Noureddine El Karoui est professeur associé de statistiques à l'Université de Californie à Berkeley. Formé à l'École polytechnique, il a soutenu sa thèse en statistiques à l'Université Stanford.

Ses recherches sont à l'intersection entre la théorie des probabilités et les statistiques. Il travaille plus précisément sur la théorie des matrices aléatoires et ses applications aux statistiques et à l'optimisation, sur les statistiques de grande dimension, et sur les méthodes de ré-échantillonnage et d'inférence statistique de type *bootstrap*.

Il est membre de l'Institut des statistiques mathématiques des Etats-Unis.

Publication sous la direction de Jean-Philippe Touffut
Illustration : NiboR / Mise en page : Sophie Otrage
Impression : Escandre-Sorel

Noureddine El Karoui

Les statistiques peuvent-elles se passer d'une théorie des probabilités ?

Prisme N°30
Octobre 2015

La Fondation et le Centre Cournot

Les statistiques peuvent-elles se passer d'une théorie des probabilités ?[1]

Noureddine El Karoui

Prisme N°30

Octobre 2015

[1] Ce texte a été traduit par Nathalie Ferron à partir de la version en anglais, publiée en décembre 2014 dans la Série *Prisme*.

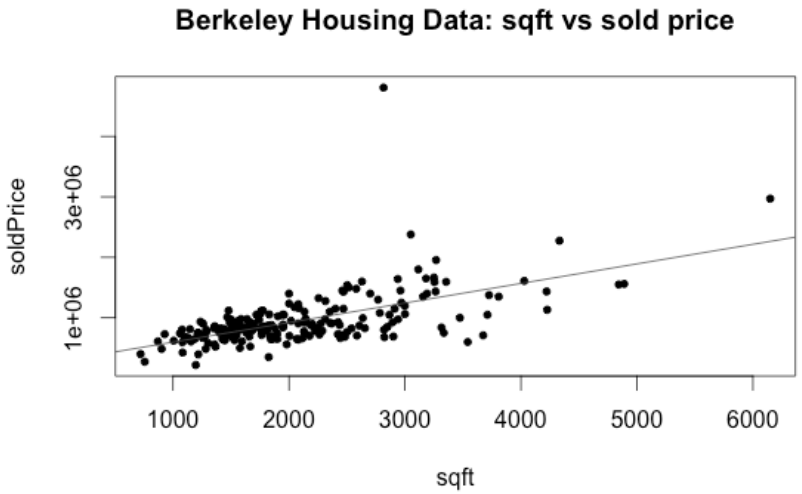
Résumé

Élaboré à partir de l'expérience quotidienne et de la pratique professionnelle d'un enseignant spécialiste des statistiques, cet article a pour but de montrer en quoi le domaine des statistiques a connu une évolution constante, en particulier dans ses rapports avec la théorie des probabilités. Deux exemples permettent d'illustrer ce propos : dans le premier, la question porte sur la possibilité, à partir de l'observation de données relatives au marché de l'immobilier, de prévoir le prix de vente d'une maison à partir de ses caractéristiques ; dans le second, l'enjeu concerne la conception d'un filtre anti-pourriel pour un compte de messagerie électronique. Nous proposons une analyse des diverses techniques statistiques, des plus classiques aux plus contemporaines, qui peuvent être utilisées pour l'analyse de ces données et en particulier, nous étudions le rôle de la théorie des probabilités dans le développement et l'utilisation de ces techniques. L'analyse fait apparaître l'évolution des rapports entre la théorie des probabilités et les statistiques.

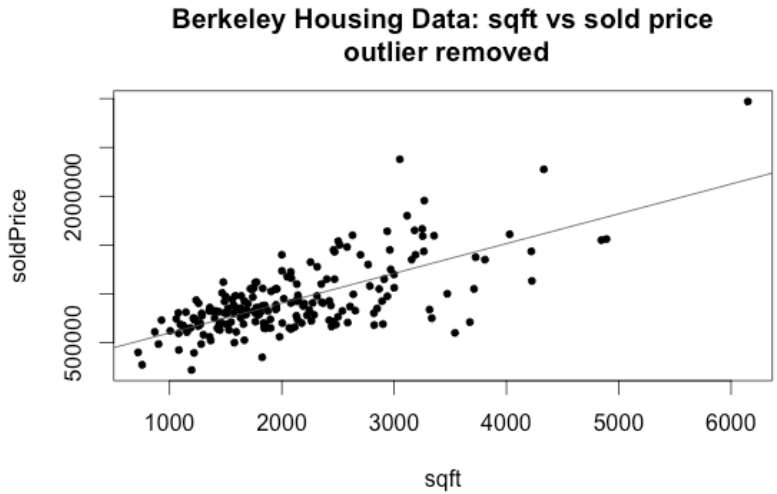
Élaboré à partir de l'expérience quotidienne et de la pratique professionnelle d'un enseignant spécialiste des statistiques, le présent article vise à mettre en évidence le caractère constamment évolutif du domaine des statistiques et en particulier de son rapport à la théorie des probabilités. Avant d'aller plus loin, il me faut préciser que ce texte a fait l'objet d'une communication lors du colloque organisé par le Centre Cournot[2] sur le recours croissant aux idées probabilistes dans de nombreux domaines scientifiques. Pour commencer par une définition des statistiques et de leur but, je propose une formulation spontanée et non technique : les statistiques consistent à interpréter, à utiliser et à exploiter des données de manière raisonnée. Deux exemples illustrent ce propos : premièrement, à partir de l'observation de données relatives au marché de l'immobilier, nous nous demandons s'il est possible de prévoir le prix de vente à venir d'une maison en fonction des caractéristiques de celle-ci. En second lieu, nous étudions la possibilité de concevoir un filtre anti-pourriel pour les comptes de messagerie électronique. Ces exemples ont été retenus pour leur accessibilité à un public non-spécialiste. On aurait pu pousser beaucoup plus loin l'exploitation statistique des données observées, mais notre but ici est de proposer une analyse des diverses techniques statistiques qui pourraient être utilisées pour l'analyse de données, en partant des plus classiques pour aller vers des techniques plus modernes, et en particulier de mettre en évidence le rôle de la théorie des probabilités dans le développement et l'utilisation de ces techniques. L'analyse fait ainsi apparaître l'évolution des rapports entre la théorie des probabilités et les statistiques.

