L'esprit des probabilités, de l'école au lycée

Érick Roser & Claudine Schwartz

Le texte ci-dessous propose des pistes pour nourrir la réflexion sur des progressions possibles, de l'école au lycée, d'un enseignement moderne de la théorie des probabilités. Il est écrit sans contrainte d'adéquation aux programmes actuellement en vigueur, avec lesquels il ne présente cependant aucune rupture. Les idées avancées n'engagent les auteurs qu'à titre personnel.

L'enseignement des probabilités et de la statistique est une composante essentielle de la formation scientifique du citoyen afin qu'il puisse :
– participer au débat démocratique (être en mesure d'utiliser à bon escient des résultats de sondages, comprendre la notion de prévision en domaine incertain et différentes notions de risques, analyser la définition et l'usage des grands indicateurs sociaux),
– comprendre la notion de preuve statistique utilisée dans différents domaines. On peut penser par exemple à la « médecine des preuves expérimentales » (*evidence*(1) *based medicine*) dont l'exigence est que toute assertion sur l'efficacité d'un traitement et toute mise en œuvre d'un protocole thérapeutique soient étayées par des preuves statistiques,
– analyser quelques éléments quantitatifs clefs des jeux de hasard les plus classiques afin d'éclairer leur pratique éventuelle.

Nous partons du principe que cette formation commence dès l'école et continue tout au long de la scolarité. Nous distinguons trois temps, celui de l'école, celui du collège et celui du lycée :

L'école élémentaire : le temps des dés.
Le collège : le temps de la simulation.
Le lycée : le temps de la théorie.

Il s'agit ici d'une vision prospective à long terme, l'école élémentaire n'étant pas aujourd'hui concernée. Cette vision permettra de s'orienter pour construire peu à peu une cohérence verticale de la formation, en évitant les répétitions sans avancées et l'apparition d'une pratique des probabilités propres à la scolarité, qui dérive loin des usages professionnels ou citoyens.

Notre propos est en grande partie fondé sur la recherche des bases à acquérir pour comprendre la statistique ; la combinatoire qui a longtemps été au centre de l'enseignement des probabilités au lycée est marginalisée dans cette approche(2). Pour ce qui est du lycée, nous proposons des sujets à répartir entre différentes disciplines :

(1) Le terme « evidence » est un faux ami et signifie « preuve ».
(2) On ne peut pas tout faire : il s'agit ici d'un choix et non d'un jugement de valeur.

si le cœur d'un enseignement des probabilités est en mathématiques, cette discipline diffuse très naturellement vers la physique et les sciences de la vie et de la terre.

Nous insistons ici plus sur les premiers pas vers l'aléatoire, c'est-à-dire sur l'école et le collège : c'est à ce niveau que le terrain est le moins défriché. Ces premiers pas sont guidés à la fois par une vision de la discipline et aussi par les objectifs à atteindre en fin de scolarité et qui sont ici simplement énumérés.

1. L'école élémentaire : le temps des dés

1.1. Où en est-on ?

On a posé à des élèves de CM1 et de CM2 de Grenoble(3), en 2006, la question suivante :

Que signifie pour vous le hasard ?

Les réponses (voir encadré ci-dessous) peuvent être classées en plusieurs catégories :
– le hasard extérieur à soi : il est caractérisé par son imprévisibilité ;
– le hasard d'une action que l'on produit est lié au « n'importe quoi, n'importe comment » ;
– le hasard a partie liée avec la chance ou la malchance, c'est-à-dire avec des évènements heureux ou malheureux dont on ne maîtrise pas le cours, des évènements « sans cause » ;
– les coïncidences.

Ces réponses reflètent une connaissance a priori qui, à ce niveau, nous semble « juste »(4).

D'autres réponses sont peu classables (après discussion avec l'enseignant, nous ne pouvons pas savoir si l'élève qui a dit que le hasard « c'est quand c'est pas rangé » a une intuition extraordinaire des liens entre hasard et désordre, ou s'il a confondu « c'est le hasard » avec « c'est le bazar »).

Réponse d'élèves de CM1-CM2 à la question : que signifie pour vous le hasard ?

