PARZYSZ BERNARD

QUELQUES QUESTIONS DIDACTIQUES DE LA STATISTIQUE ET DES PROBABILITÉS

Abstract. Some Educational Questions about Statistics and Probability. This article goes over some of the ideas which emerged in France about the teaching of probability and statistics in secondary education during the past ten years, and first of all it takes up and develops the idea of an analogy between modelling in probability and in elementary geometry, considered here from the standpoint of paradigms. Another specificity of the domain is that its teaching makes a great use of various representations: tables, varied kinds of graphs, tree diagrams, box-and-whiskers plots and so on, which are a major element of the workspace but the construction and the meaning of which are sometimes problematic. It is the same for the articulation of one register with another, which is most often considered transparent, but obviously is not without causing difficulties, even among teachers. Taking into account the notion of semantic congruence may allow working this aspect more specifically and bring to the fore the isomorphism subjacent to random experiments which seem *a priori* different, thus opening up the way to the notion of probabilistic model.

Résumé. Cet article rassemble, en leur donnant un éclairage didactique, quelques-unes des idées qui ont émergé en France sur l'enseignement des probabilités et de la statistique dans le secondaire au cours de la dernière décennie, et en premier lieu il reprend et développe celle d'une analogie entre la modélisation en probabilités et en géométrie élémentaire, vue ici sous l'angle des paradigmes. Une autre particularité du domaine est que son enseignement fait un grand usage de représentations variées : tableaux, graphiques divers, arbres, boîtes de dispersion, *etc*., qui constituent un élément majeur de l'espace de travail mais dont la construction et le sens posent parfois problème. Il en va de même de l'articulation de ces divers registres entre eux, qui est le plus souvent considérée comme transparente, mais à l'évidence n'est pas sans créer des difficultés, même chez les enseignants. La prise en compte de la notion de congruence sémantique peut permettre de travailler plus spécifiquement cet aspect et de mettre en évidence l'isomorphisme sous-jacent à des expériences aléatoires *a priori* différentes, ouvrant ainsi la voie à la notion de modèle probabiliste.

Mots-clés. Statistiques et Probabilités, paradigmes, registres, modélisation.

1. Introduction

Par rapport à d'autres domaines comme la géométrie, l'algèbre ou l'analyse, la statistique et les probabilités sont apparues très tardivement dans l'enseignement secondaire. En France, par exemple, ce n'est qu'il n'y a une quarantaine d'années − à l'occasion de la réforme dite des « mathématiques modernes », et bien que ce domaine ne fût pas en odeur de sainteté chez Bourbaki − qu'elles ont véritablement fait leur apparition dans l'enseignement général (Parzysz, 1997a), sans doute pour

ANNALES de DIDACTIQUE et de SCIENCES COGNITIVES, volume 16, p. 127 – 147.

la raison que l'axiomatique de Kolmogorov, avec l'algèbre des événements, ainsi que le traitement statistique, fournissaient des applications de la « théorie des ensembles » (organisation des données, propriétés de la probabilité). Cette place s'est ensuite trouvée confortée par le développement, dans les sciences et l'industrie, de nombreuses applications de la statistique inférentielle, développement concomitant de celui de technologies permettant la simulation des phénomènes aléatoires. En particulier, la mise en place de nouveaux programmes au lycée à partir de 2001 a permis d'installer l'aléatoire comme une branche « à part entière » des mathématiques enseignées dans le secondaire (alors qu'auparavant elle était plutôt considérée comme étant « entièrement à part »), et ce pour deux raisons principales, et non forcément indépendantes, liées aux enseignants : d'une part leur manque de formation dans ce domaine, et d'autre part le fait qu'un certain nombre d'entre eux ne considéraient pas la statistique et les probabilités comme faisant partie des sciences mathématiques. Cette relative nouveauté du domaine et la méfiance à son égard expliquent sans doute, du moins en partie, la proportion relativement faible de travaux didactiques consacrés au domaine de l'aléatoire, alors que, comme on vient de le voir, le besoin de formation était − et reste − important, comme en témoigne la multitude d'actions de formation continue mises en place par les IREM.

La mise en place des probabilités dans l'enseignement secondaire n'est pas, nous le verrons, sans présenter de fortes analogies avec celle de la géométrie « démontrée », que permettent de mettre en évidence les notions de paradigme et d'espace de travail, en soulignant dans un cas comme dans l'autre le rôle crucial des représentations. Cependant, comparées à la « figure » de géométrie, dans l'enseignement de la statistique comme dans celui des probabilités les représentations sont de types très variés, qu'il s'agisse de la statistique descriptive, dont les médias − *via* l'inventivité de leur service d'infographie − nous offrent quotidiennement des exemples plus ou moins pertinents, ou des probabilités, où l'utilisation des arbres est désormais officialisée (*cf.* programme de terminale scientifique : « *On utilisera à bon escient les représentations telles que tableaux, arbres, diagrammes, …, efficaces pour résoudre des problèmes de probabilités* » (B.O. hors série n° 4 (30 août 2001) p. 70), y compris en tant qu'outil de démonstration : « *un arbre de probabilité correctement construit constitue une preuve* » (*ibid.*).

Je voudrais, dans cet article, reprendre, en leur donnant un éclairage didactique, quelques-unes des idées qui ont émergé à ce sujet au cours de la dernière décennie, pour les relier à des cadres théoriques plus généraux.