Commençons par l'exemple de l'immobilier. En nous servant de *www.redfin.com*, un site de courtage en ligne qui collecte des informations sur les ventes immobilières aux États-Unis, on extrait des données concernant le marché à Berkeley, exactement comme si nous cherchions à acheter une maison.

[2] « Y a-t-il des limites à la probabilisation des sciences ? » Colloque organisé par la Fondation Cournot et l'université de Harvard, Harvard Medical School. Les conférences sont en ligne à l'adresse suivante : http://www.centre-cournot.org/conferences.



Les données présentées ci-dessus datent de septembre 2013. *Redfin* donne quantité d'informations sur chaque maison : prix de vente, vente à découvert (ou non), type de bien, adresse, ville, État, code postal, prix demandé par l'acheteur, superficie du bien, superficie du lot, nombre de salles de bain, etc. Pour chaque bien, environ trente-trois caractéristiques sont répertoriées. Si un nouveau bien intéressant se présente sur le marché, on a envie de savoir quel type d'offre semble appropriée. Sur le graphique ci-dessus, on voit un bien dont le prix est significativement plus élevé que celui de tous les autres. Après avoir consulté la liste des biens vendus, nous avons pu constater que la maison s'était probablement vendue à environ $500 000, et non pas $5 millions. Il s'agit sans doute d'une erreur de report, aussi avons-nous préféré exclure cette donnée de notre base, ce qui nous donne le graphique suivant :



On trace la droite correspondant aux données. Si l'on exclut le cas aberrant, on constate qu'en moyenne, une maison à Berkeley coûte environ 3032€/m^2, prix auquel il faut ajouter un prix de base de $300,000 (275.000€). À partir de cette information, on peut prévoir le prix de vente de biens de superficies diverses (celui par exemple de maisons d'une superficie allant de 165 à 195m^2). Pour cette fourchette de superficies, les biens coûtent entre $850,000 (780.000€) et $950,000 (870.000€). Nous voici en possession de certaines informations, mais nous ne savons pas dans quelle mesure notre prédiction se révèle exacte. En revanche, nous savons que le prix moyen d'une maison de 165 m^2 située à Berkeley devrait se situer entre $843,000 (770.000€) et $850,000 (780.000€). Cependant, nous pourrions chercher à avoir une idée du prix d'une maison en particulier, auquel cas, nous calculerons ce qu'on appelle un intervalle de prévision. Même si en moyenne une maison de 165m^2 coûte environ 780.000€, si l'on cherche à affiner la prévision, on constate qu'en réalité, la fourchette de prix est comprise entre $328,000 (300.000€) et $1,4 million (1.280.000€). Voilà qui ôte toute valeur à notre prédiction : la fourchette de prix est beaucoup trop étendue pour constituer une information utile. La règle à retenir ici est la suivante : en moyenne, le prix d'un bien se calcule à partir d'un prix de base auquel on ajoute une somme proportionnelle à sa superficie. Nous disposions au départ de 33 indicateurs que nous avons tous écartés, à l'exception d'un seul. Les

propriétés intrinsèques de chaque bien ou leur prix de vente entrent ainsi dans le modèle à titre « d'erreurs » aléatoires.

Dans le graphique reproduit ci-dessus, la droite correspond au prix moyen d'une maison, or on voit que tous les prix se répartissent autour de cette droite. La prise en compte du caractère stochastique de ces erreurs nous permet de prévoir le prix de vente moyen d'un bien d'une superficie donnée et de trouver la droite figurant dans ce graphique. Surtout, et c'est là que réside toute la force de l'argument stochastique, cela nous donne une idée de l'exactitude de notre mesure, laquelle nous a permis de calculer notre intervalle de prévision : en moyenne, le prix d'une maison de 165m^2 est de 780 000€, mais en réalité il peut prendre n'importe quelle valeur entre 320 000€ et 1 300 000€.

L'intervalle des prix qui apparaît dans ce graphique devrait comprendre 95% des prix de vente des biens de superficie similaire arrivant sur le marché. Cet indicateur de précision révèle que la seule connaissance du prix moyen d'un bien doté de caractéristiques données nous renseigne très peu sur les biens pris individuellement. La connaissance de l'exactitude de notre prévision nous renseigne sur sa validité, laquelle est, en l'occurrence, dénuée de valeur car nous n'avons pu en tirer que très peu d'informations. Afin de mettre en place cet indicateur d'exactitude, un certain nombre d'hypothèses probabilistes ont été formulées et entrées dans le logiciel de statistiques, ***R***, qui nous a donné le résultat communiqué plus haut. Pour l'utilisateur ignorant des hypothèses retenues, ***R*** se comporte un peu à la manière d'une boîte noire.