- Quand on ne sait pas ce qui va se passer.
- C'est une chose où on ne peut pas savoir la réponse, quelque chose que tu ne peux pas prédire/tirer au sort.
- Une désignation au pif, c'est répondre sans réfléchir ; exemple : quelqu'un te dit « combien un chat a-t-il de pattes ? » et tu lui réponds « 3 ».
- C'est quand tu réponds et tu ne sais pas la réponse.

(3) Il s'agit d'une enquête « micro-trottoir ».
(4) Il convient de respecter ces premières intuitions et, en particulier, de ne pas se lancer, sous prétexte de tester les connaissances a priori des élèves, dans des questions pièges du type :
Il y a 4 boules dans une urne, dont 3 blanches et une noire. Tu en tires une au hasard, quelle couleur paries-tu ?
Il faudra en effet longtemps (et beaucoup de tirages réels ou simulés) pour comprendre la nature du concept de prévision ou de pari (assorti d'un calcul de risque d'erreur ou d'espérance de gain), qu'on peut mettre à l'œuvre dans ce domaine : poser la question trop tôt est de nature à perturber l'intuition juste que le hasard est lié à l'imprévisibilité !

- Veut dire sans le faire exprès.
- Le hasard est une chose où on peut avoir de la chance et de la malchance.
- Une coïncidence : par exemple, on trouve un billet de 50 euros par terre.
- C'est quand c'est pas rangé.
- Le hasard n'existe pas.
- Le hasard c'est un peu de tout.
- Des évènements mystérieux de notre vie.

Rappelons que les élèves des écoles n'ont pas d'intuition de la loi des grands nombres, laquelle nécessite une bonne manipulation des proportions.

1.2. Les dés à l'école

• On considère ici des dés à 4 faces, 6 faces, 8 faces, etc.

Ce sont des objets riches à étudier au plan de la géométrie, dès le début du primaire :
Emploi des termes carré, cube, sommets, faces, arêtes, angles droits, etc.
Égalité de longueurs et de surface.
Triangles, carrés, pentagones, hexagones, polygones, polygones réguliers.
Les cinq solides de Platon (??).

• On peut découvrir une formule :

Pour un dé cubique, la somme des valeurs de faces opposées vaut 7.

On peut vérifier la formule d'Euler(5), valable pour tous les dés polyédriques :

nombre de sommets + nombre de faces = nombre d'arêtes + 2.

• Au plan du hasard, le dé permet une première prise de conscience de l'aléatoire, et la construction d'un socle sur lequel les expériences ultérieures et les savoirs spécifiques pourront venir se cristalliser.

On pourra commencer par travailler sur des termes « de la vie de tous les jours » :

– ***chances égales ou inégales*** dans des expériences simples liées aux dés (dés à 2 faces blanches et 4 faces rouges par exemple) : il ne s'agit pas de proposer des définitions, mais de travailler sur des situations variées afin de distinguer le fait d'avoir deux issues (blanc et rouge par exemple) et les « chances égales » ***vues sous l'angle du long terme expérimental*** (si on lance beaucoup de fois ce dé, on aura plus souvent rouge).

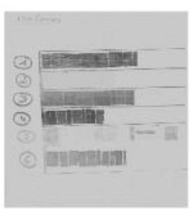
(5) Comprendre ce que signifie une formule est un enjeu important de l'enseignement des maths !

Signalons à ce propos que pour bon nombre d'élèves et d'adultes, l'emploi du terme chance est tantôt :
– synonyme de probabilité,
– lié à des notions d'effectif. Ainsi, parler d'un évènement qui a *1 chance sur 3* de se réaliser dit qu'on est face à une expérience à trois issues possibles, dont une est cet événement, sans pour autant préjuger de l'égale probabilité de chaque issue. De plus, avec ce sens là, avoir *2 chances sur 3* de tirer une boule blanche dans une urne à 3 boules « indiscernables au toucher » dont 2 sont blanches, n'est pas identique à parler de *4 chances sur 6* de tirer une boule blanche dans une urne à 6 boules « indiscernables au toucher » dont 4 sont blanches. Les probabilités sous-jacentes sont égales, mais les effectifs de boules ne sont pas les mêmes.

