PROBABILITÉS ET STATISTIQUE INFÉRENTIELLE APPROCHE SONDAGE VERSUS APPROCHE MODÈLE

Jeanne FINE[1]

TITLE

Probability and Statistical Inference – Sampling Approach versus Model Approach

RÉSUMÉ

La statistique enseignée dans le secondaire ou à l'université en France ne traite quasiment que des variables réelles et repose sur la modélisation probabiliste et l'échantillonnage i.i.d. (variables aléatoires indépendantes et identiquement distribuées) (*approche modèle*) alors que la statistique des sciences humaines et sociales concerne très souvent des variables catégorielles définies sur des populations finies, populations desquelles il est possible d'extraire des échantillons en utilisant des procédures aléatoires (*approche sondage*).

Quelques questions ont été posées à des étudiants venant d'obtenir une licence de mathématiques et souhaitant devenir professeur dans le secondaire afin de tester leurs représentations de l'aléatoire et de la probabilité et leur aptitude à raisonner de façon probabiliste et statistique. Dans la section 2, l'analyse des réponses à ce test permet de constater que les cours de probabilités et de statistique inférentielle qu'ils ont suivis à l'université ne leur donnent pas un « mode de pensée statistique ».

La théorie des sondages (échantillonnage et estimation dans des populations finies) modifie profondément les concepts étudiés en probabilités et statistique inférentielle. Dans l'approche sondage, aucune hypothèse n'est faite sur la distribution de probabilité ; l'aléatoire vient de l'échantillonnage (procédure aléatoire de génération de l'échantillon). L'approche sondage permet de sortir du paradigme de l'approche modèle dans des contextes simples, de redonner du sens au vocabulaire de la statistique (population, individus, échantillons indépendants et échantillons appariés, …), de faire la place aux variables catégorielles et à la statistique des sciences humaines et sociales. L'objectif de la section 3 est de donner quelques éléments de comparaison entre l'approche modèle et l'approche sondage et de montrer leur complémentarité.

En conclusion, nous esquissons les grandes lignes d'une nouvelle progression pour l'enseignement de la statistique et des probabilités en collège et lycée.

Mots-clés : *approche sondage, approche modèle, modélisation, simulation, mode de pensée statistique, représentation de l'aléatoire.*

ABSTRACT

Teaching statistics throughout secondary or academic courses in France deals almost only with real variables and relies on probabilistic model and i.i.d. sampling model (*model approach*) whereas social and human sciences statistics mostly refer to categorical variables defined on finite population, population in which random samples extraction is possible (*sampling approach*).

Some questions have been asked to students who have just graduated a bachelor's degree in mathematics and who want to become secondary school teachers, in order to test their random conception and their ability to think in a probabilistic and statistical way. In section 2, the answers analysis enables us to notice that probability and statistical inference courses carried out at the university fail to give them an aptitude to think in a probabilistic and statistical way.

[1] IUFM Midi-Pyrénées, Institut de Mathématiques, Laboratoire de Statistique et Probabilités, Université de Toulouse, jeanne.fine@cict.fr

The survey sampling theory (sampling and estimation in finite populations) deeply changes the concepts studied in probability and statistical inference. In the sampling approach, there is no assumption made on the distribution of probability; randomness proceeds from the sampling process. The model approach makes possible to get rid of the model approach paradigm in simple contexts, to give sense to the statistics vocabulary (population, individuals, independent and matched samples, ...) and to give way to categorical variables and social sciences statistics. The purpose of the third section is to give some elements of comparison between the model approach and the sampling approach and to show how they complement one another.

In conclusion, we outline guidelines for a new planning for statistics and probability teaching throughout secondary school.

Keywords: *sampling approach, model approach, modeling, simulation, statistical way of thinking, random conception.*

1 Introduction

En juillet 2000, un groupe d'experts présente à l'Académie des sciences un rapport sur « La Statistique » en France, faisant état d'un retard pris par la France dans cette discipline et constatant que « les citoyens n'ont pas une formation suffisante à la prise en compte du mode de pensée statistique » (Malliavin, 2000).

Un groupe d'experts travaillait déjà à la réécriture des programmes scolaires. De nouveaux programmes de mathématiques au lycée (seconde, première et terminale), incluant une part importante en statistique et probabilités, sont mis en place entre 2000 et 2002. De même en collège, de nouveaux programmes sont progressivement mis en place à partir de 2005 pour la sixième jusqu'en 2008 pour la troisième (avec une introduction aux probabilités pour la première fois au collège). Parmi les thèmes de convergence disciplinaire proposés dans le cadre de la rénovation des programmes de sciences du collège, figure « l'importance du mode de pensée statistique dans le regard scientifique sur le monde » (cf. en référence [11] l'adresse du site du ministère de l'éducation nationale où l'on peut trouver les programmes et les documents d'accompagnement).

Dans le secondaire, ce sont les professeurs de mathématiques qui sont chargés d'enseigner la statistique et les probabilités, ainsi que les TICE (Technologies de l'Information et de la Communication en Éducation). Les sujets du concours commun de recrutement des professeurs de mathématiques de l'enseignement public (CAPES) et privé (CAFEP) prennent en compte cette évolution et portent de plus en plus souvent sur cette partie du programme, à l'écrit comme à l'oral.

Les étudiants inscrits à la préparation au CAPES-CAFEP de Mathématiques à l'Institut Universitaire de Formation des Maîtres (IUFM) de Midi-Pyrénées ont généralement obtenu une licence de mathématiques dont le cursus contient un cours de statistique descriptive, de calcul des probabilités et quelques éléments de statistique inférentielle. Ils sont a priori mieux préparés pour enseigner la statistique et les probabilités que les professeurs en poste, qui se forment bien souvent à partir des programmes et des manuels scolaires. Il nous a semblé intéressant de tester ces étudiants en tout début d'année scolaire à l'IUFM sur leurs représentations de la probabilité et de l'aléatoire et sur leur aptitude à utiliser un mode de pensée statistique.

Dans la section 2, l'analyse des réponses des étudiants à ces questions introductives permet de constater que les cours de probabilités et de statistique inférentielle qu'ils ont suivis à l'université ne leur donnent pas une aptitude à raisonner de façon probabiliste et statistique ;

en effet, une grande majorité des étudiants ont été déconcertés par les questions posées et le mode de pensée statistique n'est pas du tout acquis.