Si l'on regarde les données de plus près, ces hypothèses ont toutes les chances de ne pas être respectées, en sorte qu'une plus grande prudence s'impose. Techniquement, cet indicateur d'exactitude est facile à obtenir, mais du point de vue conceptuel, il exige le recours à un raisonnement probabiliste poussé. Nous faisons l'hypothèse que le prix moyen des biens de superficie similaire se trouve sur la droite (cf. graphique ci-dessus) et que le prix de chaque bien fluctue verticalement de part et d'autre de la droite. Pour comprendre les propriétés probabilistes de la droite correspondant aux données, il faut recourir à la théorie probabiliste de la limite. L'intérêt fondamental de la théorie des probabilités consiste en l'occurrence à nous fournir un indicateur du degré d'exactitude sans lequel toute mesure statistique est

dénuée de valeur. La théorie des probabilités et les statistiques sont par conséquent très étroitement liées.

On pourrait néanmoins s'appuyer sur bien d'autres techniques pour analyser ces données. Au lieu de tracer une droite à partir des données, on pourrait s'intéresser au bien qui, au sein d'un ensemble de données, correspond le mieux à celui qui arrive sur le marché et a retenu notre attention. L'intuition est que le prix du bien arrivant sur le marché est similaire à celui du bien qui lui est le plus proche dans tel ensemble de données. Il s'agit là d'un autre type de prévision que l'on appelle « méthode des plus proches voisins ». On pourrait aussi utiliser diverses moyennes pondérées. Par exemple, on peut signaler que la maison qui nous intéresse ressemble à une maison particulière, mais qu'elle n'est pas non plus très différente des autres maisons du quartier. Dans ce cas, on peut calculer la moyenne des prix de vente de ces biens afin de prévoir le prix de vente de celle qui arrive sur le marché. Le coefficient affecté à chaque bien est proportionnel à sa ressemblance avec le bien qui nous intéresse. Une fois que l'on est en possession des données, on peut appliquer diverses méthodes. Comment choisir ? Historiquement, l'analyse probabiliste a servi à évaluer ***a priori*** les performances de diverses méthodes[3] et surtout, à concevoir des méthodes ***optimales*** en fonction des hypothèses retenues dans le cadre du modèle utilisé, « optimale » renvoyant, pour l'analyste statisticien, à « un degré d'exactitude élevé ».

Dans ce cadre, les méthodes retenues pour procéder à l'analyse des données dépendent fortement des hypothèses probabilistes formulées à propos du mécanisme générateur de données. Dans le second graphique, on forme l'hypothèse que la droite représente avec exactitude le fait que, en moyenne, le prix de vente d'un bien est proportionnel à la superficie ajoutée au prix de base, sachant que, en fonction du type d'erreurs aléatoires, nous nous orientons vers telle ou telle méthode en vue de nous approcher de la droite. Plusieurs méthodes permettent de tracer la droite correspondant aux données. Si les erreurs sont normalement et indépendamment distribuées, il paraît logique de recourir à la méthode des moindres carrés, en

[3] Voir par exemple Lehmann, Erich L. (2008), *Reminiscences of a Statistician: The Company I Kept,* New York : Springer-Verlag ; and Lehmann, Erich L. (2011), *Fisher, Neyman and the Creation of Classical Statistics,* New York : Springer-Verlag.

d'autres termes, à trouver la droite qui minimise la somme des carrés des distances verticales des points à la droite correspondant aux données. Si en revanche les erreurs se répartissent selon une distribution exponentielle double, il paraît logique de choisir la droite qui minimise la somme des valeurs absolues des résidus. Il est ainsi possible de tracer plusieurs droites à partir des mêmes données. Pour trouver une méthode permettant de choisir la bonne droite, il faut avoir une bonne compréhension des caractéristiques stochastiques des erreurs ou bien trouver le modèle de probabilité correspondant aux données. Pour revenir à notre problème initial, le recours à une méthode appropriée permet de réduire les erreurs d'estimation et d'obtenir, dans une certaine mesure, la prévision optimale.

C'est ce qui sous-tend les méthodes du maximum de vraisemblance, idée primordiale et désormais déjà ancienne. Les premières formalisations de ces méthodes remontent sans doute aux années 1920 et 30 avec les travaux de R.A. Fisher[4]. Elles reposent sur l'exploitation du mécanisme aléatoire que nous supposons à l'origine des données. Revenons à présent aux données relatives au marché immobilier et prenons, par exemple, le prix de vente d'une maison de 165m^2. Supposons qu'il s'agisse d'une variable aléatoire dont la moyenne, inconnue, est notée m, et que le prix de vente des autres biens de superficie similaire fluctue autour de cette moyenne. Ces fluctuations reflètent le fait que certaines maisons de 165m^2 possèdent un plus grand nombre de salles de bain que d'autres, qu'elles ont été rénovées récemment, etc. Ainsi, les données sont créées par un mécanisme aléatoire. Ce mécanisme aléatoire peut être désigné par m. Cette valeur, en l'occurrence le prix moyen d'une maison de 165m^2, est inconnu, mais nous cherchons à connaître sa valeur. Si on connaissait les caractéristiques probabilistes des fluctuations du prix de vente des autres maisons de 165m^2 proches de m, autrement dit la distribution normale des erreurs, m spécifierait dans sa totalité la distribution et les caractéristiques probabilistes des données.

