Des statistiques aux probabilités Un lieu privilégié pour la modélisation

Jean-Claude DUPERRET
APMEP, IREM et IUFM Champagne-Ardenne

Introduction

On est passé, en 30 ans, d'un enseignement dit de **structure** à un enseignement dit de **modélisation**, sans que cette évolution n'ait été clairement explicitée. Cela renvoie à la question bien ambitieuse de la modélisation, surtout lorsqu'on la pose dans le cadre des mathématiques. Si la plupart des autres disciplines scientifiques ont pour objet de **décrire** et de **modéliser** un point de vue du *monde réel*, point de vue différent suivant ces disciplines, comment les mathématiques peuvent-elles s'inscrire dans ce rapport au monde réel ? Les mathématiques ont-elles pour objet de *décrire* la réalité, ou ne se contentent-elles pas d'une action intellectuelle sur une réalité déjà abstraite ? Qu'est-ce qu'un modèle mathématique ? Y a-t-il unicité du modèle pour traduire une *réalité*, ou celui-ci n'est-il pas lié à *l'intention* de modélisation ? En quoi la connaissance du modèle permet-elle *d'éclairer* la réalité, voire de l'expliquer et d'avoir une attitude *opérationnelle* et *décisionnelle* ?

On peut choisir cet angle de la modélisation pour *revisiter* des notions que nous construisons dans nos classes. En effet, cette approche me paraît bien donner la philosophie de ce que devrait être un enseignement de mathématiques pour tous : donner un outil de pensée du monde dans lequel nous vivons en nous appuyant sur un processus intellectuel de description, d'investigation, d'action et de validation. C'est ce que je propose dans deux articles intitulés « *De la modélisation du monde au monde des modèles* » parus dans les numéros 484 et 486 du Bulletin Vert de l'APMEP, et que je reprends en très grande partie ici pour ce qui concerne les statistiques et les probabilités.

Mais le passage des statistiques aux probabilités est certainement le lieu privilégié dans notre enseignement pour illustrer ce qu'est un processus de modélisation, et c'est pourquoi j'ai choisi de développer plus particulièrement en fin d'article la « *situation des spaghettis* » car elle fut pour moi l'occasion de me confronter à cette question à un moment où mes connaissances et pratiques dans ce domaine étaient pour ainsi dire nulles !

Les situations que je proposerai pour éclairer mon propos vont de l'école au post-bac. Je les ai toutes vécues ou proposées, que ce soit dans le cadre de l'enseignement ou celui de la formation initiale ou continue des professeurs de mathématiques (voire professeurs des écoles pour certains).

Cet article reprend aussi en grande partie une conférence que j'ai faite lors du colloque : *« Les dés sont-ils à jeter ? »* organisé par les commissions Inter-IREM Collège et Statistique Probabilités, en juin 2008, à Périgueux (voir les actes correspondants).

1. **La modélisation**

1.1. Modélisation

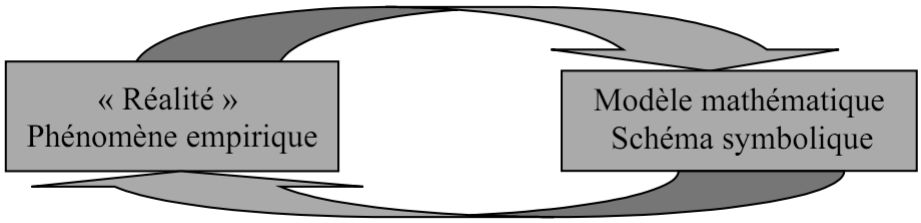
Comme je l'ai dit dans mon résumé, le mot **modélisation** a un sens lié aux différentes disciplines scientifiques. Pour ma part, dans le cadre d'un enseignement des mathématiques pour tous, j'utiliserai cette notion de modélisation comme un processus de « re-présentation » de situations d'une certaine *réalité* dans un modèle mathématique, « re-présenter » étant pris au sens de présenter cette situation avec une nouvelle description liée au modèle choisi.

Et j'attacherai à ce processus de représentation trois spécificités :

- représentation **fonctionnelle** des objets d'une certaine *réalité* par des objets *abstraits* ou *schématisés* dans un modèle où peut s'exercer un traitement théorique ;
- représentation **analogique** ou **métaphorique** : les processus naturels sont *imités* dans des conditions qui favorisent l'observation et l'étude ;
- représentation **sélective** : un travail de modélisation nécessite de *retenir* certaines caractéristiques de la situation et d'en *ignorer* d'autres.

1.2. Modélisation et modèle

Ce processus de modélisation s'illustre par le schéma ci-dessous :



Ce schéma fonctionne dans les deux sens :

- du réel vers le modèle : modèles descriptifs (transformer et interpréter des informations) ; ce sens correspond à une fonction **heuristique** ;
- du modèle vers le réel : modèles prédictifs (anticiper une action) ; ce sens correspond à une fonction **justificative**.

2. Modéliser l'information… ou « un modèle peut en cacher un autre »

2.1. Deux modèles

2.1.1. Miracle

Voici un article paru dans *Le Canard enchaîné* en juillet 2006 :

Sous le titre « Un bon cru au bac », « La République des Pyrénées » (13/7) s'extasie devant les résultats de la bonne ville de Lourdes : « 96 % de mentions très bien, bien et assez bien ». « Du jamais vu ! » Mazette ! La cité mariale serait-elle un paradis pour les surdoués ? En réalité, pour obtenir ces mirobolants 96 %, le confrère a eu un recours à un calcul simple. Il a ajouté le pourcentage du lycée public de La Serre de Sarsan (« toutes mentions confondues », 50 %), à celui du lycée privé Peyramale (« 46 % de mentions »). En additionnant ces deux nombres, il faudrait donc compter « 96 % de mentions » à Lourdes. Et ce n'est pas fini. Car un troisième lycée de la ville n'ayant pu être comptabilisé, la part de ces mentions au bac devrait, selon cette nouvelle arithmétique, dépasser largement les 100 %. Lourdes, ville de tous les miracles !

On peut dire que l'auteur de l'article incriminé maîtrise bien l'addition, mais qu'il s'est trompé de modèle !

2.1.2. Deux grands modèles dans l'enseignement

Deux grands modèles vont se construire entre l'école et le collège, avec leurs modes de traitement et de calcul spécifiques :

- le modèle **additif** : comparaison absolue ;
- le modèle **proportionnel** : comparaison relative.

Et, comme l'auteur de l'article ci-dessus, nos élèves vont faire des confusions entre ces deux modèles, comme en témoignent ces situations vécues dans mes classes.

2.1.3. Redoublants et doublements

Quand je proposais à mes élèves de sixième les résultats suivants sur le nombre de redoublants en troisième dans deux collèges de l'agglomération troyenne (chiffres fictifs) :

- *Albert Camus : 15 redoublants en 3e ;*
- *Paul Langevin : 12 redoublants en 3e.*

Leur première réaction était de dire que le collège Langevin était meilleur que le collège Camus, puisqu'il y avait moins de redoublants.

C'était une occasion de leur faire comprendre ces notions de comparaison **absolue** et **relative**, en leur proposant de calculer le taux de doublement, avec l'information ci-dessous sur les populations de référence :

- *Albert Camus : 125 élèves en 3e ;*
- *Paul Langevin : 80 élèves en 3e.*

Et ainsi de leur faire constater que leur conclusion s'inversait :

- taux de doublement à A. Camus : 12 % ;
- taux de doublement à P. Langevin : 15 %.

2.1.4. Plus de garçons ou de filles ?

Je leur proposais alors la situation suivante :

Dans une petite ville, tous les élèves de collège sont scolarisés dans l'un des deux collèges suivants, avec la proportion de garçons et filles correspondante :

- Pierre Brossolette : 45 % de garçons, 55 % de filles ;
- Gaston Bachelard : 60 % de garçons, 40 % de filles.

À ma question : « *Y a-t-il plus de garçons ou de filles scolarisés en collège dans cette ville ?* », il se trouvait toujours un certain nombre d'élèves pour répondre qu'il y avait plus de garçons, puisqu'il y en avait 105 % contre 95 % de filles !

C'était alors l'occasion de leur montrer que suivant les populations de référence, on pouvait aboutir à trois conclusions différentes :

- à Brossolette : 420 élèves ; à Bachelard : 360 élèves.

Soit 405 garçons et 375 filles.

- à Brossolette : 520 élèves ; à Bachelard : 260 élèves.

