Statistique et probabilités au lycée

Les auteurs de ce document sont les membres du groupe « Probabilités et statistique au lycée » qui a travaillé à l'Irem de Rennes de septembre 2013 à juillet 2015 [1] :

Guy CASALE

Jean-Baptiste FAURE

Laurence GAUDE

Gilbert GARNIER

Claude GUIBERT

Stéphane LE BORGNE

Isabelle LE NAOUR

Valérie MONBET

1. Le groupe a bénéficié aussi de l'expertise d'Alain Dessertaine. Nous le remercions chaleureusement. Nous remercions également l'Institut de Recherche en Mathématiques de Rennes (IRMAR) qui a pris en charge l'édition de cette brochure.

Table des matières

Introduction

Il nous semble qu'existe une demande de clarification sur les notions de probabilités et statistique aux programmes des lycées. Ce fascicule s'adresse aux professeurs de mathématiques des lycées [2]. Il comporte deux parties. La première est une présentation concise des notions de statistique et probabilités présentes dans les programmes du lycée ; nous ne nous limitons pas à ce qui est exigible des élèves ; l'objectif est de fournir aux professeurs des éléments qui leur permettront d'être à l'aise avec les notions abordées ; mais nous nous éloignons peu des programmes. La deuxième partie est une collection d'activités ou thèmes dont la plupart peuvent servir de base à des cours proposés aux élèves (elles peuvent concerner les classes de seconde à la terminale pour les séries S, ES, STMG, STL, STIDD).

Conformément aux instructions des programmes [3] nous mettrons l'accent sur l'expérimentation, la simulation. Nous avons choisi de privilégier deux types de logiciels : Algobox et le tableur. Il est possible de travailler avec R, Géogebra, Python. Nous avons préféré nous concentrer sur les deux premiers cités.

De nombreuses ressources sont disponibles pour aider les professeurs à traiter les chapitres de statistique et probabilités des programmes. Nous avons essayé de concevoir un document clair et relativement concis. Certaines des activités que nous proposons ne sont abordables au lycée que parce qu'elles peuvent donner lieu à des simulations informatiques ; le fait qu'elles ne puissent être traitées complètement mathématiquement ne nous semble pas une bonne raison pour n'en pas parler. Nous les trouvons intéressantes ; écrire des programmes de simulation nécessite d'en avoir compris certains aspects mathématiques ; les algorithmes obtenus donnent des réponses à des questions naturelles.

2. La feuille de route du groupe était : *Comment replacer dans leur contexte les principales notions de probabilités et de statistique, tant du point de vue de la théorie que du point de vue de la modélisation de situations réelles et comment les enseigner afin de motiver les élèves en mettant ceux-ci en situation d'attribuer du sens aux éléments qui leur sont présentés ? Comment penser une progression de cet enseignement de la 3e à la terminale ? Le groupe travaillera à apporter des réponses à de telles questions sur des notions du programme de terminale S ou ES de la rentrée 2012 (modèles probabilistes, loi des grands nombres, chaînes de Markov, lois à densité ?).*

3. L'objectif du groupe n'était pas de critiquer les programmes. Nous avons quand même rassemblé quelques réflexions à ce sujet en fin de première partie.

1 Point sur les notions et compléments

1.1 Statistique descriptive

Deux didacticiens des mathématiques, Yves Chevallard et Florence Wozniak parlent de l'enseignement de la statistique comme d'un problème posé à la profession [4]. L'attitude des professeurs de mathématiques à l'égard de la statistique est souvent ambiguë. Les propos suivants (cités par Chevallard et Wozniak) l'illustrent : « Étant donné que les futurs S n'en feront plus (et seront sans doute aptes à s'y mettre...) et que de toute façon les ES en refont en 1re, certains ne voyaient guère l'intérêt d'en faire vu qu'actuellement c'est la plupart du temps bâclé en fin d'année si c'est fait... Mais nous sommes gênés de demander qu'on supprime les stats alors qu'on revendique des maths citoyennes... »APMEP [Jean-Pierre Richeton], BGV, 87 (juin 1999)

Les programmes mettent beaucoup l'accent sur les liens entre statistique et probabilités et sur les propriétés asymptotiques. Ce point de vue est critiquable. Il est très intéressant bien sûr d'étudier les propriétés asymptotiques. Mais intéressant également d'étudier les données en elles-mêmes sans les considérer comme des erreurs par rapport à une limite observée à l'infini.

Nous pensons que l'enseignement de la statistique descriptive ne doit pas être sacrifié. Cela signifie qu'il ne faut pas le réduire à l'apprentissage de définitions et de calculs arithmétiques et qu'il faut sans doute accepter d'aborder des sujets non mathématiques en cours de mathématiques.

Citons quelques passages du travail de Chevallard et Wozniak : « (...) un projet d'enseignement de la statistique doit faire apparaître nettement l'objet de cette science : l'étude de la variabilité. Y renoncer implique, à terme, sinon la disparition complète de l'enseignement correspondant, du moins sa dégradation et son régime indéfiniment végétatif, tant du moins que le corpus de savoir rassemblé sous l'étiquette de statistique n'aura pas conquis une identité claire, qui le distingue des autres parties du corpus mathématique enseigné tout en permettant de le situer par rapport à elles. (...) L'un des grands obstacles à l'ambition d'un pacte sociétal et scolaire autour de l'enseignement de la statistique tient à ce fait que l'idée d'une science de la variabilité n'existe pas intensément dans la culture de la société française "éternelle" : fait massif, qui, aussi bien, touche les professeurs de mathématiques eux-mêmes. La dénégation de la variabilité est en effet la loi plutôt que l'exception. On tend ainsi à regarder toute grandeur comme constante, attitude dont le langage courant porte témoignage (on demandera facilement "combien ça pèse un lion", "combien il y a de pétales dans une rose", etc.). Or la première conquête collective à accomplir se trouve là : dans le fait d'assumer la variation. »

4. Enseigner la statistique en seconde : un Problème de la profession, http://yves.chevallard.free.fr/spip/spip/IMG/pdf/CORFEM_2007_YC_FW_actes.pdf

Le présent fascicule se veut un outil d'aide à l'enseignement des notions aux programmes. Il accorde une place importante aux rapports entre probabilités et statistique ; cela ne signifie pas que nous pensions que la statistique descriptive soit à négliger.

Les connaissances de collège :

L'étude de séries statistiques est inscrite dans les programmes du collège.

En troisième, les élèves ont appris à déterminer médiane et quartiles d'une série statistique et savent les interpréter.

Le terme " étendue " apparaît aussi en troisième et les élèves ont pu se familiariser avec la notion de dispersion lors de l'étude de la disparité des mesures d'une grandeur lors d'activités expérimentales, en sciences physiques, par exemple. Dans les classes antérieures, fréquences, moyennes et moyennes pondérées ont été étudiées.

En seconde, il s'agit de déterminer et d'interpréter des résumés d'une série statistique mais aussi de réaliser la comparaison de deux séries à l'aide d'indicateurs de position et de dispersion ou de la courbe des fréquences cumulées. Il n'y a pas à proprement parler de nouvelles connaissances (à part peut-être les effectifs cumulés et les fréquences cumulées qui ne semblent pas être au programme de collège) mais plutôt des savoir-faire assez divers :
– utiliser un logiciel ou la calculatrice
– calculer les caractéristiques d'une série donnée par effectifs ou fréquence.
– calculer effectifs et fréquences cumulés
– représenter une série statistique par différentes moyens (nuage, histogramme, courbe fréquences cumulées)

L'objectif visé est de pouvoir mener une réflexion sur des données réelles, riches et variées, de synthétiser l'information et de proposer des représentations pertinentes.

En première, l'étude et la comparaison de séries statistiques menées en classe de seconde se poursuivent avec la mise en place de nouveaux outils. En effet, c'est en première que sont introduites les caractéristiques de dispersion variance et écart-type. Un des objectifs de la première est d'utiliser de façon appropriée les deux couples d'indicateurs (moyenne-écart-type et médiane-écart interquartile) pour résumer une série statistique. Dans les séries générales et technologiques (S-ES-STG-STI2D-STL-ST2S), les différences entre les textes des programmes sont minimes :
– en STMG et ST2S, la variance n'est pas vue, de plus l'expression de l'écart-type n'est pas un attendu du programme de ces séries (on attend des élèves qu'ils sachent utiliser un tableur ou la calculatrice pour déterminer l'écart-type)
– les diagrammes en boîte sont vus dans toutes les séries sauf en STI2D et en STL (ces deux séries ont un programme commun en maths). On peut d'ailleurs noter que ces diagrammes pourraient être introduits dès la troisième ou la seconde.
– on demande explicitement en S et ES d'observer des effets de structure lors du calcul

de moyenne mais pas dans les autres séries.
– une capacité clairement attendue dans le programme de STMG est de rédiger l'interprétation d'un résultat ou l'analyse d'un graphique, ce qui n'est pas exprimé dans les autres programmes.
– il est fait référence aux spécificités des séries technologiques STI2D et STL : il est précisé dans les commentaires de ces programmes de privilégier l'étude d'exemples issus de résultats d'expériences, de la maîtrise statistique des procédés, du contrôle de qualité, de la fiabilité ou liés au développement durable. (Les premières STD2A n'ont pas de statistique à leur programme)

En terminale, les programmes ne font pas référence aux statistiques descriptives. La connaissance des notions de statistique descriptive ne sont pas évaluées au baccalauréat (encourageant les professeurs à n'y pas consacrer de temps en seconde ou première...).

La statistique descriptive est souvent perçue comme plate, vide mathématiquement et philosophiquement (ce qui est probablement une erreur). Avec les probabilités les problèmes philosophiques risquent au contraire de prendre une place démesurée...

1.2 Probabilités

1.2.1 Remarques sur la théorie des probabilités

Il peut être intéressant de rappeler quelques définitions issues de dictionnaires.

Petit Larousse

Probabilité : Vraisemblance. Examiner les probabilités. Conception scientifique et déterministe du hasard. Calcul des probabilités : ensemble des règles permettant de déterminer le pourcentage des chances de réalisation d'un événement.

Hasard (arabe, azzahr, jeu de dés) : Cause fictive des événements apparamment soumis à la seule loi des probabilités. Événement imprévu, chance bonne ou mauvaise. Le hasard d'une rencontre.

Chance (latin, cadere, tomber) : Manière dont un événement peut tourner.

Possibilité (latin, posso, pouvoir) : Ce qui peut être.

Vraisemblable : Probable.

Petit Robert

Probabilité : Caractère de ce qui est probable, vraisemblable. Grandeur par laquelle on mesure le caractère aléatoire (possible et non certain) d'un événement, d'un phénomène, par l'évaluation du nombre de chances d'en obtenir la réalisation. Apparence, indice qui laisse à penser qu'une chose est probable. Opinion fondée sur de simples probabilités.

On remarquera que le sens du mot probable a changé (cela n'arrive pas si rarement aux

mots). Par exemple on pouvait écrire au dix- septième siècle : « Un tel fait est sans doute probable, mais sans aucun doute faux. »

La notion de probabilité est souvent perçue de manière intuitive. Chercher à la préciser mène parfois à des discussions difficiles. Le lien entre probabilités et déterminisme a ainsi fait l'objet de nombreuses réflexions. Citons quelques phrases pouvant donner à penser dans cette direction.

Bachelard (Essai sur la connaissance approchée, 1927) : *En effet, si le probable se réalise, il était possible, mais il cesse de l'être pour devenir réel, s'il ne se réalise pas, il faut bien qu'il ait été impossible par certains côtés.*

Bachelier (La spéculation et le calcul des probabilités, 1938) : *Dans la théorie de la spéculation on se garde donc bien d'entreprendre l'analyse des causes qui peuvent agir sur les cours (...) C'est précisément parce que l'on veut tout ignorer qu'il est possible de savoir, c'est précisément parce que cette étude semble d'une complication inextricable qu'elle est, en réalité, d'une grande simplicité.*

Poincaré (1908) : *Une cause très petite, qui nous échappe, détermine un effet considérable que nous ne pouvons pas ne pas voir, et alors nous disons que cet effet est dû au hasard... Mais, lors même que les lois naturelles n'auraient plus de secret pour nous, nous ne pourrons connaître la situation initiale qu'approximativement. Si cela nous permet de prévoir la situation ultérieure avec la même approximation, c'est tout ce qu'il nous faut, nous dirons que le phénomène a été prévu, qu'il est régi par des lois ; mais il n'en est pas toujours ainsi, il peut arriver que de petites différences dans les conditions initiales en engendrent de très grandes dans les phénomènes finaux...*

Pour l'enseignement du calcul des probabilités (dans le cours de mathématiques) il nous semble bon d'adopter le point de vue de William Feller auteur d'un célèbre traité de calculs des probabilités (fin des années quarante du vingtième siècle). Ne pas faire comme si la notion de probabilité ne posait pas de problème philosophique ; proposer une conception scientifique dont on admet qu'elle puisse être discutée (travail qui n'est pas celui du mathématicien). Dans ce qui suit, les parties en italique de ce paragraphe sont des traductions d'extraits de l'introduction du livre de William Feller (tome 1), introduction qui porte le titre "Sur la nature de la théorie des probabilités".

Il faut se souvenir que les mathématiques traitent de modèles abstraits et que différents modèles peuvent décrire la même situation empirique avec des degrés d'approximation et de simplicité variés. La manière d'appliquer les théories mathématiques ne dépend pas d'idées préconçues et n'est pas affaire de logique : c'est une technique raisonnée (résolue, purposeful) qui dépend de l'expérience et change avec elle. Bien sûr, tous les aspects des activités humaines intéressent le philosophe et une analyse philosophique des applications des mathématiques est légitime. Cependant, une telle analyse ne fait pas partie du royaume des mathématiques, de la physique ou de la statistique. Il faut séparer la discussion philosophique des fondements des probabilités de la théorie mathématique et de ses applications

comme on sépare la discussion de notre concept intuitif de l'espace de la géométrie.

Pour illustrer son point de vue sur la confrontation entre théorie et application Feller évoque les lois de Bose-Einstein, Fermi-Dirac, Maxwell-Boltzmann. Ce sont des probabilités sur des ensembles de répartitions de n objets dans r cases. Suivant la nature des objets considérés (protons ou photons par exemple) les lois observées sont différentes. La théorie des probabilités permet de concevoir différentes lois ; seule l'expérience indique quelle loi est adaptée au phénomène concret étudié.

Savoir dans quelle mesure la logique, l'intuition et l'expérience sont liées entre elles est un problème philosophique difficile dans lequel nous n'entrerons pas. Il est certain que l'intuition peut être travaillée et développée.

L'intuition se développe avec la théorie.

Beaucoup de livres de probabilités rappellent de vieux livres de mécanique en ce qu'ils contiennent beaucoup de philosophie et de tentatives de définir des choses réelles plutôt que des relations.

On a donné une publicité malencontreuse à certaines discussions de ce qu'on appelle les fondements des probabilités, créant ainsi l'impression fallacieuse que des désaccords profonds existaient parmi les mathématiciens. En fait ces discussions ne concernent que des points mineurs et n'intéressent que quelques spécialistes.

La comparaison avec le jeu d'échec est intéressante : un débutant hésite longuement sur les règles de déplacement des différentes pièces ; un joueur chevronné saisit en un instant les positions des deux camps. Quel sens donner à la question "Quelle est la nature ultime du roi d'un jeu d'échec ?" par exemple ?

Seule une théorie mathématique générale est suffisamment souple pour fournir les outils adéquats à une telle variété de problèmes et il faut se garder de trop attacher nos notions, images ou mots à un champ d'expérience particulier.

Feller compare à plusieurs reprises les probabilités et la géométrie. Les concepts abstraits de la géométrie souvent développés pour eux mêmes sont d'une grande utilité en mécanique par exemple. On peut comparer le concept de probabilité à celui de masse. On peut utiliser la notion de masse en mécanique sans se préoccuper en permanence de la nature ultime de la masse. En probabilités on considère les issues possibles d'une expérience conceptuelle, la possibilité de réaliser concrètement l'expérience n'est pas importante en général (comme le fait de savoir si un cercle parfait existe ou non dans la nature n'est pas un problème que se pose en général le géomètre). Question : l'informatique ne permet-elle pas aujourd'hui d'approcher de la réalisation d'expériences conceptuelles ?

Le succès de la théorie des probabilités moderne a un prix : la théorie est limitée a un aspect de la notion.

C'est l'aspect qu'on pourrait nommer physique ou statistique ; on ne s'intéresse pas à ce

qui concerne le jugement.

L'approche par la théorie de la mesure n'entraîne ni paradoxe, ni difficulté et a l'énorme avantage de substituer des théorèmes à de vagues discussions de paradoxes.

Les fondements (théoriques) des choses pratiques d'aujourd'hui étaient critiqués hier, et les théories qui demain seront pratiques sont qualifiées de jeux abstraits sans valeur par les hommes pragmatiques d'aujourd'hui.

Quand dit-on qu'une chose arrive par hasard ?

Il n'est peut-être pas inutile de décrire quelques phénomènes où le hasard intervient et d'essayer de dire de quelle façon. La notion de hasard est liée à l'instabilité, à la présence d'une multiplicité d'influences difficiles à évaluer, au manque d'information. Les exemples les plus couramment cités sont la pièce lancée (suffisamment fort pour que le résultat ne soit pas prévisible (instabilité)) équilibrée, un jeu de cartes battu suffisamment longtemps et de manière suffisamment désordonnée. Certains systèmes physiques bien que considérés comme déterministes présentent une instabilité foncière (on parle de dépendance sensitive aux conditions initiales ; popularisée sous le nom d'effet papillon). Dans de tels systèmes des régularités peuvent être observées, mais elles sont de type statistique ou probabiliste.