On pourrait également supposer que la distribution normale des erreurs est inconnue et tenter de l'estimer à partir des données. Ce serait plus compliqué, mais pas insurmontable. De même, il n'est pas nécessaire que le paramètre d'indexation

[4] Fisher, R.A. (1922),"On the Mathematical Foundations of Theoretical Statistics", *Philosophical Transactions of the Royal Society of London,* Series A, 222 : pp. 309–368.

du processus générateur de données soit un seul nombre, il peut avoir deux dimensions ou plus. L'ensemble des mécanismes aléatoires possibles ayant engendré des données a pour index un paramètre que l'on nomme T. Si l'on suppose que la superficie change, pour déterminer le prix d'un bien, on représente ainsi le prix moyen par une droite, ce qui indique que le prix moyen est une fonction (linéaire) de la superficie. S'y ajoute une erreur aléatoire. Fondamentalement, le paramètre T est à la fois le coefficient directeur et l'ordonnée à l'origine de cette droite, en sorte qu'il est bidimensionnel. En supposant que les caractéristiques probabilistes des erreurs nous soient connues, si on connaît le paramètre T, on peut calculer la probabilité qu'on ait observé la base de données utilisée[5]. On peut le faire pour toutes les valeurs possibles de T, l'idée étant que, en fait, T est très certainement T_{EMV}, lequel maximise la probabilité de trouver une base de données conforme à celle que nous observons. Dans la mesure où il dépend de la base de données observée, laquelle résulte par hypothèse d'un mécanisme aléatoire, T_{EMV} est fondamentalement un objet probabiliste ou aléatoire. Ce qui signifie qu'avec une base de données différente, T_{EMV} prend une valeur différente. Curieusement, il est possible de caractériser les propriétés probabilistes de T_{EMV}, autrement dit sa loi de probabilité, à partir d'une unique base de données. On peut ainsi prévoir ***a priori*** l'évolution de T_{EMV} en cas de changement de base de données engendrée par le même mécanisme probabiliste, ainsi que dans le cas où l'expérience serait répétée à l'infini.

Cette caractérisation est assez propre à la théorie des probabilités classique. Elle est possible à condition que la base de données soit de grande dimension, c'est-à-dire que le nombre d'observations soit très important. Pour ce qui concerne nos données immobilières, on peut par exemple fournir des informations probabilistes concernant l'estimation du coefficient directeur de la droite qui nous intéresse, à condition que le nombre de biens observés soit infini. La question qui se pose habituellement concerne la taille de la base de données. À quel moment l'approximation devient-elle exacte ? Le marché immobilier évolue sans cesse, il n'est donc pas possible de s'appuyer sur des données datant de 50 ans pour tracer notre droite. Il existe donc une limite temporelle qui, pour notre base de données, implique

[5] Par souci de clarté, j'ai simplifié cette phrase qui en conséquence rend compte de l'idée mais laisse de côté les définitions mathématiques d'usage.

que nous n'obtenions que 200 points. La question qui se pose est donc la suivante : notre approximation est-elle correcte pour ces 200 points. Nous donne-t-elle des informations exactes ?

Dans ce qui suit, nous acceptons les limites de la méthode, qui nous contraint à avoir un modèle représentant le processus de génération des données, ainsi que celles de la théorie, qui nous contraint à avoir un échantillon de grande taille (une base de données importante comportant de nombreuses observations). Il est néanmoins remarquable dans ce contexte que nous puissions nous servir des propriétés probabilistes de notre estimation pour donner à la fois une estimation de T et le degré d'exactitude de cette estimation. Ce qui importe, c'est de connaître ce degré d'exactitude, lequel est lié aux fluctuations de l'estimation en question lorsque l'on répète l'expérience un grand nombre de fois. T_{EMV} – l'estimateur du maximum de vraisemblance (EMV) – est l'estimateur le plus exact de T, pour autant qu'on ait un grand nombre d'observations.[6] Ainsi, non seulement T_{EMV} est un estimateur qui a fait ses preuves, mais il est en outre optimal. Depuis 1920, il nous permet d'obtenir de bons résultats grâce à une méthode d'analyse de données et de statistiques simple qui consiste à collecter des données et éventuellement à concevoir une expérience d'avance, à leur trouver un bon modèle probabiliste, et enfin à ajuster ce modèle aux données observées à l'aide de la méthode du maximum de vraisemblance. Si le modèle est bon, il doit nous fournir une méthode quasiment optimale pour extraire de l'information à partir des données. C'est ce qui a assuré sa place et son succès à la théorie des probabilités en science : si l'on dispose d'un bon modèle probabiliste, on dispose d'un estimateur quasi-optimal pour évaluer ses paramètres.

Vers la fin des années 1940 ou au début des années 1950, Joseph Hodges, professeur à Berkeley) a inventé un « estimateur super-efficace »[7] qui surpassait l'estimateur du maximum de vraisemblance (EMV) en terme d'exactitude sur quelques points et faisait aussi bien partout ailleurs, sous l'hypothèse d'une taille de la base de données tendant à l'infini. L'EMV est fondamentalement optimal mais

[6] Ceci est vrai sous diverses conditions techniques liées au mécanisme générateur de données, mais il existe des contre-exemples.

[7] Voir Le Cam, L. (1953) : "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates", *University of California Publications in Statistics,* 1, pp. 277–330.

Hodges se rendit compte qu'on pouvait trouver mieux pour quelques valeurs dans l'espace des paramètres. Il ne s'agissait que d'un contre-exemple, mais ce fut une grande découverte dans la mesure où elle mit en évidence l'existence de problèmes conceptuels et théoriques dans les méthodes du maximum de vraisemblance.

De manière connexe, en 1956[8], Charles Stein découvrit un phénomène statistique étrange : si le paramètre est assez complexe – d'une dimension supérieure à trois – dans un contexte probabiliste aussi simple que possible pour le mécanisme générateur de données, on peut construire un estimateur plus exact que l'EMV pour toutes les valeurs du paramètre qui nous intéresse réellement, et pour une base de données de taille fixe. Tandis que l'ancienne notion d'exactitude était liée à l'amplitude des fluctuations de l'estimateur, la notion steinienne d'exactitude consistait fondamentalement à donner la mesure moyenne de cette fluctuation.