On peut, dans le cadre de l'enseignement, signaler ces deux sens et choisir, dans les murs de l'école, de n'utiliser que le premier. On pourra convenir qu'en classe, dire « il y a une chance sur 2 que je gagne au loto » est, par consensus au niveau du langage, considéré comme « faux » car la locution *une chance sur n* est réservée aux situations où toutes les possibilités ont des chances égales d'être tirées.

– ***indépendance et mémoire*** :
Les enfants découvrent seuls que si l'un d'eux lance un dé (avec gobelet pour le remuer et une piste d'atterrissage) et obtient 6, que son voisin prend le dé, le relance, obtient aussi 6, puis passe le dé à son voisin, celui-ci a les mêmes chances d'avoir 1, …6 car le dé « ne sait pas » ce qui s'est passé avant : il est sans mémoire.

Cette question d'absence de mémoire justifie de travailler en primaire sur les dés : Les élèves ne connaissent pas du tout la loi des grands nombres, ils ne sont pas face à ce grand paradoxe intuitif (nous y reviendrons) d'un processus sans mémoire, ici le lancer d'un dé cubique à chiffres, qui à long terme équilibre les fréquences de chaque face. Pour les élèves, les chances égales peuvent être traduites par le fait que dans des longues suites de lancers de dés (représentées par des histogrammes horizontaux ou verticaux comme ci-dessous), n'importe quelle face peut gagner (apparaître plus souvent que les autres) pendant un moment, puis c'est une autre (sur le dessin ci-dessous, le 5 gagne, provisoirement).



Il est important de réfléchir à l'indépendance des résultats vue sous l'angle d'absence de mémoire car :
– Les dés seront peu à peu remplacés par des « dés électroniques » : dire que le dé électronique n'a pas de mémoire relève à ce niveau de l'acte d'autorité.
– Les élèves acquièrent ultérieurement l'expérience de la convergence des fréquences des faces du dé vers 1/6, soit à travers des jeux soit par simulation en classe. S'ils ne

se sont pas déjà persuadés de l'absence de mémoire du processus de lancer du dé, ils font l'impasse sur cette propriété, impasse qui s'érige en obstacle à l'élaboration d'un mode de pensée probabiliste.

• *Le hasard c'est le royaume de l'imprévisibilité, mais cela ne dispense pas de réfléchir.*

On pourra consulter à ce propos sur le site *Statistix* un exemple d'activité menée à l'école dans le cadre d'un mémoire d'IUFM : (http://www.statistix.fr/spip.php?article35).

• *Le hasard c'est le royaume de l'imprévisibilité, mais cela ne dispense pas d'écrire.*

L'expression écrite est ici très importante, les mots (chance, hasard, hasard pur, hasard équitable) peuvent être discutés. Cet aspect des choses concerne aussi les professeurs qui n'ont pas toujours les mots justes pour argumenter.

Un exemple : L'apparition du 6 semble à beaucoup d'élèves et de professeurs plus rare que celle des autres chiffres ; cette rareté du 6 relève d'un ressenti en contradiction avec les *chances égales*. Dire qu'il s'agit d'une impression fausse peut apparaître comme un argument d'autorité ; par contre, en discutant sur des situations où il semble en être ainsi, on arrive au fait que le « 6 gagnant » de nombreux jeux est plus rare que l'ensemble des 5 autres chiffres, qui eux ne sont pas gagnants. (Sans le dire, on passe des probabilités des éléments à celle des évènements et cet exemple pourra être repris lors de l'introduction du calcul des probabilités).

2. Le collège : le temps de la simulation(6)

2.1. Un dé équilibré, c'est quoi ?

On devra à un moment donné s'interroger sur ce dont on parle quand on dit « un dé équilibré à N faces », que ce dé soit un objet réel ou électronique : c'est un procédé de simulation de la distribution de probabilité équirépartie sur 1, …, N. Cela impose en particulier le modèle à considérer pour les données, à savoir que les probabilités des N^n résultats possibles pour n lancers sont égales.

De même, les pièces de monnaie ne sont pas conçues pour la simulation, mais il se trouve qu'on peut les utiliser pour simuler la loi équirépartie : si on parle d'une pièce équilibrée, c'est qu'on la considère comme générateur possible de listes de *pile*, *face* avec équirépartition de ces deux issues.

