

QUELQUES PHÉNOMÈNES CURIEUX EN PROBABILITÉS ET STATISTIQUES

Michel ÉMERY¹
C.N.R.S. et U.L.P. Strasbourg

L'espérance n'est qu'un charlatan
qui nous trompe sans cesse.

CHAMFORT, *Maximes et pensées.*

J'ai plaisir à remercier la Mission Culture Scientifique et Technique de l'U.L.P., qui a accueilli cette conférence le 11 janvier 2001 dans le cadre du Jardin des Sciences, ainsi que L'Ouvert qui accepte d'en publier la version écrite.

Cette conférence se compose de quatre parties, consacrées à quatre questions différentes. Les trois premières peuvent être lues indépendamment, la quatrième s'appuie sur la troisième. Chacune se présente sous la forme d'un jeu, qui ne présenterait pas le moindre intérêt en tant que jeu proprement dit, mais sert à mieux mettre en évidence la situation mathématique considérée.

1 Le paradoxe de Simpson

Mis en évidence par le statisticien britannique E.H. Simpson ² voici un demi-siècle, c'est parfois une source de difficultés dans l'interprétation des tableaux de corrélations statistiques. Si j'ai choisi d'aborder ce thème, c'est que je le croyais peu connu; je remercie Richard Cabassut qui m'a détrompé et documenté : ce sujet figure au programme de mathématiques des classes de première ³, et, sous le nom d'*effet de structure*, fait l'objet d'exercices dans les manuels correspondants ⁴.

Nous allons utiliser des jeux de cartes usuels, dans lesquels nous ne garderons que des deux, des trois, des dames et des rois; les deux et les trois seront des cartes *perdantes*, et les dames et rois seront des cartes *gagnantes*. Au moyen d'un jeu de cartes à dos bleus et d'un jeu de cartes à dos noirs, formons deux paquets de onze cartes chacun, selon la composition suivante (*N* signifie dos noir, *B* dos bleu, *G* carte gagnante, *P* carte perdante) :

	<i>B</i>	<i>N</i>
<i>G</i>	1	2
<i>P</i>	3	5

Paquet n° 1

	<i>B</i>	<i>N</i>
<i>G</i>	5	3
<i>P</i>	2	1

Paquet n° 2

¹© L'OUVERT 104 (2001)

²E.H. Simpson. The interpretation of interaction in contingency tables. *J. Roy. Stat. Soc. Ser. B* **13**, 238–241 (1951).

³Il est prévu des travaux pratiques sur « les paradoxes apparents des pourcentages ».

⁴Voir les travaux pratiques n° 11 dans le manuel 1998 de mathématiques de 1^{re} ES chez Nathan; voir aussi le manuel de mathématiques de 1^{re} L chez Hachette (programme de 1993) page 19.

Étalons sur une table, dos visibles et faces cachées, les onze cartes du premier paquet, préalablement battu. Le jeu consiste à retourner l'une des cartes : si c'est une carte G , on a gagné, si c'est une P on a perdu. Les chances de gagner sont $1/4$ si l'on prend une carte à dos bleu, $2/7$ pour une carte à dos noir et $3/11$ si l'on prend une carte sans tenir compte de la couleur du dos. Puisque $\frac{1}{4} < \frac{3}{11} < \frac{2}{7}$, on a évidemment intérêt, pour maximiser les chances de gagner, à choisir une carte à dos noir.

Si l'on joue au même jeu avec le second paquet, comme $\frac{5}{7} < \frac{8}{11} < \frac{3}{4}$, la meilleure stratégie consiste, là encore, à choisir une carte noire.

Regroupons maintenant nos deux paquets en un seul jeu de vingt-deux cartes, battons-le, et jouons de même avec ce nouveau paquet. Sa composition est obtenue en additionnant termes à termes les deux tableaux précédents :

	B	N
G	6	5
P	5	6

Réunion des deux paquets

Puisque $\frac{6}{11} > \frac{11}{22} > \frac{5}{11}$, c'est cette fois-ci la première colonne du tableau qui offre les meilleures chances, et l'on a donc intérêt à choisir une carte à dos bleu, contrairement aux cas précédents où l'on jouait avec l'un des deux paquets pris individuellement.