Les méthodes du maximum de vraisemblance sont largement utilisées dans divers domaines scientifiques, mais les statisticiens ainsi que les spécialistes des théories statistiques savent pour leur part que malgré ses propriétés d'optimalité, cette méthode a ses limites. La théorie des statistiques a suscité le développement de nombreux outils probabilistes spécifiques dans la mesure où ses spécialistes travaillent sur des objets probabilistes qui ne sont peut-être pas exactement ceux de la théorie générale des probabilités. Ceci a donné lieu à un dialogue fécond entre théoriciens des probabilités et des statistiques : en leur point de rencontre, ils forgent ensemble les outils nécessaires à la compréhension du comportement d'une classe étendue d'estimateurs et de méthodes analytiques.

Reprenons à présent notre base de données d'un point de vue pratique. Comme nous venons de le montrer, performance et optimalité dépendent étroitement du modèle. On suppose en général que notre modèle est bon, mais qu'en est-il si tel n'est pas le cas ? Comment vérifier la fiabilité d'un modèle ? Pour revenir à l'exemple du marché immobilier, au moment de tracer la droite correspondant aux données relatives aux biens, on pourrait s'interroger sur le fait que la dispersion semble augmenter avec la superficie. En effet, on pourrait s'attendre à ce que la variabilité

[8] Stein, C. (1956) : "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Distribution", *Actes du colloque: Third Berkeley Symp. Math. Statistics,* 1, University of California Press : pp. 197–206.

des prix évolue en fonction du segment de marché observé. La « boîte noire » qu'on a utilisée (en utilisant le logiciel *R* de manière quelque peu naïve) ne suppose pas cependant que la variabilité des prix évolue en fonction de la superficie des biens. En fait, le degré d'exactitude de notre estimateur repose sur l'hypothèse que la variabilité des prix est la même pour tous les segments du marché immobilier observé, aussi peut-il nous induire en erreur en ce qui concerne la base de données considérée. En effet, pour qu'une analyse de données soit correcte, il faut savoir exactement quel modèle convient, avoir conscience des limites de ce modèle et être en mesure d'adapter les indicateurs intéressants lorsque les données semblent ne pas correspondre au modèle implicite du logiciel utilisé.

Un autre problème se pose : y a-t-il un moment où le modèle fournit une description exacte des données ? Pour ce qui concerne le modèle très simple que nous avons retenu pour représenter le marché immobilier, il est bien entendu que nous ne pensons pas vraiment que le prix de vente moyen d'un bien correspond à un prix de base augmenté d'une somme proportionnelle à sa superficie. Seule une fonction plus complexe pourrait décrire son comportement. Il s'agit d'un problème de spécification du modèle. Si nous ne nous fions pas au modèle, comment nous fier à l'évaluation de l'exactitude qu'il nous fournit ? Rappelons-nous que ce qui nous intéresse ici, c'est la possibilité d'obtenir non seulement une estimation, mais une estimation de l'exactitude de notre estimation. Or si nous ne nous fions pas au modèle, pourquoi nous fier à l'indicateur d'exactitude ? Par ailleurs, la théorie asymptotique – désormais pleinement intégrée au logiciel – requiert une base de données de grande dimension. La nôtre comporte 200 observations. Est-ce suffisant ? Nous en faudrait-il plutôt 1000, voire 10 milliards ? Quel est le rapport avec la complexité du paramètre T ? Par certains aspects, ces problèmes se rattachent à la théorie des matrices aléatoires (qui a été évoquée dans d'autres communications) : certains travaux issus de la recherche en statistique[9] mettent en évidence le fait que les propriétés des estimateurs du maximum de vraisemblance sont très différents selon que T est très complexe ou de grande dimension ou qu'il suit un schéma classique de dimension

[9] Voir, par exemple, El Karoui, Noureddine, Derek Bean, Peter J. Bickel, Chinghway Lim & Bin Yu (2013), "On Robust Regression with High-Dimensional Predictors", *Actes du colloque de la National Academy of Sciences,* http://www.pnas.org/content/early/2013/08/15/1307842110.full.pdf+html.

réduite. Il peut en fait se révéler fortement sous-optimal et receler d'autres propriétés indésirables.

Une grande réussite des statistiques a été de fournir à la fois des méthodes pour mesurer des quantités ainsi que l'indicateur permettant d'évaluer l'exactitude des estimations obtenues. Les sceptiques (dont l'auteur fait partie) affirment qu'ils ne se fient pas au modèle et encore moins à la théorie. Aussi se posent-ils la question suivante : est-il possible de réaliser ce programme d'estimateurs de calcul et d'évaluer leur exactitude sans avoir recours à des modèles qui ne sont pas vraiment fiables (et qui par certains aspects ne correspondent pas aux données) et sur une théorie présupposant une base de données fondamentalement infinie ? Ces objections viennent de spécialistes compétents et, en général, très expérimentés, aussi méritent-elles qu'on en tienne compte.