Quand on dit : on choisit avec remise n boules au hasard dans une urne avec N boules, on ne parle pas là non plus de comment cela peut être réalisé, mais du fait qu'on simule une loi équirépartie sur 1, …, N.

Il conviendrait, dans le cadre de l'enseignement des probabilités, de garder le terme hasard (ou « hasard pur ») pour parler de modèle équiréparti, et de se souvenir que

(6) Rappelons que tant que l'aléatoire n'est pas inscrit comme champ scientifique à travailler à l'école, et tant qu'on n'a pas la preuve que cela a été effectivement abordé par les élèves, le collège devra prendre en charge, en les adaptant, des travaux sur les lancers de dés matériels : on ne peut pas court-circuiter cette étape et passer directement à la simulation.

l'esprit humain est un piètre générateur de nombres au hasard (voir http://www.statistix.fr/spip.php?article50).

Dans l'expérience « franc-carreau » (voir http://www.statistix.fr/spip.php?article40) on dit souvent « on lance une pièce au hasard sur un plan quadrillé » : or en pratique on ne sait pas lancer un pièce de telle sorte qu'un modèle raisonnable du point où tombe le centre de la pièce soit la loi uniforme sur le carré ; il conviendrait de dire qu'on lance la pièce sans viser. Si on veut *lancer une pièce au hasard* sur un quadrillage, alors il faut utiliser un simulateur du hasard, c'est-à-dire faire une **expérience virtuelle**.

Enfin, **parler d'expériences aléatoires identiques, c'est par convention parler d'expériences relevant du même modèle probabiliste** ; les lancers de dés équilibrés sont identiques car relevant de la même distribution équirépartie sur 1, …, 6. Cette précision étant donnée, on ne s'étonnera pas d'avoir des expériences identiques donnant des résultats différents.

2.2. Lancer deux dés à N faces : un incontournable

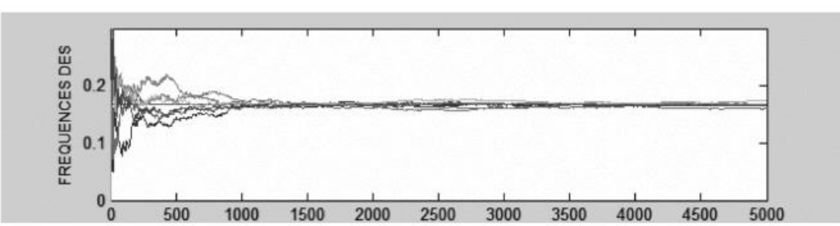
C'est l'occasion pour l'élève :
– de **réaliser lui-même** un programme de simulation avec une calculatrice de poche ou un tableur ;
– d'**estimer** une distribution de probabilités et de confronter cette estimation avec **un calcul** de cette distribution, à partir d'un tableau à double entrée.

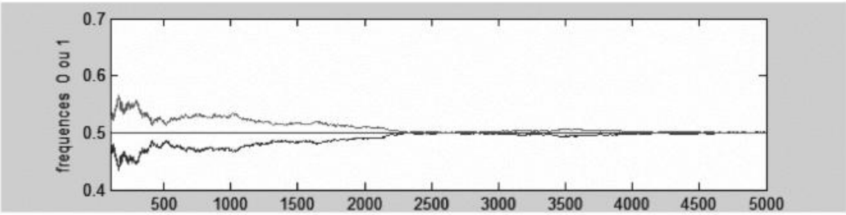
Diverses situations peuvent être construites autour de ce thème (voir par exemple http://www.statistix.fr/spip.php?article2).

2.3. Les dés électroniques et la loi empirique des grands nombres

Nous mettons sous le vocable « **dé électronique** » les programmes permettant de **simuler une loi équirépartie sur un ensemble fini d'entiers**. Parler de dés électroniques à n faces, dans la continuité d'expériences avec des dés « réels » ne pose pas vraiment de problèmes aux élèves dans la mesure où ils ont compris que la raison d'être des dés est de permettre la production de listes de nombres choisis « avec des chances égales » et indépendamment les uns des autres. Qu'un ordinateur ou une calculatrice de poche produise rapidement de telles longues listes n'est pas de nature à surprendre cette génération de « numerical natives ».