Soit 390 garçons et 390 filles.

- à Brossolette : 740 élèves ; à Bachelard : 300 élèves.

Soit 513 garçons et 527 filles.

Dans un monde d'information chiffrée comme le nôtre, pour armer nos élèves dans leur vie de futur citoyen, développer cette confrontation entre ces deux modèles me paraît fondamental.

2.1.5. Le modèle proportionnel

Le modèle proportionnel est particulièrement riche, avec la diversité des registres de représentations possibles d'une même situation :

- ***Registre numérique*** : suites proportionnelles, tableaux, règle de trois… Ce registre est celui de l'entrée dans ce modèle à l'école primaire.
- ***Registre algébrique*** : « $y = kx$ », propriétés de linéarité… On trouve ce registre dès l'école primaire avec l'utilisation en acte des propriétés de linéarité.
- ***Registre fonctionnel*** : application linéaire, traduction graphique… Ce registre est plus spécifique du collège.
- ***Registre géométrique*** : théorème de Thalès, lien entre parallélisme et proportionnalité… Là encore ce registre est présent en acte dès l'école primaire, avec par exemple le guide-âne au cycle 3 (réseau de parallèles qui permet de lire ou représenter des fractions).

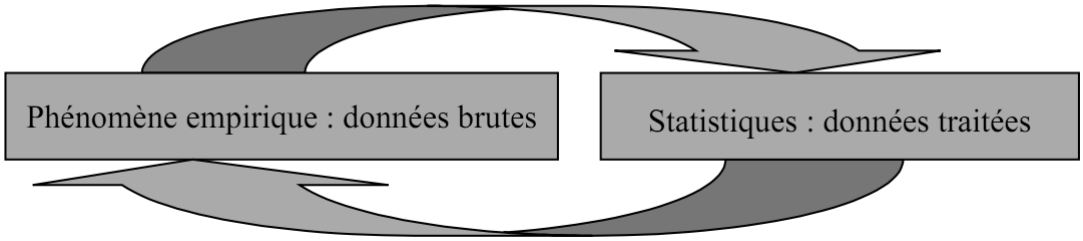
La conceptualisation, le traitement et la validation vont être spécifiques à chacun de ces registres. Notre enseignement doit à la fois travailler au maximum ces spécificités ainsi que les conversions d'un registre à l'autre pour donner des éclairages complémentaires d'une même situation.

2.2. *Les Statistiques*

2.2.1. La statistique, une pratique très ancienne !

Beaucoup d'activités humaines, comme le commerce, reposent sur l'observation de données, et en ce sens, on peut dire que la statistique est une discipline expérimentale que pratique l'homme depuis très longtemps, avec un objectif de prévision, comme la gestion de stock dans le commerce. Et, pour ce faire, l'homme a toujours essayé de « limiter » le hasard. Et pourtant ce hasard est omniprésent dans nos croyances et souvent dans nos décisions… comme nous le verrons plus loin.

2.2.2. La statistique descriptive



La statistique constitue le modèle mathématique de traitement de l'information. Cette modélisation présente, là encore, un aller-retour entre le monde réel et le monde mathématique comme l'illustre le schéma ci-dessus.

De la réalité vers les mathématiques, les statistiques vont transformer les données brutes en les représentant de façon **classée** pour pouvoir en faire des **résumés**.

En sens inverse, ces résumés vont conduire à des **interprétations** du phénomène empirique. Comprendre cette transformation synthétique des informations, pouvoir l'analyser correctement et donc prudemment, sont des enjeux d'une formation de l'individu dans la société. C'est pourquoi on caractérise souvent leur enseignement par le vocable « mathématiques du citoyen », c'est-à-dire :

- une formation à l'analyse des données et au traitement de l'information ;
- un développement des aptitudes à trier, ranger, transformer des informations, critiquer un traitement ;
- en s'appuyant sur de fréquents changements de registre : texte, tableau, graphique, résultat numérique…

De manière plus précise, il faut faire comprendre aux élèves que le problème fondamental de la statistique descriptive est de résoudre le dilemme résultant de la transformation de données *brutes* en une *synthèse* qui parvienne à concilier le mieux possible deux pôles antagonistes : la **fidélité** et la **clarté** (voir article « *Heurs et malheurs du su et du perçu en statistique* » de Bernard Parzysz, dans Repères-IREM, n°35).

2.2.3. Une question épineuse

À partir d'une question qui se pose dans le monde « réel », un premier travail sera de préciser cette question en vue de mettre en place un protocole d'observation. Un passage par les statistiques va alors permettre un traitement mathématique conduisant à proposer une réponse à cette question.

Prenons la question : « *Les garçons sont-ils meilleurs en maths que les filles ?* ». Précisons-la : « *Les garçons réussissent-ils mieux en maths que les filles ?* »

Pour répondre à cette question, on propose de faire passer un test à 700 garçons et 600 filles de troisième d'une petite ville de province (en prenant comme hypothèse que cet échantillon est

représentatif !). On leur laisse le choix de passer ce test en algèbre ou en géométrie.

Voici les résultats à ce test, c'est-à-dire le nombre d'élèves qui ont réussi (avec par exemple comme indicateur une note supérieure ou égale à 10).

	Garçons	Filles
Algèbre	$\frac{23}{200}$	$\frac{85}{500}$
Géométrie	$\frac{400}{500}$	$\frac{90}{100}$

Les filles sont meilleures que les garçons !

Pour arriver à cette conclusion, on calcule le pourcentage respectif de réussite des garçons et des filles en algèbre et en géométrie :

	Garçons	Filles
Algèbre	11,5 %	17 %
Géométrie	80 %	90 %

Les filles sont « meilleures » (c'est-à-dire : ont mieux réussi) à la fois en algèbre et en géométrie… donc elles sont meilleures en maths.

Quoique !

On regroupe maintenant les résultats pour établir le pourcentage de réussite au test :

	Garçons	Filles
Total	$\frac{423}{700}$	$\frac{175}{600}$
En %	60,5 %	29,2 %

On arrive à la conclusion contraire : les garçons sont « meilleurs en maths » (c'est-à-dire qu'ils ont mieux réussi au test).

On joue ici sur un effet de structure des sous-populations, mais le fait qu'un traitement statistique d'un même problème puisse conduire à deux réponses opposées pose à la fois la question de la complexité du modèle et celle de la fiabilité des réponses pour une personne non avertie de ces subtilités.

2.2.4. Les « nombres » dans la société

Nous vivons dans un monde d'informations baigné de pourcentages et le citoyen peut avoir beaucoup de peine à s'y repérer, à la fois par le manque de référence aux populations et aussi parce qu'avec les mêmes données, on peut arriver à deux conclusions contradictoires comme ci-dessus. Les statistiques apparaissent alors au mieux comme une science de la manipulation, au pire comme une science du mensonge, comme en témoignent les trois citations ci-dessous.

Interprétation manipulatoire des résumés du modèle :

« *Il existe trois degrés dans le mensonge : les mensonges, les affreux mensonges, et les statistiques.* »

(Benjamin DISRAELI)

Rétention d'une partie de l'information :
« *Les statistiques, c'est comme le bikini, ça donne une idée, mais ça cache l'essentiel.* »
(Louis ARMAND)

Caution intellectuelle :
« *Les statistiques sont formelles : il y a de plus en plus d'étrangers dans le monde.* »
(Pierre DESPROGES)

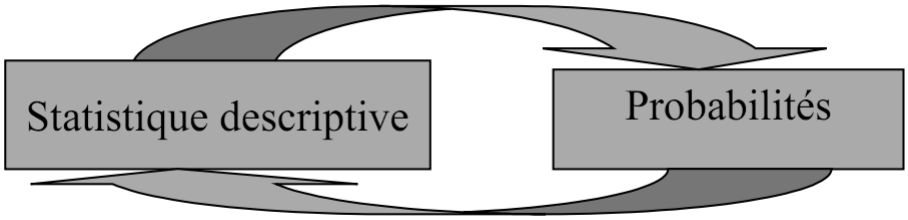
3. Le monde de l'incertitude : le modèle probabiliste

Nous vivons dans un monde empreint de hasard et d'incertitude. Comment le modéliser pour, à partir de données représentées et traitées, donner des conclusions **vraisemblables** et **probables** comme outils d'aide à la décision ? La réponse mathématique repose sur les probabilités, en nous amenant à nous interroger sur la pertinence du modèle choisi, sur la fiabilité des affirmations qu'on peut produire à partir de cette modélisation, sur l'interprétation qu'on peut en tirer.

