Probabilités et statistiques

MESURER L'ALÉATOIRE

Alors que la géométrie fut étudiée dès la Grèce Antique, la théorie des probabilités ne s'est quant à elle constituée qu'au XVII[e] siècle. Le projet était en effet mathématiquement osé : introduire des calculs pour prévoir l'issu d'expériences aléatoires ! C'est historiquement sous l'impulsion de Blaise Pascal et Pierre de Fermat, suite à un problème que leur avait soumis le chevalier de Méré qui adorait jouer aux dés, que les premiers calculs probabilistes apparaissent. Une origine de la théorie que l'on retrouve, d'ailleurs, dans l'étymologie même du mot hasard, dérivé de l'arabe *az-zahr* signifiant « dé à jouer ».

UN EXEMPLE POUR DÉMARRER

Pour commencer, imaginons simplement le jeu de dé suivant : on lance 2 dés et on parie sur la somme des valeurs qui apparaissent. La question naturelle que se pose alors le joueur est évidemment de savoir sur quelle somme il faut miser pour avoir le plus de chance de gagner. La réponse à cette question découle en fait simplement d'une analyse détaillée de l'expérience. Faisons tout d'abord un tableau décrivant toutes les issues possibles d'un lancer de deux dés.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

En supposant que le dé n'est pas truqué et que toutes les faces ont la même probabilité d'apparaître, la somme sur laquelle il convient de miser est celle qui apparaît le plus souvent à l'intérieur de ce tableau. Autrement dit ici : 7, puisque 7 apparaît six fois dans le tableau (il y a six possibilités de tomber sur une somme de 7 : en faisant 6 et 1, 5 et 2, etc.) alors que les autres sommes possibles apparaissent moins souvent (12 apparaît par exemple une seule fois, et 8 cinq fois).
Quelle chance a-t-on alors de gagner si l'on parie sur 7 ? Sachant qu'il y a 36 cases dans ce tableau, autrement dit 36 issues différentes du lancer, on a 6 chances sur 36 (c'est-à-dire une chance sur 6) de gagner quand on parie sur 7. Et si l'on parie sur 12, on a alors une seule chance sur 36 de gagner ! Attention, faire 3 avec le premier dé et 5 avec le second, ce n'est pas la même chose que de faire 5 avec le premier dé et 3 avec le second, même si la somme est égale.

CAS FAVORABLES ET CAS DES POSSIBLES

Calculer la probabilité d'un événement produit dans le cadre d'une expérience aléatoire (dans notre exemple, obtenir une somme de 7 lors du lancer de deux dés) revient donc à dénombrer tous les cas dits « favorables » correspondant à cet évènement, et tous les cas dits possibles qui peuvent arriver lors de cette expérience. Et à utiliser ensuite la formule donnant la probabilité, à savoir la division « nombre de cas favorables / nombre de cas possibles ».
Première conséquence importante : une probabilité est toujours comprise entre 0 et 1 (autrement dit entre 0 % et 100 % = 100/100). La probabilité d'un événement est nulle quand cet événement est impossible (par exemple, trouver une somme de 15 en lançant deux dés : le nombre de cas favorables est alors nul). La probabilité est égale à 1 quand cet événement est certain (par exemple, trouver une somme supérieure à 1 en lançant deux dés). Les probabilistes parlent également d'évènement contraire : par exemple, toujours avec notre lancer de deux dés, l'événement « la somme des deux dés fait 7 » admet comme événement contraire « la somme des deux dés est différente de sept ». Remarquons au passage que l'événement contraire d'un événement impossible est un événement certain, et vice versa !

DÉNOMBREMENT

À partir de là se pose quand même une question : comment faire pour compter ces cas favorables et ces cas possibles quand les expériences que l'on décrit, et les évènements dont on cherche à calculer la probabilité, sont un peu plus compliqués qu'un lancer de deux dés et une somme de deux faces ? Les méthodes de dénombrement sont fondamentalement à la base du calcul des probabilités.
Reprenons un exemple pour en faire apparaître les principales. Imaginons une urne, remplie de 5 boules rouges, 3 noires et 2 blanches, et de laquelle on tire trois boules : combien de possibilités y a-t-il pour que les trois boules soient de la même couleur ?
En l'absence de précision supplémentaire, la réponse est tout simplement impossible à donner ! Tout dépend en effet de la façon dont on tire les boules : est-ce un tirage simultané, successif avec remise de la boule dans l'urne à chaque tirage, ou successif sans remise de la boule ? Cela change tout, en effet. Dans le cas d'un tirage simultané, l'ordre n'intervient pas alors que dans un tirage successif, tirer une boule rouge puis une blanche puis une rouge, ce n'est pas la même chose que de tirer d'abord une boule rouge, puis à nouveau une boule rouge, puis une boule blanche. Le résultat du dénombrement sera donc à chaque fois différent.