Courtebras (2006), dont l'ouvrage retrace l'histoire de l'enseignement des probabilités en France, écrit :

« L'étude des premiers enseignements (fin XVIII^{e} et XIX^{e} siècles) montre que ceux-ci avaient pour but de permettre une utilisation pratique ainsi que la construction des dispositions critiques nécessaires à la constitution d'une société de citoyens scientifiquement éclairés, plus raisonnables dans leurs espérances et dans leurs craintes. L'étude des formes scolaires d'enseignement des probabilités au XX^{e} siècle fait apparaître que celles-ci ont, sous l'alibi pédagogique, déformé le savoir scientifique en le parcellisant et en l'organisant autour de la répétition d'exercices artificiels et stéréotypés : ces pratiques s'inscrivent dans l'entreprise d'assujettissement inhérente à la transmission scolaire de savoirs, qui organise et justifie la sélection des élèves autour d'activités calculatoires sans autre finalité que la seule réalisation de ces activités. »

Cette critique ne concerne pas uniquement l'enseignement des probabilités ; au collège, la mise en place du *socle commun des connaissances et des compétences* et des *thèmes de convergence disciplinaire* est une réponse à cette critique. « *Faire des mathématiques*, c'est se les approprier par l'imagination, la recherche, le tâtonnement et la *résolution de problèmes*, dans la rigueur de la logique et le plaisir de la découverte. Les mathématiques aident à structurer la pensée et fournissent des modèles et des outils aux autres disciplines scientifiques et à la technologie. » (§ II.1 de l'introduction commune aux programmes de l'enseignement de mathématiques du collège, Bulletin Officiel spécial n°6 du 28 août 2008, [11]). Pour l'enseignement de la statistique, le site Statistix ([16]) répond exactement aux objectifs des nouveaux programmes.

Dans une autre étude sur l'histoire de l'enseignement des probabilités et de la statistique, Meusnier (2006) écrit :

« L'éducation à l'aléatoire, qui passe par l'enseignement des probabilités et de la statistique, devrait avoir pour but fondamental la prise de conscience que toute décision s'accompagne d'un risque, mais que ce risque peut être évalué. »

Si les objectifs de cet enseignement sont largement partagés, il n'en est pas de même de la façon d'y parvenir. Pour l'enseignement des probabilités, de nombreux didacticiens insistent sur l'importance de prendre en compte des aspects épistémologiques et historiques. On peut citer Lahanier-Reuter (1999) parlant des élèves et des étudiants :

« Leurs conceptions du hasard, issues de leurs pratiques quotidiennes, font obstacle à la construction du concept de hasard en probabilités. »

Elle poursuit en précisant que le hasard du quotidien est un *hasard subi*, qui peut être heureux ou malheureux mais toujours inattendu, alors que le hasard en probabilités est un *hasard construit* (le plus souvent équiprobabilité quand on dit « au hasard »).

Enfin, citons Armatte (2004) en conclusion d'un article faisant référence à Adolphe Quetelet (1796-1874). A. Quetelet remarque que la distribution des tailles des conscrits a la même forme que la distribution des erreurs de mesures d'une distance astronomique que venaient de « découvrir » Laplace et Gauss. En 1846, il transpose ces découvertes aux sciences morales et politiques, en utilisant la métaphore du gladiateur : le roi de Prusse fait réaliser 1000 copies par 1000 sculpteurs différents d'une statue de gladiateur. La distribution

des tailles d'individus différents peut être assimilée à la distribution de 1000 mesures d'un même individu, ce qui suscitera plusieurs controverses sur le concept d'homme moyen et sur l'hypothèse d'homogénéité (cf. Armatte, 2004 ; Desrosières, 2002). En conclusion de son article, Armatte (2004) écrit :

« Les tentatives faites dans l'après guerre pour fusionner ces deux approches (traitement de l'erreur dans les sciences exactes et traitement de la variabilité dans les sciences de l'homme) au sein d'un même corpus rebaptisé "statistique mathématique" montrent alors leurs limites : l'utilité de ces synthèses à des fins d'unification des sciences, ou à des fins pédagogiques cache mal que les hybrides produits, cohérents d'un point de vue syntaxique, sont parfois totalement inconsistants d'un point de vue sémantique et pragmatique. »

On retrouve aujourd'hui les deux champs : d'une part, la statistique enseignée ne traite quasiment que des variables réelles et repose sur la modélisation probabiliste et l'échantillonnage i.i.d. (variables aléatoires indépendantes et identiquement distribuées) (*approche modèle*) alors que, d'autre part, la statistique des sciences humaines et sociales concerne très souvent des variables catégorielles définies sur des populations finies, populations desquelles il est possible d'extraire des échantillons en utilisant des procédures aléatoires (*approche sondage*).

La théorie des sondages (échantillonnage et estimation dans des populations finies) modifie profondément les concepts étudiés en probabilités et statistique inférentielle ; dans l'approche sondage (car il existe aussi une approche modèle dans la théorie des sondages), aucune hypothèse n'est faite sur la distribution de probabilité ; l'aléatoire vient de l'échantillonnage (procédure aléatoire de génération de l'échantillon). L'approche sondage permet de sortir du paradigme de l'approche modèle dans des contextes simples, de redonner du sens au vocabulaire de la statistique (population, individus, échantillons indépendants et échantillons appariés, …), de faire la place aux variables catégorielles et à la statistique des sciences humaines et sociales. L'objet de la section 3 est de donner quelques éléments de comparaison entre l'approche modèle et l'approche sondage et de montrer leur complémentarité.

Enfin, en conclusion, nous esquissons les grandes lignes d'une nouvelle progression de l'enseignement de la statistique et des probabilités au collège et au lycée.

2 Analyse des réponses obtenues aux sept questions introductives

Ce test a été proposé aux étudiants inscrits à la préparation au CAPES et CAFEP de mathématiques de l'IUFM Midi-Pyrénées en septembre 2007 lors de la première séance de « Dénombrement, Probabilités et Statistique ».

Il leur a été annoncé que les questions visaient à tester leurs représentations de l'aléatoire et de la probabilité et leur aptitude à raisonner de façon probabiliste et statistique, que les solutions pouvaient être rédigées sans formalisme dans la mesure où c'est l'argumentation qui nous intéressait. Enfin, les étudiants avaient une heure pour répondre aux questions posées.

Questions 1) et 2)

1) Écrivez une série de 100 « P » ou « F » (P pour pile, F pour face) qui pourrait être considérée comme le résultat du lancer d'une pièce de monnaie équilibrée 100 fois de suite.

2) On lance une pièce de monnaie équilibrée 100 fois de suite. Quelle est la probabilité d'obtenir 100 « P » ? 100 « F » ? 50 « P » puis 50 « F » dans cet ordre ? la série alternée (de taille 100) P,F,P,F,…,P,F ? la série de 100 « P » ou « F » que vous avez écrite à la question 1 ?