3.1. Des statistiques aux probabilités

La statistique descriptive est une première mathématisation et donc une première abstraction du monde. Les probabilités vont offrir une modélisation de cette « réalité abstraite » intégrant et mathématisant une dimension fondamentale du monde de l'incertitude : le hasard. Le travail de modélisation va consister en un aller-retour entre les statistiques et les probabilités, répondant aux critères suivants :

- Modéliser une expérience aléatoire, c'est lui associer une loi de probabilité.
- Les distributions de fréquences varient, mais le modèle est un invariant.
- La réalité est liée à la notion de variabilité, la modélisation a pour objectif de dégager ce qu'il y a d'intelligible et de prévisible dans cette variabilité.



Données observées	Données calculées
Résultats empiriques	Résultats théoriques
Distribution de fréquences	Loi de probabilité
Moyenne empirique	Espérance théorique
…	…

Le schéma ci-dessus traduit ce passage des statistiques aux probabilités, processus qui constitue ce qu'on appelle les statistiques **inférentielles** ou **inductives**.
Ce processus comprend trois étapes :

- On cherche, dans un premier temps, à modéliser les données expérimentales observées.
- Un modèle mathématique ayant été choisi, on peut recréer des données calculées.
- Un test peut alors permettre de mesurer l'adéquation du modèle choisi aux données observées.

3.1.1. Des statistiques aux probabilités : un problème historique

Le prince de Toscane demande à Galilée (1554-1642) pourquoi, lorsqu'on lance 3 dés, alors que les nombres 9 et 10 ont autant de décompositions (à savoir 6) en somme de 3 nombres compris entre 1 et 6, obtient-on plus souvent 10 que 9 ?

« *… bien que le 9 et le 12 se composent en autant de façon que le 10 et le 11, si bien qu'ils devraient être considérés comme ayant la même probabilité, on voit néanmoins que la longue observation a fait que les joueurs estiment plus avantageux le 10 et le 11 plutôt que le 9 et le 12.* » (Galilée, œuvres)

9	**10**
1+2+6	1+3+6
1+3+5	1+4+5
1+4+4	2+2+6
2+2+5	2+3+5
2+3+4	2+4+4
3+3+3	3+3+4

Voici comment procède Galilée pour traiter ce problème : « *Les sorties des trois dés sont au nombre de six fois 36, soit 216, toutes différentes. Mais, puisque les sommes des tirages des trois dés ne sont qu'au nombre de 16, c'est-à-dire 3, 4, 5… jusqu'à 18, entre lesquelles on a à répartir les dites 216 sorties, il est nécessaire que pour quelques-unes de ces sommes on ait beaucoup de sorties et, si nous trouvons combien on a pour chacune, nous aurons ouvert la voie pour découvrir ce que nous cherchons* ». Il considère alors le nombre de triplets présentant chacune des 3 éventualités suivantes : 3 points égaux (1 manière pour chaque cas), 2 points égaux et un 3^{e} différent (3 manières pour chaque cas), 3 points différents (6 manières pour chaque cas), et il additionne les cas correspondant à chacune des 6 décompositions de 9 et de 10. Il constate alors qu'il en obtient 25 pour 9 et 27 pour 10, ce qui conduit à 11,85 % de « chance » pour le 9, contre 12,5 % pour le 10.

Cette modélisation repose implicitement sur l'équiprobabilité de sortie des faces des dés et sur l'indépendance des jets de dés.

Mais ce qui m'a toujours fasciné, c'est le fait que le prince de Toscane ait pu émettre cette conjecture, alors que les résultats mathématiques sont si proches. Combien de fois avait-il dû jouer (ou regarder jouer) à ce jeu-là ! Et, d'une certaine manière, il pressentait la « loi des grands nombres ».

3.1.2. Des probabilités aux statistiques : anniversaires et football

Un calcul probabiliste montre que la probabilité que, dans un groupe, deux personnes aient le même jour anniversaire, devient supérieure à son contraire à partir d'un groupe de 23 personnes. Ce calcul repose sur l'équiprobabilité des jours d'une année pour les naissances.

Où trouver 23 personnes ? Sur un terrain de football, avec les 22 joueurs et l'arbitre (sans les juges de touche) !

Deux britanniques, Robert Matthews et Fiona Stones (*Teaching Statistics*, 1998), ont ainsi voulu vérifier de façon expérimentale ce résultat, en s'intéressant aux matchs de première division du Royaume-Uni joués le 19 avril 1997 : sur 10 rencontres, 6 présentaient une coïncidence (2 personnes nées le même jour de l'année) et 4 aucune.

On est ici dans la démarche inverse : contrôler un calcul probabiliste par une statistique. On mesure ici la fonction « justificative » du modèle vers la réalité et la formidable capacité d'anticipation qu'elle donne : sachant qu'il y a 365 jours dans une année, qui irait parier sur cette coïncidence dans les tribunes d'un stade ? … sinon les mathématiciens !

3.1.3. Quelle approche des probabilités ?

Les deux exemples ci-dessus montrent bien la double approche historique des probabilités.

L'approche laplacienne ou **déterministe**, qu'on trouve résumée dans le premier principe de l'« *Essai philosophique sur les probabilités* » de Laplace :

« *Le premier de ces principes est la définition même de la probabilité qui, comme on l'a vu, est le rapport du nombre des cas favorables à celui de tous les cas possibles.* »

Cela suppose évidemment les divers cas également possibles (principe 2) et renvoie à la notion d'évènements élémentaires équiprobables en nombre fini. On parle alors de probabilité *a priori*.

L'approche fréquentiste développée par Jacques Bernoulli dans son « *Ars Conjectandi* », qui repose sur l'observation stabilisée de la fréquence d'un évènement dans une longue série d'expériences répétées « à l'identique ». Celle-ci repose sur la loi des grands nombres.

On parle alors de probabilité *a posteriori*.

Les exemples que je vais développer ci-dessous essaieront de faire le lien entre ces deux approches.

J'aborderai aussi, à l'occasion d'un exemple, la théorie ensembliste des probabilités de Kolmogorov, qui repose sur la théorie de la mesure.

3.2. Des pistes pour travailler l'aléatoire au cycle 3

Que ce soit dans le domaine des statistiques ou dans celui des probabilités, notre enseignement français accuse un certain retard par rapport à d'autres pays, en particulier les pays anglo-saxons.

Et pourtant, dès l'école primaire, on peut mener des activités mettant en jeu l'aléatoire, comme en témoigne la situation ci-dessous, proposée par Claudine Schwartz et Catherine Houdement.

3.2.1. Un exemple : Qui peut le plus ?

Règles du jeu

6	**4**
3	**2**
4	**5**
141	

Les élèves ont chacun la grille vide ci-contre. Ils sont par deux et lancent un dé à tour de rôle.

Au premier lancer de dé, chacun des deux élèves choisit de placer le nombre obtenu dans une des deux cases de la première ligne (à gauche ou à droite).

Au second lancer, chaque élève place alors le nouveau nombre obtenu dans la case restée vide de la première ligne.

Ils recommencent pour la deuxième et la troisième ligne.

Ils additionnent les trois nombres à deux chiffres obtenus et mettent le résultat dans la case du bas.

Objectif

Il s'agit d'obtenir le plus grand nombre possible dans la case inférieure.

3.2.2. Du hasard aux stratégies

Nous avons pu observer des classes de CM2 sur cette activité. Si, dans un premier temps, ils placent les chiffres un peu au hasard, très vite se dégagent :

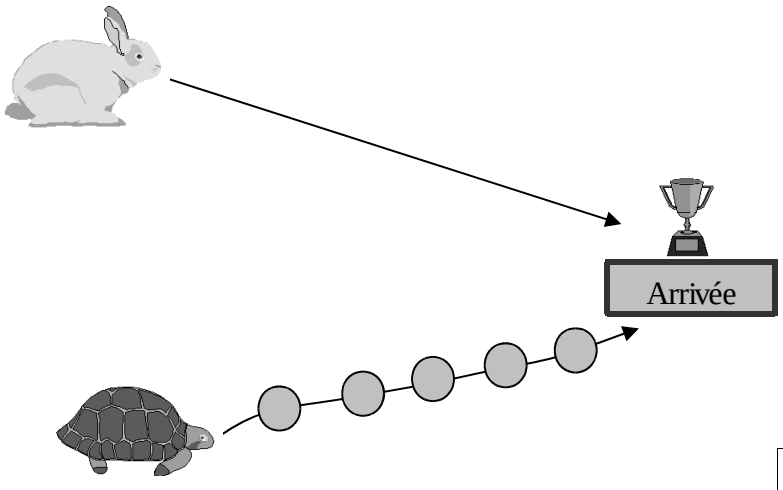
- des stratégies toujours gagnantes : le 6 à gauche et le 1 à droite (ex : le premier tirage est un 6 et il est mis à gauche).
- des stratégies fortement gagnantes : le 5 à gauche et le 2 à droite (ex : le troisième tirage est un 2 et il est mis à droite).
- des stratégies plus souvent gagnantes : le 4 à gauche et le 3 à droite (ex : le cinquième tirage est un 4 et il est mis à gauche… On ne peut pas gagner à tous les coups !