À la fin des années 1970, une approche plus moderne a été adoptée : au lieu de s'appuyer sur la théorie des probabilités et les calculs sans fin, les statisticiens se sont dit qu'ils pourraient faire un usage plus intensif des bases de données. En 1979, Bradley Efron introduisit une méthode efficace et désormais largement utilisée – le ***bootstrap***.[10] La méthode consiste à créer un grand nombre de nouvelles bases de données de manière aléatoire à partir d'une base de données initiale. On peut observer le comportement probabiliste de notre méthode d'estimation en observant la façon dont les résultats obtenus par cette méthode fluctuent au fil des nouvelles bases de données aléatoires qui ont été créés. Par la suite, on espère, et dans le meilleur des cas on prouve, que les fluctuations constatées au fil des bases de données nouvellement créées sont les mêmes que celles qu'on obtiendrait si on pouvait répéter à l'infini l'expérience initiale à l'origine de la première base de données (ce qui serait très compliqué pour la base de données relatives au marché immobilier, par exemple). Le recours au bootstrap, cependant, nous donne une idée assez précise des fluctuations de nos estimateurs sans qu'on ait besoin d'une base de données tendant à l'infini. Dans la mesure où les nouvelles bases de données ont été créées à partir d'une base initiale, la méthode devrait pouvoir tenir compte de cette complexité de manière automatique et se révéler moins dépendante des hypothèses excessivement simplifiées parfois nécessaires à l'application d'une théorie existante.

[10] Efron, Bradley (1979), "Bootstrap Methods: Another Look at the Jackknife", *The Annals of Statistics,* 7 (1): pp. 1–26.

Appliquons cette méthode à notre base de données relatives au marché immobilier : on a $n = 204$ biens, après exclusion du cas aberrant. On prend un nombre au hasard entre 1 et 204, et on répète l'opération 204 fois. À partir des 204 points initiaux, on a créé 204 nombres pris au hasard entre 1 et 204. On appelle cet ensemble I_1, c'est notre ensemble d'indices. I_1 peut comporter plusieurs fois le même nombre. À partir de la base de données initiale, on crée ensuite une nouvelle base qui ne comporte que les points indexés par les nombres figurant dans I_1. On crée ainsi une nouvelle base de données comprenant les observations initiales affectées des indices provenant de I_1. Si $I_1 = (10,3,5,3...)$, la nouvelle base de données a pour configuration ($maison_{10,}$ $maison_{3,}$ $maison_{5,}$ $maison_{3}$…). On applique ensuite notre méthode d'estimation à cette nouvelle base de données et obtenons un nouveau résultat T_1^*. Nous pouvons reproduire cette démarche 1000 fois, obtenir 1000 ensembles d'indices ($I_1,I_2...I_{1000}$), et 1000 nouvelles estimations correspondantes ($T_1^*,T_2^*...T_{1000}^*$), et par conséquent, 1000 droites différentes qui nous indiqueront en quoi la droite que nous avons choisie fluctue en fonction des variations de la base de données. On espère, et on peut parfois le prouver, que les caractéristiques probabilistes de ($T_1^*,T_2^*...T_{1000}^*$) sont similaires à celles de ($T_1,T_2...T_{1000}$), que nous aurions pu calculer si nous avions été en mesure de créer 1000 nouvelles bases de données par l'intermédiaire du mécanisme générateur de données (inconnu) à l'origine de notre base initiale. ($T_1,T_2...T_{1000}$) nous renseignerait sur les caractéristiques probabilistes de T_1 et pourrait par conséquent nous permettre de répondre à la question que nous nous sommes posée à propos de sa variabilité et de son exactitude, même si on ne peut en général le calculer dans la mesure où nous ne prétendons pas connaître parfaitement le mécanisme générateur de données. En revanche, la version bootstrap ($T_1^*,T_2^*...T_{1000}^*$) se calcule aisément et peut être utilisée à la place de ($T_1,T_2...T_{1000}$), qui est incalculable.

Le bootstrap soulève de brûlantes questions que formulent parfois les étudiants en statistiques nouvellement affranchis dans la matière : si le bootstrap marche, avons-nous encore besoin de la théorie des probabilités pour faire des statistiques ? Avons-nous besoin d'une théorie asymptotique ? Avons-nous besoin du théorème limite central, fondamental dans la théorie des probabilités, ou même de tout autre théorème propre à la théorie des probabilités, si on peut voir ou constater le même résultat à de multiples reprises grâce aux simulations numériques ? Le

bootstrap fonctionne de fait dans de nombreuses situations, cependant la théorie demeure en général asymptotique : on sait qu'il fonctionne si la taille de l'échantillon est assez grande et si les fonctions des données observées sont relativement « régulières » ou « gentilles » en un sens technique bien précis. On a alors affaire à une méthode non-paramétrique : il n'est pas nécessaire de spécifier le modèle probabiliste à l'origine des données, mais il nous faut tout de même supposer que les observations sont indépendantes, etc. En ce qui concerne le traitement de problèmes contemporains où les paramètres estimés sont de grande dimension, le bootstrap s'est révélé problématique et même défaillant en bien des occasions, même pour des fonctions très « régulières » ou des configurations statistiques très simples[11].

Dans les procédures centrées sur les données, l'étape suivante consiste à créer des méthodes non adossées à un modèle probabiliste comportant un estimateur d'exactitude intégré. Prenons l'exemple d'une base de données de pourriels. Lorsqu'un courriel arrive, on veut savoir s'il s'agit d'un pourriel. On cherche naturellement à utiliser une base de données comportant diverses caractéristiques propres aux courriels afin de savoir si chaque message entrant est un pourriel ou non. Pour résoudre ce problème, on peut essayer de concevoir un modèle probabiliste fondé sur les caractéristiques des courriels déjà arrivés, en essayant de déterminer ce qui fait qu'un message est un pourriel, par exemple en ayant recours à la méthode dite de régression logistique (option 1). On peut également appliquer une ou plusieurs méthodes, ou trouver un moyen d'agréger les méthodes (option 2) en vue d'exploiter notre base de données et faciliter le processus décisionnel.