Comment introduire le hasard en mathématiques ?

Nous conseillons de partir d'une situation symétrique dans laquelle l'équiprobabilité est facilement admise (jeu de pile ou face, dés, urnes). Cela permet de modéliser, de raisonner, de compter, de calculer. On aboutit à la fameuse définition d'une probabilité comme « nombre de cas favorables sur nombre de cas possibles ». Ces exemples élémentaires ont donné lieu à une multitude d'exercices dans lesquels certains mots sont parfois employés pour indiquer qu'il est convenu que la probabilité à considérer est la probabilité uniforme : pièce non truquée, dé non pipé, urne bien mélangée, jeu de cartes bien battu... On peut donner une justification mathématique à de tels mots : Henri Poincaré a montré que l'équiprobabilité des ordres différents de cartes apparaît comme une limite de probabilités obtenues par battage des cartes même si le batteur peut avoir des habitudes asymétriques. Dans son cours de probabilités il montre comment (et pourquoi) une roue dentée peut approcher l'équiprobabilité. Les modèles d'urne avec équiprobabilité, bien que parfois artificiels et lourds, peuvent être source d'analogies intéressantes.

Ces exemples traditionnels sont riches mais il faut sans doute pouvoir en sortir. Il est très fréquent que l'ensemble des issues possibles Ω soit inaccessible. En actuariat, mathématiques de l'assurance, les valeurs des paramètres intervenant dans les modèles utilisés seront fonctions d'observations du passé. Aucun Ω muni de l'équiprobabilité ne sera introduit. Aux observations pourront être associés des calculs de fréquences donc des rapports de cardinaux mais leur signification ne sera pas le célèbre nombre de cas favorables sur nombre de cas possibles (la preuve : ces fréquences changent avec le temps). L'utilisation de telles fréquences est basée sur une stabilité que le calcul des probabilités permet de

comprendre (grâce à loi des grands nombres dans le cadre de répétitions d'expériences indépendantes par exemple). Les hypothèses d'indépendance, la possibilité de segmenter une population ou un phénomène peuvent être discutées.

Remarque sur les très très grands nombres

On ne comprend pas mieux l'équiprobabilité sur $\{0,1\}^{1000}$ que l'indépendance de mille pile ou face successifs. Mais avoir manipulé de tels modèles avec un petit nombre de répétitions donne quelques prises sur des ensembles aussi gros.

1.2.2 Loi des grands nombres

La loi faible des grands nombres est une conséquence de l'inégalité de Bienaymé-Tchébychev.

Théorème 1.1. *(inégalité de Bienaymé-Tchébychev) : Soit X une variable aléatoire d'espérance $\mathbb{E}(X)$ et d'écart type $\sigma(X)$. Pour tout $\epsilon > 0$,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{\sigma^2(X)}{\epsilon^2}.$$

Démonstration : Nous donnons une démonstration dans le cas d'une variable discrète (d'ensemble de valeurs fini) : appelons x_i ces valeurs, i variant de 1 à n pour un certain entier naturel n, et p_i la probabilité de prendre la valeur x_i ($p_i = \mathbb{P}(X = x_i)$).

$$Var(X) = \sum_{i=1}^{n} p_i (x_i - \mathbb{E}(X))^2.$$

On note I l'ensemble des indices i tels que $|x_i - \mathbb{E}(X)| \geq \epsilon$ et J l'ensemble des indices i tels que $|x_i - \mathbb{E}(X)| < \epsilon$. On peut séparer la somme en deux en utilisant les deux ensembles I et J :

$$Var(X) = \sum_{i \in I} p_i (x_i - \mathbb{E}(X))^2 + \sum_{i \in J} p_i (x_i - \mathbb{E}(X))^2.$$

On en déduit les minorations suivantes

$$Var(X) \geq \sum_{i \in I} p_i (x_i - \mathbb{E}(X))^2 \geq \sum_{i \in I} p_i \epsilon^2 = \epsilon^2 \sum_{i \in I} p_i$$

Or $\sum_{i \in I} p_i = \mathbb{P}(|X - \mathbb{E}(X)| \geq \epsilon)$; on a ainsi

$$Var(X) \geq \epsilon^2 \mathbb{P}(|X - \mathbb{E}(X)| \geq \epsilon).$$

d'où le résultat. □

Définition : Soit (X_n) une suite de variables aléatoires réelles et soit X une variable aléatoire réelle. On dit que (X_n) converge en probabilité vers X si et seulement si

$$\lim_n \mathbb{P}(|X_n - X| \geq \epsilon) = 0.$$

Théorème 1.2. *(loi faible des grands nombres) Soit (X_n) une suite de variables aléatoires réelles de même espérance m et de même écart type σ. On suppose de plus que les X_i sont deux à deux indépendantes. On pose*

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

La suite $(\overline{X_n})$ converge en probabilité vers la constante m et on a pour tout, $\epsilon > 0$:

$$\mathbb{P}(|\overline{X_n} - m| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Ce qui a pour conséquence : pour tout $\epsilon > 0$,

$$\mathbb{P}(|\overline{X_n} - m| \geq \epsilon) \to 0$$

ou encore

$$\mathbb{P}(|\overline{X_n} - m| < \epsilon) \to 1.$$

Démonstration : L'espérance de $\overline{X_n}$ est

$$\mathbb{E}(\overline{X_n}) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}nm = m.$$

D'après l'inégalité de Bienaymé-Tchébychev, on a

$$\mathbb{P}(|\overline{X_n} - m| \geq \epsilon) = \mathbb{P}(|\overline{X_n} - \mathbb{E}(\overline{X_n})| \geq \epsilon) \leq \frac{\sigma^2(\overline{X_n})}{\epsilon^2}.$$

Calculons $\sigma^2(\overline{X_n})$: en utilisant les propriétés de la variance et l'indépendance des X_i, il vient

$$Var(\overline{X_n}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i),$$

et comme $\sum_{i=1}^{n} Var(X_i) = nVar(X_1) = n\sigma^2$, on a

$$Var(\overline{X_n}) = \frac{\sigma^2}{n},$$

d'où le résultat.

Exemple : On lance n fois un dé équilibré. Pour le i-ème lancer on considère la variable aléatoire X_i qui vaut 1 si le résultat est un multiple de 3 et 0 sinon. On pose $X = \frac{1}{n}\sum_{i=1}^{n} X_i$. Déterminer n pour que la probabilité de l'événement « X s'éloigne de son espérance de plus de 0,02 » soit inférieure à 0,1.

Il faut rechercher n tel que $\mathbb{P}(|X - \frac{1}{3}| > 0,02) < 0,1$. Comme $\mathbb{P}(|X - \frac{1}{3}| > 0,02) \leq \frac{\sigma^2(X)}{n0,02^2}$ et $Var(X_i) = \frac{2}{9}$, la majoration voulue sera satisfaite lorsque $\frac{2}{n.9.0,02^2}$ sera inférieur à 0,1, soit $n > \frac{1}{0,0018.0,1}$ ou $n \geq 5556$. Il est bon de remarquer que cette majoration est de très mauvaise qualité car le calcul donne pour $n = 5556$, $\mathbb{P}(|X - \frac{1}{3}| > 0,02) \simeq 0,0015$. Il faut en fait beaucoup moins de lancers pour obtenir la majoration souhaitée.

1.2.3 Le théorème limite central

Plusieurs démonstrations existent. Pour les variables de lois binomiales on peut utiliser la formule de Stirling dont la démonstration fait appel à la notion de développement limité. Mais l'un des aspects remarquables du théorème limite central est qu'il est valable pour toute suite de variables indépendantes de même loi de carré intégrable. Une façon élégante de le voir est de considérer la fonction caractéristique et d'appliquer le théorème de Lévy. Nous proposons ici une preuve (ou plutôt une esquisse de preuve) n'utilisant que les développements limités et quelques manipulations élémentaires. La preuve complète serait assez longue : nous n'en donnerons que les idées principales.

Théorème 1.3. *Soit une suite X_i de variables aléatoires indépendantes de même loi de carré intégrable que nous considérerons centrées pour simplifier. Notons $\sigma^2 = \mathbb{E}(X_1^2)$ la variance de chacune des variables X_i. Notons S_n la somme $X_1 + \ldots + X_n$. La probabilité $\mathbb{P}(S_n/\sqrt{n} \in [a,b])$ converge vers*

$$\int_a^b \exp(-\frac{x^2}{2\sigma^2})\frac{dx}{\sqrt{2\pi}\sigma}.$$

Idées de démonstration.

Première idée : montrer la convergence précédente revient à montrer la convergence de l'espérance de $f(S_n/\sqrt{n})$ vers l'espérance de $f(Y)$ où Y est une variable gaussienne centrée de variance σ^2 et f est l'application qui vaut 1 sur $[a,b]$, 0 sur son complémentaire.

Deuxième idée : si on arrive à montrer la même chose pour les fonctions f C^∞ à supports dans des intervalles finis, alors le théorème est vrai par approximation de l'indicatrice de $[a,b]$ par des fonctions régulières.

Troisième idée : si les variables Y_i sont indépendantes de loi gaussienne centrée de variance σ^2 alors $(Y_1 + \ldots + Y_n)/\sqrt{n}$ est gaussienne centrée de variance σ^2.

Il nous suffit donc de montrer que

$$\mathbb{E}\left(f\left(\sum_{i=1}^n \frac{X_i}{\sqrt{n}}\right)\right) - \mathbb{E}\left(f\left(\sum_{i=1}^n \frac{Y_i}{\sqrt{n}}\right)\right)$$

tend vers 0 quand n tend vers l'infini.

Quatrième idée : on écrit cette différence comme une somme de différences :

$$\begin{aligned}
&\mathbb{E}\left(f\left(\sum_{i=1}^n \frac{X_i}{\sqrt{n}}\right)\right) - \mathbb{E}\left(f\left(\sum_{i=1}^n \frac{Y_i}{\sqrt{n}}\right)\right) \\
&= \mathbb{E}\left(f\left(\sum_{i=1}^n \frac{X_i}{\sqrt{n}}\right) - f\left(\sum_{i=1}^n \frac{Y_i}{\sqrt{n}}\right)\right) \\
&= \sum_{k=1}^n \mathbb{E}\left(f\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k}^n \frac{Y_i}{\sqrt{n}}\right) - f\left(\sum_{i=1}^{k} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^n \frac{Y_i}{\sqrt{n}}\right)\right)
\end{aligned}$$

Les sommes apparaissant comme variables de f à k donné diffèrent de $\frac{X_k}{\sqrt{n}} - \frac{Y_k}{\sqrt{n}}$.

Cinquième idée : on fait des développements limités de f à k fixé au point $\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}$. On obtient :

$$\begin{aligned} f\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k}^{n} \frac{Y_i}{\sqrt{n}}\right) &= f\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right) + f'\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right) \frac{Y_k}{\sqrt{n}} \\ &\quad + f''\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right) \frac{Y_k^2}{2n} + O(n^{-3/2}) \end{aligned}$$

et

$$\begin{aligned} f\left(\sum_{i=1}^{k} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right) &= f\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right) + f'\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right) \frac{X_k}{\sqrt{n}} \\ &\quad + f''\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right) \frac{X_k^2}{2n} + O(n^{-3/2}) \end{aligned}$$

En faisant la différence on obtient :

$$\begin{aligned} &f\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k}^{n} \frac{Y_i}{\sqrt{n}}\right) - f\left(\sum_{i=1}^{k} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right) \\ &= f'\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right)\left(\frac{Y_k}{\sqrt{n}} - \frac{X_k}{\sqrt{n}}\right) \\ &\quad + f''\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right)\left(\frac{Y_k^2}{2n} - \frac{X_k^2}{2n}\right) + O(n^{-3/2}) \end{aligned}$$

Comme les variables sont indépendantes, centrées et de même variance, quand on prend l'espérance, il ne reste plus que le terme $O(n^{-3/2})$:

$$\begin{aligned} &\mathbb{E}\left(f\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k}^{n} \frac{Y_i}{\sqrt{n}}\right) - f\left(\sum_{i=1}^{k} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right)\right) \\ &= \mathbb{E}\left(f'\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right)\right)\mathbb{E}\left(\frac{Y_k}{\sqrt{n}} - \frac{X_k}{\sqrt{n}}\right) \\ &\quad + \mathbb{E}\left(f''\left(\sum_{i=1}^{k-1} \frac{X_i}{\sqrt{n}} + \sum_{i=k+1}^{n} \frac{Y_i}{\sqrt{n}}\right)\right)\mathbb{E}\left(\frac{Y_k^2}{2n} - \frac{X_k^2}{2n}\right) + O(n^{-3/2}) \end{aligned}$$

On a un tel terme pour chaque k donc n tels termes, dont on fait la somme. Reste une quantité de l'ordre de $n^{-1/2}$ qui tend vers 0. Le théorème est démontré.

1.3 Statistique inférentielle

1.3.1 Intervalle de fluctuation

Un intervalle de fluctuation donne une information sur la fluctuation du résultat d'une expérience aléatoire. Si on s'intéresse aux réalisations d'un évènement A de probabilité π $(0 < \pi < 1)$, un intervalle de fluctuation de la fréquence d'occurrence F_n de l'évènement sur n tirages mesure la fluctuation d'échantillonnage. La fréquence F_n est une variable aléatoire et un intervalle de fluctuation de niveau $1 - \alpha$ est défini comme un intervalle $[a, b]$ tel que

$$P(a < F_n < b) = 1 - \alpha.$$

α est un risque d'erreur. On cherche à connaitre la gamme des valeurs possibles de F_n avec un risque de se tromper de probabilité α.

Le plus souvent on construit des intervalles de fluctuation symétriques c'est à dire tels que la probabilité que F_n soit inférieur à a est égale à $\alpha/2$ de même que celle que F_n soit supérieur à b. Quand on construit un intervalle de fluctuation, on cherche a et b tels que

$$P(a < F_n < b) = 1 - \alpha$$

ce qui est équivalent à

$$P\left(\frac{a-\pi}{\sqrt{\pi(1-\pi)/n}} < \frac{F_n-\pi}{\sqrt{\pi(1-\pi)/n}} < \frac{b-\pi}{\sqrt{\pi(1-\pi)/n}}\right) = 1 - \alpha$$

En notant Φ la fonction de répartition de la variable aléatoire $\frac{F_n-\pi}{\sqrt{\pi(1-\pi)/n}}$

$$\Phi\left(\frac{b-\pi}{\sqrt{\pi(1-\pi)/n}}\right) - \Phi\left(\frac{a-\pi}{\sqrt{\pi(1-\pi)/n}}\right) = 1 - \alpha.$$

Si la densité associée à la loi Φ est symétrique

$$\Phi\left(\frac{a-\pi}{\sqrt{\pi(1-\pi)/n}}\right) = 1 - \Phi\left(\frac{b-\pi}{\sqrt{\pi(1-\pi)/n}}\right)$$

On a donc

$$\Phi\left(\frac{b-\pi}{\sqrt{\pi(-1-\pi)/n}}\right) = 1 - \alpha/2$$

et

$$\frac{b-\pi}{\sqrt{\pi(1-\pi)/n}} = -\frac{a-\pi}{\sqrt{\pi(1-\pi)/n}} = \Phi^{-1}(1-\alpha/2)$$

Finalement,

$$a = \pi - \Phi^{-1}(1-\alpha/2)\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}} \text{ et } b = \pi + \Phi^{-1}(1-\alpha/2)\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}$$

On peut remarquer que $\pi(1-\pi) \leq \frac{1}{4}$ et on a donc

$$P\left(\pi - \Phi^{-1}(1-\alpha/2)\frac{1}{2\sqrt{n}} \leq F_n \leq \pi + \Phi^{-1}(1-\alpha/2)\frac{1}{2\sqrt{n}}\right) > 1-\alpha$$

Soit X une variable aléatoire de loi de Bernouilli de paramètre π. Considérons différents échantillons de même effectif n tirés au sort. On montre que la loi de probabilité de nF_n est une loi binomiale de paramètres n et π. Mais si n est grand, on a aussi d'après le théorème limite central, la loi de $\frac{F_n-\pi}{\sqrt{\pi(1-\pi)/n}}$ tend vers une loi de Gauss de moyenne 0 et de variance 1. On peut utiliser l'approximation de la loi de $\frac{F_n-\pi}{\sqrt{\pi(1-\pi)/n}}$ par la loi de Gauss pour obtenir un intervalle de fluctuation asymptotique pour F_n. Dans ce cas, on a $\Phi^{-1}(1-\alpha) = 1.96$ si $\alpha = 0.05$. Et on retrouve l'approximation

$$P\left(\pi - \frac{1}{\sqrt{n}} \leq F_n \leq \pi + \frac{1}{\sqrt{n}}\right) > 1-\alpha$$

Exemple. Dans une population, la fréquence d'un facteur est de 12%. On tire au hasard un échantillon de 100 sujets. L'intervalle de fluctuation asymptotique de niveau 95% pour la fréquence observée est

$$[0.12 - 1.96\sqrt{\frac{0.12 \times 0,88}{100}}, 0.12 + 1.96\sqrt{\frac{0.12 \times 0.88}{100}}] = [0.06, 0.18]$$

1.3.2 Test d'hypothèses

De nombreux problèmes posent des questions auxquelles on répond en utilisant un intervalle de fluctuation mais il serait plus rigoureux d'utiliser des tests d'hypothèses.

Exemple (d'après le sujet de BAC 2014, métropole, ex 2, partie B) - La chaîne de production du laboratoire fabrique, en très grande quantité, le comprimé d'un médicament. Elle a été réglée dans le but d'obtenir au moins 97% de comprimés conformes. Afin d'évaluer l'efficacité des réglages, on effectue un contrôle en prélevant un échantillon de 1000 comprimés dans la production. La taille de la production est supposée suffisamment grande pour que ce prélèvement puisse être assimilé à 1000 tirages successifs avec remise. Le contrôle effectué a permis de dénombrer 53 comprimés non conformes sur l'échantillon prélevé, soit 947 comprimés conformes. Peut-on considérer que les réglages sont conformes à ce qui est attendu ?