D'un point de vue mathématique, ce phénomène n'a rien de mystérieux ; c'est simplement la remarque que la somme de deux matrices 2×2 à déterminant négatif peut avoir un déterminant positif. Pour le statisticien, qui peut être amené à interpréter ces matrices comme des tableaux de corrélations, cela peut être plus embarrassant.

Imaginez par exemple que vous vous intéressez à la corrélation entre la gravité des accidents de la route et le port ou non de la ceinture de sécurité ; supposez que vous disposez d'une statistique sur 220 accidents, répartis comme suit selon la gravité, le port de la ceinture, et le lieu de l'accident (B = ceinture bouclée, N = ceinture non bouclée, G = grave, P = pas grave) :

<table border="1" style="display: inline-table;"> <tr><th></th><th>B</th><th>N</th></tr> <tr><th>G</th><td>10</td><td>20</td></tr> <tr><th>P</th><td>30</td><td>50</td></tr> </table>		B	N	G	10	20	P	30	50	+	<table border="1" style="display: inline-table;"> <tr><th></th><th>B</th><th>N</th></tr> <tr><th>G</th><td>50</td><td>30</td></tr> <tr><th>P</th><td>20</td><td>10</td></tr> </table>		B	N	G	50	30	P	20	10	=	<table border="1" style="display: inline-table;"> <tr><th></th><th>B</th><th>N</th></tr> <tr><th>G</th><td>60</td><td>50</td></tr> <tr><th>P</th><td>50</td><td>60</td></tr> </table>		B	N	G	60	50	P	50	60
	B	N																													
G	10	20																													
P	30	50																													
	B	N																													
G	50	30																													
P	20	10																													
	B	N																													
G	60	50																													
P	50	60																													
<i>110 accidents en agglomération</i>		<i>110 accidents hors agglomération</i>		<i>220 accidents au total</i>																											

J'ai simplement recopié les trois tableaux précédents, en multipliant tous les nombres par 10 — ceci ne change pas les pourcentages mais rend ces données statistiquement plus significatives (bien qu'elles restent parfaitement fictives et tout à fait fantaisistes). Qu'observe-t-on sur ces tableaux? La comparaison des deux tableaux de gauche montre qu'en agglomération les accidents sont moins graves, mais les ceintures moins souvent attachées. Le premier tableau met en évidence un effet bénéfique du port de la ceinture en agglomération, et le second, également un effet bénéfique du port de la ceinture hors agglomération : les accidents avec ceinture sont, dans chacun des deux cas, moins souvent graves que les accidents sans ceinture. Et cependant, le troisième tableau, somme des deux premiers, manifeste un effet inverse, puisque les accidents avec ceinture y sont plus souvent graves que les accidents sans ceinture! Qu'en conclure? Faut-il la boucler?

La discussion pourrait faire intervenir l'adaptation des comportements aux dangers perçus, et, en toute généralité, la difficulté d'interpréter des corrélations en termes de causes et d'effets ⁵. Elle serait de toutes façons oiseuse, puisque nos données, fabriquées *ad hoc*, ne correspondent à aucune réalité; je ne la mènerai pas ici, passant à un autre jeu.

2 Dites : « Rouge! »

Nous allons encore utiliser des cartes à jouer, pour nous intéresser à un petit problème dû au mathématicien et prestidigitateur américain P. Diaconis; ce problème m'a été communiqué par P. Artzner, et *L'Ouvert* en a déjà fait mention (problème 46 dans le numéro 66 de mars 1992).

Prenez un jeu ordinaire de 52 cartes (26 rouges et 26 noires); battez-le, posez-le devant vous faces non visibles, puis retournez les 52 cartes une à une. À un moment que vous choisissiez à votre gré, juste avant de retourner une carte, dites : « Je parie que la prochaine carte sera rouge ». Vous gagnez si elle est effectivement rouge (cœur ou carreau), vous perdez si elle est noire (pique ou trèfle). Vous devez parier une fois et une seule durant le déroulement du jeu : si vous n'avez pas parié sur les 51 premières cartes, vous devez parier sur la dernière (et vous savez alors si vous avez ou non gagné sans avoir besoin de la retourner, puisque vous avez déjà vu passer toutes les autres). Si vous aviez le droit de choisir entre parier sur « rouge » ou sur « noir », vous gagneriez à coup sûr en attendant la dernière carte; mais la règle de Diaconis vous impose de parier sur « rouge ». Quelle est la stratégie optimale, et quelle probabilité de gain donne cette stratégie?

