Il faut pouvoir répondre à l'invasion des données

Gilbert Saporta

Gilbert Saporta, statisticien

“Il faut pouvoir répondre à l’invasion des données”

Sondages, économie, épidémiologie…, les statistiques abondent. Comment les interpréter et déjouer les biais ? En insistant sur la diversité, plaide ce chercheur.

GILBERT SAPORTA, 65 ans, est professeur titulaire de la Chaire de statistique appliquée au Conservatoire national des arts et métiers (Cnam). Il a publié *Probabilités, analyse des données et statistique* et est coauteur de *L’Analyse des données* (Que Sais-Je n° 1854). Il a été président de la Société française de statistique de 2000 à 2002, vice-président de l’Institut International de statistique de 2005 à 2007 ainsi que de l’International Association for Statistical Computing pendant la même période.

Qu’inspirent au statisticien que vous êtes les sondages qui se multiplient à l’approche de l’élection présidentielle ?
Il faut les interpréter avec circonspection. Ils portent au mieux sur 1000 personnes, plus souvent sur 800, et les instituts de sondage publient rarement les marges d’erreur *(voir Repères)*, car elles sont compliquées à calculer et à comprendre. L’un des problèmes fondamentaux est que les personnes qui répondent ne sont pas sincères. Elles sont notamment réticentes à avouer un vote extrémiste ou l’abstention. D’où la nécessité de « redresser » les résultats. Pour cela, on interroge les sondés sur leur vote précédent : si un candidat est ainsi crédité de 10 % de votes lors de la précédente élection – alors qu’il a atteint en réalité 15 % dans les urnes –, on augmente proportionnellement le poids de ses partisans pour l’élection à venir.

On dit que certains électeurs seraient influencés par les sondages. N’est-ce pas gênant lorsque l’observateur influe sur le résultat ?
On n’a jamais pu avoir de preuve que les sondages incitaient à voter pour le candidat en tête, ou au contraire décourageaient de voter pour lui. Les sondages sont un élément du choix, mais on ne sait pas dans quel sens ça marche.

Les échantillons choisis sont-ils suffisants et représentatifs ?
C’est un vieux débat. Trop souvent, on croit qu’un échantillon est représentatif si c’est une miniature de la société, respectant telle ou telle proportion. Or, pour le statisticien, l’important est de savoir si on peut extrapoler le résultat. Ainsi, on peut très bien surreprésenter certaines catégories, à condition de corriger ensuite. Le nombre ne fait pas tout, l’important est la qualité de l’échantillon, qui doit refléter la diversité de la population. Le parfait contre-exemple est ce qui s’est passé lors de l’élection présidentielle américaine de 1936, qui opposait Franklin Roosevelt à Alfred Landon. Un magazine avait consulté plus de 10 millions de personnes mais s’était trompé dans les résultats, tandis que de petits instituts de sondage, en interrogeant moins de personnes, mais mieux choisies, avaient vu juste. Il faut dire que le magazine avait pioché ses sondés dans la liste de ses abonnés ou dans l’annuaire. Il s’agissait donc de gens plus aisés que la moyenne, votant davantage républicain. La taille de l’échantillon n’est donc pas une garantie.

Il est aussi fait un grand usage des statistiques sur le chômage, la délinquance, les variations de salaires… Sont-elles biaisées ?
Tout dépend de quoi on parle, et par rapport à quoi. Autant de précisions qu’« oublient » de donner ceux qui font usage des résultats. Ainsi, en matière de salaire, on peut parler de salaire moyen, ou de salaire médian, ce qui n’est pas la même chose *(voir Repères)*. Le statisticien dit que le second est plus « robuste » que le premier, car il est moins influencé par les extrêmes. Par exemple, si on augmente les salaires des 10 % les plus riches, la médiane ne bouge pas, tandis que la moyenne augmente. Mais cet exemple montre aussi que la médiane ne suffit pas, puisqu’ici, elle ne rend pas compte de l’augmentation des inégalités due à l’augmentation des seuls hauts salaires. Il faut donc d’autres indicateurs, par exemple les déciles *(voir Repères)*. C’est l’un des points fondamentaux de la statistique : la moyenne ne suffit jamais. Ce qui importe, c’est la diversité, la variabilité… Pour ne pas être dupes des manipulations potentielles, il faudrait une formation aux statistiques dans les collèges et les lycées. Pour que l’on comprenne que ce ne sont pas x % des Français qui disent ceci ou cela, mais x % de l’échantillon, et qu’il y a une marge d’erreur.