2. La modélisation : un parallèle entre probabilité et géométrie

Je commencerai par m'intéresser à l'idée, émise naguère par M. Henry, d'une analogie entre la démarche de modélisation entreprise en géométrie dans l'enseignement secondaire et celle qui concerne la mise en place du cadre probabiliste au lycée (Henry, 1999). Tout d'abord, on peut constater que la notion de probabilité présente une dualité incontournable (Henry, 2009, Carranza, 2009). Schématiquement on peut en effet considérer, d'une part une *probabilité subjective*, degré de croyance de l'observateur relativement à la réalisation d'un événement donné (exemple : pleuvra-t-il ce matin ?), et d'autre part une *probabilité objective*, résultant de nombreuses observations de cet événement (exemple : consultation des données météorologiques) : « *Nous considérons* (…) *la probabilité comme une valeur indépendante de l'observateur, qui indique approximativement avec quelle fréquence l'événement considéré se produira au cours d'une longue série d'épreuves*. » (Renyi, 1966, p. 26). Cette idée avait déjà été émise il y a trois siècles par Jacques Bernoulli : « *Ce qu'il n'est pas donné d'obtenir a priori l'est du moins a posteriori, c'est-à-dire qu'il sera possible de l'extraire en observant l'issue de nombreux exemples semblables* ». (Bernoulli, 1713, p. 42). C'est à cette option (approche dite « fréquentiste ») que s'intéresse plus particulièrement Henry, car elle a été intégrée dans les programmes français actuels à côté de l'approche (dite « laplacienne »), seule présente dans les programmes précédents, qui consiste à définir la probabilité d'un événement associé à diverses éventualités comme le rapport du nombre de celles qui le produisent à leur nombre total. Comme il le fait remarquer : « [Laplace] *souligne la nécessité théorique de l'équiprobabilité des résultats possibles, en laissant à la subjectivité de l'observateur le soin d'en contrôler l'adéquation à la réalité observée*. » (*op. cit.* p. 17) ; c'est-à-dire que, selon la formule consacrée, la probabilité est le nombre de cas favorables divisé par le nombre de cas possibles. Cette conception est également qualifiée de « classique » (Batanero, 2001, p. 13).

Outre le fait qu'elle permet de rendre compte d'un éventail plus large de phénomènes (ceux pour lesquels « il n'est pas donné d'obtenir a priori » la probabilité, c'est-à-dire pour lesquels on ne dispose pas d'un modèle canonique laplacien), l'approche fréquentiste présente l'avantage de permettre une meilleure compréhension des phénomènes de nature aléatoire qui se présentent dans la réalité : « [l'approche fréquentiste] *est cohérente avec l'objectif d'enseigner les mathématiques en vue de donner aux futurs citoyens les capacités de résoudre de vrais problèmes ou de comprendre le fonctionnement des données scientifiques intervenant dans les choix auxquels ils auront à participer.* » (Henry, 1999, p. 33).

Je propose dans ce qui suit une relecture de l'article de Henry, en m'en tenant aux deux conceptions de la probabilité (laplacienne et fréquentiste) figurant explicitement dans les programmes français :

> « *La notion de probabilité est abordée à partir de situations familières (pièces de monnaie, dés, roues de loteries, urnes). Certaines de ces situations permettent de rencontrer des cas pour lesquels les probabilités ne sont pas définies à partir de considérations intuitives de symétrie ou de comparaison mais sont approximativement évaluées par les fréquences observées expérimentalement* (…). »
>
> (Programme français de la classe de Troisième)

2.1. Première modélisation

Dans le cas de la géométrie, il y a au départ, selon Henry, une perception de l'environnement et des actions sur celui-ci : manipulations, constructions, qui conduisent au repérage d'invariants. Il prend comme exemple emblématique une rosace médiévale (*op. cit* p. 30) mais, en référence à mes travaux (Parzysz, 2009b), je lui substituerai ici le motif central d'un panneau d'une mosaïque de Besançon (*figure 1*). On commence par observer le motif en vue d'en rechercher une procédure de construction, ce qui nécessite de repérer des segments et des arcs de cercles et de faire des hypothèses sur les égalités de longueurs et les emplacements des centres des cercles (*figure 2*).

Figure 1 : Besançon. Panneau de la mosaïque de Méduse (2^{e} moitié du 2^{e} siècle). Cliché INRAP.

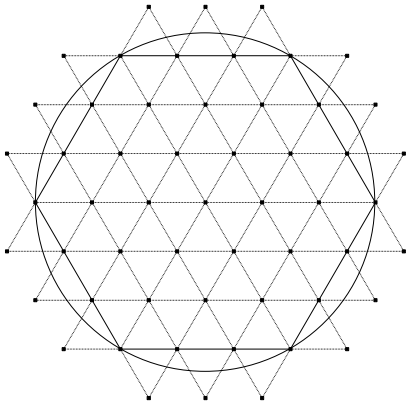


Figure 2 : Positions des centres des cercles
Les fuseaux sont remplacés par des traits droits.
Hypothèse : les triangles sont équilatéraux.

En d'autres termes, il s'agit d'entrer dans G1 (Houdement & Kuzniak, 1999, 2006, Parzysz, 2003) à partir de l'observation du réel, avec un objectif fixé au départ (la reproduction), et de recueillir des données.

On travaille ensuite dans G1 ; l'objet d'étude est maintenant un dessin épuré schématisant l'objet matériel initial, auquel on prête certaines propriétés (alignement, perpendicularité, parallélisme, *etc*.) qui sont, soit issues des observations faites sur l'objet initial, soit constatées sur le schéma lui-même. Sur la base de ces propriétés, on peut le cas échéant envisager une procédure de construction aux instruments.

Dans l'exemple de la mosaïque, le travail sera le suivant : « *Production de dessins géométriques se rapprochant de la* [mosaïque] *observée. (…) Donnée d'un programme de construction du dessin choisi pour représenter la* [mosaïque]*. Justification du programme par des arguments géométriques reprenant les hypothèses issues de l'observation.* » (Henry, 1999, p. 30). Plus précisément, l'observation initiale conduit à considérer la réunion de 72 triangles équilatéraux, dont les sommets sont les centres des cercles servant à mettre les fuseaux en place (*figure 3*).

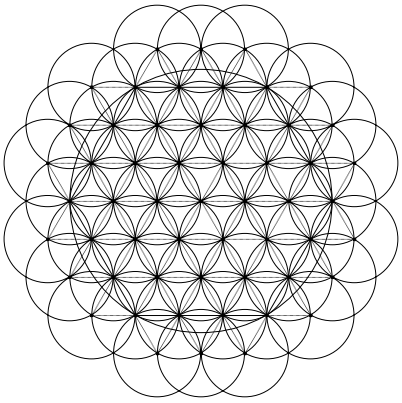


Figure 3 : Disposition des cercles.

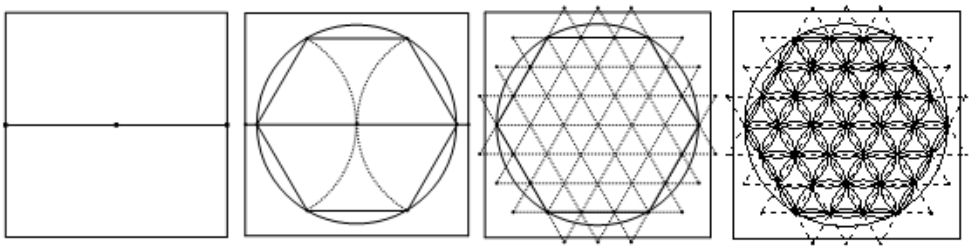
Les outils associés à cet espace de travail sont, soit ceux fournis par un logiciel de géométrie, soit − en environnement papier-crayon − la règle et le compas, et les constructions qui interviennent sont :

- cercle de centre donné passant par un point donné ;
- hexagone régulier inscrit dans un cercle donné ;
- partage d'un segment en *n* parties égales.

