les lieux d'un incendie et le montant des dégâts occasionnés !



Il serait tentant, mais hasardeux, de conclure que « les pompiers mettent de l'huile sur le feu ». Ici, il est facile de comprendre que les deux caractères étudiés sont évidemment corrélés à un troisième, à savoir : l'ampleur de l'incendie. Mais les situations concrètes ne sont pas toujours aussi évidentes... et pour montrer que même les génies peuvent parfois se tromper, on peut citer Fisher qui, face aux toutes premières études établissant un lien entre tabac et cancer, refusait obstinément de croire à la causalité, en argumentant que : *correlation is not causation* !

La variabilité : une source de malentendus

La variabilité est un principe universel souvent mal appréhendé. Sans variabilité, il n'y aurait pas de statisticien ! Il n'y aurait d'ailleurs pas grand-chose… Car, en général, la variabilité est un critère de qualité dans une population. En effet, si nous étions, par exemple, tous également et uniformément sensibles aux mêmes micro-organismes pathogènes, il y a bien longtemps que l'homme aurait disparu de la terre. Il n'y a qu'une exception à cette règle : l'industrie, où le rêve du responsable qualité serait de voir disparaître la variabilité de sa production, avec des pièces fabriquées parfaitement identiques et, bien évidemment, toutes conformes au cahier des charges !

La variabilité est la raison d'être du statisticien : il sait l'appréhender, l'analyser et la gérer, en sachant par ailleurs que la variabilité est, par essence, une entrave à l'inférence statistique. À cet égard, le statisticien sait qu'il faut se garder de toute généralisation abusive car il est bien conscient que :

- les données observées, de toute nature, sont des réalisations d'un phénomène aléatoire,
- le résultat observé est un résultat parmi « beaucoup » d'autres possibilités.

Il sait aussi qu'il y a une très grande « diversité potentielle de résultats possibles » et que ne pas en tenir compte dans l'analyse des résultats expérimentaux peut conduire à des conclusions erronées.

À cet effet, le statisticien tient à sa disposition les outils de la théorie des probabilités et de la statistique inférentielle pour tenter d'estimer les valeurs des paramètres (dont les valeurs exactes resteront à jamais inconnues) caractérisant une population, à partir des valeurs mesurées (empiriques) observées sur un échantillon (un sous-ensemble pertinent de la population étudiée). Dans le cas où les tirages sont réalisés de manière aléatoire, la théorie des probabilités permet en effet de déterminer, à partir de la valeur observée dans l'échantillon (la moyenne m, la variance s^2, la proportion x/n), un intervalle qui, avec une probabilité définie à l'avance (par exemple 95 %), contiendra la valeur inconnue du paramètre (respectivement, la moyenne μ, la variance σ^2, la proportion π) de la population.

La qualité de l'information disponible, et l'utilisation de la théorie de façon appro-

priée, nécessitent d'être très vigilants sur les caractéristiques de l'échantillon recueilli. Quelles sont ces propriétés ?
- la taille de l'échantillon : dans le cas d'un sondage d'opinion, un échantillon aléatoire de 1 000 personnes interrogées permet de construire un intervalle de confiance à 95 % avec une amplitude inférieure à 6 % (avec 4 000 personnes, l'amplitude se réduit à 3 % ; avec 9 000 personnes, elle atteint 2 %),
- la représentativité : un échantillon non représentatif est susceptible d'entraîner des biais fâcheux ; à cet égard, le tirage au hasard des individus garantit la représentativité. Il convient par conséquent d'être à la fois prudent et très exigeant dans le choix de l'échantillon, en particulier lorsque le tirage aléatoire est peu commode ou impossible ; attention aux échantillons composés de volontaires !