Les statistiques sont aussi très utilisées en santé. Quelles applications en attendez-vous ?
Nous cherchons à détecter, par exemple, des phénomènes indésirables à partir de grandes bases de données. Ce serait très utile si l’on était capable de découvrir les dangers d’un médicament tel que le Mediator plus rapidement, rien qu’en étudiant les données fournies par les caisses d’assurance-maladie ou les hôpitaux ! Autre défi : trouver les associations de médicaments néfastes, sachant qu’il y a des millions d’associations possibles. Nous pouvons aussi mettre au jour des différences de risques sanitaires selon les populations, chercher les facteurs environnementaux ou génétiques, mettre en évidence des différences de pratiques médicales afin de comprendre, par exemple, pourquoi y a-t-il plus de césariennes dans un département plutôt que dans un autre… Tout ●●●

> « Le directeur général de l'Insee est parfois soumis à des pressions »

●●● cela concourt à une meilleure utilisation des moyens en santé publique. Nous progressons, car les données sont de plus en plus disponibles.

Vous intervenez également auprès d'industriels. Quel usage font-ils de vos travaux ?
Nous intervenons à plusieurs niveaux, pour les études de marché, mais aussi pour optimiser les procédés industriels. J'ai ainsi été consulté par une grande société industrielle pour la fabrication de coussinets d'allaitement afin d'aider à choisir les matériaux, à régler les machines (pression exercée, température…), à trouver la meilleure combinaison possible pour que l'absorption soit optimale. Or, on ne peut pas essayer toutes les combinaisons de facteurs ! C'est là que le statisticien intervient, afin de définir les expériences à mener pour chercher l'optimum. Ce qui est stimulant et passionnant avec la statistique, c'est qu'on travaille aussi bien avec des chimistes que des économistes, des sociologues, des informaticiens…

Dans le domaine très sensible du nucléaire, comment calculer la probabilité d'événements présentés comme très peu probables, mais aux conséquences énormes, comme un accident ?
Ces probabilités sont la plupart du temps estimées par la méthode des « arbres de défaillance » : pour avoir un accident, il faut telle cause, puis telle autre. On estime d'abord la probabilité de chaque cause, puis celle de l'ensemble. Malheureusement pour le statisticien, mais heureusement pour la population, on a peu d'observations de tels accidents pour faire les estimations.

Néanmoins, les trois cas avérés d'accidents nucléaires, Three Mile Island, Tchernobyl et Fukushima, ont été dus à des causes non prévues. Cela remet-il en question ces calculs de probabilité ?
Il reste des choses qu'on ne peut pas anticiper. Le statisticien travaille peu sur les scénarios eux-mêmes, mais plus sur la probabilité de scénarios dans le cadre de modèles définis surtout par les ingénieurs. On utilise les données recueillies pour affiner les modèles, éventuellement les récuser. Certains statisticiens travaillent sur l'estimation des risques extrêmes, en particulier pour les amplitudes des catastrophes naturelles. C'est plus facile : il y a moins de facteurs humains que dans une centrale nucléaire, et bien plus de données historiques, par exemple des relevés de crues depuis plusieurs siècles. On peut ainsi bâtir une loi de probabilité, et estimer la probabilité de crues millénaires. Ensuite, il faut valider la loi de probabilité choisie, en vérifiant que les calculs correspondent aux données historiques. Les statistiques s'appliquent au domaine réel, on est un peu dans la situation du physicien qui applique ses lois au monde réel.