Objectif

Ces deux premières questions ont pour objectif de discuter de la notion d'aléatoire et de distinguer :

- *l'aléatoire qui provient du processus de génération des données* ; à ce titre, les 2^{100} séries qu'il est possible d'obtenir en lançant une pièce de monnaie équilibrée 100 fois de suite répondent à la question et elles ont toutes la même probabilité de réalisation $1/2^{100}$.
- *la configuration aléatoire d'une suite donnée*, concept qui vient de la théorie de l'information, lié à la longueur des algorithmes permettant d'écrire la suite ; il s'agit alors de tester non seulement l'équiprobabilité sur {P,F} mais aussi « l'indépendance » des lancers successifs : « ni ordre ni périodicité ».

Résultats

La question 2, la plus conventionnelle, a été en général bien traitée.

Pour la question 1, un tiers des étudiants a proposé une des suites proposées dans la question 2 (premier sens d'aléatoire), les deux autres tiers ont effectivement tenté de produire une suite pouvant être considérée comme aléatoire dans le deuxième sens du terme.

Cet exercice s'avère difficile, la tendance est d'alterner trop souvent les piles et les faces. Parmi les 2^{100} séries de piles ou faces de la deuxième question, 80 % contiennent au moins une série de piles ou de faces consécutifs de longueur supérieure ou égale à 6 ; 15 % seulement des étudiants ont produit une telle suite.

On pourra consulter une autre présentation de la question 1 dans Batanero (2001).

Questions 3) et 4)

Les deux questions suivantes sont rédigées de la même façon, la première concerne le lancer d'une punaise (pour tableau d'affichage), la seconde le lancer d'une pièce de monnaie équilibrée.

3) On lance une punaise (pour tableau d'affichage) 100 fois de suite. Elle est tombée sur la pointe 52 fois et sur la tête 48 fois. On se propose de la lancer une fois de plus. À votre avis, a-t-elle plus de chance de tomber sur la pointe ou sur la tête ? Argumentez votre réponse.

4) On lance une pièce de monnaie équilibrée 100 fois de suite. On a obtenu « Pile » 52 fois et « Face » 48 fois. On se propose de la lancer une fois de plus. À votre avis, a-t-on plus de chance d'obtenir « Pile » ou « Face » ? Argumentez votre réponse.

Réponses attendues et objectif

Dans le cas de la punaise, nous proposons une « *modélisation* » en notant p la probabilité qu'elle tombe sur la pointe et $1-p$ la probabilité qu'elle tombe sur la tête. L'existence d'une telle probabilité est tout à fait discutable ; il s'agit d'un modèle, simple mais suffisant ici. Puisque la punaise a été lancée 100 fois et qu'elle est tombée sur la pointe 52 fois, la loi des

grands nombres nous permet de proposer comme estimation de *p* la fréquence d'obtention de « pointe » (0.52) sur les 100 lancers. Si l'on devait parier sur l'obtention de « pointe » ou « tête », l'information donnée par les 100 lancers nous conduit à parier sur « pointe ».

Dans le cas de la pièce de monnaie « équilibrée », les observations peuvent être considérées comme une « *simulation* » de la loi équirépartie sur {P,F} 100 fois de suite ; l'adjectif « équilibrée » est traduit comme une hypothèse d'équiprobabilité qui n'est pas objet de discussion dans cet énoncé. L'écart entre la fréquence 0.52 d'obtention de « pile » sur les 100 lancers et la probabilité 0.50 vient de la fluctuation d'échantillonnage et n'apporte aucune information. Il y a donc autant de chance d'obtenir « pile » que « face ».

L'objectif de ces deux questions est donc de faire la différence entre « modélisation » et « simulation ». Dans le premier cas, la fréquence de succès observée est utilisée pour estimer la probabilité de notre modèle. Dans le second cas, la fréquence de succès observée n'apporte aucune information et « tester » l'hypothèse d'équiprobabilité n'est pas l'objet de l'énoncé.

Résultats

Sur les 60 étudiants présents, 15 donnent une réponse correcte aux deux questions. Parmi eux, 4 font référence au fait que l'on ne sait pas si la punaise est équilibrée ; ce commentaire pourrait être considéré comme bien maladroit, « pourquoi la punaise serait-elle équilibrée ? », mais pour certains, il s'agit clairement de faire ressortir la différence d'argumentation entre le lancer de la punaise et le lancer d'une pièce de monnaie équilibrée.

Est considérée également parmi les réponses correctes une réponse qui met en doute le fait que la pièce soit équilibrée au vu des résultats 52 % de « pile » et 48 % de « face ». C'est une méconnaissance de l'ampleur de la fluctuation d'échantillonnage pour un échantillon de taille 100 qui est en jeu ici et non la différence de situation entre le lancer de punaise et le lancer d'une pièce de monnaie équilibrée.

18 étudiants refusent de répondre : « on ne peut pas savoir » (même pour la pièce de monnaie), « ce n'est qu'un échantillon », « le nombre de lancers est insuffisant » ou se réfèrent à des arguments de physique : « normalement, la punaise devrait tomber sur la tête », …

27 étudiants proposent l'égalité des chances pour « pointe » et pour « tête » pour la punaise, sans argumentation pour 9 d'entre eux, avec des arguments différents pour les autres : 4 écrivent « parce qu'on a une chance sur deux que la punaise tombe sur la pointe », « parce que les deux issues sont équiprobables » ou « parce que la punaise est équilibrée », 8 « parce que les résultats des 100 premiers lancers permettent d'accepter l'hypothèse que la punaise est équilibrée », enfin 6 « parce que les lancers sont indépendants et que le résultat du 101ème lancer ne dépend donc pas des résultats précédents ».

Pour ce dernier argument, il est bien intéressant de constater que, pour certains étudiants, l'*indépendance* implique que l'on ne peut tirer aucune information du passé sur *la loi de probabilité*.

Les références à la physique sont nombreuses ; est-ce irréaliste d'écrire qu'une punaise tombe sur la pointe 52 fois et sur la tête 48 fois ? Dans le doute, lors du corrigé des réponses au test, le lancer effectif d'une punaise 100 fois de suite a donné 50 fois « pointe » et 50 fois « tête » !

Les deux questions sont corrigées ensemble car les réponses se renvoient l'une l'autre. Pour la pièce de monnaie, 18 étudiants écrivent « même raisonnement qu'à la question précédente ».

Commentaires

Le fait que, même pour la punaise, le modèle d'équiprobabilité s'impose, est riche d'enseignement. Quelles auraient été les réponses si on avait proposé 70 « pointe » et 30 « tête » ? Très probablement, les réponses auraient été différentes : la punaise aurait évidemment plus de chance de tomber sur la pointe que sur la tête et « l'hypothèse » d'une pièce équilibrée aurait été rejetée !

Se pose donc le statut de ce type d'énoncé. Le « lancer d'une pièce équilibrée » serait une expérience aléatoire « *pseudo-concrète* » (on utilise une pièce de monnaie équilibrée pour simuler la loi de Bernoulli de paramètre ½) plutôt qu'un modèle « *pseudo-concret* » (on suppose que les résultats obtenus par le lancer de la pièce équilibrée peuvent être considérés comme les observations d'une loi de Bernoulli de paramètre ½) (cf. Chaput, Girard, Henry, 2008).