LISTES

Dans le cas d'un tirage successif avec remise, une même boule peut être extraite plusieurs fois et on retrouve à chaque tirage la même situation qu'au tirage précédent. Compter alors le nombre de possibilités pour que les trois boules tirées soient blanches revient à écrire une liste de trois éléments dans laquelle chaque élément est à choisir parmi deux possibilités (les deux boules blanches de l'urne). Ou encore à imaginer que l'on a trois cases à remplir et que pour remplir chaque case, on a le choix entre deux boules. Le principe est alors dit multiplicatif : on a deux possibilités pour remplir la première case, qui se conjuguent successivement avec les deux autres possibilités de remplissage de la seconde case et à nouveau avec les deux dernières possibilités de remplissage de la troisième case, ce qui fait en tout $2 \times 2 \times 2 = 8$ possibilités.
D'où, au final, si l'on additionne les 8 possibilités de tirer trois boules blanches, les 27 ($=3 \times 3 \times 3$) possibilités de tirer deux boules blanches et les 125 ($=5 \times 5 \times 5$) possibilités de tirer trois boules rouges, on dénombre en tout 160 possibilités de tirer trois boules de la même couleur.

ARRANGEMENTS

Dans le cas d'un tirage successif sans remise, on ne peut pas avoir de répétition d'une même boule : une fois tirée, la boule n'est plus disponible pour le tirage suivant. Autrement dit, si l'on avait le courage de décrire de façon exhaustive les 160 cas précédents, il suffirait d'enlever ceux qui présentent une répétition et on aurait le résultat cherché ! Sauf qu'un bon mathématicien évitera toujours de se lancer dans une entreprise aussi pénible… d'autant qu'il existe une formule qui donne directement le résultat ! Une liste de p éléments sans répétition choisis parmi un ensemble de n éléments est appelé en mathématique un arrangement de p éléments parmi n et le nombre d'arrangements de p éléments parmi n est égal à la multiplication des nombres décroissants allant de n à (n-p+1), soit : $n(n-1)(n-2)\ldots(n-p+1)$.
Pour mieux comprendre, reprenons l'exemple précédent pas à pas : on peut déjà dire qu'il est impossible de tirer sans remise trois boules blanches puisqu'il n'y en a que deux dans l'urne. De plus, il n'y a qu'une possibilité de tirer trois noires (puisqu'il n'y a que trois boules noires dans l'urne) : jusqu'ici donc, tout va bien.
Combien y a-t-il maintenant de possibilités de tirer successivement sans remise trois boules rouges sachant qu'il y en a 5 dans l'urne ? C'est là qu'interviennent les arrangements de 3 éléments parmi 5. On reprend alors la formule avec 5 à la place de n, 3 à la place de p et $n-p+1$ valant $5-3+1=3$: le nombre total d'arrangements de 3 éléments parmi 5 vaut donc $5 \times 4 \times 3 = 60$. Conclusion : il y a 60 possibilités pour que les trois boules tirées soient rouges, et donc en rajoutant l'unique possibilité qu'elles soient toutes les 3 noires, il y a 61 possibilités pour que les trois boules tirées successivement sans remise soient de la même couleur.

PERMUTATIONS

Dans le cas enfin d'un tirage simultané, tirer trois boules de l'urne, c'est prélever une partie (un sous-ensemble) de l'ensemble des boules présentes dans l'urne. Compter les cas favorables ou possibles revient donc à dénombrer le nombre de sous-ensembles d'un ensemble donné. C'est ce que l'on appelle en mathématique les combinaisons. Vu autrement, il s'agit d'un arrangement dans lequel l'ordre ne compte plus. Là encore, une formule existe pour dénombrer des combinaisons : le nombre de combinaisons de p éléments parmi n est donné par la division du nombre d'arrangements de p éléments parmi n, par p! (factorielle p, c'est-à-dire la multiplication des nombres de p à 1, $p(p-1)(p-2)\ldots 3 \times 2 \times 1$).
Plus concrètement : combien de possibilités y a-t-il cette fois-ci pour que les trois boules tirées de l'urne soient rouges ? Pour répondre, il faut calculer le nombre de combinaisons de 3 éléments parmi 5, qui est égal au nombre d'arrangements trouvés précédemment (60) divisé par