Quelques remarques :
- la taille de la population n'intervient pas... et les mathématiques montrent qu'il vaut mieux disposer d'un échantillon aléatoire de 2 000 personnes dans une population de 60 millions que d'un échantillon de 60 personnes dans un village de 100 habitants !
- pour assurer le contrôle et la surveillance de l'évolution de la qualité de l'eau du lac Érié, qui représente pourtant une énorme masse d'eau très peu homogène, le calcul probabiliste montre qu'il suffit de prélever et d'analyser 25 échantillons d'eau bien répartis à la surface du lac !

Alors, il est permis de se demander pourquoi les sondages d'opinion sont autant, et si souvent, décriés. Et ce, malgré les remarques précédentes et la parfaite cohérence de la technique mathématique. On peut avancer quelques hypothèses :
- l'opinion est, par nature, changeante : ce qui est vrai aujourd'hui peut ne plus l'être demain,
- ne pas répondre à une question d'un sondage, ne signifie pas obligatoirement « abstention » ou « vote blanc ou nul »
- les individus ne sont pas toujours sincères, surtout si on aborde des sujets sensibles ou en rapport avec leur intimité.

Pour conclure cette partie, il faut avoir conscience qu'il est nécessaire d'avoir une formation suffisante en mathématiques, et aussi en statistique, pour ne pas se laisser abuser...

L'analyse des données

Comme nous l'avons déjà fait remarquer plus haut, le développement de l'informatique a, de nos jours, totalement banalisé la réalisation, ainsi que l'accès à de gigantesques recueils de données. On dispose en effet aujourd'hui de véritables « entrepôts de données », lesquels peuvent contenir de l'information utile, et qu'il faut savoir extraire. Or, ceci est précisément l'objet de l' « analyse des données », car il ne faut surtout pas confondre « données » et « information ».

En déclarant récemment : *Data is the new oil*, le futuriste Gerd Leonhard a comparé les données à de la matière brute qui est susceptible de contenir de l'information (comme la statue se trouve dans le bloc de marbre, comme la pépite d'or est dissimulée dans la montagne !). Il est dorénavant possible de dire que : *L'information est la matière première du 21ème siècle*. Or, sa production et son exploitation sont l'affaire des statisticiens. On perçoit donc immédiatement le rôle social du statisticien, qui est un interlocuteur privilégié des décideurs, dans tous les secteurs d'activité (politique, économique, scientifique, industriel…), et à tous les niveaux (collecte de données, conception des systèmes d'information, contrôle de

la production, analyse et restitution des données, etc.).

La statistique est une discipline essentiellement transversale, qui est largement utilisée dans de nombreux domaines, tels que : la statistique officielle (I.N.S.E.E.), la presse et les médias (même si on peut regretter que les journalistes n'aient pas toujours une formation en statistique, ou plus simplement scientifique, suffisante), les banques et les assurances, les sciences de la vie, l'environnement (foresterie, pêche …), la santé, les sciences humaines, les entreprises et l'industrie (R&D, contrôle de qualité, études de marché, management…, même si les chefs d'entreprise ne savent pas toujours que certains de leurs problèmes pourraient être appréhendés et résolus grâce à la statistique), la finance (dans ce dernier domaine, la difficulté majeure semble cependant liée aux comportements erratiques inhérents à la psychologie humaine !), etc.

On peut également citer le système statistique national (mondial), qui utilise la statistique dans la planification et la surveillance dans des domaines très variés : économie, démographie, société, santé, environnement. Pour souligner une nouvelle fois, si nécessaire, l'importance de la statistique, on peut rappeler que la *Commission de statistique de l'ONU* a décidé d'instaurer une journée mondiale de la statistique (20/10/2010), et décrété que 2013 serait l'année de la statistique mondiale. En France, il existe de nombreuses journées (journée du fromage, des gauchers...), mais pas de journée de la statistique...

On pourrait ici développer de très nombreux exemples d'utilisation de la statistique ; nous allons brièvement en exposer trois :

* Pourquoi les banques utilisent-elles la statistique ? Face aux risques financiers, il devient de plus en plus important de connaître la probabilité qu'un client rembourse son crédit. Cette probabilité peut être estimée et exprimée en fonction du montant moyen du compte courant, de la durée du crédit, du montant du crédit, du sexe, de la CSP, de la situation familiale...