On va faire l'hypothèse que les réglages sont conformes c'est à dire qu'on produit $\pi = 97\%$ de comprimés conformes. Sous cette hypothèse, la loi de nF_n, nombre de comprimés conformes, suit une loi binomiale de paramètres n et π.

Réponse par un intervalle de fluctuation.
Un intervalle de fluctuation asymptotique au risque α pour la proportion de comprimés

conformes (sous l'hypothèse d'un réglage à 97%) est donné par

$$\left[0.97 - 1.96\sqrt{\frac{0.97 \times 0.03}{1000}}, 0.97 + 1.96\sqrt{\frac{0.97 \times 0.03}{1000}}\right] = [0.959, 0.981]$$

La proportion de comprimés conformes dans l'échantillon n'est pas dans l'intervalle de fluctuation à 95%. On concluera donc, au risque 5% que les réglages ne sont pas bons.

Réponse par un test d'hypothèses.
On pose deux hypothèses alternatives. La première est l'hypothèse de référence. On l'appelle aussi *hypothèse nulle*. Elle dit que les réglages sont conformes, c'est à dire que la loi de la fréquence de pièces conformes dans un tirage de n pièces suit bien une loi binomiale de paramètres n et $n\pi$ avec $\pi = 97\%$. La seconde hypothèse est appelée *hypothèse alternative*. Ici, elle dit que les réglages ne sont pas conformes c'est à dire que $\pi \neq 97\%$.
Dans ce problème particulier, on a envie d'être un peu plus précis et de poser comme hypothèse alternative : $\pi < 97\%$.

Mais traitons d'abord le premier cas. On se propose de prendre une décision avec un risque d'erreur α sur la base d'une statistique de test qui est ici la fréquence observée F_n du nombre de pièces conformes. On détermine alors une région de rejet qui correspond à un ensemble R de probabilité α qui est tel que si $f_n \in R$ on rejette l'hypothèse selon laquelle les réglages sont corrects, avec f_n une réalisation de F_n. Ici, la forme de R est donnée implicitement par la forme de l'hypothèse alternative. Ici on a $R = \{F_n < a \text{ ou } F_n > b\}$. On cherche a et b tel que

$$\mathbb{P}_{nF_n \sim B(n,\pi)}(F_n < a \text{ ou } F_n > b) = \alpha$$

ce qui est équivalent à

$$\mathbb{P}_{nF_n \sim B(n,\pi)}(a \leq F_n \leq b) = 1 - \alpha.$$

(L'indice ajouté à $\mathbb{P}$ signifie que l'on se place sous l'hypothèse que nF_n suit la loi binomiale de paramètres n et π.) On retrouve l'intervalle de fluctuation et on conclut, avec un risque d'erreur de 5%, que les réglages sont conformes car f_n est dans le complémentaire de la région de rejet.

Revenons au second cas, qui est plus naturel : on ne peut pas reprocher à la chaîne de fabriquer trop de comprimés conformes... L'hypothèse $\pi < 97\%$ est unilatérale. On a alors $R = \{F_n < a\}$, c'est à dire qu'on rejette l'hypothèse nulle si on n'observe pas assez de comprimés conformes. On cherche donc la borne a telle que

$$P_{nF_n \sim B(n,\pi)}(F_n < a) = \alpha$$

En utilisant le théorème limite central et avec le même type de calculs que pour l'intervalle de fluctuation, on obtient

$$a = \pi - \Phi^{-1}(1 - \alpha)\sqrt{\frac{\pi(1 - \pi)}{n}}$$

Avec $\alpha = 5\%$, $\Phi^{-1}(1-\alpha) = 1.64$ et on trouve $a = 0.942$. La proportion de comprimés conforme reste supérieur à a, on conclut donc, au risque 5%, que les réglages sont bons.

Remarque. Quand on prend une décision dans le test d'hypothèses ci dessus, on a deux façons de se tromper :

1. On décide que la machine est bien réglée alors qu'elle est en réalité mal réglée.
2. On décide que la machine est mal réglée alors qu'elle est en réalité bien réglée.

La première est appelée erreur de première espèce. Elle est définie par $P_{nF_n \sim B(n,n\pi)}(F_n < a)$. C'est cette erreur qu'on controle en fixant α.
La deuxième, l'erreur de deuxième espèce, n'est pas controlée car pour la calculer on a besoin de connaître la loi de nF_n sous l'hypothèse alternative, c'est à dire le vrai π' de la machine mal réglée...

1.3.3 Remarque sur le raisonnement statistique

Dans les cours de statistique sur l'intervalle de confiance ou les tests on trouve des phrases comme les suivantes :

Un intervalle de confiance pour une proportion p à un niveau de confiance $1-\alpha$ est la réalisation, à partir d'un échantillon, d'un intervalle aléatoire contenant la proportion p avec une probabilité supérieure ou égale à $1-\alpha$. Cet intervalle aléatoire est déterminé à partir de la variable aléatoire $F_n = \frac{X_n}{n}$ qui, à tout échantillon de taille n, associe la fréquence.

Il est intéressant de démontrer que, pour une valeur de p fixée, l'intervalle $[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}}]$ contient, pour n assez grand, la proportion p avec une probabilité au moins égale à 0,95.

On passe ensuite à des énoncés du type :

On fait un sondage auprès de 21521 personnes au sujet de leur préférences OUI ou NON. Sur ces 21521 personnes, 14268 disent préférer OUI. Donner un intervalle de confiance de niveau 0,95 pour la proportion de OUI dans la population totale.

La réponse attendue est : $[\frac{14268}{21521} - \frac{1}{\sqrt{21521}}, \frac{14268}{21521} + \frac{1}{\sqrt{21521}}]$. Mais alors que dans la phrase reprise plus haut $[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}}]$ désigne un intervalle aléatoire, ici $[\frac{14268}{21521} - \frac{1}{\sqrt{21521}}, \frac{14268}{21521} + \frac{1}{\sqrt{21521}}]$ n'est pas aléatoire. C'est une réalisation de l'intervalle précédent. De deux choses l'une, soit p appartient à l'intervalle, soit il n'y appartient pas.

Quel sens a le niveau 0,95 ? La probabilité que p appartienne à l'intervalle aléatoire $[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}}]$ est 0,95. Donc quand on calcule cet intervalle à chaque fois qu'on fait le sondage, si on répète le sondage, dans 95% des cas on obtiendra un intervalle contenant p. Quand on fait le sondage une fois, on se trompe ou non.

Quand on décide de considérer que p appartient à l'intervalle trouvé pour éventuellement

prendre certaines mesures on fait donc un pari. De quel type est ce pari ? Dans 5% des cas les intervalles qu'on me donne ne contiennent pas p. Dans 5% des cas en considérant que p est dedans je me tromperai.

Imaginons qu'on ait une urne contenant cent boules : des boules rouges et des boules blanches. On ignore combien. On sait seulement que deux compositions sont possibles : une boule blanche, quatre-vingt-dix-neuf boules rouges ou bien une boule rouge, quatre-vingt-dix-neuf boules blanches. Comme d'habitude il est impossible d'ouvrir l'urne etc... et on ne peut faire qu'un chose : tirer une boule dans cette urne une seule fois. Question posée : quelle est la composition de l'urne ? Vous pouvez faire plusieurs choses sans doute (tirer à pile ou face votre réponse par exemple). Y-a-t-il une façon de procéder plus maligne ? Tirer une boule dans l'urne et regarder sa couleur. Que faire de cette information ? Elle est insuffisante pour répondre à la question à coup sûr. Vous avez très bien pu tirer la boule rouge unique parmi quatre-vingt-dix-neuf boules blanches. Avez-vous intérêt à parier sur la composition 99 rouges, 1 blanche si vous tirez une boule rouge ? En le faisant vous pariez que vous n'avez pas assisté à l'événement peu fréquent tirer une boule rouge parmi 99 blanches. Évidemment si c'est ce qui s'est produit vous vous trompez. C'est le sens du pari fait avec l'intervalle de confiance.

Émile Borel : *Une seule loi fondamentale d'interprétation des probabilités (loi unique du hasard) : si une probabilité calculée est très petite, il faut se conduire comme si l'événement est impossible.*

Supposons maintenant que les boules soient numérotées de 1 à 100. Si un ami tire une boule je suis prêt à parier qu'il ne tombera pas sur la boule numéro 1. De la même façon je suis prêt à parier qu'il ne tombera pas sur la boule numéro 2. Quel que soit n entre 1 et 100, je parierai volontiers qu'il ne tombera pas sur la boule numéro n. Pourtant il tombera bien sur l'une d'entre elles. Beaucoup d'événements sont de petite probabilité. Pourquoi choisir "tirer une boule rouge parmi quatre-vingt-dix-neuf blanches" ?

Il faut ajouter des remarques.

Imaginons que les boules soient aussi numérotées. On nous dit que dans l'une des urnes 99 boules portent le numéro 1 et une le numéro 2, dans l'autre 99 boules portent le numéro 2 et une le numéro 1. Quelqu'un tire une boule et annonce que le numéro obtenu est 1. Je peux parier au niveau 0,99 que la boule a été tirée dans celle qui contient 99 boules 1. Mais cela ne donne pas d'information sur la question des couleurs si on ne me dit rien de plus. La règle de décision qu'on se fixe doit avoir un rapport avec le problème qui nous intéresse pour être utile.

Valéry (L'homme et la coquille) : *Tout le reste, tout ce que nous ne pouvons assigner ni à l'homme pensant, ni à cette Puissance génératrice, nous l'offrons au "hasard", ce qui est une invention de mot excellente. Il est très commode de disposer d'un nom qui permette d'exprimer qu'une chose remarquable (par elle-même ou par ses effets immédiats) est amenée tout comme une autre qui ne l'est pas. Mais dire qu'une chose est remarquable,*

c'est introduire un homme, une personne qui y soit particulièrement sensible, et c'est el le qui fournit tout le remarquable de l'affaire. Que m'importe, si je n'ai pas de billet de loterie, que tel ou tel numéro sorte de l'urne ? Je ne suis pas sensibilisé à cet événement. Il n'y a point de hasard pour moi dans le tirage, point de contraste entre le mode uniforme d'extraction de ces numéros et l'inégalité des conséquences. Ôtez donc l'homme et son attente, tout arrive indistinctement coquille ou caillou ; mais le hasard ne fait rien au monde, que de se faire remarquer...

Revenons à l'expérience avec les deux urnes sans numérotation des boules. Si vous répétez l'expérience en adoptant cette stratégie (pour différentes urnes rencontrées) vous vous tromperez dans environ 1% des cas (loi des grands nombres). Si vous adoptez le pile ou face, dans environ la moitié des cas. La stratégie "statistique" semble meilleure de ce point de vue. Cela ne signifie pas qu'il n'existe pas de meilleure stratégie : par exemple ouvrir l'urne dans le cas qui nous occupe ou bien moins exigeant, mais aussi efficace ici, obtenir le droit de tirer trois boules sans remise [5].

Vous pouvez aussi décider de ne rien décider dans de telles situations. Il est très probable alors que vous ne preniez que très rarement une décision quelconque ; nous ne connaissons personne qui n'agisse qu'à coup sûr. Il est facile d'imaginer en revanche des situations où la décision est inévitable alors que les informations disponibles sont partielles. C'est même la règle. Et l'abstention est une décision, une option qui a elle aussi ses conséquences.

Imaginez que l'on vous donne le droit de répéter l'expérience de tirer une boule avec remise pour décider de sa composition. Disons deux fois. Si vous tirer deux fois une boule rouge. Là encore il se peut que vous ayez tiré deux fois de suite une boule rouge parmi 99 blanches. Cela se produit une fois sur 10000. Vous pouvez refuser de parier [6]. Trois fois une boule rouge. Il est possible d'avoir tiré trois fois de suite une boule rouge parmi 99 blanches. Cela se produit une fois sur 1000000. Vous pouvez refuser de parier. Ne seriez-vous pas de plus en plus tenté de parier ?

Il faut maintenant insister sur une différence entre l'exemple de l'urne et l'intervalle de confiance. Si vous prenez une décision à partir du pari que p appartient à $[0,25; 0,29]$, vous pouvez vous tromper. Mais ce n'est sans doute pas la même erreur si p vaut 0,2901 ou bien s'il vaut 0,56. Dans l'exemple de l'urne il y a une seule façon de se tromper. Pas pour l'intervalle de confiance. La longueur de l'intervalle de confiance dépend du niveau de confiance.

Reprenons l'exemple du sondage auprès de 21521 personnes au sujet de leur préférences OUI ou NON. Un intervalle asymptotique au niveau $1-\alpha$ pour la proportion de OUI

5. Pourquoi trois ?

6. Si on tire une boule rouge et une blanche, on se retrouve bien embêté. Amusant : on accepte facilement de parier après un tirage, le tirage suivant peut nous faire hésiter.

dans la population totale est [7]

$$\left[\hat{p} - u_\alpha \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{21521}}, \hat{p} + u_\alpha \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{21521}}\right]$$

où $\hat{p} = \frac{14268}{21521}$ et u_α tel que $\Phi(u_\alpha) = 1 - \alpha/2$ (Φ fonction de répartition de la loi normale centrée réduite). On obtient différents intervalles pour différents niveaux :

$1-\alpha$	u_α	Intervalle de niveau $1-\alpha$
0,9	1,64	[0,658 ;0,668]
0,95	1,96	[0,657 ;0,669]
0,99	2,58	[0,655 ;0,671]
0,9999	3,89	[0,650 ;0,676]
0,99999	4,42	[0,649 ;0,677]

Parier que p se trouve dans l'intervalle $[0,658; 0,668]$ c'est parier à 1 contre 10. Parier qu'il se trouve dans $[0,649; 0,677]$ c'est parier à 1 contre 100000. On peut très bien avoir tort pour le premier intervalle mais raison pour le deuxième. Imaginons que vous ne vous intéressiez qu'à la réponse à la question : "p est-t-il supérieur à 0,5 ?". Alors si vous êtes dans l'intervalle $[0,649; 0,677]$ la réponse est oui. Vous pouvez parier à 1 contre 100000 que p est supérieur à 0,5. C'est le cas aussi bien sûr si p appartient à $[0,658; 0,668]$. La situation est donc plus compliquée que dans le cas de l'urne. Ce n'est pas parce que vous ne disposez que de l'intervalle à 90% que le sondage ne donne pas des informations à des niveaux de confiance supérieurs.

Imaginons qu'un intervalle de confiance à 95% soit $[0,49; 0,61]$. Le nombre 0,5 est dans cet intervalle. Il ne faut pas en conclure qu'on ne peut rien dire sur la probabilité que p soit supérieur à 0,5.

Pour les tests d'adéquation on utilise des intervalles de fluctuation à un niveau α. On choisit souvent un intervalle centré en la probabilité théorique p. Pourquoi ? Pourquoi centré ? Pourquoi un intervalle ? De nombreux autres choix fourniraient des événements de probabilité α. Prenons un exemple idiot. On cherche à savoir si une pièce est équilibrée. On la lance 10000 fois. On utilise le théorème limite central pour trouver un ensemble de la forme $[0; 0,5 - a] \cup [0,5 + a; 1]$ qui contienne la fréquence de "pile" obtenue avec probabilité 0,95. Et on adopte la règle de décision suivante : si la proportion de pile obtenue est comprise entre $0,5-a$ et $0,5+a$ on considère que la pièce n'est pas équilibrée. Bizarre, non ? Est-ce idiot ? Pourquoi ?

7. Nous ne discutons du problème de l'approximation normale faite ici qui n'est qu'asymtotiquement vérifiée.

1.4 Chaînes de Markov

1.4.1 Notations et vocabulaire

Ligne ou colonne.

En théorie il importe peu de convenir qu'un vecteur de probabilité soit ligne ou colonne. Les habitudes d'écriture dans les cours d'algèbre linéaire et de lecture des probabilités conditionnelles font penser qu'il est préférable de faire les choix suivants.

Il est plus naturel d'appeler p_{ij} la probabilité de passer de l'état i à l'état j :

$$p_{ij} = \mathbb{P}_{(X_k=i)}(X_{k+1} = j) = \mathbb{P}(X_{k+1} = j | X_k = i)$$

La matrice de transition est la matrice fabriquée avec ces probabilités $P = (p_{ij})$. Or traditionnellement le premier indice désigne la ligne : p_{ij} est le coefficient de P situé à la i-ème ligne et la j-ème colonne.

Comme pour tout i on a

$$1 = \sum_{j=1}^{n} \mathbb{P}_{(X_k=i)}(X_{k+1} = j) = \sum_{j=1}^{n} p_{ij}$$

ce sont les sommes des coefficients des lignes de P qui font 1.

Par ailleurs, la formule des probabilités totales donne

$$\begin{aligned} \mathbb{P}(X_{k+1} = j) &= \sum_{i=1}^{n} \mathbb{P}(X_{k+1} = j, X_k = i) \\ &= \sum_{i=1}^{n} \mathbb{P}(X_{k+1} = j | X_k = i)\mathbb{P}(X_k = i) \\ &= \sum_{i=1}^{n} \mathbb{P}(X_k = i)p_{ij} \end{aligned}$$

On retrouve ici les formules définissant le produit d'un vecteur par la matrice P à condition de mettre les probabilités $\mathbb{P}(X_k = i)$ en ligne. Définissons

$$\pi_k = (\mathbb{P}(X_k = 1), \mathbb{P}(X_k = 2), \ldots, \mathbb{P}(X_k = n))$$

Les égalités obtenues plus haut s'écrivent

$$\pi_{k+1} = \pi_k P.$$

La notion d'état et les deux sens qu'elle peut avoir.