Au chapitre « choisir au hasard » du livre « Contes et Décomptes de la Statistique », C. Schwartz propose :

« Choisir 10 éléments au hasard dans un ensemble fini *E* signifie, *par convention*, que :

(1) chaque choix est fait dans *E* selon la loi de probabilité équirépartie ;

(2) les choix sont indépendants. »

Choisir au hasard 10 éléments de *E*, c'est donc dire, au niveau de la *modélisation*, que la série des résultats est un échantillon de taille 10 de la loi équirépartie sur *E*.

Dans la pratique des *sondages*, tirer au hasard et avec remise un échantillon de taille 10 dans une population finie, c'est utiliser un générateur de nombres pseudo-aléatoires permettant de *simuler* l'équiprobabilité sur l'ensemble de la population (dont la liste des individus constitue la *base de sondage*).

De fait, que l'on adopte l'approche modèle ou l'approche sondage, nous travaillerons bien sur des échantillons i.i.d..

Question 5)

La question suivante concerne un jeu national au Chili mais on retrouve les mêmes arguments pour le loto en France où il s'agit de choisir six numéros distincts parmi {1,…,49}.

5) Le jeu de loterie national TOTO 3 au Chili consiste à choisir trois numéros entre 0 et 9 (pas nécessairement distincts). Le gros lot est distribué à ceux qui ont misé sur les trois numéros tirés « au hasard » et avec remise parmi les 10 par la société organisatrice du jeu. On trouve sur Internet les statistiques de sortie des dix numéros depuis le début de ce jeu :

Tableau des résultats

Numéros	0	1	2	3	4	5	6	7	8	9
Nbre de sorties	1728	1633	1717	1670	1676	1766	1660	1647	1704	1665
Fréq. de sorties	0.1025	0.0968	0.1018	0.0990	0.0994	0.1047	0.0984	0.0977	0.1010	0.0987

On trouve également sur Internet des conseils pour le moins contradictoires :

(a) il est conseillé de jouer les numéros qui sortent le plus souvent (numéros en forme) car la fréquence théorique de sortie d'un numéro converge vers la probabilité de sortie de ce numéro (appelée probabilité expérimentale) ;

(b) il est conseillé de jouer les numéros qui sortent le moins souvent car la fréquence théorique de sortie d'un numéro converge vers la probabilité de sortie de ce numéro qui est égale à 0.1 ; les numéros de fréquence inférieure à 0.1 vont donc sortir plus souvent, et ceux de fréquence supérieure à 0.1 moins souvent, afin que les fréquences convergent toutes vers 0.1.

Que conseilleriez-vous à un joueur ?

Réponses attendues

Comme dans le cas de la pièce de monnaie, les différences entre les fréquences observées et la probabilité 0.1 sont dues aux fluctuations d'échantillonnage ; ces fréquences n'apportent aucune information. Il est indifférent de choisir un chiffre plutôt qu'un autre. On n'a donc aucun conseil à donner à un joueur concernant le choix d'un chiffre car les probabilités des dix chiffres sont toutes égales à 0.1. Les fréquences des dix chiffres sont effectivement très proches de 0.1 (un test du Khi2 ne permet pas de rejeter l'hypothèse d'équirépartition des dix chiffres).

Résultats

32 étudiants ont correctement répondu à cette question, 2 ont répondu qu'il est préférable de choisir des chiffres distincts, 2 n'ont pas répondu, 9 ont répondu (a), 10 ont répondu (b) et 5 ont répondu (a) et (b) ou hésitent entre (a) et (b) : « les études semblent dégager des conseils contradictoires mais aussi convaincants l'un que l'autre », « (b) car l'équiprobabilité de sortie des numéros semble respectée, (a) car c'est physiquement impossible que l'équiprobabilité soit exactement respectée ».

Cette question a donc déstabilisé la moitié des étudiants ayant obtenu une licence de mathématiques.

Ceux qui ont choisi (a) ont une conception bien différente de ceux qui ont choisi (b).

Pour les premiers, il existe une « *probabilité expérimentale* » (nous reviendrons sur ce vocabulaire) attachée à cette série de tirages : « je conseillerais (a), si ces numéros sont sortis plus souvent, ils continueront à sortir plus souvent », « (a) car c'est un jeu de pur hasard, les fréquences convergent vers les probabilités de sortie des chiffres ». On peut reprendre la réponse ci-dessus : « (a) car c'est physiquement impossible que l'équiprobabilité soit exactement respectée » et donc, c'est moi qui ajoute, on peut physiquement échapper à l'équiprobabilité, les probabilités expérimentales peuvent être différentes de 0.1.

Pour les seconds, c'est « *l'absence de mémoire* » qui n'est pas intégrée. C'est dès l'école élémentaire que C. Schwartz et E. Roser (2009) proposent de travailler cette notion à partir des lancers de dés. Certains étudiants réécrivent le conseil (b) : « les fréquences de sortie d'un numéro tend vers sa probabilité de sortie qui est 0.1 ; tous les numéros dont la fréquence de sortie est inférieure à 1 devraient voir leur fréquence de sortie ultérieure augmenter et ceux dont la fréquence de sortie actuelle est supérieure à 0.1 devraient sortir moins souvent », « la convergence des fréquences vers 0.1 se fera après un nombre (théoriquement) infini de tirages ».

Notons qu'un même étudiant hésite entre (b) car la convergence vers 0.1 nécessite un rattrapage des chiffres de fréquence inférieure à 0.1 et (a) car physiquement cette série de tirage va donner des probabilités limites différentes de 0.1.

Retour sur la fausse idée de « probabilité expérimentale »

Des étudiants chiliens, futurs professeurs de mathématiques, ont passé quatre mois à l'IUFM Midi-Pyrénées en 2007. Dans le cadre d'un projet, trois étudiants ont proposé de travailler sur le thème « lien entre fréquence relative et probabilité, loi des grands nombres, simulation », au programme de « tercer año medio » ([12], ce qui correspond à la classe de première en France). Dans un passage du programme chilien intitulé « orientations didactiques », on lit :

« Pour pouvoir définir la notion de *probabilité expérimentale* il sera nécessaire de définir la *fréquence relative du résultat A d'une expérience* comme le quotient du nombre de fois que le résultat *A* arrive effectivement sur le nombre de fois que l'on réalise l'expérience. On pourra dire alors que la *probabilité expérimentale* du succès *A* sera approchée par la valeur de cette fréquence relative, quand l'expérience est réalisée un grand nombre de fois. Cela permettra une bonne approximation pour une compréhension intuitive de la Loi des Grands Nombres. »

