Probabilités et statistiques dans les phénomènes réels

par Bernard Beauzamy
Société de Calcul mathématique, S.A.

août 2000

Chacun de nous utilise quotidiennement les probabilités, ne serait-ce qu'au travers d'expressions du type "il est peu probable que j'arrive à l'heure". Les statistiques sont également d'usage courant : tout le monde parle de l' "espérance de vie" d'une population. Mais ces emplois quotidiens masquent des disparités : dans certains cas, l'utilisation est légitime, dans d'autres non.

Très récemment, nous avons rencontré trois exemples : un problème de correction d'erreurs de visée (ce contrat est terminé), un problème de mesure d'effets toxiques sur des espèces vivantes (qui est en cours), un problème d'identification météorologique (qui est en projet). Chose étonnante, dans chaque cas, les méthodes que nous avons employées ou proposées vont à l'encontre de ce que l'on attendait.

Il y a d'abord une différence fondamentale entre probabilités et statistiques. Les statistiques fournissent des méthodes pour traiter des collections de nombres, pourvu qu'ils soient donnés dans les mêmes unités. Je puis faire la moyenne des prix relevés dans une rue, et en calculer l'écart type. Ces statistiques peuvent être idiotes si un boulanger voisine avec un concessionnaire automobile, mais elles sont formellement correctes. J'obtiendrai ainsi un "prix moyen" des biens vendus dans la rue en question.

Les probabilités, au contraire, reposent sur une axiomatique mathématique, et elles font référence au hasard. Elles trouvent leur origine dans la "loi empirique des grands nombres" : si un événement A est susceptible de se produire lors d'une expérience, la probabilité de A est définie comme le quotient entre le nombre de réalisations de A et le nombre de fois où l'expérience a été faite, lorsque l'expérience est répétée suffisamment longtemps. Par exemple, si je jette un dé N fois, je peux m'attendre à ce que le nombre n de fois où "1" est sorti vérifie $n/N \sim 1/6$, cette approximation étant d'autant meilleure que N est grand.

Le fondement même des probabilités -on l'oublie trop souvent- est donc la répétition d'une expérience basée sur le hasard. Il est très important de se souvenir de ces deux mots-clés : "répétition" et "hasard". Si la répétition n'est pas possible, ou si l'expérience ne dépend pas du hasard, l'usage des probabilités n'est pas licite.

Lorsqu'une expérience dépend du hasard, on cherche à déterminer ce qu'on appelle une "loi de probabilité" : quelle est la probabilité que la mesure prenne telle ou telle valeur ? C'est très facile dans les livres ; ce l'est beaucoup moins dans la réalité, parce que les contours de l'expérience ne sont jamais définis avec précision. Prenons un exemple concret : je sors de chez moi le matin ; je note la taille de la première personne que je rencontre. Le hasard intervient-il ? Oui, bien sûr. Est-ce répétable ? Ce n'est pas clair. Je puis assurément rentrer chez moi, puis ressortir, et noter une seconde personne, mais ce n'est plus la même expérience. L'espace probabilisé, cher aux théoriciens et noté traditionnellement Ω, n'est pas bien défini. Il n'empêche : je puis faire des statistiques à partir de mes relevés et conclure par exemple que, au bout d'un an, la moyenne des tailles des 320 personnes que j'ai rencontrées est de 1,68 m (peut-être y aura-t-il 45 jours où je n'étais pas chez moi, ou bien où je n'étais pas sorti). Je puis aussi faire un histogramme : tant de personnes entre 1,75 et 1,80 m, par exemple, et me faire ma propre loi de probabilité.

Imaginons que j'aie obtenu le recensement suivant :

Moins de 1,60 m :	150 personnes
Entre 1,60 et 1,70 m :	50 personnes
1,70 et 1,80 m :	60 personnes
1,80 et 1,90 m :	50 personnes
plus de 1,90 m :	10 personnes

Je puis alors conclure que les personnes de tailles ≤ 1,60 m ont une probabilité de 150/320 ~ 0,4687.

Jusque là, tout va bien, et il n'y a pas d'objection méthodologique à faire : j'ai parfaitement le droit de recenser les tailles des gens que je croise, et d'en faire des statistiques.

Mais à partir de là, je peux vouloir faire une prédiction. Je déciderai de prédire, avant de partir de chez moi, la taille de la première personne que je rencontrerai. Selon les règles en usage chez les statisticiens, la meilleure prédiction est la moyenne, soit ici 1,68 m : c'est pour elle en effet que l'erreur commise a la plus faible probabilité.

Eh bien, ce choix est idiot, et n'a aucune valeur prédictive. En effet, voici ce qui se passe : ou bien je sors de chez moi en semaine, à 7h45, pour prendre le métro, et je croise des Philippins (de très petite taille) qui vont prendre leur emploi au super marché voisin, ou bien je sors de chez moi plus tardivement, y compris le week-end où je prends quelquefois ma voiture, et alors je croise des gens "ordinaires". Ma statistique, qui ignore deux éléments essentiels (heure et distinction semaine/week-end) est donc complètement dépourvue de valeur prédictive.

Revenons maintenant aux trois exemples cités plus haut.