Il est clair que cette probabilité est au moins 50 %, qui représentent vos chances de gagner si vous décidez à l'avance de parier par exemple sur la 31^e carte (ou la première, ou la dernière) sans tenir compte des couleurs des cartes déjà retournées. La question de Diaconis est donc de comprendre comment utiliser au mieux ces informations (les couleurs des cartes) au fur et à mesure de leur arrivée, et de quantifier le gain qu'elles peuvent apporter.

⁵Les lits sont des endroits dangereux. La preuve : c'est là que survient la plupart des décès.

La réponse de Diaconis est extrêmement simple : *toutes les stratégies donnent une chance sur deux de gagner* ; vous ne pouvez pas faire mieux que les stratégies stupides décrites au paragraphe précédent — ni moins bien, d'ailleurs : aucune tentative pour parier « rouge » au moment le plus défavorable ne parviendra à réduire vos chances de gain.

Diaconis en donne une démonstration très courte, dans le cadre de la théorie des filtrations et des temps d'arrêt ; voici comment on peut s'en convaincre de façon élémentaire. Imaginons un gigantesque tableau

R♠ 9♣ 2♥ 3♣ 3♥ V♣ 6♠ 8♦ R♥ 4♣
 2♣ 8♠ R♦ 9♠ 3♣ 1♦ 10♠ 5♣ 1♥ R♠
 3♥ V♠ 1♥ 5♣ 4♠ 10♠ R♠ D♣ 7♣ 10♣

.

.
.

.

6♦ 1♠ 7♠ 5♠ 8♦ 9♦ 9♣ 10♣ D♠ 5♣
 6♣ 5♣ V♠ 9♦ 1♠ 7♦ 6♥ 5♠ 4♥ R♣

obtenu en dressant la liste de toutes les façons possibles d'ordonner les 52 cartes : chaque ligne fait apparaître les 52 cartes dans un certain ordre, et il y a autant de lignes que d'ordres possibles. Le nombre des lignes est colossal (c'est un nombre de 68 chiffres) ; l'ordre des lignes elles-mêmes n'a aucune importance, ce qui compte est seulement que chaque ligne possible (chacun des ordres possibles pour les 52 cartes) figure exactement une fois dans le tableau⁶.

Considérons maintenant une stratégie de jeu, fixée une fois pour toutes (par exemple : vous attendez la première carte qui soit un sept noir ou un quatre de cœur, puis vous pariez sur la 3^e carte qui suit). Toute règle de ce type, si compliquée soit-elle, définit une stratégie; la seule contrainte est que la décision, à chaque instant, de parier ou non, ne dépende que des cartes déjà retournées⁷.

⁶Les habitués de la théorie des probabilités ont reconnu dans ce tableau une figuration de l'univers Ω associé à l'expérience aléatoire qu'est le battage du jeu de cartes.

⁷Et non de la vitesse du vent ou de l'humeur du moment. Le résultat resterait vrai pour ces stratégies plus générales, mais le raisonnement serait plus abstrait : il faudrait remplacer notre énumération de tous les cas possibles par le formalisme des filtrations et des temps d'arrêt.

Pour chaque ordre possible des cartes, c'est-à-dire pour chaque ligne du tableau, la stratégie détermine l'instant où vous pariez que la prochaine carte sera rouge. On peut matérialiser la stratégie en insérant dans chaque ligne du tableau une barre  juste avant la carte sur la couleur de laquelle vous pariez ; cela donne quelque chose comme ça :

R♠ 9♣ | 2♥ 3♣ 3♥ V♣ 6♠ 8♦ R♥ 4♣
 2♣ 8♠ R♦ 9♠ | 3♣ 1♦ 10♠ 5♣ 1♥ R♠
 3♥ V♠ 1♥ 5♣ 4♠ 10♠ R♠ ... | .. D♣ 7♣ 10♣
 |

..... |

6♦ 1♠ 7♠ 5♠ 8♦ 9♦ 9♣ 10♣ D♠ | 5♣
 6♣ 5♣ V♠ | . 9♦ 1♠ 7♦ 6♥ 5♠ 4♥ R♣

Une ligne est gagnante (pour la stratégie considérée) si dans cette ligne la carte juste après la barre est rouge, et perdante si cette carte est noire. La probabilité de gagner en appliquant cette stratégie n'est autre que la proportion de lignes gagnantes dans le tableau.