* Et les assurances ? Le but est d'établir le juste tarif, celui qui permet de mieux résister à la concurrence, tout en respectant des règles de déontologie et d'éthique. Citons l'exemple des assurances-décès, qui nécessitent de travailler sur les tables de mortalité, sans oublier les taux d'intérêt et les frais de gestion...

* En sciences médicales, les essais cliniques (conçus pour vérifier ou comparer l'efficacité d'un ou plusieurs traitements) illustrent l'irruption de la preuve statistique en médecine. Les essais cliniques permettent en effet de s'affranchir de l'influence du jugement humain, des publicités de l'industrie pharmaceutique et des aspects psychologiques médecin-malade. C'est Ronald Fisher qui a été l'instigateur des essais randomisés. L'introduction du hasard dans l'expérience médicale, par les essais cliniques randomisés en double-aveugle qui permettent d'étudier les propriétés biochimiques et thérapeutiques des molécules, a constitué une avancée majeure, en faisant notamment abstraction des effets d'autosuggestion et d'hétérosuggestion dans la relation médecin-malade.

Pour illustrer ces propos, considérons l'exemple suivant : on veut savoir si un nouveau médicament a des effets secondaires : en l'occurrence, donne-t-il des nausées ? En d'autres termes, les patients qui prennent ce médicament ont-ils plus

de nausées que ceux à qui l'on administre un placebo ?

Les résultats de l'étude ont montré que, parmi les 50 personnes ayant suivi le traitement, 15 ont été victimes de nausées, tandis que, parmi les 50 personnes ayant absorbé un placebo, 4 seulement ont eu des nausées. Que peut-on en conclure ? Si l'on imagine que le traitement n'entraîne pas plus de nausées que le placebo, les 19 victimes de nausées auraient dû se répartir équitablement entre les deux groupes, soit 9,5 personnes à la fois chez les individus traités et chez les individus sous placebo. Or, on constate un « léger » déséquilibre dans les proportions observées. Dès lors, on peut se demander si cette différence est vraiment significative, autrement dit révélatrice d'effets secondaires du traitement, ou si elle n'est pas tout simplement une conséquence du hasard, c'est-à-dire imputable aux fluctuations d'échantillonnage ?

	Nausées	
	OUI	NON
Traitement	15	35
	9,5	**40,5**
Placebo	4	46
	9,5	**40,5**

Pour répondre à cette question, le statisticien utilise un test (ici, le test du khi-deux), dont le principe se veut très proche du raisonnement par l'absurde en mathématiques. Pour ce faire, il calcule la « distance » entre, d'une part, les observations effectuées et, d'autre part, les effectifs théoriques attendus sous « l'hypothèse nulle » (l'hypothèse selon laquelle le traitement n'occasionne pas plus de nausées que le placebo). En supposant que l'hypothèse nulle est vraie, il calcule la probabilité (probabilité critique) que la « distance » soit au moins égale à celle observée, simplement par le fait du hasard. Si cette probabilité est jugée faible, il rejette l'hypothèse nulle, en affirmant que le médicament occasionne davantage de nausées que le placebo ; dans le cas contraire, il ne la rejette pas, et conclut que la différence observée n'est pas statistiquement significative (elle peut donc être mise sur le compte du hasard). Cependant, quelle que soit sa décision finale, il s'expose inéluctablement à deux risques d'erreur :

- dire que le traitement a des effets secondaires, alors qu'en réalité il n'en a pas ; on rejette l'hypothèse nulle, alors qu'elle est vraie ; il s'agit d'une erreur de première espèce,
- dire que le traitement n'a pas d'effets secondaires, alors qu'en réalité il occasionne des nausées ; on ne rejette pas l'hypothèse nulle, alors qu'elle est fausse (l'étude n'a pas permis de déceler les effets, car ceux-ci sont peut-être trop « petits » pour émerger du « bruit de fond » dû aux fluctuations d'échantillonnage, c'est-à-dire, les effets du hasard) ; il s'agit là d'une erreur de deuxième espèce.