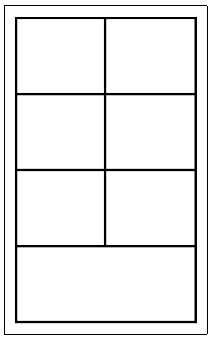
S'appuyant sur l'hypothèse de l'équiprobabilité de sortie des faces d'un dé, ces élèves dénombrent les évènements favorables. On peut donc s'appuyer sur cette conceptualisation précoce pour sensibiliser les élèves à l'entrée mathématique dans le monde de l'aléatoire.

3.3. Des pistes pour travailler l'aléatoire du collège au lycée

3.3.1. Un premier exemple : le lièvre et la tortue

Cette activité a été proposée lors des nouveaux programmes de seconde de 2000 par Claudine Schwartz. On la retrouve dans les documents d'accompagnement (Eduscol). Son adaptation au collège pose problème car elle comporte plus de deux épreuves. Pour ma part, n'ayant plus d'élèves, je l'ai menée en formation initiale (PLC2 maths) ou en formation continue.





Règle du jeu :

On jette un dé.
- Si le dé tombe sur 1, 2, 3, 4 ou 5, la tortue avance d'une case.

Elle a 5 cases à franchir avant d'atteindre l'arrivée ; la dernière case est l'arrivée, contrairement à ce que suggère le dessin ci-dessus (issu des documents d'accompagnement des programmes de 2nde de 2000).
La partie est alors terminée, la tortue a gagné.
- Si le dé donne 6, le lièvre atteint directement l'arrivée.
La partie est alors terminée, le lièvre a gagné.
Quelle est la situation la plus enviable : celle du lièvre ou celle de la tortue ?

Les élèves (professeurs stagiaires pour ma part) font un certain nombre de parties puis on regroupe les résultats. Sur un grand nombre de parties, cela donne environ 40 % de parties gagnées par la tortue (donc environ 60 % pour le lièvre).

Cette expérience permet de montrer une certaine stabilisation des résultats en cumulant les échantillons (loi des grands nombres).

On peut alors *modéliser* mathématiquement le résultat par un calcul probabiliste (produit des probabilités sous l'hypothèse d'indépendance des lancers).

$$p(T)=\left(\frac{5}{6}\right)^5\approx 0{,}40 \quad \text{d'où} \quad p(L)\approx 0{,}60$$

Lorsqu'on propose cette activité, il est toujours bon de demander les pronostics de chacun avant de commencer l'expérience. Il apparaît alors qu'un certain nombre pense que le lièvre et la tortue ont autant de chances, sachant que la tortue a 5 numéros favorables et 5 cases, et le lièvre un numéro favorable et une case. On voit que les modèles *additifs* et *multiplicatifs* se télescopent encore.

On peut alors poursuivre cette réflexion en diminuant le nombre de cases pour la tortue.

- Si la tortue a 4 étapes à franchir :

  $p(T)=\left(\frac{5}{6}\right)^4\approx 0{,}48$ d'où $p(L)\approx 0{,}52$; le lièvre est toujours favori.

- Si la tortue a 3 étapes à franchir :

  $p(T)=\left(\frac{5}{6}\right)^3\approx 0{,}58$ d'où $p(L)\approx 0{,}42$; la tortue est enfin favorite.

3.3.2. Vous avez dit « hasard »

Une première conception répandue du hasard répond à un sentiment de « justice » et, pour cela, il doit être **proportionnel**. C'est le cas de tous les jeux de hasard comme le loto. Nul n'admettrait qu'il ne soit pas régi par une loi d'équiprobabilité. Mais cela se traduit par une grande confusion entre statistiques et probabilités, via une intuition de la loi des grands nombres, comme en témoignent les journaux spécialisés : « Le 7 est en bonne forme ; le 14 devrait rattraper son retard ; le 18 est en période noire ; jouez l'outsider, le 49 ! » Cela va plus loin : qui oserait jouer 1, 2, 3, 4, 5, 6 ? Cette grille ne paraît pas « normale » au niveau du hasard. Il faut cependant noter que, si toutes les grilles ont la même chance, le gain est, lui, lié au choix de la grille des autres joueurs. On touche ici à la notion de variable aléatoire : le vrai enjeu du loto n'est pas la chance de gagner, mais de gagner beaucoup ! Les renseignements statistiques, que se refuse évidemment à donner La Française des Jeux, pourraient aider à cette finalité.

Une seconde conception du hasard répond à un sentiment de « fatalité » : c'est la reproduction d'évènements dont la probabilité est très faible, généralement appelée « loi des séries ». On retrouve cette conception dans la croyance populaire sous la forme de dictons : « Jamais deux sans trois. » « Un malheur n'arrive jamais seul. »

Ces deux extrêmes rejoignent deux conceptions antagonistes de la notion de probabilité :

- L'une *déterministe* à l'excès, tel d'Alembert qui pensait que si, dans un jeu de pile ou face, *pile* était sorti trois fois, alors la probabilité de tirer *face* devenait supérieure à $\frac{1}{2}$.
- L'autre, *aléatoire* à l'excès, qui attribue une probabilité de $\frac{1}{2}$ à chacun des évènements : lorsque je traverse une route, soit je me fais écraser par une voiture, soit non.

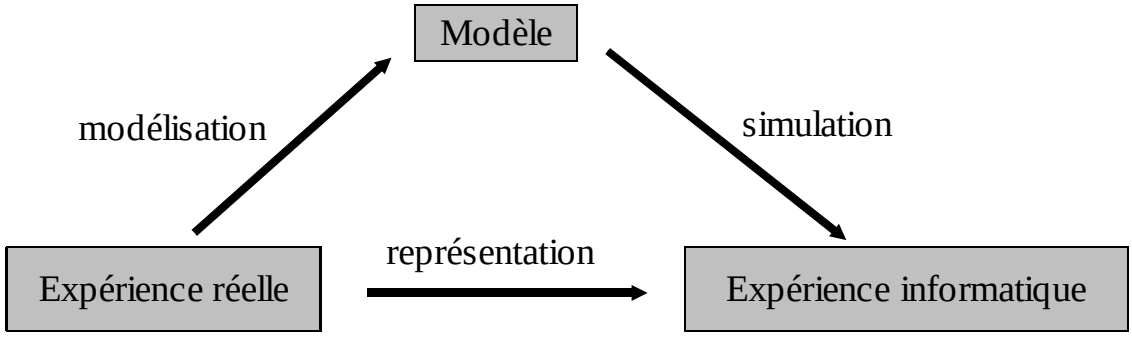
3.3.3. Le « hasard mathématique » et la place de la simulation

Le « hasard mathématique » est un modèle qui conceptualise l'équiprobabilité ou équirépartition. Exemples :

- En mathématiques, on utilise un modèle de dé qui attribue à chaque face la probabilité de $\frac{1}{6}$, de même qu'un modèle de pièce qui attribue à chaque face la probabilité de $\frac{1}{2}$.
- Pour représenter le hasard, on a créé des générateurs aléatoires qui donnent une équirépartition des nombres décimaux « formatés » comportant un nombre donné de chiffres.

La **simulation**, comme son nom l'indique, est une façon de représenter le problème de façon analogique, mais avec des outils mathématiques.

Pour comprendre la place de la simulation, j'utiliserai ce schéma ternaire que Bernard Parzysz propose dans son article « *Expérience aléatoire et simulation* », *Repères-IREM* n°66.