- Dans le cas de la correction d'erreurs de visées, ce qui est aléatoire, ce n'est pas la visée, c'est l'erreur sur la visée. Ces visées sont répétables autant qu'on le souhaite, et il est à peu près légitime de considérer que l'erreur commise est due au hasard, ou plus exactement l'erreur résiduelle, une fois que l'on a recensé toutes les causes d'erreurs déterministes et identifiables. Il est donc légitime d'utiliser les probabilités, et de dire que la probabilité que l'erreur soit comprise par exemple entre 1 degré et 1,2 degrés est de 0,231. Bien que l'appareil soit entièrement construit par l'homme (et donc on comprend comment il fonctionne), nous utilisons la théorie des probabilités.

- Dans le cas de l'influence des produits toxiques sur les espèces vivantes, nous avons une situation où la principale difficulté tient au manque de données. On pourrait imaginer d'effectuer d'innombrables expériences, où tous les toxiques seraient essayés sur toutes les espèces, en des concentrations variables et selon des durées variables. De telles expériences sont irréalisables en pratique, mais elles sont concevables en théorie : la toxicité d'un produit sur une espèce n'est pas le fruit du hasard seul. A coup sûr, si l'on répète la même expérience plusieurs fois, on n'obtiendra pas le même résultat : il y a une forte variabilité, due en particulier au fait que le protocole d'expérience doit être défini plus précisément (température, milieu, composition de la population, etc). Dans ce cas, nous proposons une approche à base de statistiques, et en particulier de classifications, regroupements, etc, qui permettent de discerner des grandes lignes : telle classe de toxiques est plus dangereuse pour telle catégorie d'espèces. En revanche, nous n'utilisons pas d'approche probabiliste, parce qu'il y a explication à la toxicité, et qu'elle n'est pas le seul fruit du hasard.

- Le projet relatif à la météorologie consiste en ceci : à partir des relevés pris par différents capteurs (vitesse du vent, température, hygrométrie, etc, une dizaine en tout), peut-on identifier de manière automatique le temps qu'il fait sans intervention humaine ? A priori, le temps qu'il fait est le fruit du hasard. On s'attendrait donc à une approche de type statistique, puis probabiliste : recenser les situations où il pleut, et, pour chacune, les valeurs possibles des différents indicateurs, et, à partir de là, construire des histogrammes puis des lois de probabilité. Au contraire, nous avons proposé une approche complètement déterministe : si on travaille avec dix capteurs, la mesure est donc dans un espace de dimension 10 ; à l'intérieur de cet espace, chercher à déterminer les régions qui correspondent à la pluie, celles qui correspondent à la neige, etc. Pourquoi procéder ainsi ? Tout simplement parce que l'approche probabiliste est incorrecte. Le lien qu'il y a entre, par exemple, la température de l'air et la précipitation (pluie ou neige) est un lien déterministe, lié à une loi physique, et il n'a rien d'aléatoire.

Il est amusant d'observer que, dans chacun des cas cités, la méthode que nous avons employée ou proposée était à l'opposé de celle qu'on attendait : déterministe pour la météorolo-

gie, statistique pour l'écotoxicologie, probabiliste pour la vision. Quelques règles très simples peuvent être dégagées ; elles aideront à comprendre le choix qui doit être fait.

- Il ne faut pas confondre la mesure et l'erreur sur la mesure. Le résultat d'une mesure tient en général à une cause déterministe, qu'il faut rechercher. Le hasard n'intervient généralement que dans l'erreur sur la mesure.

- Il est dangereux d'imputer au hasard ce qui tient en réalité à notre ignorance. En procédant ainsi, on s'interdit de rechercher les véritables lois, qui existent en général.

- En adoptant d'emblée un point de vue probabiliste, on est tenté d'utiliser les outils fondamentaux de la théorie des probabilités, à savoir les lois des grands nombres.

La plus importante est le "théorème central limite" : la moyenne de variables aléatoires, prise sur un nombre suffisamment grand d'expériences, après normalisation, se comporte comme une gaussienne centrée réduite. Or ceci n'est vrai que si les variables sont indépendantes. J'ai omis le mot à dessein dans l'énoncé précédent. En pratique, dans les phénomènes réels, cette hypothèse d'indépendance est rarement satisfaite. Par exemple, elle ne l'est manifestement pas lorsque j'observe des tailles à la sortie de chez moi, parce que c'est plus ou moins le même échantillon qui revient. Elle peut l'être, s'il s'agit d'erreurs dues à un phénomène de haute fréquence que l'on observe à basse fréquence. Mais dans chaque cas, il faut une réflexion soigneuse et approfondie sur la question de l'indépendance, sans quoi le résultat est absolument faux et revient à remplacer arbitrairement la loi réelle par une gaussienne.

Comme chacun sait, on peut faire dire n'importe quoi aux statistiques : explications fausses sur le passé, prédiction fausses sur l'avenir. Il suffit pour cela de changer quelques pondérations ici, quelques coefficients là. On n'a pas la même liberté avec les probabilités : le hasard, le vrai hasard, a ses vraies loi, qui sont incroyablement rigides et précises. On peut distinguer une suite de 0 et de 1 donnée par le hasard d'une suite construite à la main par un faussaire. Mais à prétendre que les phénomènes réels sont régis par les lois du hasard, on obtient des résultats encore plus faux que ceux qu'exhiberait le plus chevronné des statisticiens.