On peut alors réaliser des représentations graphiques telles que la suivante, relative aux fréquences de chaque face d'un dé après n lancers, n = 1, …, 5 000, ou à une expérience à deux issues :





On voit ainsi la distribution des fréquences « converger vers la distribution des *fréquences théoriques* ou *distribution de probabilités* » : il est intéressant à ce niveau de quitter le vocabulaire « *une chance sur 6* » pour employer une terminologie plus spécifique (ce qui n’interdit pas pour autant de revenir au langage des chances dans des cas où cela n’introduit pas d’ambiguïtés).

Le terme **distribution de probabilités** semble plus compréhensible, au niveau scolaire, que le terme de loi de probabilité.

À ce niveau, on observe une convergence des fréquences : il s’agit d’une *loi expérimentale* (ou empirique(7)) des grands nombres.

2.4. Protocole et outils de description

Si on veut que les simulations s’inscrivent dans une progression de la pensée, il faut en garder des traces : celles de la question posée, de la simulation mise en œuvre, de la description des résultats. À ce propos, la pratique de la simulation est très adaptée pour apprendre à élaborer et interpréter des représentations graphiques, à travailler sur la moyenne, la médiane, l’intervalle interquartile ; on peut imaginer que l’entraînement à la description de données non simulées relève essentiellement des autres disciplines (SES, biologie, physique, géographie).

Pour pratiquer la simulation, il faut établir un protocole expérimental dans lequel on fixe la taille de la simulation. Par exemple, pour le lancer de deux dés, on peut envisager que des binômes d’élèves simulent chacun sur tableur 100 (ou 1 000) valeurs de la somme des faces : chaque binôme peut fournir à la classe ses résultats (distribution des fréquences des nombres de 2 à 12), ce qui rend apparente la fluctuation d’échantillonnage. Les données sont ensuite mutualisées, ce qui affine l’estimation qu’on peut proposer de la loi de probabilité. Dans cette situation, la loi de probabilité étant calculable, on pourra le faire dans un second temps et cela permettra d’observer la qualité des estimations faites à partir des simulations.

2.4. Simulations d’expériences

De nombreuses activités mathématiques en collège peuvent utiliser des simulations de diverses situations ; on en trouvera quelques exemples sur le site

(7) Le terme empirique en statistique traduit un lien direct avec des expériences. Ainsi, la moyenne empirique est une moyenne calculée sur des données expérimentales et non une moyenne évaluée « pifométriquement » ; le terme empirique se pose en contraste avec le terme théorique : la moyenne théorique ou espérance relevant, elle, du choix d’un modèle. On rajoute de même souvent l’adjectif *empirique* devant le nom *fréquence* pour rappeler que celle-ci est calculée sur des données d’expériences réelles ou virtuelles.

statistix, à la rubrique collège (par exemple http://www.statistix.fr/IMG/doc/Un_jeu_de_l_oie.doc et http://www.statistix.fr/IMG/pdf/rumeur.pdf).

3. Le lycée : le temps de la théorie

3.1. Les lois des grands nombres

Les graphiques tels ceux du paragraphe 2.3 permettent de faire connaissance avec une loi expérimentale de *convergence* des fréquences vers les probabilités.

Mais comment « comprendre » qu'un processus sans mémoire (lancers successifs d'un dé ou d'une pièce) « converge », donc qu'à long terme, il y ait compensation relative (i.e. au niveau des fréquences) ? Ce phénomène est pour l'intuition paradoxal.

L'explication de cette loi empirique est un théorème qui porte aussi le nom de loi des grands nombres : dans le cadre strict de la théorie des probabilités, ce théorème montre que les fréquences *convergent presque sûrement* vers les probabilités correspondantes. Il convient de dire aux élèves qu'un tel théorème existe ; on peut se contenter de dire que les fréquences *convergent* vers les probabilités, la notion de convergence presque sûre n'étant ici pas mentionnée(8).