Pour simuler, il est nécessaire d'avoir modélisé le problème, c'est-à-dire d'en avoir fait une représentation qui permette de travailler avec des *générateurs aléatoires*, c'est-à-dire des modélisateurs du hasard. Mais cela ne veut pas dire qu'on connaisse une loi mathématique qui explique le phénomène étudié.

Ces générateurs aléatoires doivent avoir pour qualité essentielle l'équirépartition des nombres, comme ceux que j'ai cités ci-dessus.

Réaliser de « vraies » expériences aléatoires avant de les simuler est certes indispensable. Mais lancer les dés dans une classe de troisième ou de seconde peut avoir un côté infantilisant et, pour le moins, être une source de bruit et d'agitation difficile à contenir !

C'est pourquoi il peut être intéressant d'utiliser assez vite d'autres procédés de simulation ; les plus communément utilisés de nos jours étant les générateurs aléatoires des ordinateurs et des calculatrices (random).

Pour simuler un problème, on peut :

- soit faire des échantillons (par exemple de 100 tirages) et cumuler les résultats ;
- soit programmer et laisser tourner l'ordinateur, et constater une certaine stabilisation de la fréquence.

Les probabilités traitent avec le même modèle ces deux approches (voir mon article « *L'apprenti fréquentiste* » paru dans *Repère-IREM* n°21).

3.3.4. Un autre exemple : politique nataliste

Cette situation est aussi issue du document d'accompagnement de seconde de 2000 (Eduscol).
Supposons qu'une politique nataliste soit mise en place à partir de la règle suivante.
Les naissances au sein d'une famille s'arrêtent :

- *soit à la naissance du premier garçon ;*
- *soit lorsque la famille comporte quatre enfants.*

Quelle est l'influence d'une telle politique sur la répartition des sexes ?
Quelle est l'influence d'une telle politique sur la composition des familles ?
L'hypothèse de travail est l'équiprobabilité de naissance d'un garçon ou d'une fille.

Cette situation serait bien délicate à réaliser « en temps réel » dans le « monde réel », d'où la nécessité de *simuler les naissances*, après les avoir modélisées (équiprobabilité d'avoir un garçon ou une fille, indépendance des naissances). On a alors une grande panoplie de possibilités pour cette simulation : pièce (pile pour garçon, face pour fille), dé, touche random (avec par exemple le choix : impair pour les garçons, pair pour les filles), etc.

Là encore, il peut être intéressant de sonder les représentations *a priori* sur la répartition des sexes que peut entraîner une telle politique. Il apparaît souvent qu'il y aura davantage de filles, à cause du fait qu'on peut aller jusqu'à 4 dans une famille, alors que la naissance d'un garçon « termine » la famille.

L'activité peut se dérouler dans une classe en demandant à chaque élève de simuler 20 familles, ce qui conduit à noter par exemple : G/FFG/FG/G/FFFG/G/FFFF/G/FG/G/…

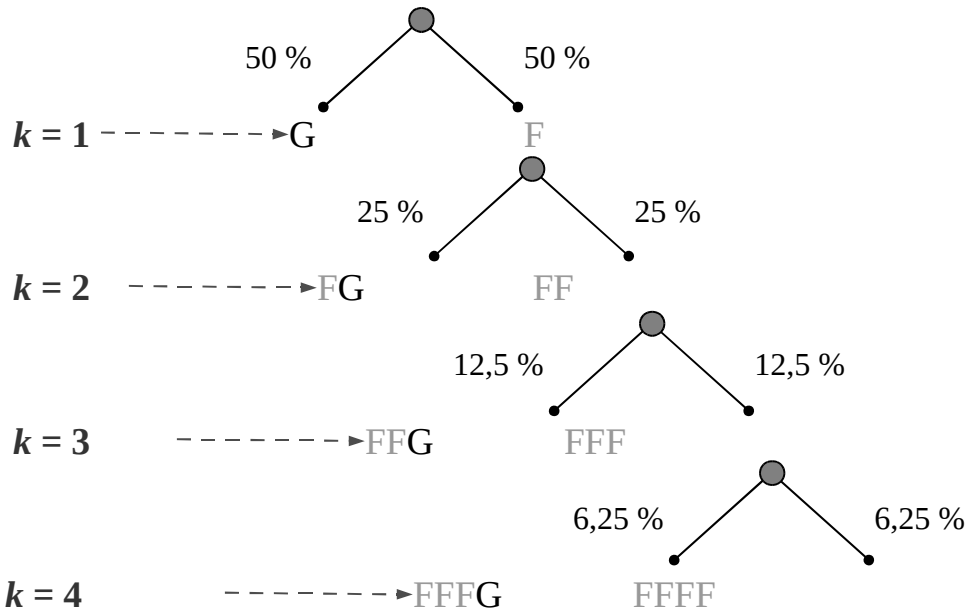
On cumule alors les résultats et on obtient « en moyenne » :

Proportion de filles (ou de garçons) : autour de 50 % ; ce qui était à prévoir et qui conforte que le générateur aléatoire est « bon », au sens de l'équirépartition !

Proportion de familles

- de 1 enfant : autour de 50 % ;
- de 2 enfants : autour de 25 % ;
- de 3 enfants : autour de 12 % ;
- de 4 enfants : autour de 12 %.

L'intérêt de cette situation est qu'on peut la modéliser mathématiquement en utilisant un outil important des probabilités : les arbres.



3.4. D'autres pistes qui ne sont plus pour le collège

3.4.1. Les bouteilles

Cette situation est due à Maurice Glaymann. Elle a été développée par François Huguet dans le numéro 36 de *Repères-IREM*. Elle exemplifie bien l'approche expérimentale par la simulation.

Un fabricant de bouteilles en verre dispose de 100 kg de verre liquide.
Avec 1 kg de verre liquide, on peut fabriquer une bouteille.
Dans les 100 kg de verre liquide, il y a 10 pierres ou impuretés que l'on ne peut pas enlever et qui sont réparties de manière aléatoire.
Le fabricant ne s'intéresse qu'à la fabrication de bouteilles de « haute qualité », c'est-à-dire sans impureté.
Si une bouteille contient au moins une pierre, elle est mise au rebut !
Combien peut-il espérer obtenir de bouteilles de haute qualité ?

Simulons cette expérience

Lorsque je mène cette activité avec des professeurs stagiaires, c'est l'occasion de réfléchir sur des procédures de simulation. Pour simuler cette expérience, on peut par exemple construire un tableau 10×10 avec repérage (voir ci-dessous) qui représente les 100 bouteilles. On procède alors à un tirage de 100 couples de chiffres, en utilisant un générateur aléatoire. Chaque couple de chiffres tiré représente une impureté dans la bouteille repérée par ce couple.

Le tableau ci-dessous est la réalisation d'une telle expérience : on trouve 37 bouteilles sans impureté.

Si on répète un certain nombre de fois cette simulation expérimentale, les résultats oscillent entre 33 et 41, de façon très stable (la modélisation mathématique que nous allons établir explique cette stabilité).

Si on prend l'angle d'approche **fréquentiste**, on peut donc dire, à partir de ces expériences, que la probabilité d'avoir une bouteille sans impureté est d'environ 37 % (probabilité *a posteriori*).

9	xx	x			x		x		x	xx
8	x		xxx	xx		xx		x		x
7	x	xx		x	xxx		xx	x		xxx
6		x	x	x		x			x	xx
5	xxx		xx	xx		x	xx	x		x
4	x		xx		x			xx		x
3		xx			xx					
2	x		xxx		x	x	xx		xxx	
1	xx		x	xx	x	x	x	xx	x	xx
0	x	x	xx	x		xxx	xx	x	xx	x
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

Modélisons mathématiquement cette situation

Soit I l'ensemble des impuretés.
Soit B l'ensemble des bouteilles.
Soit $F(I,B)$ l'ensemble des applications de I dans B.

$F(I,B)$ représente l'ensemble des « cas possibles ».

Soit b_i une bouteille, *quelle est la probabilité que b_i soit sans impureté ?*

Soit $B_i = B - \{b_i\}$, $F(I,B_i)$ représente l'ensemble des « cas favorables » à l'évènement « b_i est sans impureté ».

Utilisons alors le calcul probabiliste déterministe de Laplace (probabilité *a priori*).

La probabilité que b_i soit sans impureté est :

$$p = \frac{\text{card}(F(I,B_i))}{\text{card}(F(I,B))} = \frac{99^{100}}{100^{100}} \approx 0{,}3665$$

Curiosité : $\frac{1}{0{,}3665} \approx 2{,}72\ldots$

Et… à quoi peut bien faire penser 2,72… ? … !