Somme de deux dés

la somme du tirage est de 6

le nombre de tirages possibles est :

6 x 6 = 36

somme des 2 dés	nombre de cas favorables	probabilité
2	1	1/36 = 0,0028
3	2	2/36 = 0,056
4	3	3/36 = 0,085
5	4	4/36 = 0,111
6	5	5/36 = 0,139
7	6	6/36 = 0,167
8	5	5/36 = 0,139
9	4	4/36 = 0,111
10	3	3/36 = 0,085
11	2	2/36 = 0,056
12	1	1/36 = 0,028

Naissance d'une science

XVI[e] siècle

Jérôme Cardan écrit un traité « de Ludo Aleae » relatif au jeu de dés.

1771

Abraham de Moivre applique les techniques de dénombrement aux probabilités. Deux ans plus tard, les frères Jacques et Daniel Bernoulli découvrent la loi des grands nombres.

1812

Simon de Laplace publie sa « Théorie analytique des probabilités » et formalise le calcul du nombre de cas favorables sur le nombre de cas possibles.

XIX[e] siècle

Naissance de la statistique. À la fin du siècle, le Russe Tchebychev utilise les travaux de son compatriote Markov et rend la théorie plus rigoureuse.

1933

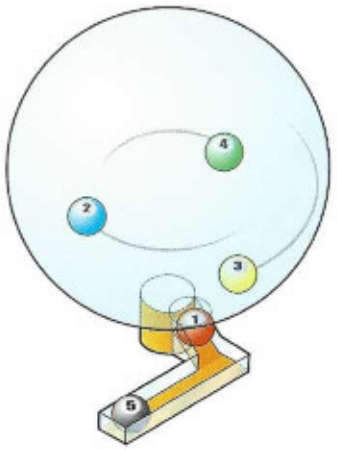
Le Russe Kolmogorov met en place une axiomatique très abstraite qui fonde la théorie des probabilités. C'est sous cette forme que la théorie est aujourd'hui enseignée à l'Université.

Les premiers calculs de probabilité

Réalisés par Pierre de Fermat et Blaise Pascal

en **1654**

Combinaison de 3 éléments parmi 5



tirages possibles

$$C_n^p = \frac{n \times (n-1) \times \ldots \times (n-p+1)}{p!}$$

$$C_5^3 = \frac{5 \times 4 \times 3}{3 \times 2} = 10$$

3! = 3 x 2 x 1 = 6. Résultat : 60/6 = 10. Il y a donc 10 possibilités différentes pour que les trois boules tirées soient rouges, et donc en rajoutant à nouveau l'unique possibilité pour que les trois boules soient noires, il y a 11 possibilités pour que les trois boules tirées simultanément soient de la même couleur.

CALCUL DES PROBABILITÉS

On l'aura compris : calculer des probabilités nécessite de connaître les formules de dénombrement de listes, arrangements et combinaisons. Mais les probabilités répondent elles-mêmes à certaines propriétés calculatoires importantes et très utiles dans la pratique. En premier lieu, celle-ci : si vous considérez un événement A et son contraire, la somme de la probabilité que A se produise et de la probabilité que le contraire de A se produise (autrement dit que A ne se produise pas) est toujours égale à 1 (100 % de chance) : c'est le bon sens qui parle ! De là, la méthode de calcul suivante : imaginons que l'on cherche la probabilité d'obtenir au moins un six en quatre lancers successifs d'un dé. Plutôt que de calculer la probabilité d'obtenir un 6, ajoutée à celle d'obtenir deux « 6 », etc., il est plus simple de calculer la probabilité de ne pas obtenir de 6 du tout, et d'utiliser la formule précédente. La probabilité de ne pas obtenir de 6 en un lancer étant 5/6 (5 chances sur 6 que ça se produise), la probabilité de ne pas obtenir de 6 en quatre lancers est 5/6 x 5/6 x 5/6 x 5/6 = 5/64 soit à peu près 0,48 (48 % de chance que le 6 ne sorte jamais en quatre lancers). De sorte que la probabilité que le 6 sorte au moins une fois en quatre lancers est 1 – 0,48 = 0,52 (52 % de chance).
Dans bien des cas, on cherchera en outre à calculer des probabilités un peu plus complexes, comme par exemple celle que l'événement [A ou B] se produise, ou que l'événement [A et B] se produise. Il suffit dans ce cas de savoir que si A et B sont indépendants l'un de l'autre, la probabilité P(A et B) est égale à P(A) x P(B), et que si A et B sont exclusifs l'un de l'autre, P(A ou B) = P(A) + P(B). Dans le cas où A et B ne sont pas exclusifs et peuvent se produire en même temps, on a la formule : P(A ou B) = P(A) + P(B) – P(A et B).
Connaître par exemple la probabilité de tirer un 6 en jetant deux dés revient à calculer la probabilité de tirer un 6 avec le premier dé ou de tirer un 6 avec le second dé, sachant bien sûr que ces deux événements ne sont pas exclusifs puisque l'on peut tirer un double 6. On calcule alors la probabilité de tirer un 6 avec le premier dé et n'importe quoi avec le second (6/36 : 6 cas favorables sur 36 possibles), la probabilité de tirer un 6 avec le second dé (même chose : 6/36). On ajoute ces deux nombres (12/36) et on retranche la probabilité de tirer un double 6 (1/36), d'où un résultat final cherché de 11/36 soit environ 0,30 c'est-à-dire 30 % de chance de tirer un 6 en jetant deux dés.