(Notons au passage que le mosaïste antique réalisait vraisemblablement cette dernière construction par approximations successives − c'est-à-dire par report de longueur −, qui est une technique propre à G1, beaucoup plus efficace − car sans calculs ni constructions − et économe en temps.)

Une procédure de construction possible fait intervenir sept phases dans l'ordre suivant, en partant du « cadre » carré cernant le panneau (*figure 4*) :

1) Médiane « horizontale » du carré.
2) Milieu de cette médiane.
3) Cercle centré en ce point.
4) Hexagone inscrit dans ce cercle (la médiane tracée étant une diagonale).
5) Partage des côtés de l'hexagone en trois et prolongement des côtés d'une subdivision de chaque côté (N.B. : Ces points extrêmes sont théoriquement nécessaires pour le tracé des arcs extérieurs à l'hexagone, mais il est fort possible − sinon vraisemblable − que le mosaïste antique les ait tracés à main levée, par symétrie d'arcs déjà tracés par rapport aux côtés de l'hexagone).
6) Parallèles aux côtés de l'hexagone passant par les points de subdivision.
7) Cercles et arcs de cercles.



phases 1 et 2 phases 3 et 4 phases 5 et 6 phase 7

Figure 4 : Une procédure de construction.

Dans l'approche fréquentiste de l'aléatoire, et de façon analogue, il y a au départ une perception du hasard (expériences réelles), ainsi que des actions : jeux de hasard, relevés statistiques, conduisant au repérage de stabilités. Pour renouveler l'exemple désormais emblématique de la punaise de bureau, je prendrai ici un exemple fort populaire dans les jeux africains : le cauri. Il s'agit d'un petit coquillage (*cypraea moneta*), jadis utilisé comme monnaie à travers toute l'Afrique (Doumbia & Pil, 1992), dont le dos bombé a été retaillé à plat (*figure 5*) et dont le lancer peut par conséquent produire deux résultats, « fente » (ouverture naturelle du coquillage) et « dos » (face bombée retaillée), selon la face qui apparaît au-dessus (pour une étude plus complète de ce générateur aléatoire, voir Parzysz, 2011).

Figure 5 : Cauris. Cliché B. Parzysz.

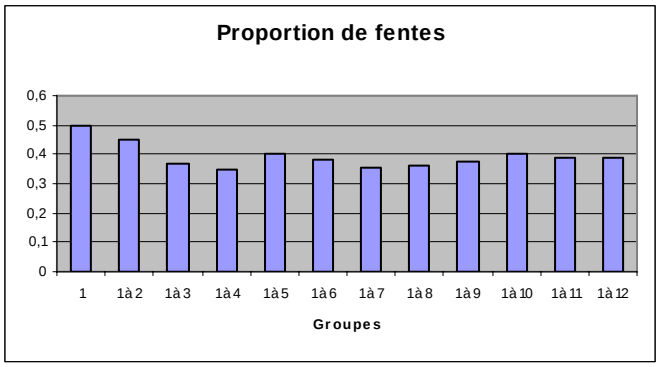
On peut observer les variations de la proportion de « fentes » par groupe de 10 lancers (*tableau 1*), et la stabilisation de cette proportion lorsqu'on regroupe les échantillons (*tableau 2* et *graphique 1*). Il s'agit donc ici d'entrer dans la statistique descriptive (SD) à partir de l'observation de faits, avec un objectif donné (Quel résultat a-t-on le plus souvent ?), et de recueillir des données ayant pour but de fournir une réponse à cette question.

Groupe	1	2	3	4	5	6	7	8	9	10	11	12
Fente	5	4	2	3	6	3	2	4	5	6	3	4
Dos	5	6	8	7	4	7	8	6	5	4	7	6
Proportion de fentes	.5	.4	.2	.3	.6	.3	.2	.4	.5	.6	.3	.4

Tableau 1 : Recueil des données par groupe.

Groupes	1	1 à 2	1 à 3	1 à 4	1 à 5	1 à 6	1 à 7	1 à 8	1 à 9	1 à 10	1 à 11	1 à 12
Fentes	5	9	11	14	20	23	25	29	34	40	43	47
Dos	5	11	19	26	30	37	45	51	56	60	67	73
Prop. fentes	.50	.45	.37	.35	.40	.38	.36	.36	.38	.40	.39	.39

Tableau 2 : Regroupement progressif des données.



Graphique 1 : Stabilisation de la proportion de fentes par regroupement progressif.

Il s'agit également ici, en s'intéressant cette fois aux expériences aléatoires, de formaliser les constatations faites en définissant, pour une expérience donnée, un protocole expérimental garantissant qu'on pourra la répéter « dans les mêmes conditions » et en émettant des hypothèses en termes probabilistes. Dans le cas du lancer de cauri, on précisera les conditions requises pour le lancer (façon de procéder, hauteur minimale, force, nature du sol, *etc*.). Et, au vu des résultats des tableaux 1 et 2, on pourra conclure que la probabilité d'obtenir fente est plus faible que celle d'obtenir dos, voire la quantifier approximativement (à peu près 2 chances sur 5).

On passe ici de la statistique descriptive (SD) à un premier paradigme probabiliste (P1).

L'espace de travail (Kuzniak, 2009) est constitué :

- de l'expérience « pseudo-concrète », définie par le protocole et les modalités retenues ;
- des essais réalisés (espace réel et local) ;
- des outils mathématiques qui interviennent dans le calcul (moyenne, pourcentage, comparaison de nombres) et la représentation graphique (diagramme en bâtons) ;
- éventuellement, d'instruments permettant la réalisation des actions nécessaires (calculatrice, tableur-grapheur) ;
- de l'ensemble **R** des nombres réels et du plan cartésien (référentiel théorique).

Quant à l'horizon théorique de cette expérience, il s'agit du modèle d'urne de Bernoulli.