Généralisons le problème

On passe à *n* bouteilles et *n* impuretés. Le calcul précédent devient :

$$p = \frac{(n-1)^n}{n^n} = \left(1 - \frac{1}{n}\right)^n$$

Or $\lim\limits_{n\to\infty}\left(1-\frac{1}{n}\right)^n = \frac{1}{e}$ et $\frac{1}{e} \approx 0{,}3678\ldots$

3.4.2. Probabilité que deux entiers naturels (non nuls) pris au hasard soient premiers entre eux

Si on prend ces entiers naturels entre 1 et 10, on a le tableau :

10	x		x				x		x	
9	x	x		x	x		x	x		x
8	x		x		x		x		x	
7	x	x	x	x	x	x		x	x	x
6	x				x		x			
5	x	x	x	x		x	x	x	x	
4	x		x		x		x		x	
3	x	x		x	x		x	x		x
2	x		x		x		x		x	
1	x	x	x	x	x	x	x	x	x	x
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

On constate que la proportion de couples premiers entre eux est de 63 %. On peut donc dire que la probabilité que deux entiers naturels compris entre 1 et 10, pris au hasard, soient premiers entre eux est 0,63.

Comment définir de manière générale cette probabilité « p » que deux entiers naturels non nuls soient premiers entre eux ?

Si on prend comme ensemble de référence les sous-ensembles $[1,n]\times[1,n]$ de $\mathbb{N}^*\times\mathbb{N}^*$, on peut déterminer le nombre de couples premiers entre eux. Ceci donne, comme ci-dessus, la probabilité p_n que deux entiers pris entre 1 et n soient premiers entre eux.

Si on passe à $\mathbb{N}^*\times\mathbb{N}^*$, la définition laplacienne des probabilités ne peut fonctionner, elle conduirait à un rapport d'infinis !

On définit donc la probabilité « p » cherchée comme la limite de p_n lorsque n tend vers l'infini… si elle existe !

En utilisant cette démarche, on obtient que la probabilité qu'un nombre entier soit pair est $\frac{1}{2}$, qu'il soit multiple de 3 est $\frac{1}{3}$, etc.

Montrer l'existence de « p » n'est pas simple ; cela se fait avec du matériel mathématique un peu plus lourd ! Nous l'admettrons ici.

Supposons donc l'existence de « p ».

En utilisant la théorie de Kolmogorov et la remarque ci-dessus sur la probabilité qu'un entier soit un multiple de n, on obtient :

Soit $A_1=\{(x,y)\in\mathbb{N}^*\times\mathbb{N}^*/\, x\wedge y=1\}$ donc $p(A_1)=p$

Soit $A_2=\{(x,y)\in\mathbb{N}^*\times\mathbb{N}^*/\, x \text{ et } y \text{ divisibles par } 2 \text{ et } \frac{x}{2}\wedge\frac{y}{2}=1\}$ donc $p(A_2)=\frac{p}{4}$

…

Soit $A_n=\{(x,y)\in\mathbb{N}^*\times\mathbb{N}^*/\, x \text{ et } y \text{ divisibles par } n \text{ et } \frac{x}{n}\wedge\frac{y}{n}=1\}$ donc $p(A_n)=\frac{p}{n^2}$

Or $\bigcup_{n=1}^{\infty} A_n=\mathbb{N}^*\times\mathbb{N}^*$, où les A_n sont deux à deux disjoints.

D'où : $1=p(\mathbb{N}^*\times\mathbb{N}^*)=p(\cup A_n)=\sum_{n=1}^{+\infty} p(A_n)=\sum_{n=1}^{+\infty}\frac{p}{n^2}=p\times\sum_{n=1}^{+\infty}\frac{1}{n^2}$

Or $1=\sum_{n=1}^{+\infty}\frac{1}{n^2}=\frac{\pi^2}{6}$ donc p est l'inverse de $\frac{\pi^2}{6}$ d'où $p\approx 0{,}607\ldots$

Probabilité qu'une fraction soit irréductible

Le calcul ci-dessus montre donc que si on prend une fraction au hasard, elle a 61 % de chances d'être irréductible.

Que veut alors dire prendre un échantillon de fractions au hasard. Si on prend la page d'un manuel de collège intitulée « *simplification des fractions* », elle a bien peu de chances d'être représentative… ou alors l'auteur du manuel s'est fourvoyé dans les exercices qu'il propose !

Tirer « au hasard » ???

On voit dans l'exemple ci-dessus que cela doit être précisé… comme je vais le montrer avec mes « spaghettis ».

3.5. Une histoire de spaghettis

J'ai commencé ma carrière d'enseignant de mathématiques au temps des « mathématiques modernes » ! À l'époque, ne se posait pas la question du rapport des mathématiques au réel : les mathématiques étaient un magnifique édifice qui se construisait de façon purement interne. C'était essentiellement un enseignement de structures.

Au contact des IREM, j'ai commencé à me poser la question de la pertinence de ces mathématiques modernes, à la fois dans leur rôle de sélection, mais aussi dans leur capacité de construction d'un vrai outil scientifique à la disposition des élèves et des autres disciplines.

Toutes ces questions fortement posées par différents instituts et associations, dont l'APMEP, ont conduit aux nouveaux programmes de 1986, où les mots-clés sont devenus, pour le collège « activités », et pour l'école « situations-problèmes », mettant en avant les problèmes concrets, quotidiens, issus du monde réel, et prônant une démarche expérimentale. Le mot de « modélisation » ne figure pas dans ces programmes.

Cette période fut pour moi une formidable bouffée d'air frais en tant qu'enseignant, et me donna la chance de pouvoir développer un travail en équipe aussi bien au niveau de mon collège qu'au niveau de la commission Inter-IREM Premier Cycle (maintenant : Collège) investie dans les « suivis scientifiques », commission dont je fus alors le responsable.

Des spaghettis réels…

Dans le cadre de ces nouveaux programmes, j'essayais au maximum de mettre les élèves en situation d'activité (versant parfois dans l'activisme), et pour introduire l'inégalité triangulaire en quatrième j'eus une idée que je trouvais *a priori* géniale : j'apportais des spaghettis en classe, j'en donnais quelques-uns à chaque élève, et leur demandais de les « casser » en trois morceaux « au hasard ». Ils devaient alors essayer de faire un triangle avec ces trois morceaux. Je leur demandais de mesurer la longueur de chacun des morceaux et de conjecturer à partir de cette mesure une règle qui permette de discriminer les cas où ils obtenaient des triangles. L'état de la classe à la fin de l'heure m'a déterminé à ne pas reconduire une telle expérience !

… aux spaghettis mathématiques

Dans notre collège, nous suivions les classes de quatrième en troisième. J'ai voulu revenir sur cette expérience pas très heureuse des spaghettis et, pour ce faire, j'ai inventé le « spaghetti mathématique ». C'était un spaghetti de longueur **1**, avec équiprobabilité de cassure (ce qui est évidemment inconcevable avec un spaghetti réel !). Et, pour faire ces cassures, j'ai utilisé la simulation. J'expliquais donc aux élèves ce nouveau contexte et leur proposais de faire ces cassures avec leur calculatrice en utilisant la touche « random » qui leur donnait, à l'époque, un nombre compris entre 0 et 1 avec 3 chiffres après la virgule. Avec 3 tirages aléatoires (ex : 0,167 **;** 0,534 **;** 0,435), ils simulaient la cassure de 3 spaghettis mathématiques. Pour donner un sens tangible à l'expérience, je leur proposais de multiplier par 100 chacun des nombres obtenus, ce qui leur donnait 3 mesures de longueur en mm. Ils pouvaient ainsi vérifier par construction s'ils avaient ou non « tiré » un triangle.

Vous aurez noté que cette nouvelle situation ne reproduit pas l'expérience précédente où je cassais un spaghetti en 3, alors que là je casse 3 spaghettis en 2 !

L'objectif de la séance était d'arriver à se passer de l'expérience physique pour décider simplement avec les 3 tirages si on obtenait un triangle ou non, via l'inégalité triangulaire immédiatement traduite par : « il ne faut pas qu'un des nombres soit plus grand que la somme des

deux autres ».