LES STATISTIQUES

Tout les calculs que l'on a faits jusqu'à présent reposent néanmoins sur un résultat que nous n'avons pas démontré : celui consistant à affirmer que les tirages des différentes faces d'un dé sont tous équiprobables et ont chacun une chance sur 6 de se produire. Que se passerait-il si ce n'était pas le cas ? Quelle valeur autre que 1/6 prendrait-on alors comme probabilité de tirer la face « 1 » ? ou la face « 6 » ? Pour résoudre ce problème, les mathématiciens ont mis en place des méthodes dites statistiques : ils considèrent que la probabilité résulte de la connaissance détaillée et précise d'un très grand nombre de phénomènes analogues.
Le principe est simple et s'appuie sur la loi dite des grands nombres qui peut s'énoncer de la façon suivante : étant donné un événement ayant une probabilité donnée (comme par exemple la sortie d'un six dans le jeu de dés, dont la probabilité est de 1/6 si le dé n'est pas pipé), plus le nombre de tentatives est grand (dans notre exemple, plus de fois on lance le dé), plus l'écart entre le nombre de fois où l'événement se vérifie effectivement et le nombre prévu théoriquement par la probabilité est petit. Autrement dit, si le dé n'est pas pipé, pour calculer la probabilité d'obtenir 1, il suffit de lancer un très grand nombre de fois le dé et et de calculer la fréquence d'apparition du 1 (c'est-à-dire le nombre de fois où l'on a obtenu le résultat 1 divisé par le nombre de lancers réalisés) : plus on réalisera de lancers, plus cette fréquence se rapprochera de la probabilité cherchée (d'où la nécessité de réaliser, ou de simuler sur ordinateur, plusieurs milliers de lancers de dés pour avoir un résultat précis). Et si le dé est pipé, on s'en apercevra tout de suite et on aura les nouvelles valeurs d'apparition de chaque face !
Formalisée au tout début du XVIIIe siècle par les frères Bernoulli, cette loi est essentielle. C'est sur elle que reposent par exemple les sondages : elle permet en effet, en interrogeant un nombre suffisamment important de personnes, de connaître l'opinion de la population entière (avec une marge d'erreur).
La statistique est donc une branche mathématique qui a pour but de déterminer des propriétés caractéristiques d'une population donnée, impossible à étudier dans son intégralité, uniquement à travers l'étude de ce qui se passe sur un petit échantillon. On sépare la statistique en deux branches : la statistique descriptive, correspondant à l'étude des caractéristiques du petit échantillon tiré, et la statistique inductive, c'est-à-dire la déduction, à partir de l'échantillon, d'informations approchées concernant la population totale. Et dans le domaine, une chose essentielle est à garder en mémoire : connaître avec certitude la population entière à partir de la seule étude d'un échantillon est impossible. Tout résultat statistique est un peu faux et ne veut rien dire sans l'estimation de l'erreur que l'on commet à le considérer comme juste.