Les statistiques ont été très décriées pour leur rôle dans la crise financière. Est-ce justifié ?
Beaucoup de mathématiques entrent en jeu dans les marchés financiers, notamment les modèles dits stochastiques, qui essaient de prévoir les variations de cours des actifs. Or, selon le modèle choisi, les prévisions changent beaucoup. S'ils fonctionnent, c'est uniquement parce que les différents opérateurs utilisent tous les mêmes modèles, et les marchés finissent donc par se comporter comme prévu. C'est le problème des prévisions « autoréalisatrices » : les modèles prédisent une remontée du marché, donc tout le monde achète, et donc le marché monte effectivement ! Jusqu'à ce que des événements non prévus détraquent ce beau système, les modèles créant des bulles financières. Ce n'est pas un échec des statistiques, c'est un échec des modèles. On a donné plus de rôle aux modèles qu'aux données. Or, le statisticien commence toujours par regarder les données.

Pour autant, comment s'assurer de la qualité des données ?
En s'appuyant en principe sur des données fiables, comme celles de l'Institut national de la statistique et des études économiques (Insee). Reste que parfois, le directeur général de l'Insee est soumis à des pressions. Beaucoup d'agents de l'Insee sont détachés dans les ministères, ils ont un code d'éthique. Du côté des ministères, les services statistiques sont coordonnés par le Conseil national de l'information statistique (Cnis), qui établit les programmes d'enquête. Le ministère de l'Intérieur n'y participe que marginalement et ne possède pas de service statistique unique, mais des cellules rattachées à diverses directions. Celle sur le thème de l'immigration et de l'intégration comporte en tout et pour tout 19 statisticiens, à comparer aux 68 du ministère de la Justice ou aux… 486 du ministère de l'Agriculture.

De quand datent les statistiques ?
Ce sont Blaise Pascal et Pierre de Fermat au XVII[e] siècle qui ont posé les bases de la théorie des probabilités. On peut citer ensuite Abraham de Moivre, Jacques et Daniel Bernoulli au XVIII[e] siècle, puis Siméon Denis Poisson au début du XIX[e] siècle. Mais probabilité n'est pas statistiques : celles-ci commencent à partir du moment où on recueille des données pour les analyser. Au début du XIX[e] siècle, Pierre-Simon de Laplace est ainsi l'un des premiers à vouloir tirer des échantillons de communes pour ne pas être obligé de faire des recensements dans chacune d'entre elles, et il avait trouvé des formules pour calculer les marges d'erreurs. Mais cela a été abandonné. Il faudra attendre le début du XX[e] siècle pour que la statistique émerge. On travaillait alors sur des petits nombres : plus de 30 observations étaient considérées comme un grand échantillon. Aujourd'hui, cela fait rire mes étudiants ! Puis, dans les années 1960, la révolution informatique a permis d'effectuer des calculs sur des milliers d'observations. On a enfin pu faire les calculs que les mathématiciens avaient imaginés.

Les statistiques évoluent-elles sous l'impulsion des progrès mathématiques ou des demandes de la société ?
Les deux, mais on est surtout tirés par la demande. Actuellement, l'enjeu peut même paraître paradoxal : face au déluge de données, on essaie d'inventer des méthodes pour répondre à l'invasion ! Comment peut-on les analyser dans un temps très court sans les stocker ? Comment traiter par exemple toutes les données relevées *via* les téléphones portables (géolocalisation…), qui arrivent par milliards tous les jours ? De même, les ingénieurs d'ERDF travaillent actuellement sur les meilleures façons d'analyser les données qui seront fournies par les futurs « compteurs intelligents » d'électricité Linky. Il faut en effet trouver de nouveaux algorithmes pour traiter sans stocker, mais sans non plus oublier le passé.