Les étudiants proposent comme projet un travail d'une heure et demi sur le jeu TOTO 3, à faire réaliser par les élèves deux par deux dans une salle informatique. Le rapport, d'une dizaine de pages, détaille les objectifs, les pré-requis, la suite des différentes activités proposées, les hypothèses didactiques, les réponses attendues. Ils insistent sur le fait que les développements qui suivent ne concernent que les jeux de *pur hasard* et non les jeux comme le poker ou le tiercé qui font appel à l'adresse du joueur ou à une connaissance des chevaux. La première activité consiste à proposer aux élèves le tableau de fréquences de sorties des dix chiffres depuis le début du jeu TOTO 3 (celui présenté à la question 5) et à les faire jouer en commentant leur choix. Des simulations sont ensuite proposées afin d'observer les fluctuations d'échantillonnage. La session se termine par un résultat encadré qui présente le concept à institutionnaliser (il s'agit d'un énoncé vulgarisé de la loi des grands nombres) et par un retour des élèves sur la première activité, « quels chiffres choisiraient-ils à présent ? ».

N'étant pas suffisamment sollicitée lors de la préparation de ce projet, c'est lors de la soutenance qu'il est apparu que c'est la réponse (a) de la question 5 qu'ils cherchaient à institutionnaliser. Suite à une demande de leur part concernant la bibliographie sur « la notion de *probabilité expérimentale* », j'avais répondu que cette notion n'existe pas en probabilité et que l'introduire à des fins didactiques présente des dangers. Je n'avais pas réalisé que leur interprétation de la loi des grands nombres était erronée.

Questions 6) et 7)

Les deux dernières questions sont les plus difficiles. L'objectif est de voir si les étudiants ont acquis un *mode de pensée statistique* (cf. le thème de convergence au collège) ou s'ils restent sur des raisonnements purement déductifs et certains.

6) Une urne opaque contient *N* boules indiscernables. On tire « au hasard » et simultanément 40 boules de l'urne, on fait une marque sur chacune des boules tirées et on les remet dans l'urne.

On tire de nouveau « au hasard » et simultanément 40 boules de l'urne ; 8 d'entre elles sont marquées. Que peut-on dire du nombre *N* de boules de l'urne ?

7) Une urne opaque contient *N* boules indiscernables au toucher numérotées de 1 à *N*. On tire « au hasard » et avec remise une boule de l'urne quatre fois de suite. Les numéros tirés sont les suivants : 512, 987, 355 et 1200. Que peut-on dire du nombre *N* de boules de l'urne ?

Réponses espérées (plutôt qu'attendues)

Pour la question 6, lors du 2ème tirage l'urne est composée de *N* boules dont 40 marquées. On tire 40 boules et 8 sont marquées : la proportion de boules marquées dans l'échantillon ($1/5$) est une bonne estimation de la proportion de boules marquées dans l'urne ($40/N$). On estime donc ponctuellement *N* par 200.

En utilisant les cours de probabilité de licence, lors d'un tirage *avec remise* de 40 boules, le nombre de boules marquées sur les 40 tirées suit une loi binomiale de paramètres 40 (nombre de boules tirées) et $40/N$ (proportion de boules marquées dans l'urne) et son espérance mathématique est égale à $1600/N$.

Lors d'un tirage *sans remise* ou d'un tirage *simultané* de 40 boules, le nombre de boules marquées sur les 40 tirées suit une loi hypergéométrique de paramètres 40 (nombre de boules tirées de l'urne), $40/N$ (proportion de boules marquées dans l'urne) et *N* (nombre de boules de l'urne) et son espérance mathématique est encore égale à $1600/N$.

On déduit de la loi de probabilité du *nombre* de boules marquées dans l'échantillon la loi de probabilité de la *proportion* de boules marquées dans l'échantillon (fréquence d'échantillonnage) ; quel que soit le mode de tirage, l'espérance mathématique est égale à la proportion de boules marquées dans l'urne d'où l'estimation de *N* en posant $1/5 = 40/N$.

Pour préciser la réponse, on peut estimer *N* par intervalle de confiance ; on est dans les conditions de l'approximation normale. La proportion de boules marquées parmi les 40 tirées suit approximativement une loi normale de moyenne $40/N$ et d'écart-type $\sqrt{0.2\times 0.8\times 0.8/40}$ (éventuellement $\sqrt{0.2\times 0.8/40}$ si l'on utilise l'approximation de la loi binomiale plutôt que de la loi hypergéométrique). Une réponse utilisant le mode de pensée statistique peut être : on estime *N* à 200 et *N* appartient à l'intervalle $[129\ ;\ 444]$ (ou $[123\ ;\ 526]$) à 95 % de confiance.

Pour la question 7, on peut estimer la moyenne $(N+1)/2$ de la loi uniforme sur $\{1,\ldots,N\}$ par la moyenne des quatre observations 763.5 ce qui permet d'estimer *N* à 1526. Pour préciser la réponse, on peut chercher à majorer *N* avec un risque d'erreur fixé à l'avance, par exemple 5 %. Si l'on note $(X_i)_{i=1,\ldots,4}$ quatre variables aléatoires indépendantes et de même loi uniforme sur $\{1,\ldots,N\}$ et *Y* leur maximum, alors 1200 est une observation de *Y* et la probabilité que *Y* soit inférieur ou égal à 1200 est égale à $\left(\frac{1200}{N}\right)^4$. Cette probabilité est supérieure ou égale à 0.05 si et seulement si *N* est inférieur ou égal à 2537. Une réponse utilisant le mode de pensée statistique peut être : *N* est supérieur à 1200 avec certitude, on l'estime à 1526 et *N* est inférieur ou égal à 2537 à 95 % de confiance.

Pour les deux questions, les estimations par intervalle de confiance peuvent être améliorées mais il s'agit d'un premier pas vers une réponse de nature statistique.

Résultats

Pour la question 6, un seul étudiant propose d'estimer *N* par 200, tous les autres écrivent que le nombre *N* de boules de l'urne est supérieur ou égal à 72. L'idée qu'un échantillon aléatoire simple à probabilités égales de taille suffisante soit un modèle réduit de la population dont il est extrait n'est pas du tout acquise à la fin de la licence.

Pour la question 7, tous écrivent que le nombre *N* de boules de l'urne est supérieur ou égal à 1200. Les étudiants restent sur la production d'un résultat sûr et obtenu à partir d'une déduction mathématique.

Ces deux dernières questions sont connues des professeurs de statistique et probabilités avec un « habillage concret », nombre de poissons dans un bassin pour le premier, on se réfèrera aux expérimentations de Lahanier-Reuter (1999), nombre de taxis dans une ville pour le second (Engel, 1987). Les « habillages » pouvaient expliquer l'échec des étudiants. Une rédaction enlevant toute ambiguïté sur les problèmes permet de mieux constater la difficulté des étudiants à utiliser leurs connaissances en statistique et probabilités récemment acquises.