Sens 1 : un système peut avoir plusieurs positions et peut évoluer au cours du temps en sautant avec certaines probabilités d'un état à un autre.

Sens 2 : évolution des parts de marchés. L'état est ici la répartition en part de marché donc l'équivalent des probabilités de présence à chaque état d'un client pris au hasard. La position d'un client suit l'évolution définie par la chaîne de Markov. Pour un tel client, les états sont les différents marchands à qui il choisit de s'adresser. Mais l'état du marché est le découpage en parts de marchés qui s'identifie avec les probabilités de présence d'un client aux différents états.

On a ainsi deux notions d'états : les différentes positions possibles des clients d'une part, le découpage en parts de marchés d'autre part. Si les probabilités de transition sont toutes positives la trajectoire d'un client donné passera par tous les états (plus ou moins rapidement suivant les valeurs des probabilités de passage), alors que les parts de marché se stabiliseront. Au sens 2 il y a convergence du système vers un état d'équilibre (pas au sens 1).

Convergence des lois dans le cas où toutes les probabilités de transitions sont positives

On veut montrer que les vecteurs π_k convergent lorsque k tend vers l'infini, c'est-à-dire que les suites définies par les coordonnées convergent. D'après la relation obtenue plus haut, il s'agit de voir que

$$\pi_k = \pi_0 P^k$$

converge. Nous allons voir que la suite de matrices (P^k) converge. Pour cela nous considérons la multiplication de P par les vecteurs colonne à droite.

Remarquons d'abord que le vecteur colonne composé uniquement de 1 est invariant (c'est une conséquence du fait que la somme des coefficients de P sur une ligne vaut 1) :

$$1 = \sum_{j=1}^{n} p_{ij}.1$$

Considérons maintenant un vecteur colonne V à coefficients positifs. On a

$$(PV)_i = \sum_{j=1}^{n} p_{ij}.V_j.$$

La somme sur j des p_{ij} valant 1, cela signifie que les coefficients de PV sont des barycentres des coefficients de V. Introduisons quelques notations :

$$\alpha = \min\{p_{ij} \ / \ i, j = 1 \dots n\},$$

et pour un vecteur colonne U

$$U^* = \max\{U_i \ / \ i = 1 \dots n\},$$

$$U_* = \min\{U_i \ / \ i = 1 \dots n\}.$$

Pour un certain i on a $(PV)_* = (PV)_i$. Pour cet i, on a alors

$$(PV)_* = (PV)_i = \sum_{j=1}^{n} p_{ij}.V_j.$$

Dans cette somme figure V^* multiplié par un nombre plus grand que α et tous les autres termes sont plus grands que (ou égaux à) V_*. On en déduit l'inégalité

$$(PV)_* \geq \alpha V^* + (1-\alpha)V_* > V_*.$$

Par ailleurs, pour un autre indice i' on a

$$(PV)^* = (PV)_{i'} = \sum_{j=1}^{n} p_{i'j}.V_j.$$

Dans cette somme figure V_* multiplié par un nombre supérieur à α. On obtient l'inégalité

$$(PV)^* \leq \alpha V_* + (1-\alpha)V^* < V^*.$$

Des deux inégalités précédentes on déduit

$$0 \leq (PV)^* - (PV)_* \leq \alpha V_* + (1-\alpha)V^* - \alpha V^* - (1-\alpha)V_* = (1-2\alpha)(V^* - V_*).$$

On obtient ainsi les propriétés suivantes sur les suites $((P^kV)^*)$ et $((P^kV)_*)$:
– $((P^kV)^*)$ est décroissante,
– $((P^kV)_*)$ est croissante,
– pour tout k, $((P^kV)_*) \leq ((P^kV)^*)$,
– $(P^kV)^* - (P^kV)_* \leq (1-2\alpha)^k(V^* - V_*)$, donc tend vers 0.
Cela entraîne que ces deux suites tendent vers une limite commune. Comme par définition, les coordonnées de P^kV sont entre $(P^kV)_*$ et $(P^kV)^*$, cela entraîne que toutes les coordonnées de P^kV tendent vers la même limite.

En prenant pour V successivement tous les vecteurs colonne constitués d'un seul 1 et des 0 (vecteurs de la base canonique) on voit ainsi que les vecteurs colonne de P^k convergent vers des vecteurs ayant toutes leurs coordonnées égales. Cela signifie que P^k converge vers une matrice de la forme

$$\begin{pmatrix} \rho_1 & \rho_2 & \cdots & \rho_n \\ \rho_1 & \rho_2 & \cdots & \rho_n \\ \vdots & \vdots & \cdots & \vdots \\ \rho_1 & \rho_2 & \cdots & \rho_n \end{pmatrix}.$$

Du coup la suite $(\pi_k = \pi_0 P^k)$ converge vers

$$\pi_0 \begin{pmatrix} \rho_1 & \rho_2 & \cdots & \rho_n \\ \rho_1 & \rho_2 & \cdots & \rho_n \\ \vdots & \vdots & \cdots & \vdots \\ \rho_1 & \rho_2 & \cdots & \rho_n \end{pmatrix} = \begin{pmatrix} \rho_1 & \rho_2 & \cdots & \rho_n \end{pmatrix},$$

et nous avons démontré le résultat souhaité.

Nous pouvons décrire ce qui se produit en utilisant la notion de valeur propre. Ce que nous avons montré est que la matrice P admet 1 pour valeur propre (associée au vecteur propre ayant tous ses coefficients égaux) et que toutes les autres valeurs propres sont de modules strictement inférieurs à 1. Quand on décompose l'espace en sous-espace propres (ou caractéristiques) et que l'on décompose un vecteur en utilisant cette décomposition la partie dans la direction du vecteur associé à la valeur propre 1 ne bouge pas, les autres coordonnées sont contractées.

1.5 Les probabilités et statistique au lycée

Première STI2D (B.o du 17 mars 2011) Même programme que celui des STMG , à quelques détails près : pas de diagramme en boîte pour les STI2D des commentaires parfois formulés différemment. Par exemple dans la partie statistiques descriptives, on doit privilégier l'étude d'exemples issus de résultats d'expériences, de la maîtrise statistique des procédés, du contrôle de qualité, de la fiabilité ou liés au développement durable. Ou encore dans la partie échantillonnage, on peut traiter quelques situations liées au contrôle en cours de fabrication ou à la réception d'une production.

PROGRAMMES DE PREMIERE EN PROBABILITES ET STATISTIQUES

NOTIONS		Série S (B.o du 30/09/2010)	Série ES (B.o du 30/09/2010)	Série STMG (B.o du 09/02/2012)
Statistiques descriptives **Analyse de données**	**Contenu**	Caractéristiques de dispersion : Variance, écart-type. Diagramme en boîte		Caractéristiques de dispersion : Écart-type, écart interquartile. Diagramme en boîte *(l'expression de l'écart-type n'est pas un attendu du programme)*
	Capacité	• Utiliser les 2 couples usuels (moyenne/écart-type et médiane/écart-interquartile) pour résumer une série statistique • Etudier une série ou mener une comparaison entre 2 séries à l'aide de la calculatrice ou d'un logiciel (tableur)		
Probabilités (1) **Variable aléatoire**	**Contenu**	Variable aléatoire discrète et loi de probabilités - Espérance **En S** : variance et écart-type		*Voir ci-dessous : uniquement dans le cadre d'un schéma de Bernoulli*
	Capacité	• Déterminer et exploiter la loi d'une variable aléatoire • Interpréter l'espérance comme valeur moyenne dans le cas d'un grand nombre de répétitions		
Probabilités (2) **Loi binomiale**	**Contenu**	Modèle de la répétition d'expériences identiques et indépendantes à 2 ou 3 issues. Epreuve de Bernoulli, loi de Bernoulli. Schéma de Bernoulli et loi binomiale. Coefficients binomiaux. Espérance de la loi binomiale. **En S** : Triangle de Pascal + Variance et écart-type de la loi binomiale		Schéma de Bernoulli Variable aléatoire associée au nombre de succès dans un schéma de Bernoulli Loi binomiale Espérance de la loi binomiale
	Capacité	• Représenter la répétition d'expériences identiques et indépendantes par un arbre pondéré • Utiliser cette représentation pour déterminer la loi d'une variable aléatoire • Reconnaître les situations relevant de la loi binomiale • Calculer une probabilité dans le cadre de la loi binomiale • Utiliser l'espérance d'une loi binomiale dans des contextes variés • **En S:** représenter graphiquement la loi binomiale		Représenter un schéma de bernoulli par un arbre pondéré et simuler un schéma de Bernoulli à l'aide d'un tableur et de la calculatrice. Connaître, utiliser les notations $P(X=k)$ $P(X<k)$ Reconnaître les situations relevant de la loi binomiale Calculer une probabilité dans le cadre de la loi binomiale à l'aide de la calculatrice ou du tableur. Représenter graphiquement la loi binomiale par un diagramme en bâtons. Déterminer l'espérance de la loi binomiale et l'interpréter comme valeur moyenne dans le cas d'un grand nombre de répétitions
Echantillonnage et prise de décision	**Contenu**	Utilisation de la loi binomiale pour une prise de décision à partir d'une fréquence		Intervalle de fluctuation d'une fréquence Prise de décision
	Capacité	Exploiter l'intervalle de fluctuation à un seuil donné, déterminé à l'aide de la loi binomiale, pour rejeter ou non une hypothèse sur une proportion		Déterminer à l'aide de la loi binomiale un intervalle de fluctuation, à environ 95% , d'une fréquence. Exploiter un tel intervalle pour rejeter ou non une hypothèse sur une proportion.

PROGRAMMES DE TERMINALE EN PROBABILITES ET STATISTIQUE

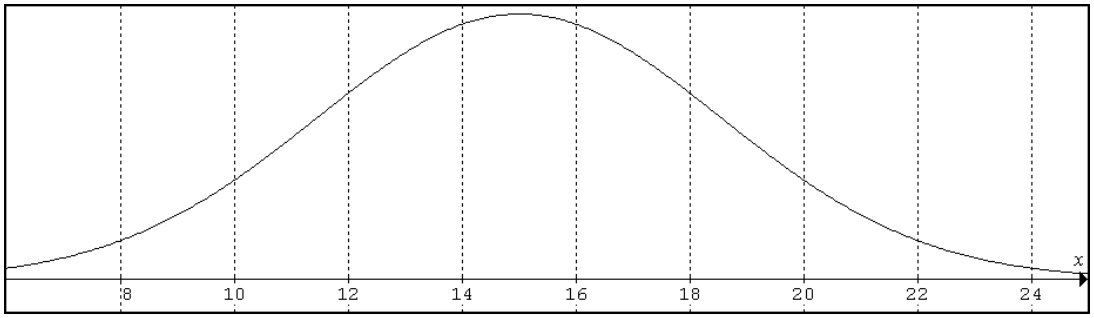
NOTIONS	Série S	Série ES	Série STMG
conditionnement	Formulations identiques pour contenu, capacités attendues et commentaires		
	Indépendance de deux événements.		
Statistiques à deux variables			Etude de séries de données statistiques quantitatives à deux variables. Nuage de points. Ajustement affine.
Lois à densité	Définition de loi à densité. Loi uniforme : fonction de densité et espérance.		
	Loi exponentielle et espérance		
Loi normale N(0,1)	Connaître la fonction de densité et sa représentation graphique		
	Théorème de Moivre Laplace. Montrer l'existence de u_α tel que $P(-u_\alpha < X < u_\alpha) = 1-\alpha$. Connaître les valeurs approchées $u_{0,05} \approx 1,96$ et $u_{0,01} \approx 2,58$. Calcul de l'espérance ; variance admise.	Connaître une valeur approchée de la probabilité de l'événement $-1,96 < X < 1,96$.	
Loi normale N(μ, σ^2) d'espérance μ et d'écart type σ	Utilisation de la calculatrice ou du tableur pour les calculs ; lien avec la loi normale N(0,1)	Utilisation de la calculatrice ou du tableur pour les calculs ; approche intuitive de l'espérance.	Utilisation de la calculatrice ou du tableur pour les calculs. (à partir de la loi binomiale ; allure de la courbe de densité et sa symétrie.
	Connaître une valeur approchée de la probabilité des événements suivants : $\{X \in [-\sigma\ ,\ +\sigma]\}$, $\{X \in [-2\sigma\ ,\ +2\sigma]\}$ et $\{X \in [-3\sigma\ ,+3\sigma]\}$.		
Intervalle de fluctuation	Démontrer $\lim\limits_{n \to +\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1-\alpha$ ... Connaître, pour *n* assez grand, l'intervalle de fluctuation asymptotique (*) au seuil de 95 % . Prise décision.	Connaître, pour *n* assez grand, l'intervalle de fluctuation asymptotique (*) au seuil de 95 % . Prise décision.	Connaître et interpréter graphiquement une valeur approchée de la probabilité de l'événement $\{X \in [\mu - 2\sigma\ ,\ \mu + 2\sigma]\}$ lorsque *X* suit la loi normale d'espérance μ et d'écart type σ.
Intervalle de fluctuation d'une fréquence			Connaître un intervalle de fluctuation à au moins 95 % d'une fréquence de taille *n* : $\left[p - \frac{1}{\sqrt{n}}\ ;\ p + \frac{1}{\sqrt{n}}\right]$; prise de décision.
Estimation Intervalle de confiance (*). Niveau de confiance.	• Estimer par intervalle une proportion inconnue à partir d'un échantillon. • Déterminer une taille d'échantillon suffisante pour obtenir, avec une précision donnée, une estimation d'une proportion au niveau de confiance 0,95.		
Estimation Intervalle de confiance d'une proportion			Estimer une proportion inconnue par l'intervalle $\left[f - \frac{1}{\sqrt{n}}\ ;\ f + \frac{1}{\sqrt{n}}\right]$ où *f* est la fréquence obtenue sur un échantillon de taille *n*.

Proposition de progression en terminale

Il est recommandé de ne pas reléguer les chapitres consacrés à l'étude des probabilités et de la statistique en fin d'année. Mais nous savons d'expérience que ceci est difficile à tenir et de plus nous avons besoin de l'année entière pour traiter le programme. Dans ces conditions, il faut bien finir par quelque-chose... En revanche il nous est apparu plus pertinent de nous interroger sur le découpage du programme de terminale. La difficulté qu'on a à réinvestir ces notions dans d'autres champs peut être compensée par leur étalement sur toute l'année. Nous sommes arrivés à cette proposition de progression.
1 . Loi uniforme et première approche de la loi normale.
2 . Probabilités conditionnelles et indépendance.
3 . Lois exponentielle et normale.
4 . Estimation.

La première partie, qui peut se faire dès septembre, pourrait se dérouler comme suit. La loi uniforme sur $[a, b]$ s'introduit en se basant sur l'intuition. On peut aussi la simuler sur tableur par la fonction $ALEA * (b - a) + a$. Elle se traduit ensuite en terme d'aire d'un rectangle.

A partir de cette interprétation par une aire, on peut glisser vers la loi normale en remplaçant la fonction constante sur $[a, b]$ par une fonction représentée par une courbe en cloche sur $\mathbb{R}$. On impose à cette courbe d'être symétrique par rapport à la droite verticale passant par le sommet et de délimiter un domaine dont l'aire est 1. On décide alors de s'intéresser à la variable aléatoire X définie par sa loi de probabilité : $\mathbb{P}(X \in [a, b])$ est l'aire sous la courbe entre a et b.



Diverses activités sont alors possibles dans lesquelles il s'agira de déterminer des probabilités par lecture graphique. Par exemple :

Si l'axe de symétrie est la droite d'équation $x = 15$ et si on sait que $\mathbb{P}(12 < X < 15) = 0,35$, on peut chercher $\mathbb{P}(X > 15)$; $\mathbb{P}(X < 12)$. Pour quelle valeur de t a-t-on $\mathbb{P}(X < t) = 0,85$? ... Enfin il semble utile d'étudier la fonction f définie sur $\mathbb{R}$ par $f(x) = exp(\frac{-x^2}{2})$ dans le cadre du chapitre consacré à la fonction exponentielle.

1.6 La question de la qualité des nouveaux programmes

Quelles notions faut-il enseigner aux élèves ? Les programmes le disent. Ils sont relativement précis mais laissent évidemment place à interprétation. Le volume des probabilités au lycée a beaucoup augmenté dans les programmes de 2010 (entrés en vigueur en terminale en septembre 2012). Cela a fait l'objet de polémiques [8] Voici quelques extraits de réactions critiques...

Un enseignement rigoureux de la statistique me semble poser un vrai problème sans un bagage mathématique suffisant. S'il s'agit juste de laisser entendre que fréquence et probabilité sont en gros la même chose, cela ne pose pas de problème. Par contre, si on veut quantifier la différence, ce que l'on est en droit d'attendre d'un traitement mathématique, on se heurte à de vraies difficultés. Je n'ai toujours pas réussi à comprendre ce que signifie exactement " la proportion p est élément de l'intervalle $[f - 1/\sqrt{n}, f + 1/\sqrt{n}]$ *avec un taux de confiance de plus de 95%, où f désigne la fréquence observée sur un échantillon de taille n " (cet énoncé semble être le but de tout l'enseignement mathématique en section scientifique au lycée). Je n'ai pas de mal à saisir ce que cela peut vouloir dire d'un point de vue intuitif, mais c'est la signification mathématique qui m'échappe (un comble !) : il me semble qu'en l'absence de renseignements concernant la distribution potentielle de p, cette phrase n'a aucun sens. Du coup, j'ai des doutes sur ce qu'on peut vraiment faire en classe préparatoire (surtout avec le niveau actuel de l'enseignement au lycée, en partie du fait de la présence d'un gros module de statistique), mais on pourrait se donner comme but de démontrer certains cas du théorème de la limite centrale sur lequel beaucoup de choses reposent (cela reste du domaine des probabilités mais peut préparer le terrain pour un enseignement en école d'ingénieurs). Il ne faut pas oublier que l'esprit des classes préparatoires est assez particulier et que les élèves sont déjà traumatisés par l'existence d'une infinité de pièges potentiels dans des questions qui n'en comporteraient pas dans un enseignement classique.* Pierre Colmez, Images des mathématiques.