Les statisticiens ont-ils un rôle citoyen à jouer lorsque des statistiques sont déformées ?
Récemment, le ministre de l'Intérieur Claude Guéant a mentionné dans les médias des chiffres sur l'échec scolaire des enfants d'immigrés, qui ont été démentis par le directeur général de l'Insee, sous la pression de ses syndicats. C'est aussi le rôle des organisations telles que la Société française de statistique (SFdS) que d'intervenir dans le débat. Il existe aussi des instances internationales, par exemple l'Institut international de statistique, qui a réagi lorsque le gouvernement argentin a voulu manipuler l'indice des prix, lorsque le gouvernement canadien a tenté de réduire drastiquement le contenu du questionnaire de recensement. En Grande-Bretagne, la Royal Statistical Society avait mené une campagne contre la dégradation de la confiance dans les statistiques publiques, à la suite des coupes budgétaires du gouvernement Thatcher. Ces instances sont souvent prudentes et lentes à réagir, mais elles sont généralement bien entendues.

Propos recueillis par Cécile Michaut
Photos : Céline Anaya Gautier pour Sciences et Avenir

REPÈRES

MARGE D'ERREUR : lorsqu'on interroge un échantillon de n personnes pour estimer l'opinion d'un bien plus grand nombre de gens, le résultat n'est pas exact. Par exemple, ce ne sont jamais strictement *« 15 % des Français qui… »*, mais *« entre 13 et 17 % »* comme on devrait l'écrire : la marge d'erreur est ici de plus ou moins 2 %. On doit en outre préciser le niveau de confiance de cette marge d'erreur. Dans l'exemple précédent, un niveau de confiance de 95 % signifie qu'il y a 5 % de chances que le résultat soit au-dessous de 13 % ou au-dessus de 17 %. La marge d'erreur décroît en fonction de n, mais assez lentement car elle est inversement proportionnelle à la racine carrée de n. L'exemple numérique précédent concernait un sondage simple de 1000 unités.

MOYEN : pour faire une moyenne, on additionne la somme de toutes les données, et on divise par le nombre de données. Ainsi, le salaire moyen est la somme de tous les salaires, divisée par le nombre de salariés.

MÉDIAN : la médiane est la valeur qui sépare un échantillon en deux parties égales. Par exemple, le salaire médian est le salaire au-dessous duquel se situent 50 % des salaires. Autrement dit, la moitié des salariés gagnent plus que ce salaire, l'autre moitié gagne moins.

DÉCILE : lorsqu'on veut faire des statistiques plus précises, on divise la population en dix parties égales. Toujours dans l'exemple des salaires, le premier décile correspond au salaire maximum que touchent les 10 % de gens les moins payés, le dixième au salaire minimum des 10 % les mieux payés, et on mesure les inégalités en regardant l'écart.

BERNARD MARTINEZ

MA BIBLIOTHÈQUE ÉGOÏSTE

« Une interrogation sur l'existence même des statistiques »

« Les ouvrages qui m'ont marqué sont des livres pointus dans mon domaine. Tout d'abord, une véritable encyclopédie en trois volumes *The Advanced Theory of Statistics* , de Maurice George Kendall et Alan Stuart, parue en 1966. Grâce à elle, j'ai enfin compris les statistiques ! En quelques lignes, tout était expliqué. J'aime beaucoup le livre de Bruno de Finetti, *Theory of Probability*, qui commence par la phrase *« La probabilité n'existe pas. »* On a trop souvent l'impression que la probabilité est quelque chose d'objectif. Or, la probabilité dépend à la fois des circonstances et de l'observateur. Bruno de Finetti a ensuite développé toute une théorie sur la probabilité subjective. Trop souvent, les livres de probabilité sont des livres de maths, où on ne se pose même pas la question de l'existence des probabilités. Enfin, le meilleur livre depuis vingt ans, *The Elements of Statistical Learning*, de Trevor Hastie, Robert Tibshirani et Jerome Friedman, couvre l'état actuel de la statistique. C'est le livre que j'aurais aimé écrire. »

* www.sciencesetavenir.fr/sante/20120117.OBS9003/decryptage-y-a-t-il-un-exces-de-leucemies-autour-des-centrales-nucleaires.html