Il est important dans la formation des élèves à ce niveau de faire l'expérience de ce tempo d'une démarche scientifique : ***l'explication d'une loi empirique réside non dans des arguments interprétatifs, mais dans la démonstration d'un théorème.***

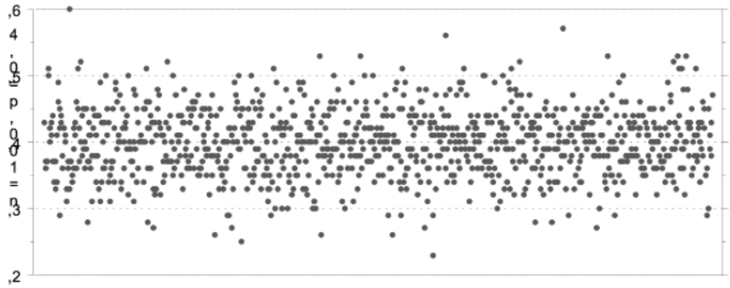
On notera que l'expérience et le théorème cité concernent l'évolution temporelle d'un processus : on fait une expérience, puis deux, … puis n, … et on regarde comment évolue la distribution des fréquences.

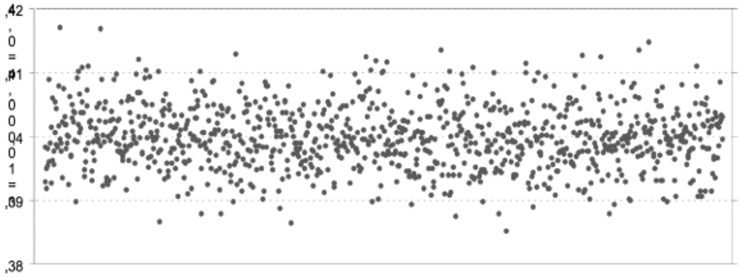
Dans le même mouvement que précédemment, c'est-à-dire en passant d'une loi empirique à un théorème mathématique, on peut compléter ce qui a été dit par une vision qui mélange le spatial (expériences en parallèle) et le temporel : on mène en parallèle N séries de n simulations et on observe la fluctuation entre les N distributions de fréquences.

On se limite ici à une expérience à deux issues, codées 0 et 1, la probabilité sur $\{0, 1\}$ étant alors complètement déterminée par la probabilité de 1, que nous noterons p. Les graphiques ci-dessous sont faits avec $p = 0{,}4$, $\mathrm{N} = 1\,000$ et $n = 100$ et $10\,000$. On observe (c'est l'occasion de bien lire les échelles des représentations graphiques !) que « la plupart » des 1 000 données sont dans [0,3 ; 0,5] pour $n = 100$, et dans [0,39 ; 0,41] pour $n = 10\,000$. D'autres simulations permettent de dégager une loi empirique sous une forme vague : « pour des valeurs de p pas trop proches de 1 ou 0 et pour n assez grand, environ 95% des fréquences calculées sur des échantillons de taille n sont dans l'intervalle $\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]$.

(6) L'ensemble des « trajectoires », c'est-à-dire les « listes infinies » de résultats qui ne réalisent pas cette convergence est de probabilité nulle, mais cette notion est trop difficile pour l'enseignement secondaire. Pour autant, il est souhaitable de ne pas insister sur un aspect propre aux probabilités sur un ensemble fini, à savoir qu'une issue de probabilité nulle est « impossible ».

Cela permet de rendre plus concret un mode de convergence des fréquences vers les probabilités : pour tout intervalle centré en p et de longueur $2a$, pour n suffisamment grand (ici $n > \frac{1}{a^2}$), environ 95% des données sont dans cet intervalle).





On pourra alors énoncer une application du théorème central limite qui explique cette loi empirique, à savoir :

Proposition : La probabilité que la fréquence calculée sur un échantillon de taille n soit dans $\left[p - t_\alpha \frac{\sqrt{p(p-1)}}{\sqrt{n}}, p + t_\alpha \frac{\sqrt{p(p-1)}}{\sqrt{n}}\right]$, où t_α désigne le quantile d'ordre 0,975 de la loi normale centrée réduite ($t_\alpha \approx 1{,}96$), tend vers 0,95 quand n tend vers l'infini.