Les cours s'arrêtent vendredi 06 Juin ; combien de classes de terminales S n'ont pas encore traité la partie fluctuation/ échantillonnage ? Un nombre non-négligeable donc il va falloir aller " à l'essentiel " comme on dit pudiquement. J'invite enfin à consulter sur le site de l'APMEP par exemple les sujets du bac S/ES cru 2014 (de moins en moins distincts d'ailleurs ; même les sujets de ES sont moins décevants). Tous les énoncés ne sont pas accessibles en Latex donc je ne peux pas faire un copié-collé des extraits des énoncés et c'est bien dommage car ils sont tellement pauvres et redondants. Voici la dernière question du petit exercice 1 de l'épreuve de Pondichéry. L'entreprise A annonce que le pourcentage de

8. On pourra consulter les sites suivants : http://lemonde-educ.blog.lemonde.fr/2011/04/26/terminale-s-deminents-mathematiciens-contestent-les-nouveaux-programmes/, http://images.math.cnrs.fr/Faut-il-arreter-d-enseigner-les-1321.html, http://images.math.cnrs.fr/Pourquoi-enseigner-les.html, http://www.irem.univ-mrs.fr/Quelques-exercices-d-analyse-en, http://images.math.cnrs.fr/Mathematiques-post-modernes.html

moteurs défectueux dans la production est égal à 1 %. Afin de vérifier cette affirmation 800 moteurs sont prélevés au hasard. On constate que 15 moteurs sont détectés défectueux. Le résultat de ce test remet-il en question l'annonce de l'entreprise A ? Justifier. On pourra s'aider d'un intervalle de fluctuation. On n'échappe pas aux lois normales, intervalles de fluctuation ou confiance mais c'est traité de manière complètement systématique et dépouillée. Karen Brandin, Images des mathématiques.

J'ai été présidente du groupe de travail chargé d'introduire vers les années 2000 de l'aléatoire à partir de la classe de seconde. Les nouveautés introduites n'avaient pas vocation à être scellées dans la pierre sous la forme donnée à cette époque et les programmes étaient bien sûr destinés à évoluer. Aujourd'hui, je constate qu'il y a vraiment problème : les débats sont peu ou prou toujours les mêmes. Depuis bientôt 15 ans ! Je n'ai bien sûr pas de solution toute faite, mais après avoir visité pas mal de classes, écouté les enseignants, je pense qu'on fait fausse route et que la statistique devrait être prise en charge par ceux qui ont à la fois les questions et la production de données. C'est-à-dire la physique (incertitude des mesures), la biologie, les SES. En mathématiques, on ferait des probabilités, de l'expérimentation numérique, (un peu de processus, on essaye de faire comprendre le théorème central limite). Evidemment, on va parler de l'incontournable interdisciplinarité. Est-ce que cela change les termes du débat ? Claudine Schwartz, Images des mathématiques.

En vérité, au niveau du lycée, il ne reste plus guère que la partie calculatoire de la géométrie. Alors, bien sûr, le calcul c'est important et ce livre en est une illustration permanente, mais ce n'est pas tout ! C'est mutiler la géométrie que de la réduire au calcul. Je ne peux que vouer aux gémonies les auteurs de ces programmes désastreux, mais cela relativise évidemment la portée pratique de ce que je vais proposer ci-dessous : pour q'un discours sur l'enseignement de la géométrie ait un intérêt, encore faut-il que cet enseignement subsiste. Daniel Perrin, dans la postface de son livre de géométrie.

Comme le dit Karen Blandin, le temps passé sur les notions difficiles d'intervalles de fluctuation et de confiance n'est la plupart du temps pas suffisant pour que ces notions soient assimilées. Il nous semble que les exercices proposés au baccalauréat encouragent un enseignement mécanique aboutissant à une résolution facile d'exercices stéréotypés. Le fait que les programmes soient à peu près les mêmes en terminale S et terminale STI2D ou STMG constitue sans doute un signal aux professeurs les encourageant à aller dans ce sens. En pratique le programme de terminale STMG permettra aux élèves de cette série de faire les exercices de baccalauréat de la série S. Ce n'est pas un problème en soi. Il n'y a rien de choquant à ce que l'on demande aux élèves de séries différentes des choses semblables sur un point particulier. Cependant la motivation et le niveau moyen (en mathématiques) des élèves de la série STMG étant très inférieurs à ceux des élèves de terminale S, on en déduit que les exigences mathématiques de cette partie du programme doivent être modestes, et qu'il est préférable de n'y consacrer que peu de temps.

Le choix de ces notions a sans doute été guidé par leur importance dans la vie sociale,

politique, économique,... et leur utilité pour le citoyen. C'est tout à fait compréhensible. On n'a pas choisi par exemple de reparler de probabilités et modélisations liées aux problèmes de dénombrement (sans doute pour éviter les problèmes et situations artificiels : cartes, boules,...). Cela aboutit à un appauvrissement conceptuel et scientifique de la filière scientifique : la géométrie a été sacrifiée, la combinatoire est évitée,... Il faut y ajouter la démathématisation de la physique...

Le raisonnement mathématique peut prendre de nouvelles formes ou s'appliquer à de nouveaux objets (autres que ceux des programmes du vingtième siècle). Mais il faut un contenu clair, consistant, intéressant, propice à l'apprentissage de la rigueur,... À notre avis, les nouveaux programmes constituent un recul de ce point de vue[9].

Citons encore Chevallard et Wozniak : « Le danger que fait courir la proximité "culturelle" (dans l'enseignement des mathématiques) de la notion de probabilité à l'existence d'un enseignement de la statistique comme science de la variabilité est liée à deux contraintes massives, l'une propre à la statistique, l'autre beaucoup plus large. La première est le refoulement de la variation, qui pousse à mettre en avant le "passage à la limite". »

D'autre part, ne serait-ce pas anachronique ? Il est aujourd'hui facile d'avoir des intervalles de fluctuations exacts pour la loi binomiale pour d'assez grands échantillons.

Enfin, et c'est très important, il faudrait préparer le corps enseignant aux nouvelles orientations que l'on souhaite donner aux programmes.

9. Ajoutons que, même si la nouveauté est bienvenue, certaines compétences classiques de calculs algébriques, de raisonnements logiques sont indispensables à un scientifique et qu'assurer la transmission de ces compétences classiques est un des objectifs majeurs au lycée.

2 Activités et thèmes d'étude

2.1 Statistique descriptive

2.1.1 Activités sur les communes

Le but de cette activité réalisée en seconde est de travailler sur un grand nombre de données, trouvées sur le site de l'Insee. Elle permet aussi de se familiariser avec (et d'entretenir ses connaissances sur) le tableur. Les données étaient fournies aux élèves. Cette activité peut être prolongée en première avec la comparaison des données de deux départements (diagrammes en boîte, écart type).

Un tableau contenant la liste des communes du Morbihan par ordre alphabétique et leurs nombres d'habitants est donné. On pose les questions suivantes :
1. Quel est le nombre de communes du Morbihan ?
2. Quelle est la commune la moins peuplée, la plus peuplée ?
3. En colonne D et E recopier les données de la plage A3 :B263 et les trier de la moins peuplée à la plus peuplée.
4. Calculer la population moyenne par commune.
5. Déterminer la valeur médiane du nombre d'habitants par commune et interpréter concrètement le nombre choisi.
6. Déterminer le premier quartile du nombre d'habitants par commune et interpréter le nombre trouvé.
7. Déterminer le troisième quartile du nombre d'habitants par commune et interpréter le nombre trouvé.
8. Créer le diagramme bâtons du nombre de personnes par communes pour les communes de moins de 1000 habitants.
9. Calculer la part(en colonnes H en % avec quatre décimales) en nombre d'habitants par rapport au total des habitants du département.
10. Calculer en colonne I les parts cumulées croissantes en %.
11. Interpréter le nombre 50% obtenu en colonne I.

2.1.2 L'espérance de vie

L'espérance de vie d'une population n'est pas calculée en faisant la moyenne des âges des personnes mourant dans l'année. On utilise une autre méthode de calcul apparemment plus compliquée. Commençons par décrire cette méthode. Notons p_i la proportion des personnes ayant l'âge i au premier janvier qui sont encore vivantes le 31 décembre.

L'espérance de vie d'une population est égale à

$$\sum_{k=0}^{130}\prod_{i=0}^{k} p_i.$$

Expliquons un peu.

D'abord le 130. Les humains vivent très rarement plus de 120 ans. Imaginons qu'en France le doyen âgé de 118 ans meure dans l'année. Alors $p_{118} = 0$, et tous les produits suivant sont nuls. La somme précédente vaut alors

$$\sum_{k=0}^{117}\prod_{i=0}^{k} p_i.$$

Il suffit qu'il existe un âge tel que toutes les personnes de cet âge meurent dans l'année pour que les produits soient nuls à partir d'un certain rang. Cela ne se produit pas à coup sûr. On peut imaginer par exemple que pour tous les âges représentés une personne au moins survive. Il faudrait alors dire ce que vaut p_i quand aucune personne d'âge i n'existe au premier janvier (p_i est alors une proportion de la quantité nulle). Si on prend 0 (pourquoi pas ?) alors par exemple comme personne en France n'a atteint l'âge de 130 ans $p_{131} = 0$ et on a

$$\sum_{k=0}^{\infty}\prod_{i=0}^{k} p_i = \sum_{k=0}^{130}\prod_{i=0}^{k} p_i.$$

On pourrait trouver que cela pose un problème. Par exemple si le doyen des français a trois ans de plus que celui le plus vieux qui le suit (par exemple le doyen a 121 et le suivant 118) alors comme $p_{119} = 0$ on ne prend pas en compte le doyen. Remarquons qu'on ne le prend pas en compte non plus si tous ceux qui avaient 118 ans meurent mais pas lui. Si on considère que $p_i = 1$ quand l'âge i n'est pas représenté alors on considère que toutes les personnes d'âge i atteignent l'âge $i + 1$; or si une année tous les âges représentés ont des survivants alors on considère que si un certain âge est atteint alors on vit indéfiniment. La somme

$$\sum_{k=0}^{\infty}\prod_{i=0}^{k} p_i$$

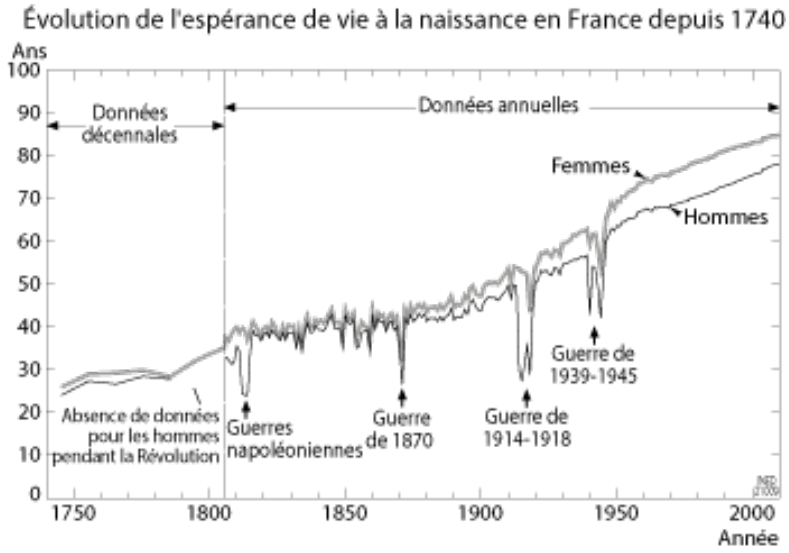
sera alors infinie. On considérera donc que $p_i = 0$ si l'âge n'est pas représenté. Cela fait qu'on ne tient pas compte des personnes dont l'âge dépasse 115 ou 120 ans. Elles sont si peu nombreuses que l'influence sur le résultat sera très faible quelle que soit la manière acceptable qu'on pourrait trouver de les prendre en compte.

Voici la définition que donne l'INSEE de l'espérance de vie : *L'espérance de vie à la naissance (ou à l'âge 0) représente la durée de vie moyenne - autrement dit l'âge moyen au décès - d'une génération fictive soumise aux conditions de mortalité de l'année. Elle caractérise la mortalité indépendamment de la structure par âge.*

On considère donc une population fictive (disons de 1000 personnes) et on lui applique à chaque âge les taux de mortalité de la population française constatée en 2012 par exemple. Au bout d'un certain temps il ne reste plus personne (les 1000 personnes sont mortes). On fait alors la moyenne des âges de décès constatés sur cette population fictive : on obtient l'espérance de vie à la naissance en France en 2012. Lien sur une animation créée par l'Ined : http://www.ined.fr/fr/tout_savoir_population/animations/esperance_vie/.

Pourquoi fait-on ça ? On obtient ainsi un résultat qui donne une moyenne si les taux de mortalité étaient les mêmes que ceux de l'année 2012. Cela permet de comparer les années : qu'une avancée médicale viennent modifier le taux de mortalité des nourrissons alors l'espérance de vie est modifiée, l'âge moyen de décès des morts de l'année beaucoup moins. D'autre part, comme affirmé dans la définition de l'Insee cela donne un résultat indépendant de la structure par âge. Imaginons par exemple qu'existe une classe démographique creuse dans une population. Alors quand cette classe atteindra un grand âge les décès de personnes très âgées seront sous représentés par rapport à celle de plus jeunes.

Quel est l'effet des guerres ? L'espérance chute brutalement pendant la période de guerre. Elle remonte presque aussi brusquement après la guerre.



Quel rapport entre ce qu'on trouve et la moyenne des âges de décès constatés dans l'année ? L'espérance de vie en 2012 donne la moyenne des âges de décès d'une population si tout se passe comme en 2012. Si tout se passait comme en 2012 pendant disons 250 ans et que chaque année le nombre d'enfants naissant était le même alors au bout d'un certain temps (disons 100 ans) les deux quantités coïncideraient. Remarquez que l'âge moyen des morts d'une année se comporte de manière similaire à l'espérance de vie pour les périodes de guerre. Pour tenir compte des guerres dans l'espérance de vie il faut faire une moyenne sur plusieurs années. On peut aussi calculer l'âge moyen de décès des personnes nées au cours d'une année donnée : par exemple personnes nées en 1895, personnes nées en 1943.

Pourquoi l'espérance de vie est-elle donnée par la formule

$$\sum_{k=0}^{\infty}\prod_{i=0}^{k} p_i ?$$

Notons X la variable aléatoire "durée de vie". On a

$$\mathbb{P}(X \geq 1) = p_0.$$

De même

$$\mathbb{P}(X \geq i+1 | X \geq i) = p_i.$$

Ceci donne

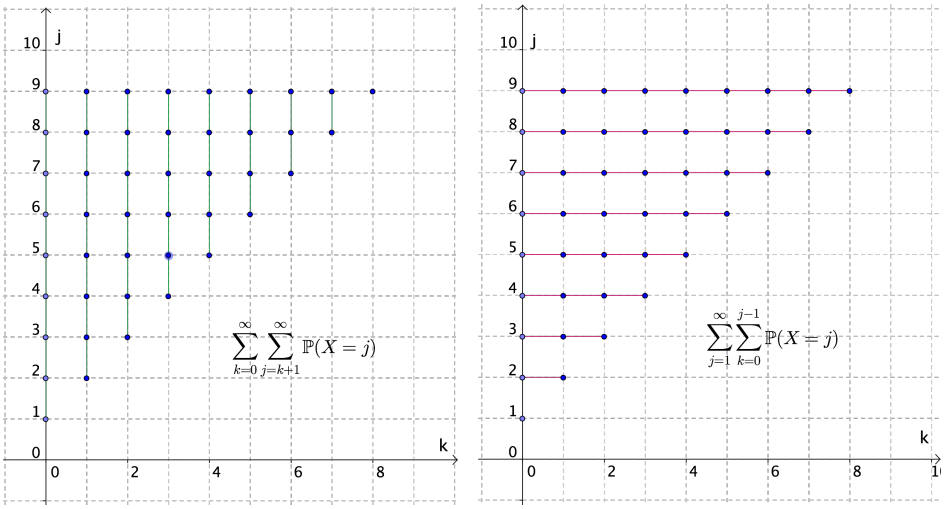
$$\mathbb{P}(X \geq k) = \mathbb{P}(X \geq k | X \geq k-1)\mathbb{P}(X \geq k-1) = p_{k-1}\mathbb{P}(X \geq k-1 | X \geq k-2)\mathbb{P}(X \geq k-2).$$

On obtient ainsi

$$\mathbb{P}(X \geq k) = \prod_{i=0}^{k-1} p_i.$$

Écrivons maintenant la somme

$$\begin{aligned}
\sum_{k=0}^{\infty}\prod_{i=0}^{k} p_i &= \sum_{k=0}^{\infty} \mathbb{P}(X \geq k+1) \\
&= \sum_{k=0}^{\infty}\sum_{j=k+1}^{\infty} \mathbb{P}(X = j) \\
&= \sum_{j=1}^{\infty}\sum_{k=0}^{j-1} \mathbb{P}(X = j) \\
&= \sum_{j=1}^{\infty} j\mathbb{P}(X = j) \\
&= \mathbb{E}(X)
\end{aligned}$$



2.1.3 Calculs d'indices

Exercice 1

Dans une unité non précisée, les revenus d'une population sont uniformément répartis entre 1 et 3. Autrement dit pour a et b deux nombres compris entre 1 et 3 tels que $a < b$, la proportion de la population dont les revenus sont compris entre a et b est donnée par

$$\int_a^b \frac{1}{2}\, dt.$$

Le but de l'exercice est de calculer l'indice de Gini correspondant.