À toute ligne ℓ du tableau, associons une autre ligne ℓ' obtenue à partir de ℓ en échangeant la carte juste après la barre et la dernière carte de la ligne. Par exemple, si ℓ est la deuxième ligne

2♣ 8♠ R♦ 9♠ | 3♣ 1♦ 10♠ 5♣ 1♥ R♠

du tableau de tout à l'heure, ℓ' sera la ligne

2♣ 8♠ R♦ 9♠ R♠ 1♦ 10♠ 5♣ 1♥ 3♣;

et si ℓ est l'avant-dernière ligne

6♦ 1♠ 7♠ 5♠ 8♦ 9♦ 9♣ 10♣ D♠ | 5♣,

où la barre est juste avant la dernière carte, alors ℓ' sera la ligne ℓ elle-même.

Remarquons que toute la partie de la ligne ℓ située avant la barre n'est pas modifiée par cette opération ; donc, puisque la stratégie décide à chaque instant de parier ou non au vu des seules cartes déjà retournées, la barre est placée au même endroit dans les deux lignes ℓ et ℓ' (juste après le 9♠ et la D♠ dans les deux exemples qui précèdent). En conséquence, *si l'on répète l'opération à partir de la ligne obtenue ℓ' , on retrouve la ligne initiale ℓ* . Ainsi, nous pouvons associer les lignes par paires, deux lignes d'une même paire étant échangées par l'opération décrite ci-dessus ; seules les lignes pour lesquelles on parie sur la dernière carte, telles l'avant-dernière du tableau, ne sont pas membres d'une paire : ces lignes célibataires seront considérées comme associées à elles-mêmes.

Il ne reste qu'à observer que si une ligne est gagnante, la ligne qui lui est associée finit par une carte rouge et réciproquement ; et si une ligne est perdante, la ligne qui lui est associée finit par une carte noire et réciproquement. En raisonnant séparément sur les lignes célibataires et les autres, on en déduit qu'il y a exactement autant de lignes gagnantes que de lignes finissant en rouge. La proportion des lignes gagnantes dans la liste est donc aussi la proportion des lignes finissant par une carte rouge, et la probabilité de gagner est ainsi égale à la probabilité pour que la dernière carte tirée soit rouge. Comme cette probabilité-là vaut évidemment un demi, la probabilité de gagner vaut aussi un demi. Nous avons fait le raisonnement pour une stratégie particulière ; il est bien entendu valable pour toutes les stratégies : la probabilité de gain est toujours égale à un demi.

3 Deux enveloppes (première version)

Nous venons de voir un jeu où, en dépit des apparences immédiates, la probabilité de gagner ne dépasse pas 50 %. Voici maintenant, à l'inverse, une situation où, bien qu'il semble à première vue impossible de faire mieux que de gagner une fois sur deux, on peut néanmoins y parvenir.

J'écris deux nombres différents (positifs ou non, entiers ou non) chacun sur une feuille de papier, je les mets sous enveloppe sans vous les montrer, et je vous remets les deux enveloppes. À vous de deviner laquelle des deux contient le nombre le plus grand. Une stratégie évidente, n'utilisant qu'une pièce de monnaie, vous donne une chance sur deux de gagner : choisissez au hasard celle des deux enveloppes sur laquelle vous pariez.

Pour rendre le jeu plus intéressant, modifions la règle : vous avez maintenant le droit, avant de parier, de prendre connaissance du contenu de l'une des deux enveloppes, celle que vous voulez. Cette nouvelle règle vous est-elle plus favorable que la précédente ? En d'autres termes, pouvez-vous tirer parti de la connaissance de l'un des deux nombres pour augmenter, si peu que ce soit, vos chances de gagner ?

Un argument de symétrie (ou de simple bon sens) semble dire que non : celui des deux nombres dont vous prenez connaissance a exactement une chance sur deux d'être le plus grand ; le connaître ne vous donne donc aucun avantage supplémentaire. Cet argument, que nous développerons un peu dans la remarque 1 plus loin, est faux : nous allons voir qu'en réalité vous pouvez mettre à profit l'information supplémentaire qu'est la valeur numérique du nombre lu.