L'utilisation pratique peut être obtenue en majorant t_α par 2 et $\sqrt{p(p-1)}$ par 1/2, en « cadrant p » et en déterminant alors numériquement des valeurs de n pour lesquelles la limite est atteinte à une précision de 0,01.

On peut retenir que : Dans les usages pratiques, pour $p \in [0{,}2 ; 0{,}8]$ et $n > 25$ et pour des échantillons de taille n, la probabilité que la fréquence soit dans $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]$ est approximativement 0,95, c'est-à-dire :

$$\text{Prob}\left(p - \frac{1}{\sqrt{n}} < \mathrm{F_n} < p + \frac{1}{\sqrt{n}}\right) \approx 0{,}95.$$

Pour n fixé, l'intervalle $\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]$ est appelé **intervalle de dispersion, ou de fluctuation de F_n, au niveau 0,95**.

Si F_n est dans un intervalle centré en p de longueur $2a$, alors p est dans un intervalle centré en F_n et de même longueur $2a$. Donc, si on ne connaît pas p mais qu'on cherche à l'estimer par F_n, on pourra dire que l'intervalle aléatoire $\left[F_n-\frac{1}{\sqrt{n}}, F_n+\frac{1}{\sqrt{n}}\right]$ a une probabilité 0,95 de contenir p. Cet intervalle s'appelle **intervalle de confiance** de p (ou **fourchette** dans le cas d'un sondage) au niveau 0,95. On peut aussi dire qu'en estimant p par F_n la précision est $\frac{1}{\sqrt{n}}$ avec une probabilité 0,95.

On peut voir cet intervalle de confiance sous deux angles :

– pour une fréquence observée f_n, toutes les lois sur $\{0 ; 1\}$ telles que p appartient à l'intervalle $\left[f_n-\frac{1}{\sqrt{n}}, f_n+\frac{1}{\sqrt{n}}\right]$ sont « compatibles » avec les données, au sens où l'intervalle de fluctuation associé à de tels p au niveau 0,95 contient f_n.

– si on simule des séries de n données à partir d'une valeur p, alors pour environ 95% d'entre elles, l'intervalle $\left[f_n-\frac{1}{\sqrt{n}}, f_n+\frac{1}{\sqrt{n}}\right]$ contiendra p[(9)].

Un exemple : **la parité, c'est quoi ?**

Deux entreprises A et B recrutent dans un bassin d'emploi où il y a autant de femmes que d'hommes, avec la contrainte du respect de la parité. Dans l'entreprise A, il y a 100 employés dont 43 femmes ; dans l'entreprise B, il y a 2 500 employés dont 1 150 femmes (soit 46%). Or 46% est plus proche de 50% que 43% : les chiffres parlent d'eux-mêmes, pourrait-on dire, et l'entreprise B respecte mieux la parité que l'entreprise A. Si on admet que la parité, c'est exactement 50% de femmes, il est vrai que B en est plus proche que A. Mais une telle définition, à l'unité près, de la parité n'aurait ici pas de sens.

La parité, cela signifie que l'identité sexuelle n'intervient pas au niveau du recrutement, que ce soit directement ou indirectement à travers des paramètres qui y seraient liés. Cela signifie donc qu'au niveau de la variable identité sexuelle, les résultats observés pourraient être obtenus par choix au hasard des individus dans la population. Dans ce cas, d'après ce qu'on vient de voir, on a une probabilité 0,95 que :

(9) Ayant observé une valeur $f_n = 0{,}7$ de la fréquence des 1 pour $n = 100$, on dira que $\left[f_n-\frac{1}{\sqrt{n}}, f_n+\frac{1}{\sqrt{n}}\right] = [0{,}6 ; 0{,}8]$ est l'intervalle de confiance de p au niveau de confiance 0,95. On évite de dire que l'intervalle $[0{,}6 ; 0{,}8]$ a une probabilité 0,95 de contenir p, car, stricto sensu, cette phrase ne veut rien dire : p est ou n'est pas dans cet intervalle. D'où la terminologie « niveau de confiance ».