Forts de cette règle, les élèves effectuèrent alors 10 tirages et, sans avoir vraiment préparé ce passage aux statistiques, j'ai proposé de voir quel était le pourcentage des triangles obtenus… Devant le résultat (48 %), les élèves me demandèrent : « c'est bon ? ; c'est ça ? ; c'est juste ? », comme s'ils pensaient que je connaissais « ce résultat ». Leur questionnement pouvait être traduit par : existe-t-il un modèle mathématique qui me permette d'affirmer que ce résultat est « vraisemblable » ? Et j'étais bien incapable de leur répondre, sinon en faisant tourner mon ordinateur et en constatant qu'il y avait une certaine stabilisation de la fréquence autour de 50 %. Je crois que ce fut mon premier **vrai** contact avec la modélisation…

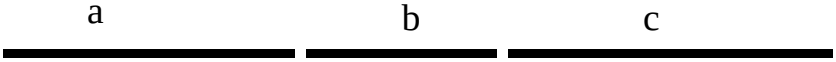
3.6. Une modélisation de cette expérience des spaghettis

Mes spaghettis de quatrième

Mon expérience malheureuse de quatrième consistait à couper un spaghetti en trois « au hasard » pour voir si les trois morceaux obtenus pouvaient « former » un triangle. Je suis bien incapable de modéliser cette expérience réelle. Mais elle va m'éclairer pour simuler. En observant les élèves, on constate deux grandes façons de faire : soit ils essaient de couper le spaghetti *d'un seul coup* en trois, avec peu de chances d'obtenir trois morceaux, soit ils font une première cassure, puis recassent l'un des deux morceaux obtenus.

Pour simuler cette expérience, je prends le « spaghetti mathématique » que j'avais proposé en troisième (de longueur **1**, avec équiprobabilité de cassure) et je décris en langage informatique chacune des deux expériences.

Couper un spaghetti « au hasard » en 3 morceaux « simultanément »



x = rnd ; y = rnd
a = min (x, y) ; b = max(x, y) – a ; c = 1 – (a+b)
Test : max {a, b, c} $< \frac{1}{2}$ *(la somme des trois longueurs étant 1)*

On peut alors :

- soit faire des « échantillons » (par exemple de 100 tirages) et on trouve comme fréquences de triangles : 0,26 ; 0,23 ; 0,25 ; 0,27 ; 0,24…
- soit programmer, laisser tourner l'ordinateur et constater une certaine stabilisation de la fréquence dans l'intervalle [0,24 ; 0,26]…

… en s'appuyant sur le fait que les probabilités traitent avec le même modèle ces deux approches (voir plus haut).

Couper un spaghetti « au hasard » en 2 morceaux, puis « le morceau restant » en 2

x= rnd ; y = rnd
a = x ; b = (1 – x)y ; c = 1 – (a+b)
Test : max {a, b, c} $< \frac{1}{2}$

- Le tirage d'échantillons donne comme fréquences de triangles : 0,22 ; 0,18 ; 0,19 ; 0,20 ; 0,21 ; 0,17 ; 0,17…

- La suite obtenue en faisant tourner l'ordinateur a une certaine stabilisation dans l'intervalle [0,18 ; 0,21].

Au-delà de montrer que la mesure de la fréquence est ici un intervalle, cette double expérience met en évidence que « au hasard » mérite d'être précisé. Dans les deux cas, on a coupé un spaghetti au hasard, mais ce sont les conditions de l'expérience qui permettent de modéliser ce hasard (cf. le paradoxe de Bertrand).

Modélisation géométrique

Comment alors trouver un modèle mathématique qui permette de calculer « **la** » probabilité d'obtenir un triangle dans les conditions d'expérience ci-dessus ? Il faut passer du modèle discret au modèle continu, et de l'équiprobabilité à la probabilité uniforme.

Chacun des tirages donne un couple (x ; y) qui peut être représenté par un point dans un repère. La probabilité cherchée est donc le rapport du nombre de points satisfaisant à l'obtention d'un triangle par rapport au nombre de points possibles.

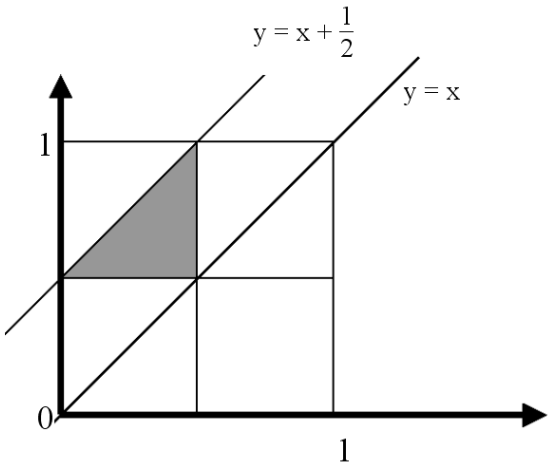
En plongeant dans le modèle continu, cela va se traduire par le rapport de l'aire de la surface où se trouvent ces points-solutions, à l'aire totale possible qui est ici celle du carré [0,1] × [0,1], c'est-à-dire **1** (la notion d'ouvert ou de fermé n'ayant pas d'importance compte tenu de la modélisation).

Couper un spaghetti « au hasard » en 3 morceaux « simultanément »

Distinguons les deux cas : $x < y$ et $x > y$.

Si $x < y$, le test $\max\{a,b,c\} < 1/2$ donne :

- $x < 1/2$
- $y - x < 1/2$
- $1 - y < 1/2$
-



L'aire de la surface « solution » et donc la probabilité est : $\frac{1}{8}$.

Si $x > y$

On trouve comme surface solution le triangle symétrique de celui ci-dessus par rapport à la diagonale du carré (y = x), donc de nouveau une probabilité de $\frac{1}{8}$.

Ces deux cas étant exclusifs, la probabilité cherchée est donc :

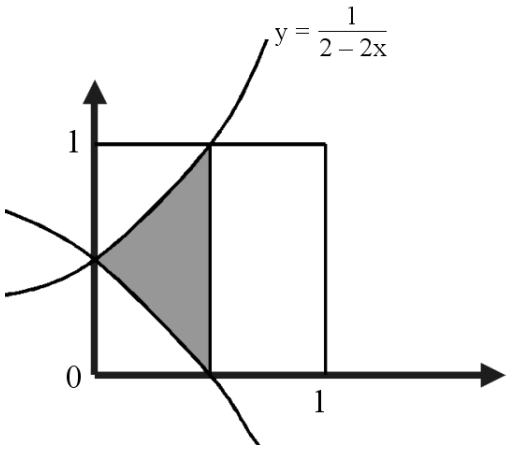
$$p = \frac{1}{4} \text{ , soit } 0{,}25$$

ce qu'on aurait pu pronostiquer compte tenu des fréquences obtenues !

Couper un spaghetti « au hasard » en 2 morceaux, puis « le morceau restant » en 2

On traduit comme ci-dessus le test :
max {a,b,c} < $\frac{1}{2}$, ce qui donne :

- x < 1/2
- (1 – x)y < 1/2
- 1 - [x + (1 - x) y] < 1/2



Le calcul de l'aire de la surface solution donne, via le calcul intégral, la probabilité :
$p = \ln 2 - \frac{1}{2}$, soit environ 0,193 1
Là, c'était beaucoup plus difficile de pronostiquer un tel résultat à partir des fréquences !

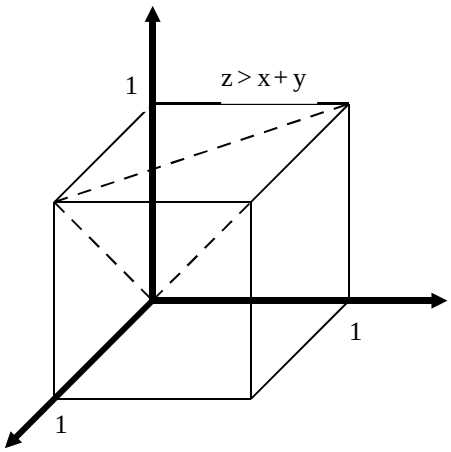
Et mes spaghettis de troisième ?