2.2. Seconde modélisation

En géométrie, on a ensuite une phase de transition de G1 à G2 : « *du dessin à la figure signifiée, regard sur la figure éclairé par un savoir théorique* » (*ibid.* Kuzniak, 2009) ; on commence à donner des définitions et des axiomes, on énonce des théorèmes et on les démontre. Dans le cas de la mosaïque, on s'intéresse aux « *propriétés de l'hexagone régulier inscrit dans un cercle. Construction des arcs de la* [mosaïque]*, propriétés géométriques de la figure abstraite* » (*ibid.* Kuzniak, 2009).

On travaille ensuite uniquement dans G2, en s'appuyant néanmoins sur les représentations figurales comme aide à la recherche de propriétés et à l'élaboration de conjectures (allers-retours G1 ↔ G2). Par exemple (*figure 6*) : la droite (BC) est-elle parallèle à la droite (AD) ? La droite (BF) est-elle perpendiculaire à la droite (AD) ? Si oui, est-elle la médiatrice du segment [OA] ? *Etc*.

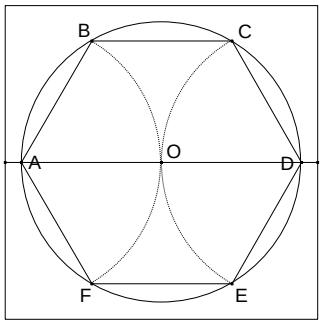


Figure 6 : Propriétés de l'hexagone régulier.

Pour l'aléatoire, on jette un « regard probabiliste » sur l'expérience, on définit l'expérience aléatoire générique, la probabilité mathématique, et on commence à étudier ses propriétés. On passe donc du paradigme précédent (P1) à un autre, plus théorique (P2), dans lequel sont mises en oeuvre des méthodes mathématiques. Ainsi, dans le cas du cauri, on introduit sous une forme quelconque la dichotomie du modèle binomial pour étudier la probabilité d'obtenir *k* fentes à l'issue de *n* lancers, …

On travaille ensuite uniquement dans P2, en s'appuyant éventuellement sur des expériences réelles ou des simulations informatiques (Parzysz, 2007), par exemple pour élaborer une conjecture, ou pour confirmer/infirmer un modèle. Ainsi, lorsqu'on lance deux fois successivement une pièce de monnaie « équilibrée », la probabilité d'obtenir deux fois pile est-elle 1/3 ou 1/4 ?

L'espace de travail est maintenant constitué de l'algèbre des événements et des propriétés de la probabilité, avec des modèles « classiques » (notamment les modèles d'urne : tirages avec ou sans remise, tirages simultanés) et des principales lois (binomiale, géométrique, *etc*.) Les outils sont la démonstration mathématique, les techniques de calcul, ainsi que divers registres de représentation (diagramme ensembliste, arbre de probabilité (*figure 7*), tableau à double entrée, …).

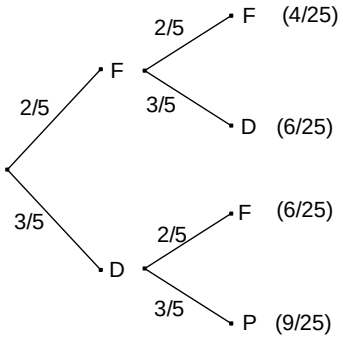


Figure 7 : Arbre associé à deux lancers successifs du cauri (F = fente, D = dos).

Il faut également y inclure les outils de la statistique descriptive (calculatrice, tableur-grapheur), mais ils sont ici utilisés de façon différente, dans le but de produire et d'étudier des expériences aléatoires simulées. On peut aussi réinvestir la statistique descriptive en explicitant l'analogie entre les notions de base de SD et de P2, comme :

fréquence ↔ probabilité
moyenne ↔ espérance mathématique
variance ↔ variance, *etc*.

ce qui permettra la « transposition » de leurs propriétés. La probabilité peut alors être considérée comme une fréquence théorique, et les connaissances en SD peuvent en quelque sorte servir de passerelles pour des apprentissages nouveaux dans P2. Par exemple : **«** *La notion de probabilité conditionnelle peut être introduite en traitant d'abord des calculs de fréquences. Par exemple on lance trois dés ; parmi les lancers dont la somme est 12, quelle proportion de lancers contiennent le nombre 2 ? (…) Le passage à la définition de « la probabilité d'avoir un 2 sachant que la somme est 12 » comme quotient de deux probabilités est plus compréhensible que si une telle définition est posée ex nihilo.* » (Schwartz & Roser, 2009, p. 10). Enfin, le référentiel théorique est constitué par le concept d'espace probabilisé (Ω, T, P), et plus généralement l'axiomatique de Kolmogorov.

À propos de simulation, le point essentiel est que ce qui est simulé est un *modèle* probabiliste, ce qui pose théoriquement problème au début de l'enseignement, avant la mise en place de cette notion. Cette difficulté peut néanmoins être contournée si l'on recourt à des simulations qui présentent une congruence sémantique avec l'expérience réelle et en explicitant les hypothèses probabilistes sous-jacentes (Parzysz, 2009a). Considérons par exemple le cas de deux tirages successifs sans remise d'une boule d'une urne à 4 boules (une blanche et trois noires). Pour étudier la distribution de probabilité du nombre de boules blanches obtenues, on commence par préciser l'hypothèse qu'à chaque tirage les boules présentes dans l'urne ont la même probabilité d'être tirées (boules « indiscernables au toucher ») et on introduira deux aléas successifs plutôt qu'un seul aléa, qui correspondrait à une autre expérience, à savoir le tirage simultané de deux boules de l'urne (*figure 8*). Certes, cette seconde expérience, plus simple à simuler, est théoriquement équivalente à l'expérience initiale pour le problème posé, mais elle ne l'est pas pour les élèves ; seule la comparaison les convaincra que c'est un même modèle qui est sous-jacent aux deux.

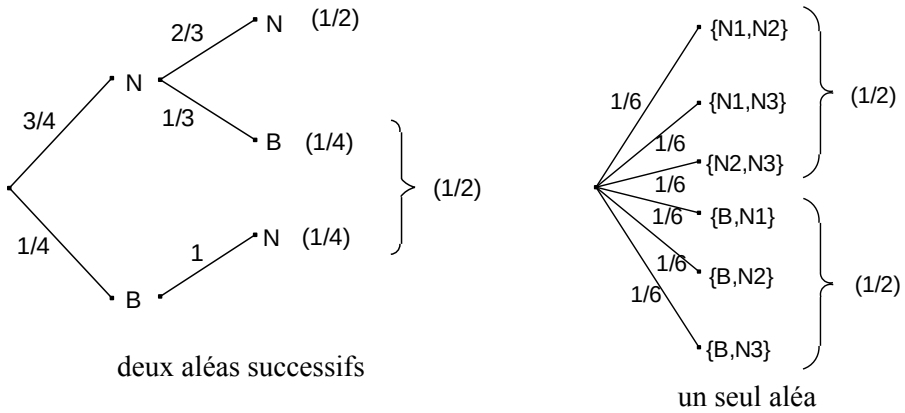


Figure 8 : Les deux modèles.