3 Probabilités et statistique inférentielle, approche modèle versus approche sondage

3.1 Introduction des probabilités dans un cadre de modélisation

Les probabilités sont bien souvent introduites, à l'université comme dans le secondaire, dans un cadre de modélisation : on part de la notion d'expérience aléatoire, qui n'a pas de définition mathématique (le mot « aléatoire » est utilisé ici dans le sens commun et annonce la modélisation probabiliste qui va suivre) et on définit l'*espace probabilisé associé*, appelé dans certains cours le *modèle probabiliste associé*.

Pour le choix du « bon modèle » probabiliste (ce qui laisse entrevoir la nécessité de discuter de la qualité d'un modèle), deux approches sont proposées (à ne pas confondre avec l'approche sondage et l'approche modèle discutées dans ce texte) :

– l'*approche laplacienne,* dite approche classique, pour la modélisation des expériences aléatoires de référence menant à l'équiprobabilité : lancer d'une pièce de monnaie « équilibrée », lancer d'un dé cubique « non pipé » dont les faces sont marquées de 1 à 6, tirage « au hasard » d'une boule d'une urne contenant *N* boules numérotées de 1 à *N* ;

– l'*approche fréquentiste* pour la modélisation des expériences aléatoires répétées dans les mêmes conditions, exemple du lancer de punaise ; les probabilités sont définies à partir des fréquences observées.

Relevons quelques difficultés.

C'est en même temps que l'on introduit le *processus de modélisation*, passage du « réel » à la structure mathématique qui permet de le représenter, et la structure mathématique elle-même, ici l'*espace probabilisé*.

Autre difficulté déjà bien repérée : les deux approches de la probabilité, approche classique et approche fréquentiste, présentent des cercles vicieux. Dans la première, on définit la probabilité d'un événement à partir de l'*équiprobabilité* des événements élémentaires, dans

la seconde, on s'appuie sur la loi faible des grands nombres qui est un théorème de la théorie des probabilités : si la *probabilité* d'un événement lié à une expérience aléatoire est *p*, alors la fréquence de réalisation de l'événement sur *n* expériences modélisées par des lois identiques et indépendantes en *probabilité* converge en *probabilité* vers *p* lorsque *n* tend vers l'infini.

La troisième difficulté a déjà été évoquée précédemment : mettre en parallèle les expériences aléatoires de référence menant à l'équiprobabilité et le lancer d'une punaise entretient la confusion. Le lancer d'une pièce de monnaie « équilibrée » ou d'un dé « non pipé », le tirage « au hasard » d'une boule dans une urne, sont des périphrases pour dire « on génère une observation de la loi uniforme sur $\{P, F\}$, sur $\{1,\ldots,6\}$ ou sur $\{1,\ldots N\}$ ». On est du côté de la *simulation* de loi uniforme (ou équiprobabilité) et non de la *modélisation* d'expériences aléatoires. C'est cette confusion entre simulation et modélisation que nous avons retrouvée dans les réponses des étudiants et que nous avons déjà discutée sans prétendre clore la discussion. La confusion est encore présente dans le document d'accompagnement du programme de probabilité au collège. Il est proposé de faire jouer les élèves au jeu de Franc Carreau, jeu qui relève des probabilités dites « géométriques » introduites par Buffon (1707-1788) : « On peut proposer la situation du jeu de Franc Carreau, en cherchant à déterminer approximativement *la* probabilité de gagner … Cette situation présente l'avantage que l'on peut déterminer *cette* probabilité à l'aide de considérations géométriques sans que cette valeur soit connue au départ (comme c'est le cas avec le jeu de pile ou face avec une pièce « équilibrée » ou le jeu du lancer d'un dé cubique « non truqué ») ». Alors que la solution probabiliste repose sur une modélisation du jeu de Franc-Carreau, le jeu est présenté ici comme une simulation du modèle probabiliste. Pour le lien simulation modélisation des expériences de Buffon, on lira avec intérêt le chapitre 10 du livre de l'IREM de Grenoble coordonné par C. Schwartz (2006).

3.2 Introduction des probabilités à partir de simulations

La réforme du lycée en France entre 2000 et 2002 proposait d'étudier « simulation et fluctuation d'échantillonnage » en seconde avant d'introduire les probabilités en première. Ce choix entretenait la même confusion. La simulation du lancer d'un dé équilibré ou d'un dé non pipé était présentée en parallèle comme simulation d'expériences aléatoires, la stabilisation de la distribution de fréquences des issues sur un grand nombre de simulations permettant d'approcher la distribution de probabilités des issues.

L'année suivante, les professeurs de première découvrent dans les nouveaux programmes, après l'introduction des probabilités : « simuler une expérience consiste à simuler un modèle de cette expérience ».

Dans le document d'accompagnement des programmes de première, est précisé :

« *Modéliser* consiste à associer un modèle à des données expérimentales, alors que *simuler* consiste à produire des données à partir d'un *modèle prédéfini*. Pour simuler une expérience, on associe d'abord un modèle à l'expérience en cours, puis on simule la loi du modèle. »

Les professeurs qui pensaient que la simulation de l'expérience permettait de connaître le modèle ont été déstabilisés : « Comment simuler un modèle que l'on ne connaît pas ? » ou « Pourquoi simuler un modèle que l'on connaît ? »

Ils n'avaient pas conscience, lors de l'enseignement en seconde, qu'ils simulaient des lois équiprobables pour approcher la loi de probabilité de variables aléatoires définies sur un ensemble muni de l'équiprobabilité. On sait que, lors du lancer simultané de deux dés cubiques équilibrés indiscernables, la loi de probabilité de la somme des points marqués n'est pas l'équiprobabilité sur les 21 résultats « visibles » mais doit être déduite de l'équiprobabilité sur les 36 résultats que l'on aurait obtenus si on avait lancé un dé après l'autre. Dans plusieurs manuels scolaires, on lit que la bonne réponse serait donnée par simulation sur ordinateur alors que l'on ne sait simuler sur ordinateur que le lancer d'un dé après l'autre.

L'introduction des probabilités en troisième et la réécriture des programmes de seconde pour la rentrée 2009 devraient donner davantage de cohérence au statut de la simulation.

3.3 Introduction de la statistique inférentielle dans le cadre de la modélisation

La statistique inférentielle est bien souvent introduite en supposant que les valeurs de l'échantillon sont des observations de variables aléatoires indépendantes et identiquement distribuées (i.i.d.), même lorsqu'elles ont été obtenues par tirage aléatoire dans une population finie. Il s'agit bien du processus d'unification de la statistique décrit par Armatte (2004) et cité en introduction.