1. Pour x compris entre 1 et 3, quelle est la proportion de la population dont les revenus sont inférieurs à x ?
2. Soit u un nombre compris entre 0 et 1. Exprimer en fonction de u le revenu $x(u)$ tel que la proportion de la population dont les revenus sont inférieurs à $x(u)$ est u.
3. Le revenu total de la population est donné par

$$N \int_1^3 \frac{t}{2}\, dt,$$

   où N est la taille de la population. Calculer le revenu total T de la population (en fonction de N).
4. Pour x compris entre 1 et 3, calculer la somme S des revenus perçus par la partie de la population dont les revenus sont inférieurs à x (en fonction de N et de x).
5. Pour u compris entre 0 et 1, exprimer en fonction de u la part des revenus perçue par la proportion u de la population ayant les revenus les plus faibles (autrement dit exprimer S en fonction de u plutôt que de x). On note $F(u)$ cette part divisée par T (autrement dit la proportion des revenus perçus par la proportion u de la population ayant les revenus les plus faibles). Le résultat à établir est : $F(u) = \frac{u(u+1)}{2}$.
6. Le graphe de la fonction F obtenue dans la question précédente (de $[0, 1]$ dans $[0, 1]$) est la courbe de Lorenz. Dessiner le graphe de F et calculer l'indice de Gini. On rappelle que l'indice de Gini est défini à partir d'une certaine aire qu'on pourra faire apparaître sur le dessin avant de la calculer.

Exercice 2

Tracer la courbe de Lorenz et calculer l'indice de Gini correspondant aux répartitions données dans le tableau suivant.

Déciles	Revenus mensuels perçus	Patrimoine (en euros)
1er	20	0
2e	516	1267
3e	968	7661
4e	1352	42968
5e	1658	88785
6e	1948	114937
7e	2293	150505
8e	2791	211167
9e	3786	384091

Les tableaux de l'Insee donnant les répartitions de revenus par déciles se présentent comme le précédent. Pour calculer l'indice de Gini il manque le dixième décile (donné par le plus haut revenu). Les répartitions des hauts revenus et des gros patrimoines étant très inégalitaires, il est préférable d'entrer dans les détails et de passer aux centiles et même à un découpage plus fin. Les données de l'exercice sont reprises de Landais, Piketti et Saez.

Le 95ème centile des revenus se situe à 5098, le 99ème centile à 10303, que 995 personnes sur 1000 gagne moins que 14104 euros mensuels, 999 sur mille moins de 30185 euros par mois.

Pour le patrimoine le 95ème centile est 710 311, le 99ème 1 934 824, 995 patrimoines sur 100 sont inférieurs à 3 180 679, 999 sur mille inférieurs à 9 161 107, 9995 sur dix mille inférieurs à 13 400 000, 9999 sur dix mille inférieurs à 32 900 000, 99 995 sur cent mille inférieurs à 54 500 000, 99 998 sur cent mille inférieurs à 123 000 000.

Pour tracer la courbe de Lorenz il faudra calculer les proportions de revenus perçus par l'ensemble des personnes dont les revenus sont inférieurs au premier décile, deuxième, etc... Pour ce faire, il faudra faire une hypothèse de répartition (raisonnable) des revenus entre deux déciles. Même chose pour le patrimoine.

Exercice 3

Ce qu'on appelle le développement d'un pays est fréquemment mesuré uniquement par le PIB ou le PNB par habitant. D'autres mesures existent qui permettent aussi d'évaluer le niveau d'un pays de façon différente : espérance de vie, taux de mortalité infantile, taux de survie à cinq ans... L'indice de développement humain (IDH) est un indice créé pour tenir compte de différents facteurs importants. On dit que l'IDH est un indice agrégé.

	Mesure	Valeur min.	Valeur max.
Longévité	Espérance de vie à la naissance	20 ans	83,5 ans
Éducation	Durée moyenne de scolarisation	0 an	13,1 ans
Éducation	Durée attendue de scolarisation	0 an	19 ans
Niveau de vie	PNB par hab.	100	108831

Dans chaque domaine on calcule l'indice en posant

$$Indice = \frac{valeur - valeur\ min.}{valeur\ max. - valeur\ min.}.$$

L'IDH est calculé de la façon suivante. On utilise la formule précédente pour les quatre critères pris en compte. Attention, pour l'indice de niveau de vie, on applique la formule non au PNB par habitant mais au logarithme népérien du PIB par habitant. On fait ensuite la moyenne géométrique des deux indices concernant l'éducation. On applique la formule à cette moyenne géométrique (avec l'indice maximal 0,994 et 0 pour l'indice minimal). On obtient ainsi trois indices. L'IDH est obtenu en prenant leur moyenne géométrique.

1. Quel est l'indice de longévité du pays ayant la plus grande espérance de vie à la naissance ? Quel est l'indice d'éducation le plus grand ?
2. Pourquoi prendre le logarithme des revenus ?
3. Pourquoi considérer des moyennes géométriques plutôt qu'arithmétiques ?
4. Au Portugal, l'espérance de vie à la naissance est 77,9 ans, la durée attendue de scolarisation 16 ans, la durée moyenne de scolarisation 7,7 ans, le PIB (PPA) par habitant 21558 dollars. Calculer l'IDH du Portugal.

Exercice 4 Proposer un calcul d'espérance de vie de la population suivante (fictive mais ressemblant à la France de 2014).

Âges	Taille de la population correspondante	Taux de mortalité (pour mille ; par an)
0-9	7 806 516	0,5
10-19	7 799 891	0,2
20-29	7 621 255	0,5
30-39	7 945 037	0,8
40-49	8 813 198	2
50-59	8 349 101	4,8
60-69	7 361 738	9,6
70-79	4 520 161	21
80-89	3 046 554	63,4
90-99	644 705	200
100-109	20 452	250

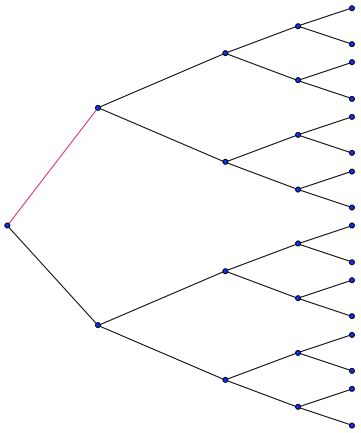
2.2 Probabilités

2.2.1 Manipulation des coefficients binomiaux sans les factorielles

Les coefficients binomiaux sont désormais introduits au lycée comme désignant des nombres de chemins dans des arbres. Les valeurs des coefficients binomiaux sont obtenues grâce

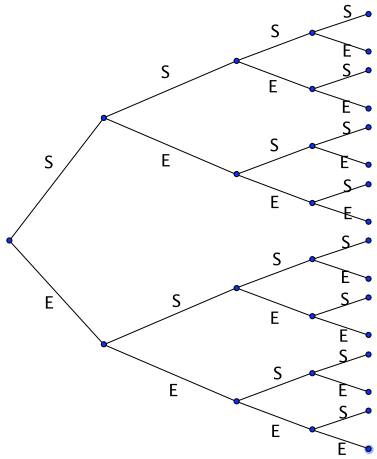
à la calculatrice ; les formules utilisant la notation factorielle ne sont plus d'actualité. Le message des lignes qui suivent est que ces choix n'entraînent pas nécessairement qu'il faille abandonner les raisonnements combinatoires, que la définition comme nombre de chemins dans un arbre permet de faire de tels raisonnements.

Les coefficients binomiaux sont définis comme nombre de chemins dans un arbre comme le suivant :



Le nombre $\binom{n}{k}$ est le nombre de chemins dans un tel arbre de longueur n dont k arêtes sont dirigées vers le haut.

Cette définition suffit à comprendre et presque démontrer la formule du binôme. Commençons par voir les choses de manière légèrement différente. Imaginons que que les branches de l'arbre dirigées vers le haut soient étiquetées par un S (comme succès) et les branches dirigées vers le bas par un E (comme échec).



Alors en suivant toutes les branches de l'arbre on obtient tous les mots de longueur n (n vaut 4 dans l'exemple considéré ci-dessus) écrits avec les deux lettres S et E. Le nombre

$\binom{n}{k}$ est le nombre de mots de longueur n écrits en utilisant k fois la lettre S et $n-k$ fois la lettre E.

Venons-en alors à la formule du binôme. Il s'agit de donner une expression de $(S+E)^n$ où E et S sont deux nombres réels. Regardons ce qui se passe pour de petites valeurs de n en développant sans prendre en compte la commutativité de la multiplication, ni l'écriture des puissances.

$$(S+E)^2 = (S+E)(S+E) = SS + SE + ES + EE$$

$$\begin{aligned}(S+E)^3 &= (S+E)(S+E)^2 = (S+E)(SS+SE+ES+EE)\\ &= SSS+SSE+SES+SEE+ESS+ESE+EES+EEE\end{aligned}$$

$$\begin{aligned}(S+E)^4 &= (S+E)(S+E)^3\\ &= (S+E)(SSS+SSE+SES+SEE+ESS+ESE+EES+EEE)\\ &= SSSS+SSSE+SSES+SSEE+SESS+SESE+SEES+SEEE\\ &+ ESSS+ESSE+ESES+ESEE+EESS+EESE+EEES+EEEE\end{aligned}$$

Que montrent ces développements effectués de manière idiote (pas si idiote) ? Ils montrent que $(S+E)^n$ est la somme des produits définis par tous les mots de longueur n que l'on peut écrire avec les deux lettres S et E. Mais lorsqu'on tient compte de la commutativité de la multiplication et qu'on utilise la notation puissance on voit que de nombreux mots donnent le même produit. Par exemple $ESEE$ et $EESS$ donnent tous les deux S^2E^2. À quelle condition un mot de longueur n donne-t-il un produit S^kE^{n-k} ? À condition qu'il comporte k S et $n-k$ E. Combien y a-t-il de mots de longueur n écrits avec k S et $n-k$ E ? Nous l'avons vu : $\binom{n}{k}$.

Conclusion :

$$(S+E)^n = \sum_{k=0}^{n} \binom{n}{k} S^k E^{n-k}.$$

C'est bien la formule du binôme.

Comment peut-on démontrer des formules telles que

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1},$$

$$n\binom{n-1}{k-1} = k\binom{n}{k},$$

en n'utilisant que cette définition ? Le jeu est donc d'éviter d'utiliser l'expression des coefficients binomiaux à partir des factorielles. Pour la première relation c'est assez facile. Si on veut le nombre de chemins dans l'arbre qui compte $k+1$ branches vers le haut,

on peut décomposer ces chemins en deux parties : ceux qui commencent par une branche vers le haut (il y en a $\binom{n}{k}$: le nombre de chemins comptant k branches vers le haut dans l'arbre obtenu en coupant la première branche vers le haut) et ceux qui commencent par une branche vers le bas (il y en a $\binom{n}{k+1}$: le nombre de chemins comptant $k+1$ branches vers le haut dans l'arbre obtenu en coupant la première branche vers le bas).

Une fois qu'on a la première relation, on peut obtenir la deuxième par récurrence. Allons-y. Soit à démontrer : pour tous $k \leq n$, $n\binom{n-1}{k-1} = k\binom{n}{k}$.

Hypothèse de récurrence $\mathcal{H}_n$: pour tout $1 \leq k \leq n$, pour $n\binom{n-1}{k-1} = k\binom{n}{k}$.

$\mathcal{H}_1$ est vraie : $1\binom{0}{0} = 1 = 1\binom{1}{1}$.

Supposons que $\mathcal{H}_j$ soit vraie pour tout $j \leq n$. Montrons qu'alors $\mathcal{H}_{n+1}$ est vraie. Soit k un entier $1 \leq k \leq n+1$. Si $k \geq 2$, on a :

$$\begin{aligned}
(n+1)\binom{n}{k-1} - k\binom{n+1}{k} &= (n+1)\left(\binom{n-1}{k-1} + \binom{n-1}{k-2}\right) - k\left(\binom{n}{k} + \binom{n}{k-1}\right) \\
&= \left(n\binom{n-1}{k-1} - k\binom{n}{k}\right) + \binom{n-1}{k-1} \\
&\quad + \left(n\binom{n-1}{k-2} - (k-1)\binom{n}{k-1}\right) \\
&\quad + \binom{n-1}{k-1} + \binom{n-1}{k-2} - \binom{n}{k-1} \\
&= 0 + 0 + 0,
\end{aligned}$$

les deux premiers 0 par hypothèse de récurrence, le troisième par la relation du triangle de Pascal. Si $k = 1$, on a :

$$(n+1)\binom{n}{0} - \binom{n+1}{1} = (n+1) - (n+1) = 0.$$

Bien. Mais cette démonstration ne donne pas d'explication.

Voici une autre façon de faire. On s'intéresse aux chemins dont on colorie une branche disons en rouge, les autres étant noires. On compte le nombre de chemins de longueur n dans l'arbre dont k branches sont vers le haut dont la branche rouge. Une fois que vous avez un chemin de longueur n avec k branches vers le haut, vous pouvez fabriquer k chemins colorés différents : on peut colorier en rouge la première des branches vers le haut, la deuxième..., la $k$ème. Le nombre des chemins coloriés est donc

$$k\binom{n}{k}.$$

Mais on peut compter ces chemins d'une autre façon. On commence par choisir quelle branche sera coloriée en rouge (et vers le haut) : on a n possibilités. Ce choix étant fait, il

y a toujours $\binom{n-1}{k-1}$ façons de compléter le chemin pour obtenir un chemin avec k branches vers le haut (dont une seule rouge). Le nombre de tels chemins est donc :

$$n\binom{n-1}{k-1}.$$

D'où l'égalité souhaitée.

Cette égalité permet de répondre à des exercices du type suivant. Selim et Chloé font partie d'une classe de trente-cinq élèves. Dans cette classe on tire au hasard sept élèves pour constituer un jury. Quelle est la probabilité que Sélim fasse partie du jury ? Que Chloe en soit membre mais pas Sélim ? L'ensemble des résultats possibles (que l'on supposera équiprobables) est l'ensemble des jurys possibles. Il y en a $\binom{35}{7}$. D'autre part il y a $\binom{34}{6}$ jurys dont Sélim est membre. La probabilité que Sélim fasse partie du jury est donc

$$\frac{\binom{34}{6}}{\binom{35}{7}} = \frac{7}{35},$$

(appliquer l'égalité précédente avec $k = 7$ et $n = 35$).

2.2.2 Justification du fait que le tri suivant ALEA() donne la probabilité uniforme sur les parties

On suppose que la répétition de la commande ALEA() du tableur fournit une réalisation de tirage au hasard de manière indépendante de nombres compris entre 0 et 1 pour la loi uniforme sur $[0, 1]$ [10]. La question posée est la suivante : supposons que l'on dispose de 125 objets (par exemple des noms de communes) disposés dans la colonne 1 d'un tableur ; dans la colonne 2 on tire la fonction ALEA() ; on trie ensuite suivant cette deuxième colonne et on prend les 50 premiers objets obtenus. En procédant de la sorte, tire-t-on au hasard suivant la loi uniforme, une partie à 50 éléments de l'ensemble des 125 objets ?

L'ensemble des parties à 50 éléments de nos 125 objets a $\binom{125}{50}$ éléments. Il s'agit de montrer que chaque partie a une probabilité de $\frac{1}{\binom{125}{50}}$ d'être tirée.

Regardons la probabilité que les 50 premiers objets soient sélectionnés. La deuxième colonne est une réalisation du tirage au sort de 125 variables aléatoires indépendantes de loi uniforme sur $[0, 1]$. Nous abandonnons 50 et 125 et les remplaçons par k et n ; ce sera plus court et plus général. Les k premiers objets sont sélectionnés par notre procédure si

$$\max(X_1, \ldots, X_k) \leq \min(X_{k+1}, \ldots, X_n).$$

10. C'est évidemment faux en toute rigueur puisque l'ordinateur a une précision limitée. La question de la fabrication d'une bonne fonction ALEA() est intéressante mais n'est pas discutée ici.