La simulation informatique peut jouer un rôle important dans l'acquisition de la notion de modèle probabiliste. En effet, la comparaison des procédures de simulation associées à diverses expériences aléatoires, ainsi que des tableaux qu'elles produisent, peut faire apparaître des similitudes conduisant à considérer comme légitime la substitution d'une expérience à une autre et à dégager l'idée d'un schéma d'expérience commun, sur laquelle pourra s'élaborer la notion de modèle (*Parzysz, 2009a,* p. 101), selon le schéma suivant (*figure 9*) :

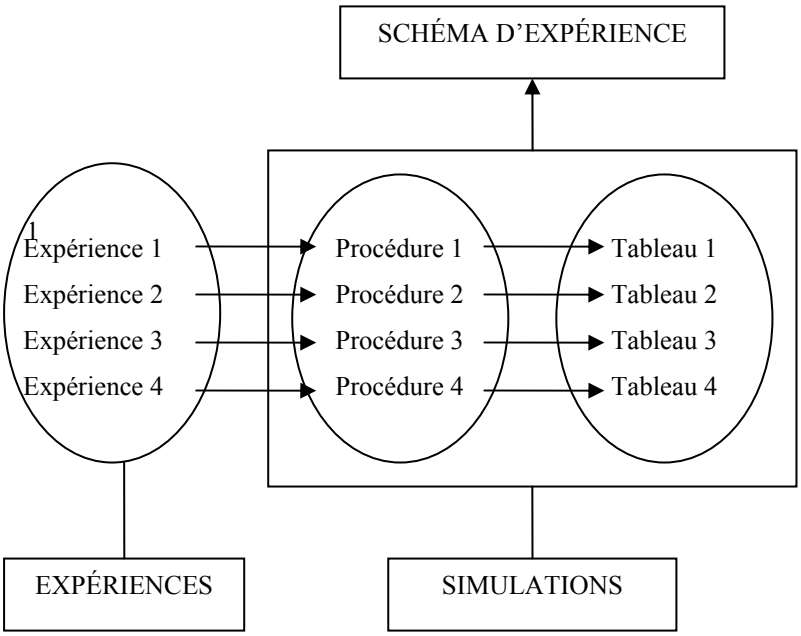


Figure 9 : De l'expérience aléatoire vers le modèle probabiliste.

3. Les registres et leur articulation dans l'enseignement de la statistique et des probabilités

Comme on en a vu un exemple plus haut, les registres de représentation sont à même de jouer un rôle fondamental dans les espaces de travail, comparable à celui des « figures » en géométrie. Le domaine de la statistique et des probabilités met en jeu des représentations variées qui constituent non seulement des *illustrations* des situations étudiées, mais surtout peuvent − à l'instar des dessins en géométrie − devenir des *outils* de résolution des problèmes, à condition de les munir de règles leur conférant un caractère opératoire. Cette variété pose, de façon peut-être plus

cruciale qu'en géométrie, non seulement la question de la *traduction* de la situation (concrète ou pseudo-concrète) dans la théorie, mais aussi la double question du *traitement* (transformation au sein d'un registre donné) et de la *conversion* (passage d'un registre à un autre) (Duval, 1995). Je voudrais maintenant montrer comment il est possible d'améliorer l'opérationnalité de certains registres de représentation, à savoir les graphiques en statistique et les arbres et tableaux en probabilités.

3.1. En statistique descriptive : élaboration et gestion des graphiques

Un problème fondamental de la statistique descriptive est l'élaboration, sur la base d'un corpus donné, d'une synthèse qui arrive à concilier les deux objectifs antagonistes que sont la *fidélité* et la *clarté* (Pichard, 1992 ; Parzysz, 1999) : fidélité à l'ensemble des données récoltées à la suite de l'enquête statistique et clarté de la synthèse qui permet de rendre compte et de présenter les résultats de cette enquête, et d'en tirer des conclusions. Certes, la technologie actuelle permet le plus souvent de préserver la totalité de l'information recueillie sous forme de fichier informatique (tableau), mais il est ensuite nécessaire de traiter ce fichier, avec pour objectif d'apporter une réponse − ou, à défaut, des éléments de réponse − aux questions initiales qui ont motivé l'enquête. C'est pour cette raison qu'ont été définis divers paramètres statistiques (de position, de dispersion, …), calculés à partir de l'ensemble des données, qui sont destinés à les synthétiser. La distribution sera alors « résumée » grâce à des valeurs « typiques ». Et, plus le nombre de ces valeurs sera réduit, plus la vue qui en résultera sera synthétique, mais plus on risquera de perdre de l'information.

3.1.1. Quel graphique ?

Les représentations graphiques élaborées, soit directement à partir du corpus, soit à partir des paramètres calculés[1], constituent un autre mode de représentation de l'information recueillie, destiné essentiellement à être appréhendé visuellement. La question de l'encodage (« écriture ») et du décodage (« lecture ») de ces représentations s'avère essentielle par le fait qu'elles ont tendance, non seulement à accompagner, mais aussi à se substituer aux données initiales. D'autre part, les tableurs-grapheurs proposent indifféremment à l'utilisateur une multitude de types de représentations graphiques mais ne prennent pas en compte la question de leur adéquation. Il est donc essentiel que celui-ci soit à même d'identifier le(s) type(s) de graphique adaptés à ses données statistiques et à ce qu'il veut mettre en évidence. Les principaux critères de choix reposent sur le type de variable (caractère), le type de situation et, pour ce qui est de la « visibilité » des résultats, sur le type de graphique.

[1] Comme par exemple la boîte de dispersion, qui visualise médiane, quartiles et étendue.

a) On définit classiquement divers types de caractères statistiques, que je me contenterai de rappeler brièvement :
   - un caractère est dit *quantitatif* lorsque ses modalités possibles sont des valeurs mesurables ; sinon, il est dit *qualitatif* ;
   - un caractère quantitatif peut être *discret* ou *continu*, selon la nature de l'ensemble de ses valeurs possibles (*possibles* et non *observées*, car celles-ci constituent toujours un ensemble discret) ;
   - un caractère qualitatif peut être *ordinal* (lorsque ses modalités sont ordonnées de façon « naturelle ») ; sinon, il est dit *nominal*.