On peut résumer la situation décisionnelle par le tableau suivant :

Décision / Situation réelle	Pas d'effets secondaires	Effets secondaires
Pas d'effets secondaires	Décision correcte	**Décision erronée Erreur de 1ère espèce Risque fournisseur**
Effets secondaires	**Décision erronée Erreur de 2ème espèce Risque client**	Décision correcte

Nous aurions pu également citer, et illustrer l'utilisation de la statistique, par des exemples concrets dans les champs disciplinaires suivants :

- les sciences de l'environnement : surveillance de la qualité d'un milieu hydrique, estimation de l'effectif d'une population animale ou végétale, toxicologie de l'environnement, analyse du risque, épidémiologie environnementale, prévision des phénomènes extrêmes (vagues, cyclones…), prévision des épisodes de pollution atmosphérique, estimation des effets du changement climatique, etc. Le lecteur intéressé trouvera une carte de France, illustrant une estimation du nombre de jours/an avec des températures supérieures à 35°C sur la période 2090-2099, sur le site de météo France,
- la santé-biologie,
- la biométrie,
- l'imagerie médicale,
- la génétique (trouver les gènes « coupables »),
- l'épidémiologie,
- la recherche de preuves en sciences forensiques (médecine légale).

Pour terminer, citons l'exemple d'une utilisation pour le moins insolite de la statistique en… poésie.

En 1985, on a retrouvé un sonnet composé de 429 mots, que certains experts ont été tentés d'attribuer à Shakespeare. Dès lors, comment répondre à l'inéluctable question : *Shakespeare en est-il vraiment l'auteur ?* Ce sont des statisticiens qui, en utilisant les outils de l'analyse discriminante textuelle, ont permis d'apporter une réponse claire à cette question : le sonnet a bien été écrit par Shakespeare.

Les nouveaux défis de la statistique

Les 50 dernières années ont été marquées par des bouleversements conceptuels et technologiques majeurs. En effet, de nouveaux champs disciplinaires sont apparus, les techniques de mesure ont considérablement évolué (apparition de l'imagerie médicale), les moyens de stockage, toujours plus grands, ont généré des bases de données gigantesques (astrophysique, génétique, populations…), et les moyens informatiques (capacité de calcul, logiciels) ont permis d'envisager de nouvelles méthodes statistiques, qui exploitent très largement la puissance de l'ordinateur. Mais toutes ces mutations qui, incontestablement, renforcent le rôle de la discipline, notamment par l'amélioration du potentiel des méthodes du traitement des données, ne changent cependant pas les principes fondamentaux de la pensée statistique !

Et, en guise de conclusion, quelques nouvelles citations mais, cette fois, elles émanent de statisticiens

La statistique publique rend le citoyen plus puissant, et les autorités plus responsables (Hans Rosling)

*Le statisticien est une personne qui préfère les vrais doutes aux fausses certitudes (*Hans Rosling*)*

À méditer :

Le hasard est la somme de nos ignorances (Pierre-Simon de Laplace)

Le hasard, cet hôte indésirable de la pensée humaine… (Gérald Bronner)

Comment nos esprits parviennent-ils à inférer autant à partir de si peu ? (Josh Tenenbaum)

Nous terminerons, en répondant à cette dernière question : peut-être tout simplement grâce… à la statistique !

NDLR : Les principaux événements ayant jalonné l'histoire de la statistique peuvent être visionnés directement sur le diaporama de la conférence des Journées de Metz, où le lecteur trouvera une frise historique résumant la chronologie des faits marquants (http://www.apmep. asso.fr/Les-conferences,4803).