Voici un exemple d'énoncé selon cette approche modèle :

« Un jour donné, une machine produit N objets, chacun pouvant être bon ou mauvais. Les qualités des N objets produits sont supposées indépendantes et de même loi caractérisée par la probabilité p pour chacun des objets d'être bon. En fin de journée, on tire au hasard un échantillon de n objets dans l'ensemble des objets produits et on observe k bons et $n-k$ mauvais ; n étant petit devant N, on admettra que les n observations sont indépendantes. Estimer par intervalle de confiance à 95 % la probabilité p. »

D'un point de vue sémantique, nous avons en fin de journée des objets bons et des objets mauvais ; comment admettre qu'un objet a une probabilité p d'être bon ? S'il s'agit de tirer un échantillon aléatoire dans une population de femmes et d'hommes adultes, comment admettre qu'une personne choisie au hasard a une probabilité p d'être une femme ? C'est ce type de modélisation qui laisse peut-être penser que les individus sont des membres aléatoires d'une population (cf. citation de Kyburg, 1974, dans Batanero *et al.*, 2004).

Dans l'effort d'interpréter l'intervalle de confiance à 95 % d'une proportion, on lit dans certains articles ou manuels : « l'intervalle de confiance a une probabilité de 95 % de contenir la proportion ». Alors que l'intervalle de confiance contient la proportion ou ne la contient pas. La « procédure de construction de l'intervalle de confiance à 95 % » conduit à ce que 95 % des échantillons contiennent la proportion et 5 % ne la contiennent pas (dans le cas d'un ensemble fini d'échantillons équiprobables).

Il en est de même pour les objets bons ou mauvais, c'est la « procédure de tirage de l'échantillon aléatoire » qui conduit à un échantillon d'objets bons et d'objets mauvais selon une certaine proportion. L'approche sondage permet de réécrire l'énoncé :

« Un jour donné, une machine produit N objets, chacun pouvant être bon ou mauvais. En fin de journée, on tire à probabilités égales et sans remise un échantillon de n objets (n négligeable devant N) dans l'ensemble des objets produits et on observe k objets bons et $n-k$

mauvais. Estimer par intervalle de confiance à 95 % la proportion *p* d'objets bons dans la production de la journée. »

Dans ce dernier énoncé, il n'y a aucune hypothèse de nature probabiliste ; en revanche, la procédure aléatoire de tirage de l'échantillon est bien précisée.

3.4 Les échantillons de la théorie des sondages

Dans une population finie de taille *N*, l'échantillonnage aléatoire simple à probabilités égales *avec remise* (resp. *sans remise*, resp. *simultané*) de taille *n* n'est autre que l'équiprobabilité sur les N^n (resp. A_N^n, resp. $\binom{N}{n}$) échantillons possibles.

Si la sous-population d'intérêt *A* est de cardinal *K*, donc en proportion $p = K/N$, la proportion ou fréquence de *A* dans l'échantillon s'écrit $F = X/n$ avec *X* binomiale de paramètres *n* et *p* (resp. hypergéométrique de paramètres *N*, *n* et *p*). La loi hypergéométrique peut être approchée par une loi binomiale de paramètres *n* et *p* dès que le taux de sondage n/N est négligeable, dans la pratique inférieur à 1/10). Les deux lois binomiale et hypergéométrique peuvent être approchées par une loi normale dès que *n* est suffisamment grand (dans la pratique supérieur à 30).

Dans la pratique des sondages, lorsque la liste des individus de la population, c'est-à-dire la *base de sondage*, est saisie sur fichier informatisé, c'est à partir de générateurs de nombres pseudo-aléatoires que les échantillons sont définis. Il faut veiller à ce que les algorithmes utilisés assurent bien l'équiprobabilité sur l'ensemble des échantillons.

3.5 Deux liens entre fréquence et probabilité, entre fréquence conditionnelle et probabilité conditionnelle

Dans le cas où l'échantillon est de taille 1, il s'agit de choisir un individu selon la loi équiprobable sur la population (« expérience de base de la théorie des sondages »). Si la sous-population d'intérêt *A* est de cardinal *K*, donc en proportion $p = K/N$, la probabilité d'observer un individu de la catégorie *A* est *p* (nombre de cas favorables sur nombre de cas possibles, car tous les cas possibles ont la même probabilité). Si l'on définit un caractère (qualitatif ou quantitatif) $\mathcal{X}$ sur cette population et que l'on note *X* la valeur observée sur notre échantillon de taille 1, alors *X* est une variable aléatoire dont *la distribution de probabilité n'est autre que la distribution de fréquence du caractère sur la population*. On ne fait là aucune hypothèse sur la loi de probabilité de *X*. C'est le *premier lien entre fréquence et probabilité* et il s'agit d'un *lien d'égalité.*

Si l'on définit sur la population deux caractères qualitatifs $\mathcal{X}$ et $\mathcal{Y}$ (supposons les, pour simplifier, à deux modalités chacun et notons $\{A, \overline{A}\}$ et $\{B, \overline{B}\}$ les partitions engendrées par $\mathcal{X}$ et $\mathcal{Y}$ respectivement), les distributions de fréquence conjointe et marginales se transposent immédiatement dans le cadre probabiliste en distributions de probabilité conjointe et marginales grâce à l'expérience de base de la théorie des sondages ; de même les distributions de fréquence conditionnelles se transposent en distribution de probabilité conditionnelles et le théorème de Bayes tant redouté s'impose sans difficulté dans ce contexte. C'est le *premier lien entre fréquence conditionnelle et probabilité conditionnelle.*

Le *deuxième lien entre fréquence et probabilité* est l'*approche fréquentiste* de la probabilité. Elle consisterait ici à répéter *n* fois « l'expérience de base » dans les mêmes conditions (tirage à probabilités égales et avec remise d'un échantillon de taille *n*) ; si n_A est le nombre de fois où l'on observe un individu de *A*, alors la fréquence $f_A = n_A/n$ est l'observation d'une variable aléatoire $F = X/n$ avec *X* binomiale de taille *n* et de paramètre *p* ; l'espérance mathématique de *F* étant égale à *p*, f_A est une *estimation* de la probabilité *p* (c'est-à-dire de la proportion K/N). Il s'agit du *deuxième lien entre fréquence et probabilité* et il s'agit d'une *approximation* (estimation). La connaissance de la loi de probabilité de *F* permettra d'estimer *p* par intervalle de confiance. De même si $n_{A\cap B}$ est le nombre de fois où l'on observe un individu appartenant simultanément à *A* et à *B* dans l'échantillon de taille *n*, et $f_{A\cap B} = n_{A\cap B}/n$ la fréquence correspondante, le rapport $n_{A\cap B}/n_A$ égal à $f_{A\cap B}/f_A$ est une estimation de la probabilité conditionnelle $P_A(B)$. C'est le *deuxième lien entre fréquence conditionnelle et probabilité conditionnelle.*

4 Conclusion

La démarche en Statistique commence par le recueil des données en fonction du problème posé ; à ce titre la « *théorie des sondages* » et les « *plans d'expériences* » sont deux parties de la statistique essentielles pour le recueil de données et pour une bonne compréhension des théories et techniques de l'échantillonnage, mais malheureusement très peu enseignées dans les universités françaises.