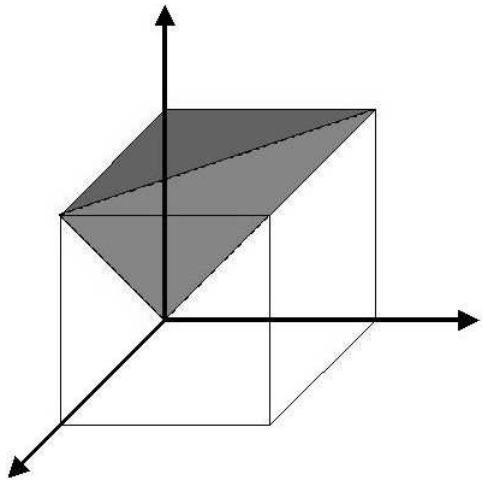
Je rappelle l'expérience : trois tirages aléatoires indépendants et un test pour savoir si on peut faire un triangle avec les trois longueurs obtenues. Il m'a fallu un certain temps pour quitter le plan et comprendre que je travaillais avec trois spaghettis indépendants, et que les trois cassures pouvaient être représentées par un triplet (x, y, z), coordonnées d'un point de l'espace. La modélisation géométrique donne alors comme solution le rapport du volume du solide solution sur le volume du solide possible, ici le cube [0,1] × [0,1] × [0,1], soit **1**.

Pour trouver ce solide solution, il suffit d'enlever le solide non-solution. Celui-ci se découpe en trois solides élémentaires du cube, donnés par les conditions :

$$z > x+y \; ; \quad y > x+z \; ; \quad x > y+z \; .$$

Le cas $z > x+y$





On trouve un tétraèdre de volume $\frac{1}{6}$.

Les deux autres cas donnent chacun un tétraèdre de volume $\frac{1}{6}$ et n'ont pas de parties solides communes.

La probabilité de non-solution est donc : $\frac{1}{6}+\frac{1}{6}+\frac{1}{6}$, soit $\frac{1}{2}$.

Donc la probabilité d'obtenir un triangle dans mon expérience de troisième est :

$p=1-\frac{1}{2}$, soit $p=\frac{1}{2}$.

Lorsque j'ai trouvé cette modélisation et ce résultat, qui confirmait mon approche fréquentiste, j'étais heureux ! Ma pensée était devenue libre par rapport à ce problème et c'est avec confiance que je regardais tourner mon ordinateur ou répétais cette expérience par échantillon avec mes élèves.

Ce qu'apportent les mathématiques, c'est cette merveilleuse compétence d'anticipation et de contrôle !

4. À la recherche d'une loi modèle : la loi de Benford

Supposons une situation qui nous donne une grande quantité de nombres qui nous apparaissent tout à fait aléatoires et que l'on nous pose la question suivante :

« Prenons le premier chiffre de chacun de ces nombres : quelle est la répartition des 1, des 2, … , des 9 ? ».

En l'absence de toute autre connaissance, notre réflexe sera l'équiprobabilité, c'est-à-dire que nous supposerons que chaque chiffre a une probabilité d'apparition de $\frac{1}{9}$.

4.1. Trois références de données

La situation suivante nous a été proposée par Claudine Schwartz lors d'une réunion de la CREM (Commission de Réflexion sur l'Enseignement des Mathématiques).

Le tableau ci-dessous donne la fréquence d'apparition du premier chiffre de nombres pris respectivement :

- colonne 2 : 1000 nombres *du Monde* daté du vendredi 23 avril 1999 ;
- colonne 3 : 914 nombres d'un historique de compte de la société Gilibert ;
- colonne 4 : nombres d'habitants de 1229 communes obtenus lors du recensement de 1992.

Premier chiffre	Le Monde	Gilibert	Commune
1	0,322	0,317	0,321
2	0,151	0,161	0,168
3	0,108	0,142	0,133
4	0,099	0,088	0,081
5	0,073	0,070	0,087
6	0,081	0,061	0,067
7	0,055	0,070	0,055
8	0,065	0,040	0,045
9	0,046	0,050	0,044

Deux constats s'imposent :

- on est bien loin de l'équiprobabilité (qui est pourtant notre premier réflexe) ;
- les trois expériences donnent des résultats vraiment proches.

4.2. La loi de Benford

Claudine Schwartz, en s'appuyant sur le constat que les résultats étaient invariants par changement d'échelle, a modélisé ces situations en utilisant la loi de Benford : *« La probabilité que le premier chiffre à gauche dans l'écriture en base 10 soit i = 1, … , 9 est log (1 + 1/i) (logarithme décimal) ».*

La dernière colonne du tableau ci-dessous donne les fréquences théoriques obtenues par calcul avec cette loi. La modélisation par cette loi apparaît comme très bonne d'un point de vue qualitatif.

Premier chiffre	Le Monde	Gilibert	Commune	Loi de Benford
1	0,322	0,317	0,321	0,301
2	0,151	0,161	0,168	0,176
3	0,108	0,142	0,133	0,125
4	0,099	0,088	0,081	0,097
5	0,073	0,070	0,087	0,080
6	0,081	0,061	0,067	0,067
7	0,055	0,070	0,055	0,058
8	0,065	0,040	0,045	0,051
9	0,046	0,050	0,044	0,046

Mais avoir modélisé mathématiquement nous donne-t-il le sens profond du phénomène ?
Cette question m'a conduit à deux pistes de réflexion.

4.3. Comment Benford a-t-il eu l'idée d'une telle loi ?

Benford a établi cette loi en 1938, à la suite d'étude de nombreuses données. Il s'appuyait sur les travaux d'un astronome américain, Simon Newcomb, qui en avait donné les prémisses en 1881, en s'appuyant sur un constat : la forte usure des premières pages des tables de logarithmes ! Dans le système à base 10, les logarithmes décimaux des nombres sont uniformément distribués, ce qui peut se traduire par le fait qu'un nombre a autant de chances d'être entre 100 et 1000 (log 2 et log 3) qu'entre 10 000 et 100 000 (log 4 et log 5). Cette répartition va s'appliquer aux phénomènes de type exponentiel.

4.4. Comment donner du sens à cette loi ?

Nous sommes devant des phénomènes *évolutifs*. Pour donner du sens à cette modélisation, j'ai essayé d'imaginer une simulation (qui ne peut reposer sur le tirage au hasard de nombres) : on écrit la suite des entiers naturels en déclenchant un chronomètre ; le chronomètre s'arrête au hasard et je fais mes comptes ! Il y a donc bien du hasard là-dedans, mais pas là où on croit !

4.5. Connaître le bon modèle, ça sert !

Tout cela, me direz-vous, n'est que jeu de mathématicien ! Ceux qui se sont fait « épingler » par le fisc, qui utilisait cette loi pour vérifier leur comptabilité, ne seront certainement pas d'accord !

5. Conclusion

En guise de conclusion 1…

J'avais été sollicité par l'IREM de Montpellier pour un séminaire sur les « probas/stat ». La conférence de clôture avait été assurée par un professeur de médecine du CHU de Montpellier et j'avais été très sensible à sa conclusion :

> **« Nous sommes tous les enfants du hasard. »**
>
> **« Le jour où le hasard n'existera plus, c'est l'homme qui n'existera plus, car c'est le hasard génétique qui sauve les espèces. »**

Cette conclusion illustre parfaitement la spécificité du « modèle mathématique du hasard » que j'ai développée dans ce texte, par rapport au « hasard » de la nature. Et, en vertu des éléments que j'ai donnés sur la perception des nombres par notre société, elle me conduit à affirmer :

> ***« La formation à la pensée statistique, ça n'est pas l'école du mensonge, c'est celle de l'humilité* ! »**

En guise de conclusion 2…

Les statistiques ont fait leur apparition dans le programme de collège en 1986. Il a fallu attendre 2008 pour que les probabilités y trouvent une première place. Et pourtant, dès 1812, Laplace affirmait :

> **« Et si l'on observe ensuite que dans les choses qui peuvent ou non être soumises au calcul, la théorie des probabilités apprend à se garantir des illusions, il n'est pas de science qu'il soit plus utile de faire entrer dans le système de la fonction publique. »**

Cela traduit, pour le moins, un grand retard dans notre enseignement dans ce domaine, beaucoup plus développé dans d'autres systèmes scolaires, comme les pays anglo-saxons.

En guise de conclusion 3…

Les mathématiques sont, au regard de l'histoire, un formidable outil intellectuel pour penser le monde qu'a créé l'homme, qu'il a enrichi au fil des siècles et des civilisations ! Et, si nous pouvions persuader nos élèves de cela, peut-être notre enseignement produirait-il moins « d'écorchés vifs des mathématiques » !

Nous avons le devoir de transmettre ce patrimoine de l'humanité ; Joseph Fourier résume bien cela en disant des mathématiques qu'elles sont ***« une faculté de la raison humaine, destinée à suppléer à la brièveté de la vie et à l'imperfection des sens ».***