– pour l'entreprise A la proportion de femmes soit dans [0,4 ; 0,6] ;
– pour l'entreprise B la proportion de femmes soit dans [0,48 ; 0,52].
La valeur 43% est donc dans l'intervalle de fluctuation, alors que 46% ne l'est pas.
Si on convient que respecter la parité, c'est avoir une proportion de femmes située dans l'intervalle de confiance à 0,95, alors seule l'entreprise A respecte la parité.

La proposition ci-dessus indique qu'il est délicat de comparer des pourcentages si on ne connaît pas les tailles des séries de données en jeu. Ainsi, dans une expérience de Bernoulli avec $p = 0{,}5$, observer 42% de 1 pour $n = 100$ n'est pas « exceptionnel » ; observer 48% de 1 pour $n = 10\,000$ sera jugé exceptionnel, bien que 48% > 42%. La statistique fait grand usage de ce genre de considérations.

On pourra bien sûr travailler sur des exemples ne relevant pas d'une expérience de Bernoulli.

3.2. Probabilités conditionnelles

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 calcul sur des données simulées consistera à diviser la fréquence des triplets contenant un 2 et de somme 12 par la fréquence des triplets de somme 12.

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.

On pourra admettre les deux notations P(A et B) et $P(A \cap B)$; de même on pourra employer indifféremment P(A ou B) et $P(A \cup B)$: l'enseignement des probabilités doit se prémunir d'une tendance à une rhétorique superflue.

3.3. Variables aléatoires, distributions de probabilités

Au niveau du lycée, on pourrait utiliser indifféremment distributions ou lois de probabilité.

Lois discrètes : lois équiréparties, lois binomiales, lois de Poisson (l'approximation de certaines lois binomiales par une loi de Poisson peut être en partie justifiée en terminale).
Lois continues(10) : loi uniforme sur un segment, un rectangle, un pavé ; lois exponentielles (loi sans vieillissement, utile pour étudier la radioactivité en physique), lois de Gauss(11)(on peut d'ailleurs imaginer que la loi de Gauss soit principalement traitée en cours de physique, dans la mesure où d'autres lois continues auront été étudiées en mathématiques).

(10) Comme nous l'avons déjà dit, tout n'est pas à traiter en cours de mathématiques.
(11) La loi de Gauss justifie d'introduire la notion d'écart-type dont l'introduction en dehors de la connaissance de ces lois ne se justifie pas vraiment au lycée.

3.4. Processus, états d'équilibre dynamiques

Avec les lancers successifs de dés, on a un processus sans mémoire. Quelques processus avec mémoire simples peuvent ensuite être abordés, notamment les chaînes de Markov.

En guise de conclusion :

Les pistes indiquées permettent dans le même temps de traiter une grande partie de la statistique (représentations graphiques, paramètres de tendance centrale et de dispersion) qu'on pourrait aborder dans l'enseignement secondaire dans différentes disciplines, et qui nécessite d'être complétée par l'étude de quelques grands indicateurs (en démographie par exemple)

C'est la réflexion sur la statistique construite (indices et indicateurs variés) et la réflexion sur ses usages possibles pour gouverner, administrer et agir, complétée par une formation en probabilité, qui fonde la statistique en tant que mode de pensée de la variabilité. L'enjeu et le défi pour les professeurs actuellement en poste est de former une nouvelle génération d'hommes et de femmes qui, partageant une culture commune de l'aléatoire, pourront plus facilement prendre une part active aux débats démocratiques.

La formation en statistique a donc vocation à s'adresser à tous les élèves, des voies générale, technologique et professionnelle, tant, comme nous l'avons dit, pour être en mesure d'exercer leur citoyenneté que pour d'éventuelles applications professionnelles, et cela dans tous les domaines (industriel, tertiaire, économique, médical, etc.). La statistique fait aujourd'hui partie du corpus de connaissances indispensables à tous et que l'école obligatoire se doit d'enseigner.

Enfin, les thèmes introduits ci-dessus pourront guider les enseignants pour leur formation dans le champ des « proba-stats », et éclairer leur réflexion sur les enjeux de leur enseignement. On pourra alors plus facilement tenir des débats pédagogiques sans que la crainte de l'inconnu prenne la place de l'argumentation.