Il vous suffit pour cela de disposer d'une machine à fabriquer du hasard, qui produise un nombre aléatoire x ayant les deux propriétés suivantes :

(1) si a est n'importe quel nombre donné à l'avance, la probabilité pour que x soit exactement égal à a est nulle;

(2) si a et b sont deux nombres différents donnés à l'avance, la probabilité pour que x tombe entre a et b n'est pas nulle.

Dans les propriétés (1) et (2), l'expression « donné à l'avance » signifie précisément : connu avant que la machine ne commence à fabriquer x . Par exemple, la propriété (2) serait évidemment impossible à satisfaire si l'on s'autorisait à prendre $a = x + 1$ et $b = x + 2$.

Un nombre aléatoire x vérifiant (1) et (2) pourrait par exemple être obtenu en demandant à la touche RANDOM d'une calculatrice un nombre aléatoire y entre 0 et 1, puis en posant

$$x = \frac{2y - 1}{y(1 - y)}, \quad \text{ou bien encore} \quad x = \tan(\pi y);$$

mais en raison de la précision limitée de la calculatrice, les deux propriétés ne seront satisfaites qu'approximativement : la calculatrice choisit en fait y parmi un ensemble *fini* (bien que très grand) de nombres, chacun avec probabilité non nulle, et les chances qu'a y de tomber strictement entre deux de ces nombres sont rigoureusement nulles. Pour les lecteurs soucieux de rigueur, nous reviendrons plus bas sur ce point (remarque 2). Notons pour les spécialistes qu'une variable aléatoire tirée selon une loi normale (ou une loi de Cauchy comme c'est le cas pour $\tan(\pi y)$, etc.) vérifie les deux propriétés requises.

Voici une recette permettant d'accroître vos chances de gain :

- Tirez au sort l'une des deux enveloppes et prenez connaissance du nombre z qu'elle renferme.
- Actionnez ensuite votre machine à fabriquer un nombre au hasard pour obtenir un nombre x ; ce nombre x est différent de z en raison de la propriété (1).
- Si z est plus grand que x , pariez que z est aussi plus grand que le nombre caché dans l'autre enveloppe.
- Si au contraire z est plus petit que x , pariez que z est aussi plus petit que le nombre caché dans l'autre enveloppe.

L'estimation de vos chances de gain si vous suivez cette recette n'est pas difficile (en tous cas pas difficile pour moi, qui connais les deux nombres cachés). Appelons a le plus petit des deux nombres mis sous enveloppe, et b le plus grand. Trois cas sont possibles, selon que le nombre x produit par votre machine est plus petit que a , tombe entre a et b , ou est plus grand que b ; appelons p , q et r leurs probabilités respectives (elles vérifient $p + q + r = 1$ et, en raison de la propriété (2), $q > 0$). En tenant compte des deux éventualités que z soit égal à a ou à b , cela fait six possibilités en tout, résumées dans le tableau ci-dessous.

	$x < a$ (probabilité : p)	$a < x < b$ (probabilité : q)	$b < x$ (probabilité : r)
$z = a$ (probabilité : $\frac{1}{2}$)	Vous perdez (probabilité : $\frac{p}{2}$)	Vous gagnez (probabilité : $\frac{q}{2}$)	Vous gagnez (probabilité : $\frac{r}{2}$)
$z = b$ (probabilité : $\frac{1}{2}$)	Vous gagnez (probabilité : $\frac{p}{2}$)	Vous gagnez (probabilité : $\frac{q}{2}$)	Vous perdez (probabilité : $\frac{r}{2}$)

La probabilité de chacun des six cas est obtenue en multipliant la probabilité du comportement de z (égale à $\frac{1}{2}$) par celle du comportement de x (égale à p , q ou r). Ceci exprime l'indépendance entre les deux tirages au sort (d'abord, choix de l'enveloppe avec une chance sur deux, ensuite, choix de x avec probabilités p , q et r pour les trois cas). Puisque les six cases épuisent toutes les possibilités, votre probabilité de gagner est la somme des probabilités des quatre cases gagnantes, c'est-à-dire $\frac{p}{2} + \frac{q}{2} + \frac{q}{2} + \frac{r}{2}$, ou encore $\frac{1+q}{2}$. Et puisque q n'est pas nulle, c'est mieux qu'une chance sur deux.