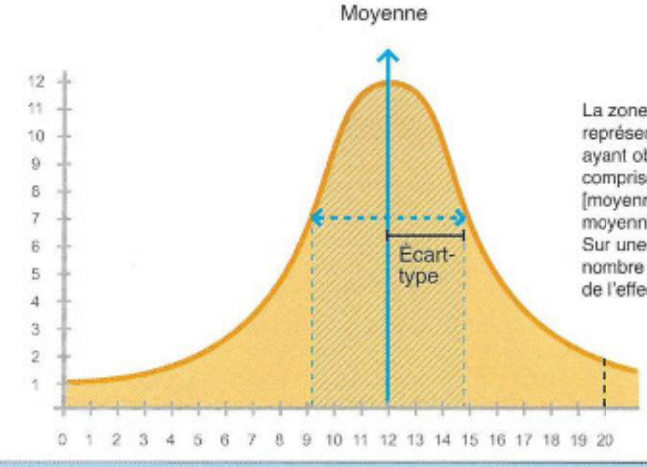
STATISTIQUES DESCRIPTIVES

En guise d'illustration, prenons un exemple concret de statistiques descriptives. Imaginons un échantillon de 25 élèves ayant eu les notes suivantes à une épreuve : 12, 10, 14, 15, 12 ,13, 11, 17, 12, 12, 9, 10, 12, 14, 13, 12, 15, 11, 13, 12, 19, 14, 13, 11, 8.
Les paramètres statistiques ont pour but de résumer, à partir de quelques nombres clés, l'essentiel de l'information relative à l'observation d'une variable, ici les notes obtenues. Certains sont dits de tendance centrale, car ils représentent une valeur numérique autour de laquelle les observations sont réparties. La moyenne est par exemple un paramètre de tendance centrale très facile à calculer : il suffit d'additionner les notes des 25 élèves et de diviser par le nombre d'élèves. On trouve ici 12,56. Ou encore de multiplier chaque note par le nombre d'élève qui l'ont obtenu, d'additionner tous les résultats, et de diviser par le nombre d'élèves : ici (8 x 1 + 9 x 1 + 10 x 2 + 11 x 3 + 12 x 7 + 13 x 4 + 14 x 3 + 15 x 2 + 17 x 1 + 19 x 1)/25 = 12,56. Sauf que la moyenne donne une information assez faible sur le résultat de la classe : elle ne permet pas de savoir, par exemple, si tous les élèves ont eu à peu près 12,5 ou si certains ont eu de très bonnes notes et d'autres de très mauvaises. C'est pour cela que l'on calcule aussi des paramètres dits de dispersion qui résument le plus ou moins grand étalement des observations de part et d'autre de la tendance centrale. L'écart-type est par exemple un paramètre de dispersion souvent utilisé. Il vaut 0 si tous les élèves ont la même note (pas de dispersion du tout). À l'inverse, plus l'écart-type est grand, plus la dispersion est importante. Ici la formule (un peu complexe) qui permet de calculer l'écart-type donne 2,33. Sur des échantillons plus importants, la distribution des notes suit généralement des courbes appelées « gaussiennes » présentant des propriétés remarquables.

STATISTIQUES INDUCTIVES

En matière de statistiques inductives, l'exemple type est le sondage. Et dans ce domaine, répétons-le, il est essentiel de garder en mémoire que l'erreur reste inévitable. D'ailleurs, des formules permettent même de la quantifier et donc d'en déduire si le résultat statistique que l'on trouve est fiable ou non. Par exemple, si l'on réalise un sondage auprès de 1 000 personnes et qu'on mesure le pourcentage de personnes souhaitant voter pour Monsieur X, la théorie dit qu'on pourra estimer le pourcentage réel de personnes souhaitant voter pour X comme égal à celui calculé sur l'échantillon, à plus ou moins 3,2 % près. La précision est d'importance ! Lors du premier tour des élections présidentielles de 2002, le dernier sondage effectué par l'Institut BVA sur 1 000 électeurs prévoyait que Jacques Chirac obtiendrait 19 %, Lionel Jospin 18 %, Jean-Marie Le Pen 14 %. Cela voulait dire que le pourcentage réel de gens projetant de voter pour Jacques Chirac était compris entre 19 % – 3,2 % = 15,8 % et 19 % + 3,2 % = 22,2 %. En faisant les calculs pour Lionel Jospin et Jean-Marie Le Pen, on en déduit que le sondage réalisé voulait dire que le pourcentage réel dans la population française de personnes souhaitant voter pour Jacques Chirac était compris entre 15,8 % et 22,2 %, pour Lionel Jospin entre 14,8 % et 21,2 % et pour Jean-Marie Le Pen entre 10,8 % et 17,2 %. Des fourchettes qui ne permettaient même pas de prévoir la base : l'ordre d'arrivée des candidats !

Moyenne et écart-type



La zone hachurée en bleu représente le nombre d'élèves ayant obtenu une note comprise dans l'intervalle [moyenne – écart-type ; moyenne + écart type]. Sur une gaussienne, ce nombre est d'environ 68 % de l'effectif total.