Lorsque le nombre des modalités observées est important, une première synthèse consiste généralement − notamment dans le cas d'une variable quantitative − à les présenter dans un tableau à double entrée dans lequel elles sont groupées en classes (plusieurs questions se posent alors, et notamment celui du choix du nombre de classes et de leurs limites). On peut noter que, dans ce cas, les paramètres calculés à partir du tableau risquent de différer − de façon parfois assez sensible − de ceux calculés sur les données brutes. Il ne s'agit plus alors de *perte* d'information (puisqu'on peut toujours calculer les paramètres) mais de *modification* de l'information.

b) Le tableau statistique fournit des couples dont le premier élément est la modalité du caractère considéré et le second un effectif (ou une fréquence), cet effectif (ou cette fréquence) pouvant être :
   - soit celui de la modalité en question (exemple : répartition des individus selon leur taille) ;
   - soit celui d'un second caractère au sein de cette modalité (exemple : taux de natalité en fonction du département) ;

Dans le premier cas, J.-F. Pichard parle de situation de *partition* (les diverses modalités déterminent une partition de la population statistique), et dans le second de situation de *fonction* (la fréquence du second caractère est fonction de la modalité du premier) (Pichard, 1992, 77–80).

C'est sur la base de ce tableau que vont ensuite pouvoir être élaborées une ou plusieurs représentations graphiques, destinées à visualiser la distribution. Les tableurs-grapheurs usuels en proposent une multitude de types (pour Excel, par exemple : histogramme, barres, courbes, secteurs, nuage de points, aires, anneau, radar, surfaces, bulles, boursier, chacun de ces types comportant lui-même de 2 à 7 sous-types). Et, d'autre part, les infographistes s'en donnent à cœur joie pour les personnaliser. Un objectif de l'apprentissage doit donc être de permettre à l'élève de choisir, pour une enquête donnée, entre les divers types de graphiques *a priori* possibles, car tous ne se valent pas.

c) En ce qui concerne maintenant la typologie des graphiques, on peut en distinguer trois principaux types (Parzysz, 1999, p. 96) :
   - *orthogonal* (selon le couple de directions horizontale-verticale) : diagrammes en bâtons, en barres, histogramme, polygone ;
   - *circulaire* : diagramme circulaire (« camembert ») ou semi-circulaire ;
   - *analogique* : de loin le plus varié (exemple : carte de France où la couleur d'un département est fonction de son taux de natalité).

3.1.2. Qu'est-ce qu'un « bon » graphique ?

Pichard distingue trois qualités principales pour un graphique statistique : la lisibilité, la fidélité et l'auto-suffisance (Pichard, 1992, 76–77). Je ne parlerai pas ici de la dernière de ces qualités, qui me semble relever du simple bon sens. En ce qui concerne la première, A. Jelinski a montré que la relation d'ordre sur les fréquences est mieux perçue sur un diagramme orthogonal que sur un diagramme circulaire, et que l'évaluation quantitative d'une fréquence est également plus facile sur un diagramme orthogonal (Jelinski, 1993). Inversement, elle conclut que le diagramme circulaire « *est plus utile et même irremplaçable dans le cas de données représentant une entité, un tout* » (*op. cit.* p. 56), ce qui est aussi le cas de l'histogramme.

La deuxième qualité, déjà évoquée plus haut, concerne le rapport entre les données et leur représentation et se rapporte plus directement à la notion de congruence sémantique (Duval, 1995), question qu'il importe de se poser au moment de la conversion du tableau au graphique qu'on se propose d'utiliser pour le représenter, et qui est étroitement liée, non seulement au type de situation, mais aussi au type de caractère ; c'est ainsi qu'un diagramme en barres sera indiqué pour représenter la variation d'un caractère qualitatif, et un diagramme en bâtons pour un caractère quantitatif discret.

Il convient aussi de tenir compte de deux dangers inverses : d'une part occulter des éléments importants de la situation, et d'autre part introduire des éléments non pertinents. C'est par exemple le cas lorsqu'on représente l'évolution de la fréquence d'un caractère quantitatif continu par un diagramme en bâtons (qui discrétise le caractère) plutôt que par un « polygone » (en fait, une ligne brisée). Même si celui-ci ne représente pas exactement la distribution, il préserve la continuité du caractère, et en conséquence tout point de la courbe pourra être interprété (ce qui n'est pas le cas lorsqu'on utilise un polygone pour représenter un caractère discret). Remarquons encore que, pour un caractère continu, l'histogramme et le polygone sont deux types de représentation possibles, mais l'histogramme est plus facile à encoder et à décoder, en raison de sa meilleure congruence avec le tableau qui a servi à le créer.

Enfin, il va de soi qu'on ne trompera pas le destinataire du graphique. Je me contenterai ici d'un exemple qui concerne le « camembert 3D en perspective ». La même distribution (*tableau 3*) est ci-dessous représentée deux fois par le tableur-grapheur, avec une simple modification de l'ordre des modalités (*figure 10*).

Modalité	Fréquence
A	0,1
B	0,2
C	0,3
D	0,4

Tableau 3

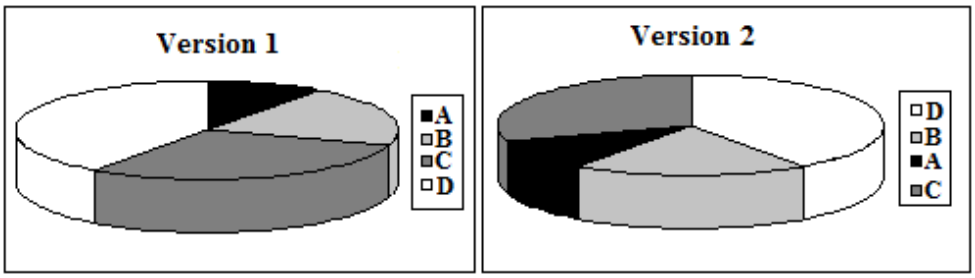


Figure 10

La différence saute aux yeux. Plus précisément, lorsqu'on calcule la proportion de l'aire visible correspondant à chacune des modalités par rapport à l'aire visible totale, on obtient les résultats suivants (*tableau 4*) :

Modalité	Version 1	Version 2
A	6 %	13 %
B	14 %	35 %
C	48 %	20 %
D	32 %	32 %

Tableau 4

On peut constater que, mis à part la modalité D (dont les représentations dans les deux versions sont symétriques), les différences sont sensibles, et on voit le parti que pourraient en tirer des gens peu scrupuleux pour favoriser telle ou telle modalité.