Dans la seconde option, l'exactitude est mesurée de façon globale, par exemple en fonction de la fraction d'erreurs commises, plutôt qu'en fonction des performances de la méthode pour chaque message pris individuellement. Classiquement, pour évaluer l'exactitude, on fractionne les données de manière aléatoire. Par exemple, on garde 90% des données pour « apprendre » la méthode – c'est-à-dire qu'on va calculer divers paramètres ayant une pertinence relativement

[11] Voir par exemple El Karoui, Noureddine et Elizabeth Purdom (2015), "Can We Trust the Bootstrap in High- dimension?", Technical report 824, UC Berkeley Department of Statistics, February. En cours de soumission auprès de *JASA*.

à notre méthode pour cette partie de nos données, comme nous l'avions fait avec la droite à propos des données relatives au marché immobilier – puis nous évaluons l'exactitude de notre méthode en l'appliquant aux 10% restants. Nous n'avons donc plus besoin de théorie, il nous suffit d'observer les performances de notre méthode. Cette technique, ainsi que toutes ses variantes, dont celle de la validation croisée, sont actuellement parmi les plus efficaces pour évaluer le degré d'exactitude d'une méthode.

Étudions à présent le cas de la base de données de pourriels : 4601 messages sont marqués « pourriel » ou « non-pourriel » en fonction de 58 caractéristiques (dont le nombre de majuscules, de points d'exclamation, de symboles monétaires, etc. contenus dans le message). On fractionne les données selon un partage 2/3+1/3, et on commence par le groupe des 2/3 qu'on appelle « l'ensemble d'entraînement » ; le dernier tiers, l'ensemble « test », est laissé de côté jusqu'à la dernière étape. On fait appel à la méthode des « forêts aléatoires » introduite par Leo Breiman in 2001[12] : si on prélève un sous-groupe aléatoire dans l'ensemble d'entraînement et un autre sous-groupe aléatoire constitué de caractéristiques propres aux données de l'ensemble, cela nous permet de construire un « arbre de décision » (voir figure ci-dessous). On peut élaborer cet arbre en suivant la méthode CART[13] : à chaque nœud, il suffit d'appliquer le test proposé afin de décider si l'on va déplacer tel courriel vers la droite ou la gauche en descendant dans l'arbre. Au bout d'un moment, on arrive aux nœuds du bas qu'on appelle « feuilles » et qui nous donnent une estimation quant à la probabilité que tel message soit un pourriel.

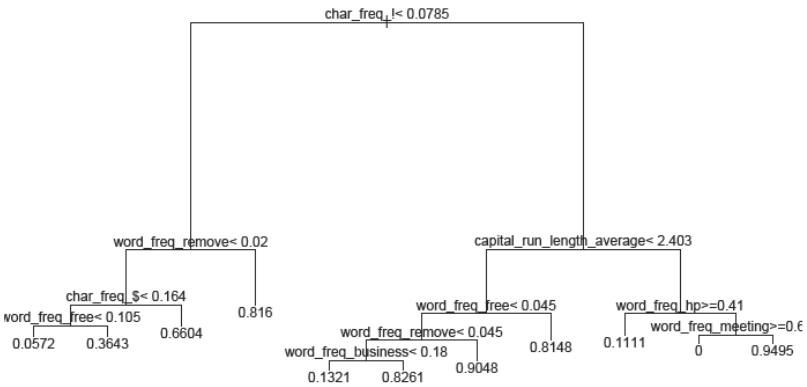
On peut répéter ces opérations un grand nombre de fois. Dans le cas présent, nous avons procédé à 1000 segmentations aléatoires et obtenu 1000 arbres de classification – autrement dit une forêt d'arbres de classification – et donc 1000 classifieurs. On fait ensuite passer chacun des messages de l'ensemble test à travers la forêt et on obtient une « probabilité » empirique quant à la nature des messages.

---

[12] Breiman, Leo (2001), "Random Forests", *Machine Learning*, ,45 (1), pp. 5–32, http://link.springer.com/article/10.1023%2FA%3A1010933404324

[13] Breiman, Leo, Jerome H. Friedman, Richard A. Olshen et Charles J. Stone (1984), *CART: Classification and Regression Trees*, Wadsworth Advanced Books and Software, Wadsworth Statistics/Probability Series.

Même en cas de méconnaissance des données, appliquer cette méthode à un ensemble de messages électroniques ne prend que quelques minutes et permet d'obtenir un classificateur fiable à 95% (c'est-à-dire que 95% des messages de l'ensemble test sont correctement classés en tant que courriel ou pourriel). La portée pratique de cette méthode est immense. Elle permet de se passer de tout modèle de probabilité, si ce n'est à travers l'idée qu'il vaut mieux utiliser des classifieurs en quelque sorte non corrélés entre eux pour les agréger ensuite plutôt que des classifieurs très similaires (cette idée d'une nécessaire décorrélation entre classifieurs explique pourquoi les 1000 classifieurs sont construits à partir de 1000 sous-ensembles aléatoires tirés de l'ensemble d'apprentissage, en utilisant à chaque fois un ensemble aléatoire de caractéristiques différent).



La base de données permet ici de faire deux choses en même temps : ajuster la méthode aux données et mesurer la fiabilité de la procédure pour les données à venir. Cette technique permet aux spécialistes de se passer de la modélisation probabiliste dans la mesure où ils savent ***a priori*** qu'il peuvent évaluer la performance de la méthode sur les données disponibles et sur les données à venir, dans l'hypothèse où les nouvelles données ont des caractéristiques similaires aux données de la base actuelle (hypothèse faible). Essentiellement, la seule chose qui importe, c'est la base de données elle-même. Il n'existe pas de théories sur elle.