Quelle est la probabilité de cet événement ? Par hypothèse, on a :

$$\begin{aligned}
&\mathbb{P}(\max(X_1,\dots,X_k) \leq \min(X_{k+1},\dots,X_n)) \\
&\quad = \int_{[0,1]^n} \mathbf{1}_{\max(x_1,\dots,x_k)\leq\min(x_{k+1},\dots,x_n)}\ dx_1\dots dx_n \\
&\quad = \int_{[0,1]^{n-k}} \left(\int_{[0,1]^k} \mathbf{1}_{\max(x_1,\dots,x_k)\leq\min(x_{k+1},\dots,x_n)}\ dx_1\dots dx_k \right) dx_{k+1}\dots dx_n \\
&\quad = \int_{[0,1]^{n-k}} \min(x_{k+1},\dots,x_n)^k dx_{k+1}\dots dx_n.
\end{aligned}$$

On note C_i l'ensemble

$$C_i = \{(x_{k+1},\dots,x_n)\ /\ \forall j = k+1,\dots,n,\ x_j \leq x_i\}.$$

On a :

$$\begin{aligned}
&\mathbb{P}(\max(X_1,\dots,X_k) \leq \min(X_{k+1},\dots,X_n)) \\
&\quad = \sum_{i=k+1}^{n} \int_{C_i} \min(x_{k+1},\dots,x_n)^k dx_{k+1}\dots dx_n \\
&\quad = \sum_{i=k+1}^{n} \int_{C_i} x_i^k dx_{k+1}\dots dx_n \\
&\quad = \sum_{i=k+1}^{n} \int_0^1 x_i^k \left(\int_{[x_i,1]^{n-k}} dx_{k+1}\dots \hat{dx_i} dx_n \right) dx_i
\end{aligned}$$

le $\hat{dx_i}$ signifiant que dx_i n'apparaît pas dans la liste. Cela donne :

$$\begin{aligned}
&\mathbb{P}(\max(X_1,\dots,X_k) \leq \min(X_{k+1},\dots,X_n)) \\
&\quad = \sum_{i=k+1}^{n} \int_0^1 x_i^k (1-x_i)^{n-k-1} dx_i \\
&\quad = (n-k) \int_0^1 x^k (1-x)^{n-k-1} dx,
\end{aligned}$$

car les intégrales obtenues ont la même valeur pour tous les i. Reste donc à calculer ces intégrales. Cela peut se faire par parties ; on dérive la puissance de x et on intègre celle

de $(1-x)$ (par exemple).

$$\begin{aligned}
\int_0^1 x^k(1-x)^l dx &= \left[\frac{x^{k+1}}{k+1}(1-x)^n\right]_0^1 + \frac{l}{k+1}\int_0^1 x^{k+1}(1-x)^{l-1}dx \\
&= 0 + \frac{l}{k+1}\int_0^1 x^{k+1}(1-x)^{l-1}dx \\
&= \frac{l}{k+1}\frac{l-1}{k+2}\cdots\frac{2}{k+l-1}\int_0^1 x^{k+l-1}(1-x)dx \\
&= \frac{l}{k+1}\frac{l-1}{k+2}\cdots\frac{2}{k+l-1}\frac{1}{k+l}\frac{1}{k+l+1} \\
&= \frac{l!k!}{(k+l+1)!}
\end{aligned}$$

On obtient donc :

$$\begin{aligned}
&\mathbb{P}(\max(X_1,\ldots,X_k) \leq \min(X_{k+1},\ldots,X_n)) \\
&= (n-k)\frac{(n-k-1)!k!}{n!} = \frac{(n-k)!k!}{n!}
\end{aligned}$$

ce qui est bien

$$\mathbb{P}(\max(X_1,\ldots,X_k) \leq \min(X_{k+1},\ldots,X_n)) = \frac{1}{\binom{n}{k}}$$

Autre façon de voir les choses : on range les nombres $x_1,\ldots,x_n$. Quelle est la probabilité que les k premiers restent aux m premières places ? Réponse :

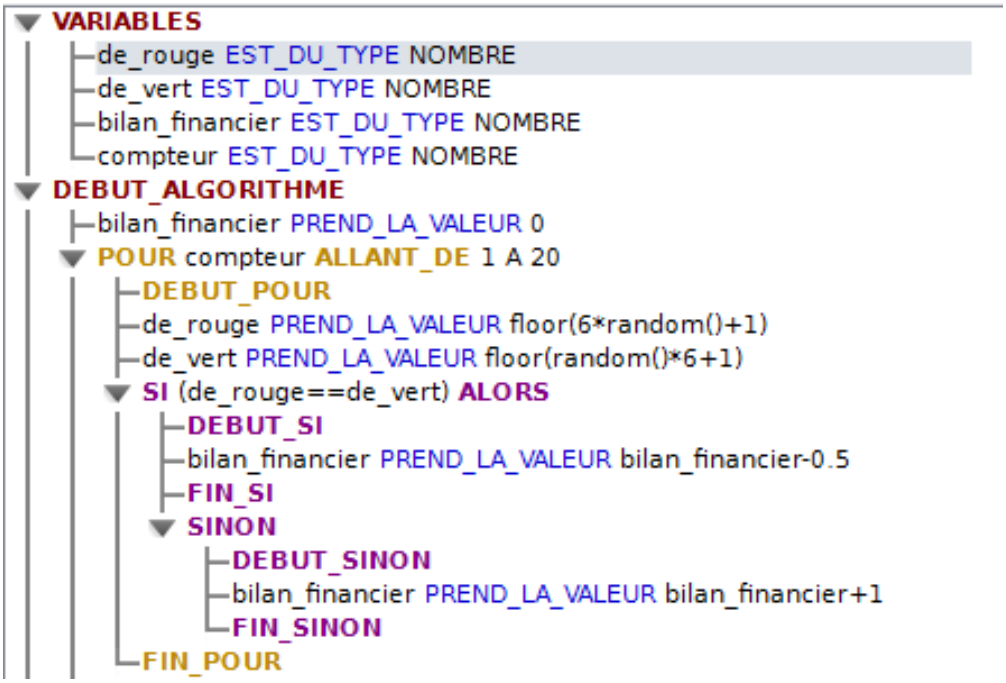
$$\frac{(n-k)!k!}{n!}.$$

2.2.3 Jeu et loi des grands nombres

Un jeu de dés : on lance deux dés cubiques numérotés de 1 à 6. Le joueur gagne 0,5 ? si les deux numéros sont identiques. Sinon, il verse 1 ? à l'organisateur du jeu. Pensez-vous que l'organisateur du jeu gagnera assez d'argent ? Combien peut-il espérer gagner en moyenne par partie quand un grand nombre de joueurs participe ?

Pour répondre à ces questions, nous allons simuler à l'ordinateur le lancer d'un dé, simuler le lancer de deux dés 20 fois puis 1200 fois et calculer le bilan financier de l'organisateur lorsque 20 personnes puis 1200 participent.

Avec Algobox, on propose de simuler un lancer de deux dés. Avec Algobox, une boucle "pour" permet de répéter un traitement à plusieurs reprises, le nombre de répétitions étant connu à l'avance. Une boucle « POUR compteur ALLANT DE 1 à 20 »a été utilisée dans l'algorithme suivant :



1. Comprendre que la formule "floor(random()*6+1)" simule un lancer de dé et donne de façon aléatoire soit le nombre 1, soit le 2, le 3, le 4,le 5 ou le 6.
2. Taper puis faire fonctionner 20 fois l'algorithme. Recueillir les résultats dans un tableau. Calculer alors le gain moyen du joueur par jeu, lors de ces 20 parties successives. Combien l'organisateur du jeu aura-t-il gagné en tout, lors de ces 20 jeux ?
3. Améliorer l'algorithme pour que l'utilisateur puisse entrer le nombre de parties jouées (déclarer une variable N et noter lire N au début de l'algorithme). Le faire fonctionner à 10 reprises avec le même effectif total : 1200 joueurs et noter les 10 valeurs obtenues du bilan financier. Calculer alors la moyenne de ces 10 bilans financiers. En déduire le bilan financier moyen par partie que peut espérer gagner l'organisateur quand un grand nombre de joueurs participe. Déposer ce second algorithme sur les abeilles Pédagogiques.
4. Combien peut-il espérer gagner en moyenne par partie quand un grand nombre de joueurs participe ?

2.2.4 Le pari de Pascal

Extrait du fragment 397 des pensées. Suivez l'argumentation de Pascal en gardant en tête la notion d'espérance mathématique associée à un jeu de pile ou face ou de dé.

Examinons donc ce point, et disons : "Dieu est, ou il n'est pas." Mais de quel côté pencherons-nous ? La raison n'y peut rien déterminer : il y a un chaos infini qui nous sépare. Il se joue un jeu, à l'extrémité de cette distance infinie, où il arrivera croix ou pile. Que gagerez-vous ? Par raison, vous ne pouvez faire ni l'un ni l'autre ; par raison, vous ne pouvez défendre nul des deux. Ne blâmez donc pas de fausseté ceux qui ont pris

un choix ; car vous n'en savez rien.

- "Non ; mais je les blâmerai d'avoir fait, non ce choix, mais un choix ; car, encore que celui qui prend croix et l'autre soient en pareille faute, ils sont tous deux en faute : le juste est de ne point parier."

- Oui ; mais il faut parier. Cela n'est pas volontaire, vous êtes embarqué. Lequel prendrez-vous donc ? Voyons. Puisqu'il faut choisir, voyons ce qui vous intéresse le moins. Vous avez deux choses à perdre : le vrai et le bien, et deux choses à engager : votre raison et votre volonté, votre connaissance et votre béatitude ; et votre nature a deux choses à fuir : l'erreur et la misère. Votre raison n'est pas plus blessée, en choisissant l'un que l'autre, puisqu'il faut nécessairement choisir. Voilà un point vidé. Mais votre béatitude ? Pesons le gain et la perte, en prenant croix que Dieu est. Estimons ces deux cas : si vous gagnez, vous gagnez tout ; si vous perdez, vous ne perdez rien. Gagez donc qu'il est, sans hésiter.

- "Cela est admirable. Oui, il faut gager ; mais je gage peut-être trop."

- Voyons. puisqu'il y a pareil hasard de gain et de perte, si vous n'aviez qu'à gagner deux vies pour une, vous pourriez encore gagner ; mais s'il y en avait trois à gagner, il faudrait encore jouer (puisque vous êtes dans la nécessité de jouer), et vous seriez imprudent, lorsque vous êtes forcé de jouer, de ne pas hasarder votre vie pour en gagner trois, à un jeu où il y a pareil hasard de perte et de gain. Mais il y a une éternité de vie et de bonheur. Et cela étant, quand il aurait une infinité de hasards, dont un seul serait pour vous, vous auriez encore raison de gager un pour avoir deux ; et vous agiriez de mauvais sens, en étant obligé à jouer, de refuser de jouer une vie contre trois à un jeu où d'une infinité de hasards il y en a un pour vous, s'il y avait une infinité de vie infiniment heureuse à gagner Mais il y a ici une infinité de vie infiniment heureuse à gagner, un hasard de gain contre un nombre fini de hasards de perte, et ce que vous jouez est fini. Cela ôte tout parti : partout où est l'infini, et où il n'y a pas infinité de hasards de perte contre celui du gain, il n'y a point à balancer, il faut tout donner. Et ainsi, quand on est forcé à jouer, il faut renoncer à la raison pour garder la vie, plutôt que de la hasarder pour le gain infini aussi prêt à arriver que la perte du néant. Car il ne sert de rien de dire qu'il est incertain si on gagnera, et qu'il est certain qu'on hasarde, et que l'infinie distance qui est entre la certitude de ce qu'on s'expose, et l'incertitude de ce qu'on gagnera, égale le bien fini, qu'on expose certainement, à l'infini, qui est incertain. Cela n'est pas ; aussi tout joueur hasarde avec certitude pour gagner avec incertitude ; et néanmoins il hasarde certainement le fini pour gagner incertainement le fini, sans pécher contre la raison. Il n'y a pas infinité de distance entre cette certitude de ce qu'on s'expose et l'incertitude du gain ; cela est faux. Il y a, à la vérité, infinité entre la certitude de gagner et la certitude de perdre. Mais l'incertitude de gagner est proportionnée à la certitude de ce qu'on hasarde, selon la proportion des hasards de gain et de perte. Et de là vient que, s'il y a autant de hasards d'un côté que de l'autre, le parti est à jouer égal contre égal ; et alors la certitude de ce qu'on s'expose est égale à l'incertitude du gain : tant s'en faut qu'elle en soit infiniment distante. Et ainsi,

notre proposition est dans une force infinie, quand il y a le fini à hasarder à un jeu où il y a pareils hasards de gain que de perte, et l'infini à gagner. Cela est démonstratif ; et si les hommes sont capables de quelque vérité, celle-là l'est.

L'intérêt de jouer quand l'espérance est positive est lié à la possibilité de répéter le jeu. Cet intérêt est-il le même lorsque la répétition est impossible ?

Prenons un exemple.

Supposez que vous et votre famille disposiez d'un revenu annuel de 50 000 euros après impôts (selon l'INSEE le revenu moyen en France en 2007 avant impôts par ménage est de 39 000 euros environ, 54 000 pour les familles ayant deux enfants) et qu'aucune autre source de revenu ne soit imaginable. Et que je vous propose le jeu suivant. Vous misez 500 000 euros. Avec probabilité 1/100 je vous donne 101 fois votre mise, avec probabilité 99/100 vous perdez 500 000 euros.

L'espérance du gain est-elle positive ? Jouez-vous ?

Si je vous faisais crédit pendant autant de parties que vous voulez vous auriez peut-être intérêt à jouer. Mais je ne fais pas crédit. Vous devez déposez l'argent sur mon compte et je n'accepte de jouer qu'une fois. Il vous faut emprunter l'argent. Disons que vous l'empruntez à 3%. Faites le calcul (utilisant les suites géométriques) des sommes que vous devrez rembourser par mois pendant 40 ans si vous perdez (et pendant 20 ans). Jouez-vous ?

2.2.5 L'expérience de Luria et Delbrück

Une citation de Lamarck :

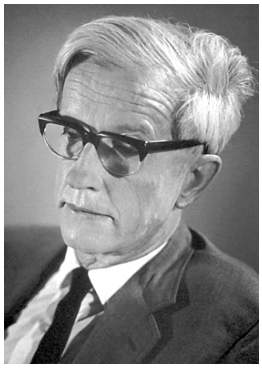
Relativement aux habitudes, il est curieux d'en observer le produit dans la forme particulière et la taille de la girafe (camelo-pardalis) : on sait que cet animal, le plus grand des mammifères, habite l'intérieur de l'Afrique, et qu'il vit dans des lieux où la terre, presque toujours aride et sans herbage, l'oblige de brouter le feuillage des arbres, et de s'efforcer continuellement d'y atteindre. Il est résulté de cette habitude, soutenue, depuis longtemps, dans tous les individus de sa race, que ses jambes de devant sont devenues plus longues que celles de derrière, et que son col s'est tellement allongé, que la girafe, sans se dresser sur les jambes de derrière, élève sa tête et atteint à six mètres de hauteur (près de vingt pieds).

Lamarck est souvent opposé à Darwin. Les lois de l'hérédité de Mendel n'étaient connues ni de l'un ni de l'autre. Comment fonctionne l'évolution ?

Lamarck écrivit "adaptation" au lieu de "sélection". Comme le dit Benzecri il faut sans doute se garder d'exagérer les différences entre Lamarck et Darwin et de voir (avec nos yeux d'aujourd'hui) le premier comme le vaincu, le second comme le héros d'un combat scientifique reconstruit de façon précisément à fabriquer un héros. Nous associerons quand

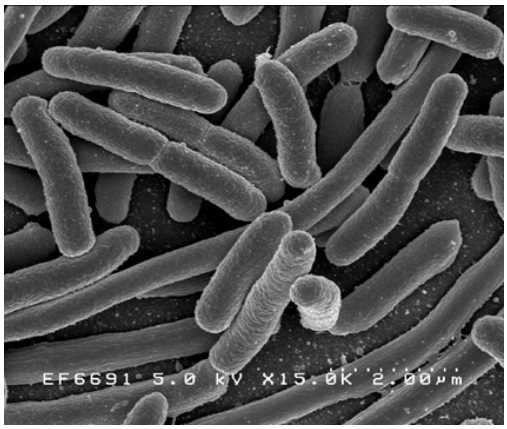
même Lamarck à l'idée d'adaptation et Darwin à la sélection (pariant que cette opinion courante traduit une certaine vérité). Certaines des idées de Lamarck pouvaient donner lieu à une vérification expérimentale : elles se sont révélées (essentiellement mais pas complètement) incorrectes ([4]). Est-il possible de *réfuter* l'idée d'adaptation de Lamarck et de montrer que des mutations bénéfiques ont lieu *avant* que se produise le changement du milieu dans lequel elles se révèlent bénéfiques ? Vous ne pouvez le faire en observant les restes fossilisés d'animaux éteints.

En 1942, Max Delbrück et Salvador Luria montrent que la résistance manifestée par des bactéries aux infections par les virus n'est pas due à une adaptation des bactéries mais à des mutations aléatoires. Ce travail, publié en 1943, peut être considéré " comme le tournant décisif dans l'approche du vivant "[23]. Delbrück et Luria apportent non seulement la preuve statistique de l'existence de gènes chez les bactéries mais ils proposent aussi une méthode de calcul de leur taux de mutation. L'importance de ce travail tient aussi à l'utilisation faite des mathématiques comme outil de prédiction quantitative.



Le prix Nobel de médecine a été attribué à Delbrück et Luria (et à Hershey, dont nous ne dirons rien) en 1969 pour leurs découvertes concernant le mécanisme de division et la structure génétique des virus.

Escherichia coli ou colibacille est une bactérie très présente dans les intestins des mammifères (elle compose 80% de notre flore intestinale). Elle est très utilisée par les biologistes pour des expériences. C'est un hôte commun du microbiote intestinal (anciennement appelé microflore commensale intestinale) de l'Homme et des animaux homéothermes. Son établissement dans le tractus digestif s'effectue durant les premières heures ou journées qui suivent l'accouchement. E. coli constitue alors tout au long de la vie de l'hôte l'espèce bactérienne dominante de la flore aérobie facultative intestinale. E. coli est sans doute l'organisme vivant le plus étudié à ce jour : en effet, l'ancienneté de sa découverte et sa culture aisée (division cellulaire toutes les 20 minutes à 37 degré Celsius dans un milieu riche) en font un outil d'étude de choix. Elle est en général détruite par le bactériophage T1.



On soumet au bactériophage une colonie d'un million de bactéries. En général, toutes les bactérie sont détruites. Dans 2% des cas il reste des bactéries vivantes. Ces bactéries survivent car elles n'ont pas les mêmes gènes que les bactéries normales. Une mutation leur a apporté une propriété P qui protège du bactériophage.