Il ne s'agit pas de proposer de remplacer l'approche modèle par l'approche sondage mais de proposer différentes approches en fonction des données. Les deux approches sont complémentaires.

L'approche sondage est très concrète à condition de définir correctement les notions de statistique descriptive sur une population finie (effectifs, fréquences, distributions d'effectifs, distributions de fréquences et, dans le cas de deux caractères, distributions conjointe et marginales d'effectifs, distributions conjointe et marginales de fréquences et distributions de fréquences conditionnelles). Chacun des deux caractères peut être qualitatif ou quantitatif discret, ou encore quantitatif continu dont les valeurs sont regroupées en classes. On ne retient du caractère que la partition de la population qu'il engendre.

Le tirage avec équiprobabilité d'un individu dans la population permet de passer du registre de la statistique descriptive au registre des probabilités et de résoudre dans ce cadre bon nombre de problèmes de probabilités relevant du domaine des sciences humaines et sociales.

Malheureusement, l'étude des « tableaux à double entrée » (tableaux d'effectifs ou de fréquences de deux caractères qualitatifs) n'est abordée qu'en première, pour pratiquement toutes les sections sauf la section S. Les mots « tableaux à double entrée » figurent bien dans le programme de sixième mais les notions d'effectifs et de fréquences n'apparaissent qu'en cinquième.

L'étude des tableaux d'effectifs ou de fréquences de deux caractères permet d'introduire les diagrammes de Venn et le vocabulaire des ensembles, de faire attention à l'ambiguïté de la langue française en distinguant la proportion par rapport à *A* de *B*, la proportion par rapport à

B de *A*, la proportion par rapport à la population de *A* et *B*, d'introduire les arbres de fréquences et fréquences conditionnelles, …

Pour éviter les cercles vicieux des approches classique et fréquentiste de la probabilité, on peut définir une distribution de probabilité finie comme un ensemble fini, pondéré de poids positifs de somme 1. On définit la probabilité d'un sous-ensemble et la règle d'additivité. On a équiprobabilité lorsque les poids sont égaux.

Tirer « au hasard » dans une urne composée de trois boules numérotées 1, 2 et 3 revient à simuler l'équiprobabilité sur l'ensemble {1,2,3}, c'est-à-dire {(1,1/3),(2,1/3),(3,1/3)}. Si la boule 1 est rouge (notation R) et les deux autres blanches (notation B) et qu'on ne s'intéresse qu'à la couleur de la boule tirée, on simule la loi de probabilité {(R,1/3),(B,2/3)}.

Avant de simuler des échantillons de taille *n*, les ordinateurs permettent, pour des petites valeurs de *n* et de *N*, de donner les *distributions de probabilité exactes* des fréquences ou moyennes d'échantillonnage (plutôt que des approximations par simulation).

C'est ensuite par un grand nombre de simulations que l'on peut vérifier la loi des grands nombres : la distribution de fréquence observée sur un grand nombre de simulations est proche de la composition de l'urne, et d'autant plus proche que le nombre de simulations est plus grand.

On peut alors se servir de ce résultat pour approcher par simulation des lois de probabilités difficiles à obtenir par le calcul formel, enfin, aborder la modélisation d'échantillons de données réelles.

Références

[1] Armatte, M. (2004), *La théorie des erreurs (1750-1820) : enjeux, problématiques, résultats*. In Barbin, E. et J.-P. Lamarche (Ed.), *Histoires de Probabilités et de Statistiques*, IREM – Histoire des Mathématiques, Ellipses, Paris, 141-160.

[2] Batanero, C. (2001), Didáctica de la Estadística, *http://www.ugr.es/~batanero/*, Publicaciones, Formación de profesores.

[3] Batanero, C., J.D. Godino, and R. Roa (2004), Training teachers to teach probability, *Journal of Statistics Education*, **12**. Retrieved August 31, 2006 from *http://www.amstat.org/publications/jse/* .

[4] Chaput, B., J.-C. Girard et M. Henry (2008), Modeling and simulations in statistics education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference. Teaching Statistics in School Mathematics – Challenges for Teaching and Teacher Education, Mexico 2008, *http://www.stat.auckland.ac.nz/~iase/publications.php?show=rt08* .

[5] Courtebras, B. (2006), *À l'école des probabilités. Didactiques-Mathématiques*, Presses Universitaires de Franche-Comté.

[6] Desrosières, A. (2002), Adolphe Quételet, *INSEE, Courrier des Statistiques*, **104** (*http://www.insee.fr/fr/ffc/docs_ffc/cs104a.pdf*).

[7] Engel, A. (1990), *Les certitudes du hasard*, ALEAS Ed., Lyon.

[8] Lahanier-Reuter, D. (1999*), Conceptions du hasard et enseignement des probabilités et statistiques*, Education et Formation, Recherches Scientifiques, Presses Universitaires de France.

[9] Académie des Sciences (sous la direction de P. Malliavin) (2000), *La statistique. Rapport sur la science et la technologie n°8*, Éditions Tec et Doc, Paris.

[10] Meusnier, N. (2006), Sur l'histoire de l'enseignement des probabilités et des statistiques, *Journal Électronique d'Histoire des Probabilités et de la Statistique, http://www.jehps.net/Decembre2006/Meusnier.pdf* .

[11] Site du ministère de l'éducation de France, programmes de mathématiques des collèges et lycées et documents d'accompagnement : *http://eduscol.education.fr/D0015/* .

[12] Site du ministère de l'éducation du Chili, programmes de l'enseignement secondaire : *http://www.curriculum-mineduc.cl/curriculum/programas-de-estudios/* .

[13] Schwartz, C. (2003), *Contes et décomptes de la statistique. Une initiation par l'exemple*, Vuibert, Paris.

[14] Schwartz, C. (Coord.) (2006), *Pratiques de la statistique. Expérimenter, modéliser et simuler*, Vuibert, Paris.

[15] Schwartz, C. et E. Roser (2009), L'esprit des probabilités, de l'école au lycée. Dossier du numéro 13 de la revue MathemaTICE : Les probabilités / statistiques et les TICE, *http://revue.sesamath.net/spip.php?article193* .

[16] Statistix, Centre de ressources, lieu de partage et de mutualisation pour l'enseignement de la statistique pour les enseignants des écoles, des collèges et des lycées de toutes les disciplines, *http://www.statistix.fr/* .