Quelle que soit la méthode utilisée, il n'est point nécessaire de recourir à une quelconque hypothèse de modélisation, il suffit de vérifier si elle fonctionne ou pas pour notre base de données.

Les techniques issues des théories du maximum de vraisemblance fournissent des méthodes pour calculer des estimations : pour créer des données, ce qui était leur raison d'être initiale, elles n'ont intrinsèquement aucun lien avec un quelconque modèle probabiliste. On peut les appliquer à la base de données disponible sans que n'entre en jeu la notion de probabilité. On peut bien entendu appliquer la méthode même si les données ne sont pas conformes au modèle initialement prévu pour mettre en place la procédure. Ainsi, les beaux résultats probabilistes mentionnés plus haut (lesquels supposent que les données suivent un modèle spécifié par l'utilisateur) ont laissé place à des investigations numériques. La pensée probabiliste demeure très présente dans cette méthodologie, mais sous une forme très différente de celle qui était la sienne entre les années 1920 et le milieu des années 1970. La théorie des probabilités est très peu présente même si on recourt amplement à la randomisation.

En résumé, il va de soi que les statistiques et les probabilités sont étroitement liées. L'utilisation « subtile » de la théorie des probabilités a joué un rôle crucial dans la création de nombreux outils statistiques visant notamment à évaluer les performances des différentes méthodes. Le domaine des statistiques accorde cependant une place de plus en plus prépondérante à la méthodologie : on compte de plus en plus de bases de données complexes à analyser, il devient parfois difficile de créer des modèles probabilistes adéquats, du moins cela prend-t-il du temps, et il arrive parfois que ce soit impossible. L'influence des idées probabilistes se fait encore sentir dans les méthodes, mais beaucoup moins dans les pratiques. En fait, au-delà des principes de base servant à formuler les problèmes, on trouve très peu de probabilités dans les manuels actuels de statistiques appliquées[14]. La statistique bayésienne, dont je n'ai pas parlé ici, s'appuie bien sûr largement sur des notions probabilistes pour poser ses problèmes, mais d'un point de vue pratique, elle se sert

[14] Voir par exemple le très bon manuel, largement utilisé, de Hastie, T., R. Tibshirani and J. Friedman, *The Elements of Statistical Learning,* Springer Series in Statistics, New York : Springer-Verlag.

relativement peu de la théorie des probabilités une fois le problème posé[15]. Je renvoie le lecteur à la communication proposée par Erwin Bolthausen lors du colloque annuel des spécialistes des statistiques pour une perspective probabiliste sur des problèmes similaires[16].

D'intéressants problèmes théoriques continuent de se poser au confluent des statistiques, des probabilités et de l'optimisation lorsque les paramètres que l'on cherche à estimer sont de très grande dimension. Comme on l'a mentionné plus haut, il est possible de montrer que l'interprétation classique du bootstrap ou des méthodes du maximum de vraisemblance peut être source d'erreurs, et que dans le contexte actuel, ces méthodes sont susceptibles de connaître de sérieuses défaillances (échec et résultat largement sous-optimal, respectivement). Ainsi, il semble qu'on n'ait d'autre choix que de s'appuyer sur la théorie des probabilités si l'on veut produire des inférences statistiques valides.

Les idées probabilistes ont par ailleurs aujourd'hui une influence décisive sur les algorithmes ; dans les statistiques notamment, on recourt massivement à l'algèbre linéaire, laquelle intègre une part toujours plus grande d'aléatoire : étant donné la dimension des ensembles, la randomisation permet en effet d'accélérer considérablement l'algorithme, au prix d'une légère perte d'exactitude.

La théorie des probabilités et les statistiques ont toujours été étroitement liées. Comme nous avons tenté de le montrer brièvement ici, la pensée probabiliste est essentielle à la démarche statistique. Par ailleurs, de nombreux résultats largement exploités dans les statistiques s'appuient sur des résultats obtenus en théorie des probabilités. Or ces résultats sont désormais intégrés au logiciel de statistiques indispensable à l'analyse de données, en sorte que l'importance et la visibilité de la théorie des probabilités soient beaucoup moins prégnantes aujourd'hui dans la formation et la pratique des statisticiens (et des étudiants en début de cursus) qu'il y a quelques décennies. Les progrès de la puissance de calcul ont également mené à la création de nombreuses méthodes qui ont fait leur preuve

[15] Voir par exemple le manuel *Bayesian Data Analysis* de Gelman, A., J.B. Carlin, H.S. Stern, D. Dunson, A. Vehtari et D.B. Rubin, Boca Raton, FL, USA : Chapman and Hall/CRC Press.

[16] Bolthausen, Erwin, IMS Presidential Address at 2015 JSM, Seattle, "Some Thoughts about the Relations between Statistics and Probability Theory".

en matière d'analyse de données et ont remplacé les résultats issus de la théorie des probabilités et obtenus par simulations ciblées sur ordinateur. Comme nous l'avons expliqué, ces méthodes reposent bien sur des idées et un cadre probabilistes, mais très peu cependant sur la théorie des probabilités telle que la conçoivent les probabilistes d'aujourd'hui. Et pourtant, cette théorie commence à prendre une importance primordiale au plan pratique, notamment dans le développement d'algorithmes statistiques aléatoires, rapides et dont l'exactitude est vérifiable, et elle demeure un instrument-clé dans de nombreux secteurs des statistiques théoriques modernes.