Évidemment, vous qui connaissez la valeur de z mais pas les valeurs de a et b , vous ne connaissez pas non plus q et vous n'avez aucune idée de vos chances $\frac{1+q}{2}$. Vous pourrez les connaître *a posteriori*, une fois le jeu terminé et la seconde enveloppe ouverte, mais leur valeur aura alors perdu toute signification : le jeu sera fini, vous aurez gagné ou perdu, vos chances de gain seront devenues 1 ou 0, et estimer rétrospectivement ce qu'elles étaient au moment du pari sera dénué d'intérêt. Mais, bien que vous ne puissiez donc pas estimer quantitativement vos chances de gagner⁸, il reste le fait brut que le jeu vous est favorable. Cette affirmation a d'ailleurs un contenu empirique (sans quoi elle n'aurait pas de sens) : imaginez un très grand nombre d'opérateurs, chacun dans une salle équipée d'un générateur aléatoire du même type que le vôtre, et recevant chacun deux enveloppes qui contiennent les mêmes nombres a et b . Chacun, utilisant sa machine, va parier, et gagner ou perdre ; ils ne peuvent pas estimer leurs chances de gain mais ils savent qu'*ils seront plus nombreux à gagner qu'à perdre* pourvu qu'ils soient en nombre suffisant. C'est en ce sens que le jeu vous est favorable. (Et moi qui connais q , je peux en plus prédire la proportion des gagnants $\frac{1+q}{2}$.)

Remarque 1. — Si les deux nombres que j'ai écrits sont

$$a = 173\,261\,372\,804 \times 10^{766\,146\,459\,801} + 0,109\,085\,028\,871 \times 10^{-172\,754\,358\,235}$$

et

$$b = 173\,261\,372\,804 \times 10^{766\,146\,459\,801} + 0,109\,085\,028\,872 \times 10^{-172\,754\,358\,235},$$

il est fort vraisemblable que la machine à hasard que vous choisirez donnera à x une probabilité q évanescence de tomber entre a et b . Le bénéfice apporté par le droit d'ouvrir une des deux enveloppes est donc homéopathique, et la discussion ci-dessus

⁸Sauf à adopter un point de vue bayésien en mettant une loi de probabilité *a priori* sur la façon dont j'ai choisi mes deux nombres. Nous reviendrons sur cet aspect des choses dans la quatrième partie.

purement académique, vous ne pourrez en tirer aucun avantage pratique. Son seul intérêt, tout théorique, réside dans le fait que, si infinitésimal que soit votre gain, sa non-nullité réfute l'argument de symétrie évoqué plus haut. Une fois formalisé, cet argument de symétrie dit ceci : pour vous empêcher de gagner, je pourrais moi aussi recourir à un générateur aléatoire, et fabriquer mes deux nombres au hasard, indépendamment l'un de l'autre, de façon telle que, pour tout a donné à l'avance, chacun des deux ait une chance sur deux d'être plus grand que a et une chance sur deux d'être plus petit. Si j'opérais ainsi, une fois la première enveloppe ouverte il y aurait exactement une chance sur deux pour que l'autre nombre soit plus grand que le nombre lu, et vous n'auriez aucun avantage. Bien sûr, cet argument ne tient pas, car il n'existe aucune loi de probabilité sur la droite possédant la propriété requise (pour n'importe quel nombre $a > 0$, mon nombre aléatoire devrait avoir 50 % de chances d'être plus grand que a et 50 % de chances d'être plus petit que $-a$, donc aucune chance d'être entre $-a$ et a).

Remarque 2. — La recette pour améliorer vos chances nécessite un générateur aléatoire; et nous avons remarqué tout à l'heure qu'utiliser brutalement les nombres aléatoires fournis par une calculatrice n'est pas pleinement satisfaisant. Voici, parmi mille autres, une méthode possible pour engendrer un nombre aléatoire x vérifiant les propriétés (1) et (2).