3.2. En probabilités : arbres et tableaux

Comme je l'ai rappelé au début, l'arbre pondéré est désormais, en France, un outil officiellement admis. Plusieurs études ont montré (Parzysz, 1997, Pluvinage, 2005) que, pourvu qu'on le nantisse de règles de traitement et de conversion précises, il possédait les caractéristiques d'un registre de représentation qui de plus, étant commun à la statistique et aux probabilités, pouvait constituer un point d'articulation entre ces deux domaines selon une transition effectifs → fréquences → probabilités (voir plus haut). Transition qui − à condition que le passage de l'un à l'autre soit dûment explicité − est susceptible de constituer un bon accès à la notion de probabilité conditionnelle, *via* celles d'effectif et de fréquence relatifs. Signalons cependant le danger de renforcer chez les élèves une conception « cardinaliste » de la probabilité conditionnelle (Totohasina, 1994), ce qui, semble-t-il, peut être évité, d'une part en passant rapidement des effectifs aux fréquences, et d'autre part en choisissant des situations dans lesquelles le recours aux effectifs est impossible ou malaisé.

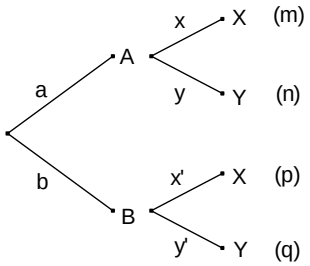
La nécessité de construire un arbre qui soit aussi proche que possible de l'expérience décrite (congruence sémantique) a déjà été évoquée plus haut ; je n'y reviendrai donc pas ici. La première difficulté de l'utilisation de l'arbre de probabilité est celle des conversions énoncé → arbre et vice versa, selon le schéma ci-dessous (*tableau 5*). Notons que l'énoncé est le plus souvent en langage naturel, mais qu'il peut aussi faire intervenir d'autres registres, tels le registre symbolique ou celui des tableaux de contingence.

	situation-→ problème	stratégie → de résolution	résolution
Domaine concerné	« réalité »	probabilités	« réalité »
Registres associés	langagier	langagier symbolique arbres tableaux	langagier

Tableau 5

Dans le tableau 5, les guillemets de « réalité » signifient que le substrat de l'énoncé est généralement un « pseudo-concret » (voir plus haut), c'est-à-dire une réalité simplifiée par la description qui en est faite. Il ne s'agit pas d'un véritable modèle, mais le plus souvent un modèle particulier est sous-entendu, voire fortement suggéré (cas des dés « non pipés » ou « homogènes », des boules « indiscernables au toucher » et des pièces de monnaie « bien équilibrées », par exemple) (Parzysz, 1980).

Par rapport au tableau de contingence (Dupuis & Rousset-Bert, 1996), un intérêt non négligeable de l'arbre est que l'on peut y faire figurer tous les éléments. Comparons par exemple l'arbre et le tableau correspondant ci-dessous (*figure 11*).



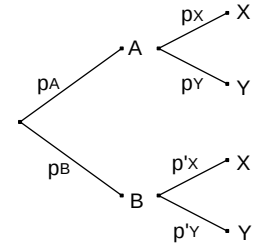
	A	B	Somme
X	m	p	m+p
Y	n	q	n+q
Somme	m+n	p+q	1

Figure 11 : Arbre de probabilité et tableau de contingence.

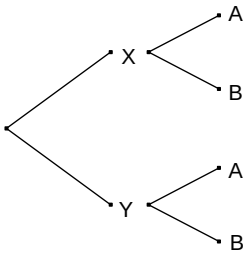
On constate immédiatement que les données visibles ne sont pas les mêmes. En particulier, les probabilités conditionnelles (x, y, x', y') n'apparaissent pas dans le tableau.

Cependant, pour la résolution des problèmes de probabilité conditionnelle, et notamment de ceux qui font intervenir la formule de Bayes, un avantage du tableau est que l'on peut l'utiliser dans les deux sens, en raison de l'interchangeabilité des lignes et des colonnes : la probabilité de A (soit m+n) et celle de X∩A (soit m) étant données, pour trouver P(X|A) il suffit de diviser la seconde par la première ; de même, si on a la probabilité de X (soit m+p) et celle de X∩A, P(A|X) s'obtient comme quotient de m par m+p. L'arbre, lui, nécessite d'être « retourné ». Les règles de traitement régissant ce retournement sont relativement simples (Parzysz, 1997b, p. 235) (*figure 12*), mais elles sont un peu plus « coûteuses » (ce sont elles qui justifient la phrase du programme rapportée au début de cet article).

Pour chaque situation, et en fonction des connaissances et de la familiarité des élèves, la question sera donc finalement celle du choix entre la lisibilité de l'arbre et la versatilité du tableau.



Arbre initial (pondéré)



Arbre retourné (à pondérer selon les règles de traitement)

Figure 12 : Retournement de l'arbre de probabilité.

4. Conclusion

Il n'était pas possible, dans le cadre restreint de cet article, d'envisager de façon exhaustive tous les problèmes didactiques qui se posent dans l'enseignement de la statistique et des probabilités. Je me suis donc contenté d'en évoquer quelques-uns, en renvoyant à des publications antérieures pour plus de détails. Ce domaine souffre en effet de deux handicaps : d'une part sa « jeunesse » en tant que domaine d'enseignement, et d'autre part la dualité de la notion de probabilité, qui dans les programmes français a mis en avant tantôt une conception et tantôt l'autre.

On a pu voir que la question de la modélisation du hasard en probabilités n'est pas sans présenter des analogies avec celle de la modélisation de l'espace en géométrie, avec notamment la nécessité de gérer des rapports ambigus avec l'expérience sensible et l'existence de plusieurs paradigmes ayant pour horizon commun une théorie mathématique de type axiomatique. En vertu de ces analogies, la notion d'espace de travail peut être étendue *mutatis mutandis* au domaine de l'aléatoire, et met en évidence le rôle des registres de représentation.