Jouez à pile ou face de façon répétée jusqu'à obtenir pour la première fois pile; ceci nécessite un certain nombre $n \geq 1$ de lancers. Tirez maintenant au sort, d'abord un signe $+$ ou $-$, puis, au moyen d'une roulette à dix cases numérotées de 0 à 9, une série de chiffres que vous écrivez à droite du signe, en plaçant une virgule après le n -ième chiffre. C'est tout! Si vous souhaitez connaître x exactement, cette opération demandera un temps infini, et vous avez sans doute mieux à faire; mais rappelez-vous qu'au moment où vous fabriquez x , vous connaissez déjà z et que la seule chose qui vous importe est de savoir si x est plus petit ou plus grand que z ; vous pouvez donc arrêter la procédure au bout d'un nombre fini (aléatoire) d'étapes. (Connaître n , le signe, et le premier chiffre non nul sera suffisant dans l'immense majorité des cas.) Bien entendu, tous ces tirages au sort pourraient être simulés par une calculatrice, et la procédure est entièrement automatisable; mais je n'en vois pas l'intérêt : tout ceci n'est qu'une expérience de pensée, un tel « jeu » n'amuserait évidemment personne.

4 Deux enveloppes (seconde version)

Voici pour finir un autre jeu à deux enveloppes, beaucoup plus connu ⁹, semble-t-il, que le précédent.

Je mets sous enveloppe deux chèques à votre nom, le montant de l'un étant le double de l'autre. Vous tirez au sort l'une des deux enveloppes, vous l'ouvrez, puis vous devez choisir celui des deux chèques que vous allez toucher. Vous ne connaissez donc que le montant m de l'un des chèques, et vous savez que l'autre vaut le double ou la moitié de m . Quel choix faites-vous?

⁹Il est également discuté par J.-P. Delahaye dans l'intéressant article Désespérante espérance, *Pour la Science* 205 (1994).

La présentation faite de ce problème est souvent accompagnée de la discussion suivante, en deux points :

1. Appelons c le chèque de valeur m que vous avez trouvé en ouvrant l'enveloppe, et i le chèque de montant inconnu $2m$ ou $\frac{m}{2}$ qui est dans l'autre enveloppe. Si vous choisissez i , vous avez une chance sur deux d'encaisser le montant $2m$ et une chance sur deux d'encaisser $\frac{m}{2}$. Votre espérance de gain est alors $\frac{1}{2} \times 2m + \frac{1}{2} \times \frac{m}{2}$, c'est-à-dire $\frac{5}{4}m$. Cette quantité est toujours supérieure à m ; ainsi, en choisissant i plutôt que c , vous augmentez votre espérance de gain, ceci quelle que soit la valeur de m lue.
2. La conclusion ci-dessus est manifestement absurde. La stratégie préconisée au point précédent — tirer au sort l'une des enveloppes, l'ouvrir, puis, quel que soit son contenu (donc, en fait, sans même avoir à l'ouvrir!), choisir systématiquement l'autre enveloppe — n'est qu'un procédé un peu détourné pour tirer au sort, avec chances égales, celle des enveloppes qui sera choisie. Mais c'est aussi ce que fait la stratégie qui consiste à toujours choisir c ; la stratégie proposée en 1, systématiquement choisir i , n'est donc finalement pas meilleure que de choisir systématiquement c .

Ces deux arguments démontrent une chose et son contraire. Où est le vrai? À quel endroit y a-t-il une erreur?

Tout d'abord, une remarque s'impose à propos du premier argument. Il y est implicitement admis qu'entre deux stratégies, celle qui procure la plus grande espérance de gain est préférable. Non seulement ceci n'a rien d'évident, mais c'est en fait contredit par de nombreuses théories ou observations. Comme ce qui nous intéresse ici est l'aspect mathématique, nous allons esquiver cette difficulté en décidant *a priori* du critère de choix et en remplaçant la question initiale « Quel choix faites-vous? » par « Quel choix maximise votre espérance de gain? » Malheureusement l'énoncé perd alors en signification ce qu'il gagne en précision : prononcer le mot *espérance* ne suffit pas à lui donner un sens. L'espérance est obtenue en pondérant les gains possibles par les probabilités de ces gains; on est ainsi ramené à la question du sens à donner, dans l'argument 1, aux probabilités $\frac{1}{2}$ et $\frac{1}{2}$ de trouver $2m$ et $\frac{m}{2}$ à l'ouverture de la seconde enveloppe. La discussion du jeu précédent a justement mis en évidence (dans une situation différente, mais c'est vrai ici aussi) qu'une fois connu le contenu d'une enveloppe, on ne peut plus affirmer que l'autre a une chance sur deux d'être plus grand.