Un aspect prégnant de la statistique comme des probabilités est en effet l'usage qu'elles font des registres de représentation comme outils de résolution de problèmes : « figures » en géométrie, graphiques divers en statistique, tableaux à double entrée et arbres en probabilités. Ceci nécessite la prise en compte, dans l'espace de travail, de la façon dont sont mis en œuvre ces outils, avec pour objectif de les optimiser (on rejoint ici la notion de genèse instrumentale). Dans le domaine des probabilités, les registres en jeu constituent en effet un élément essentiel de l'espace de travail ; il en résulte que les articulations doivent être prises en compte et travaillées, que ce soit en statistique descriptive (type de variable ↔ type de graphique) ou en probabilités (type de problème ↔ type de registre).

Ainsi, en statistique descriptive, la grande variété des types de représentations graphiques et le fait que les outils technologiques les présentent comme *a priori* également acceptables peuvent conduire à des incompréhensions et à des dérives. Et les enjeux sociaux sont tels qu'on ne peut pas faire l'économie d'une réflexion en profondeur sur cet aspect souvent jugé mineur − parce que transversal ? − de l'enseignement. De même, la notion de congruence sémantique s'avère fondamentale, que ce soit pour choisir le registre le mieux adapté au contexte (situation, élèves, …), pour contrôler les conversions de registres ou pour guider l'apprentissage des concepts.

Les travaux didactiques dans ce domaine sont encore relativement peu nombreux, mais on peut espérer que, en lien avec la montée en puissance du domaine et le développement d'une culture de l'aléatoire chez les professeurs, les chercheurs vont s'intéresser davantage à son enseignement, ainsi qu'en témoignent des travaux comme la récente thèse de P. Carranza (Carranza, 2009).

Bibliographie

BATANERO, C. (2001), *Didáctica de la estadística,* Universidad de Granada.

BERNOULLI, J. (1713), *Ars Conjectandi*, Traduction N. Meusnier, IREM de Rouen 1987.

CARRANZA, P. (2009), *La dualité de la probabilité dans l'enseignement de la statistique. Une expérience en classe de BTS*, Thèse de doctorat, Université Paris-Diderot.

DOUMBIA, S. & PIL, J.-C. (1992), *Les jeux de cauris*, Institut de Recherches Mathématiques d'Abidjan.

DUPUIS, C. & ROUSSET-BERT, S. (1996), Arbres et tableaux de probabilité : analyse en termes de registres de représentation, *Repères-IREM*, **22**, 51–72.

DUVAL, R. (1995), *Semiosis et pensée humaine*, Peter Lang, Berne.

HENRY, M. (1999), L'introduction des probabilités au lycée : un processus de modélisation comparable à celui de la géométrie, *Repères-IREM*, **36**, 15–34.

HENRY, M. (2009), Émergence de la probabilité : de la définition classique à l'approche fréquentiste, quelle introduction en classe de troisième ?, *Repères-IREM*, **74**, 67–89.

HOUDEMENT, C. & KUZNIAK, A. (1999), Quelques éléments de réflexion sur l'enseignement de la géométrie, de l'école primaire à la formation des maîtres, *Petit x,* **51**.

HOUDEMENT, C. & KUZNIAK, A. (2006), Paradigmes géométriques et enseignement de la géométrie, *Annales de Didactique et de Sciences Cognitives*, **11**, 175–193, IREM de Strasbourg.

JELINSKI A. (1993), Diagramme circulaire ou orthogonal ? Une efficacité différente des images graphiques dans la transmission de l'information, *Les Sciences de l'Education*, **1-3**, 39–56.

KUZNIAK, A. (2009), Sur la nature du travail géométrique dans le cadre de la scolarité obligatoire, *Actes de la 14e École d'été de Didactique des Mathématiques* (Bloch I. & Conne F., éds), La Pensée Sauvage, Grenoble.

PARZYSZ, B. (1980), L'équiprobabilité, est-ce que cela va sans dire ? *Chantiers de pédagogie mathématique*, **47**, 2–4, APMEP.

PARZYSZ, B. (1997a), Les probabilités et la statistique dans l'enseignement secondaire, d'hier à aujourd'hui, *Enseigner les probabilités au* lycée (M. Henry, éd.), 17–38, APMEP / ADIREM,

PARZYSZ, B. (1997b), L'articulation des cadres et des registres en probabilités : le cas des arbres et des tableaux, *Enseigner les probabilités au lycée* (M. Henry, éd.), 225–238, APMEP / ADIREM.

PARZYSZ, B. (1999), Heurs et malheurs du su et du perçu en statistique, des données à leurs représentations graphiques, *Repères-IREM*, **35**, 91–112.

PARZYSZ, B. (2003), Articulation entre perception et déduction dans une démarche géométrique en PE1, *Concertum. Dix ans de formation des professeurs des écoles en mathématiques,* Tome 2 (Démarches et savoirs à enseigner) chap. 1 (Espace et géométrie), 107–125, ARPEME.

PARZYSZ, B. (2007), Expérience aléatoire et simulation : le jeu de croix ou pile. Relecture actuelle d'une expérimentation déjà un peu ancienne, *Repères-IREM*, **66**, 27–44.

PARZYSZ, B. (2009a). Des expériences aléatoires au modèle, *via* la simulation, *Repères-IREM*, **74**, 91–103.

PARZYSZ, B. (2009b). À la recherche des espaces de travail géométrique des mosaïstes antiques, *Chypre et France. Recherche en Didactique des Mathématiques* (A. Gagatsis, A. Kuzniak, E. Deliyanni & L. Vivier, éd.), 287–305, Université de Chypre.

PARZYSZ B. (2011), Un générateur aléatoire de pile ou face venu d'ailleurs, *Bulletin de l'APMEP*, **494**, 309–314.

PICHARD, J.-F. (1992), Représentations graphiques en statistiques, *Bulletin inter-IREM, « Des chiffres et des lettres »*, 75–101, IREM de Rouen.

PLUVINAGE, F. (2005), Árboles de transiciones etiquetadas en cálculo de probabilidades, *Relime*, **8-1**, 91–99.

RENYI, A. (1966), *Calcul des probabilités*, Dunod, Paris.

SCHWARTZ, C. & ROZER, E. (2009), L'esprit des probabilités, de l'école au lycée. *MathémaTice*, **13** (revue en ligne).

TOTOHASINA, A. (1993), Introduction du concept de probabilité conditionnelle. Avantages et inconvénients de l'arborescence, *Repères-IREM*, **15**, 93–117.

Bernard PARZYSZ
Université d'Orléans
& Laboratoire de Didactique André Revuz
Université Paris-Diderot
Paris, France
parzysz.bernard@wanadoo.fr