Dès lors, deux attitudes sont possibles. Vous pouvez déclarer que les probabilités en cause ne sont pas quantifiables, que maximiser l'espérance est donc vide de sens, et que vous ferez votre choix selon d'autres critères (rationnels ou non). C'est une position parfaitement tenable, parce qu'il n'est pas possible de donner un sens objectif à ces probabilités. C'est l'attitude sécuritaire : vous êtes dans une position retranchée, très facile à défendre contre toute critique.

Plus hasardeuse (c'est le cas de le dire!), mais aussi plus constructive, l'autre attitude permet de faire avancer la discussion, par le calcul d'une espérance de gain qui vous servira de critère de choix. C'est la méthode appelée *bayésienne*; elle consiste à admettre que le choix fait par moi des montants inscrits sur les

chèques est assimilable au résultat d'une expérience aléatoire, dont vous connaissez les probabilités des résultats. Concrètement, avant même d'ouvrir la première enveloppe, vous dressez la liste ¹⁰ de tous les montants couplés $(x, 2x)$ que je pourrais avoir choisis, vous affectez à chacun une probabilité ¹¹ $p(x, 2x)$ de façon que la somme des probabilités sur toute la liste fasse 1, et vous admettez que mes chèques sont rédigés par un générateur aléatoire qui écrit x et $2x$ avec probabilité $p(x, 2x)$. Il ne reste qu'à résoudre le petit exercice suivant : une paire aléatoire $(x, 2x)$ étant produite par ce générateur, on tire ensuite au sort entre x et $2x$ pour obtenir un montant (doublement) aléatoire m . Une fois observée la valeur de m , quelles sont les probabilités ¹² p' et p'' pour que l'autre nombre de la paire soit respectivement $\frac{m}{2}$ ou $2m$? Le calcul est élémentaire et la réponse bien intuitive :

$$p' = \frac{p(\frac{1}{2}m, m)}{p(\frac{1}{2}m, m) + p(m, 2m)} ; \quad p'' = \frac{p(m, 2m)}{p(\frac{1}{2}m, m) + p(m, 2m)}$$

(en posant $p(x, 2x) = 0$ si $(x, 2x)$ ne figure pas dans la liste, ce qui se produit par exemple lorsque x est le plus grand ou $2x$ le plus petit des montants admissibles). En reprenant l'argument 1 ci-dessus avec les probabilités p' et p'' au lieu de $\frac{1}{2}$ et $\frac{1}{2}$, votre espérance de gain se calcule immédiatement, et, en la comparant à m , il est finalement avantageux, indifférent, ou désavantageux de choisir i selon que $p(m, 2m)$ est supérieur, égal, ou inférieur à $\frac{1}{2}p(\frac{1}{2}m, m)$. La stratégie optimale ainsi obtenue fait explicitement intervenir le montant observé m . Dans cette situation, le dilemme entre les arguments 1 et 2 s'évanouit : il n'est pas vrai que choisir i soit toujours préférable à choisir c , comme le prétendait l'argument 1 ; et l'argument 2, qui prenait comme prémisse cette conclusion, n'a plus de raison d'être, rien ne permettant d'affirmer que les deux choix i et c soient indifférents — cela n'est vrai que lorsque le montant m observé se trouve vérifier l'égalité $p(m, 2m) = \frac{1}{2}p(\frac{1}{2}m, m)$. Le point de vue bayésien est puissant et séduisant ; sa faiblesse réside dans la nécessité de choisir préalablement les $p(x, 2x)$, choix dont dépendra ensuite la stratégie suivie, et qui prête le flanc à une critique : pour calculer les probabilités p' et p'' , on part d'autres probabilités $p(x, 2x)$, que l'on a au préalable estimées. Mais tant qu'à estimer des probabilités, on pourrait attendre d'avoir lu le premier chèque, et estimer à ce moment là directement les chances p' et p'' pour que l'autre chèque soit moitié ou double. L'avantage de la méthode bayésienne sur l'estimation directe de p' et p'' est qu'elle évite tout paradoxe du type arguments 1 et 2 ci-dessus, ce qui n'est pas le cas de l'estimation directe (en prenant par exemple $p' = p'' = \frac{1}{2}$ sans tenir compte de la valeur de m , on retomberait évidemment en plein dans le paradoxe précédent).

¹⁰Cette liste est finie parce que le montant de mes chèques est un nombre entier de cents ou de centimes, et que mon compte en banque n'est pas illimité.

¹¹Appelée probabilité a priori.

¹²Appelées probabilités a posteriori.