Les générations successives de bactéries sont obtenues par division cellulaire (c'est une reproduction asexuée). Ces divisions ne se produisent pas toutes aux mêmes instants. Le plus simple pour décrire la croissance de la taille d'une population de bactéries est d'adopter un modèle continu (la taille de la population est exprimée par une formule de type $C\exp(\gamma t)$). C'est ce formalisme qu'ont utilisé Luria et Delbrück. Nous allons supposer ici que les divisions donnant les générations successives de bactéries se produisent toutes aux mêmes instants.

Par ailleurs nous ferons aussi l'hypothése que la mutation vers la propriété P se produit uniquement au moment d'une division cellulaire : une bactérie normale donnant naissance à une bactérie normale et une bactérie P [11]

Peut-on imaginer des expériences permettant de trancher la question "Les mutations surviennent-elles au hasard ou sont-elles provoquées par les circonstances ?" ?

L'expérience de Salvador Luria et Max Delbrück (1943) a été inventée pour tester deux hypothèses concurrentes (mutations adaptatives ou aléatoires) pour les bactéries.

Pour obtenir un million de bactéries à partie d'une seule il faut donc 20 générations. On peut mettre en culture des bactéries pendant un temps correspondant à 20 générations et soumettre les cultures obtenues au bactériophage. Savoir s'il reste des bactéries est facile. Il suffit de laisser la culture et d'observer si une population de bactéries se développe ou pas.

11. Les hypothèses que nous faisons ne sont pas toutes réalistes. Elles peuvent être discutées, comme leur éloignement de la réalité. Nous souhaitons faire comprendre le sens, le principe de l'expérience, l'argument central sur lequel elle est basée. De ce point de vue dire que les divisions se produisent aux mêmes instants est irréaliste, mais ne change rien au cœur de l'argument. En pratique, de nombreuses discussions, adaptations, modifications techniques peuvent et doivent être apportées.

Nous opposons schématiquement deux hypothèses.

Une hypothèse que nous dirons darwinienne : les mutations peuvent se produire de manière aléatoire à chaque division cellulaire avec une certaine probabilité α (à la première génération aussi bien qu'à la dernière).

Une hypothèse que nous dirons lamarckienne : les mutations se produisent en réponse à l'agression par le bactériophage (à la vingtième division cellulaire) avec une probabilité β.

Dans les deux cas nous supposerons que les mutations se produisent de manière indépendante.

Éxaminons le cas de l'hypothèse darwinienne. Le nombre total de mutations est

$$1 + 2 + 4 + 8 + \ldots + 2^{19} = 2^{20} - 1.$$

Admettons que le gène P soit la seule raison qui puisse assurer la survie d'une bactérie. La probabilité qu'aucune bactérie ne survive est alors égale à la probabilité qu'aucune mutation vers P n'ait lieu. Comme on observe qu'il ne reste des survivantes que dans 2% des cas nous obtenons l'égalité

$$(1 - \alpha)^{2^{20}-1} = 1 - 0,02.$$

Cela donne

$$(2^{20} - 1)\ln(1 - \alpha) = \ln(0,98)$$

d'où on tire une valeur approchée de α

$$\alpha \simeq -\frac{\ln(0,98)}{(2^{20} - 1)} \simeq 2.10^{-8}.$$

Sous l'hypothèse lamarckienne les mutations peuvent se produire au moment de la vingtième génération (quand on introduit le bactériophage[12]). Le nombre de divisions concernées est 2^{20}. Un raisonnement analogue donne alors

$$(1 - \beta)^{2^{20}} = 0,98$$

d'où l'on tire

$$\beta \simeq -\frac{\ln(0,98)}{(2^{20})} \simeq 2.10^{-8},$$

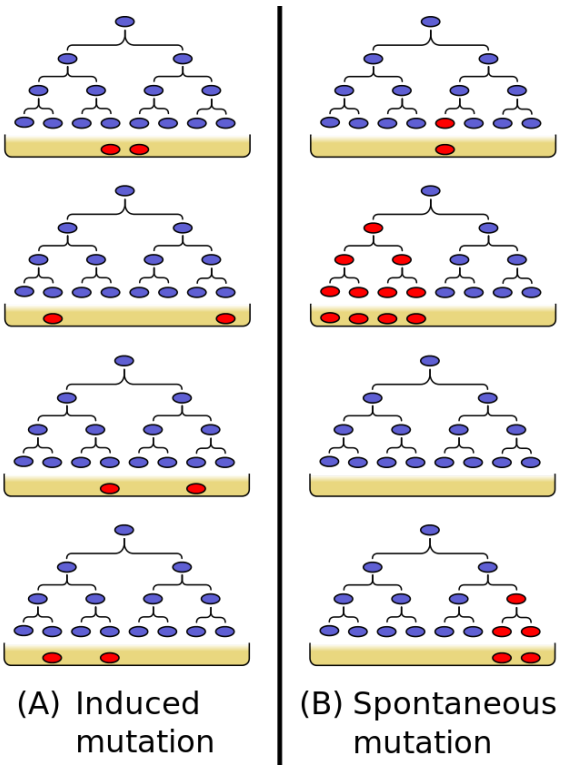
comme dans l'autre cas.

12. C'est un inconvénient de notre modèle : on peut se poser des questions sur l'instant exact d'introduction du bactériophage, des conséquences que cela peut avoir sur la division cellulaire... Certaines de ses questions ne se posent pas quand on adopte un modèle en temps continu. De ce point de vue notre simplification apparaît plutôt comme une complication.

Dans les deux cas nous ajustons la probabilité de mutation (α ou β) pour que la probabilité d'avoir des survivantes soit conforme à l'observation (en recommençant un grand nombre de fois l'expérience on constate qu'il y a des survivants dans 2 cas sur 100).

Peut-on faire une différence entre les deux hypothèses ? Les deux peuvent expliquer de façons différentes le fait que dans 2% des cas on ait des bactéries survivantes. Quelle différence peut-on espérer entre les deux mécanismes imaginés ?

L'article wikipédia consacré à l'expérience de Luria et Delbrück affirme : « Si la résistance au virus chez les bactéries est causée par une réponse de la bactérie, alors les nombres de colonies résistantes devraient être assez proches dans chaque boîte. Or ce n'est pas ce qu'ont observé Luria and Delbrück. En fait, les nombres de colonies résistantes varient de manière importante. » La dispersion du nombre de bactéries survivantes doit être plus grande sous l'hypothèse darwinienne. Cette affirmation est accompagnée du schéma suivant qui illustre les différences de dispersion devant être observées.



Nous allons voir les choses de manière légèrement différente. La probabilité qu'il y ait plus d'une bactérie survivante sous l'hypothèse lamarckienne est beaucoup plus faible que sous l'hypothèse darwinienne.

Le cas de l'hypothèse lamarckienne est le plus simple : le nombre de bactéries survivantes suit la loi binomiale de paramètre $(\beta, 2^{20})$. Appelons S ce nombre. On a

$$\mathbb{P}(S = k) = \binom{2^{20}}{k} \beta^k (1-\beta)^{2^{20}-k}.$$

Cette expression peut contenir de grands nombres (par exemple $\binom{2^{20}}{45}$)et de petits nombres (par exemple β^{45}) difficiles à manipuler. Nous souhaitons avoir un ordre de grandeur de

$$\mathbb{P}(S>1)=\sum_{k=2}^{2^{20}}\binom{2^{20}}{k}\beta^k(1-\beta)^{2^{20}-k}.$$

Il est possible d'avoir une idée de la taille des nombres $\binom{2^{20}}{k}\beta^k(1-\beta)^{2^{20}-k}$ lorsque k est de l'ordre de quelques unités. Pour $k=0$, on a $\mathbb{P}(S=0)=0,98$ (nous avons ajusté β pour que ce soit le cas). Pour $k=1$, on a

$$\mathbb{P}(S=1)=\binom{2^{20}}{1}\beta^1(1-\beta)^{2^{20}-1}\simeq 2^{20}.2.10^{-8}(1-\beta)^{2^{20}-1}\simeq 0,02.(1-\beta)^{2^{20}-1}.$$

On a vu que $(1-\beta)^{2^{20}}$ vaut 0,98, on obtient

$$\mathbb{P}(S=1)=\binom{2^{20}}{1}\beta^1(1-\beta)^{2^{20}-1}\simeq 2^{20}.2.10^{-8}(1-\beta)^{2^{20}-1}\simeq 0,02.0,98.(1-\beta)^{-1},$$

ce qui donne

$$\mathbb{P}(S=1)\simeq 0,02.0,98=0,0196.$$

Remarquons qu'on a

$$\mathbb{P}(S=0)+\mathbb{P}(S=1)=0,9996$$

soit presque 1. On obtient ainsi que

$$\mathbb{P}_{S>0}(S>1)\leq\frac{4.10^{-4}}{0,02}=0,02,$$

autrement dit, quand il y a des bactéries survivantes, dans plus de 98 cas sur 100 il n'y en a qu'une.

Faisons maintenant quelques calculs sous l'hypothèse darwinienne. La probabilité qu'aucune mutation n'ait lieu avant la dernière division (la dix-neuvième) est :

$$(1-\alpha)^{2^{19}-1}\simeq(1-\alpha)^{2^{19}}=(1-\alpha)^{2^{20}/2}=\sqrt{(1-\alpha)^{2^{20}}}=\sqrt{0,98}\simeq 0,99,$$

car le nombre de divisions avant la dernière est $1+2+4+8+\ldots+2^{18}=2^{19}-1$. De la même façon la probabilité qu'aucune mutation n'ait lieu avant l'avant dernière division est :

$$(1-\alpha)^{2^{18}-1}\simeq(1-\alpha)^{2^{18}}=(1-\alpha)^{2^{20}/4}=\left((1-\alpha)^{2^{20}}\right)^{1/4}=0,98^{1/4}\simeq 0,995.$$

ou avant la dix-septième

$$(1-\alpha)^{2^{17}-1}\simeq(1-\alpha)^{2^{17}}=(1-\alpha)^{2^{20}/8}=\left((1-\alpha)^{2^{20}}\right)^{1/8}=0,98^{1/8}\simeq 0,9975.$$

On en déduit que la probabilité qu'une mutation se produise lors de la dernière division est environ 0,01, lors de l'avant dernière 0,005, lors de la dix-septième 0,0025. Mais si une division se produit lors de la dix-septième division alors (le gène P se transmettant) le nombre de bactéries survivantes est au moins quatre. Sous l'hypothèse darwinienne on a donc

$$\mathbb{P}_{S>0}(S > 3) \simeq \frac{0,0025}{0,02} = 0,125.$$

Autrement dit si l'hypothèse lamarckienne est vraie alors on n'observera très difficilement un nombre de survivant supérieur à 1, alors que ce sera relativement fréquent si l'hypothèse darwinienne est vraie [13].

Ce qui se produit est conforme à l'hypothèse darwinienne. Comment compter de petit nombre de bactéries ? Après avoir appliqué le bactériophage, diluer le substrat obtenu puis l'étaler sur un milieu nutritif ; on observera autant de taches (colonies de bactéries issues des survivantes par divisions) qu'il y avait de survivantes suite à l'attaque du bactériophage.

2.3 Statistique inférentielle

2.3.1 Analyse de quelques exercices de baccalauréat

Année 2013

Dans tous les sujets, un exercice de probabilité est à traiter avec des lois continues et/ou la loi binomiale sauf en métropole où seule la loi binomiale apparaît. Pour la loi normale, les calculs de probabilité se font surtout à partir d'extraits de la table de la fonction de répartition de la loi normale utilisée dans le texte ou de la loi centrée et réduite après avoir incité l'élève à l'utiliser. Sauf dans le sujet de Liban 2013, où une table de valeurs de $\mathbb{P}(-\beta \leq Z \leq \beta)$ est donnée.

Amérique du Nord mai 2013. *« Une boulangerie industrielle utilise une machine pour fabriquer des pains de campagne pesant en moyenne 400 grammes. Pour être vendus aux clients, ces pains doivent peser au moins 385 grammes. Un pain dont la masse est strictement inférieure à 385 grammes est un pain non commercialisable, un pain dont lamasse est supérieure ou égale à 385 grammes est commercialisable. La masse d'un pain fabriqué par la machine peut être modélisée par une variable aléatoire X suivant la loi normale d'espérance $\mu = 400$ et d'écart-type $\sigma = 11$. Les probabilités seront arrondies au millième le plus proche.*
Partie A. On pourra utiliser le tableau suivant dans lequel les valeurs sont arrondies au

13. Là encore notre modèle discret en arbre est gênant. Dans ce cas les nombres de bactéries survivantes possibles sont des puissances de 2 (la probabilité d'occurrence de deux mutations est très faible). Les instants de division se décalant, ce n'est pas ce qu'on observe. Les calculs que nous avons faits donnent une bonne idée de ce qui se produit.

millième le plus proche

x	380	385	390	395	400	405	410	415	420
$P(X \leqslant x)$	0,035	0,086	0,182	0,325	0,5	0,675	0,818	0,914	0,965

1. Calculer $\mathbb{P}(390 \leq X \leq 410)$ *(...) »*

Pondichéry avril 2013. *« (...) On désigne par X la variable aléatoire qui donne le nombre de salariés malades une semaine donnée.*
b. On admet que l'on peut approcher la loi de la variable aléatoire par la loi normale centrée réduite c'est-à-dire de paramètres 0 et 1. On note Z une variable aléatoire suivant la loi normale centrée réduite. Le tableau suivant donne les probabilités de l'évènement $Z < x$ *pour quelques valeurs du nombre réel* x*.*

x	−1,55	−1,24	−0,93	−0,62	−0,31	0,00	0,31	0,62	0,93	1,24	1,55
$P(Z < x)$	0,061	0,108	0,177	0,268	0,379	0,500	0,621	0,732	0,823	0,892	0,939

Calculer, au moyen de l'approximation proposée en question b., une valeur approchée à 10^{-2} *près de la probabilité de l'évènement : « le nombre de salariés absents dans l'entreprise au cours d'une semaine donnée est supérieur ou égal à 7 et inférieur ou égal à 15 ».*

Liban mai 2013. *« L'entreprise Fructidoux fabrique des compotes qu'elle conditionne en petits pots de 50 grammes. Elle souhaite leur attribuer la dénomination « compote allégée ». La législation impose alors que la teneur en sucre, c'est-à-dire la proportion de sucre dans la compote, soit comprise entre 0,16 et 0,18. On dit dans ce cas que le petit pot de compote est conforme.*
Partie B 1. On note X la variable aléatoire qui, à un petit pot pris au hasard dans la production de la chaîne F1, associe sa teneur en sucre. On suppose que X suit la loi normale d'espérance $m_1 = 0,17$ *et d'écart-type* $\sigma_1 = 0,006$*. Dans la suite, on pourra utiliser le tableau ci-dessous.*

α	β	$P(\alpha \leqslant X \leqslant \beta)$
0,13	0,15	0,000 4
0,14	0,16	0,047 8
0,15	0,17	0,499 6
0,16	0,18	0,904 4
0,17	0,19	0,499 6
0,18	0,20	0,047 8
0,19	0,21	0,000 4

Donner une valeur approchée à 10^{-4} *près de la probabilité qu'un petit pot prélevé au hasard dans la production de la chaîne F1 soit conforme (...) »*

Pour l'utilisation d'un intervalle de fluctuation ou d'un intervalle de confiance (un seul sujet concerné : Antilles Guyane juin 2013), une indication est donnée. Peu d'initiative à prendre pour le candidat !

Centres Etrangers juin 2013. *« (...) Partie C L'industriel affirme que seulement 2 % des vannes qu'il fabrique sont défectueuses. On suppose que cette affirmation est vraie, et l'on note F la variable aléatoire égale à la fréquence de vannes défectueuses dans un échantillon aléatoire de 400 vannes prises dans la production totale.*
1. Déterminer l'intervalle I de fluctuation asymptotique au seuil de 95 % de la variable F.
2. On choisit 400 vannes au hasard dans la production, On assimile ce choix à un tirage aléatoire de 400 vannes, avec remise, dans la production. Parmi ces 400 vannes, 10 sont défectueuses. Au vu de ce résultat peut-on remettre en cause, au seuil de 95 %, l'affirmation de l'industriel ? (...) »

Année 2014

Dans tous les sujets (sauf Antilles Guyane septembre 2014) les candidats doivent traiter un exercice de probabilité et sont testés sur la loi normale. Les tables ne sont plus fournies pour la loi normale, les calculs de probabilité et la détermination de sigma peuvent se faire à l'aide de la calculatrice ou à l'aide des valeurs remarquables pour les intervalles à 1,2, 3 sigmas. On trouve tout de même des questions où seul l'usage de la calculatrice est possible.

Liban 2014. *« (...) Lorsqu'il utilise le vélo, on modélise son temps de parcours, exprimé en minutes, entre son domicile et son lycée par une variable aléatoire T qui suit le loi normale d'espérance $\mu = 17$ et d'écart-type $\sigma = 1,2$. 1. Déterminer la probabilité que l'élève mette entre 15 et 20 minutes pour se rendre à son lycée. (...) »*

Que teste-t-on ? L'usage d'une calculatrice ?

Les questions portant sur l'intervalle de fluctuation ou de confiance (un seul sujet Amérique du nord mai 2014) reste guidées.

Amérique du Nord mai 2014. *« (...) Partie B : Campagne publicitaire Une association de consommateurs décide d'estimer la proportion de personnes satisfaites par l'utilisation de cette crème. Elle réalise un sondage parmi les personnes utilisant ce produit. Sur 140 personnes interrogées, 99 se déclarent satisfaites. Estimer, par intervalle de confiance au seuil de 95 %, la proportion de personnes satisfaites parmi les utilisateurs de la crème. (...) »*

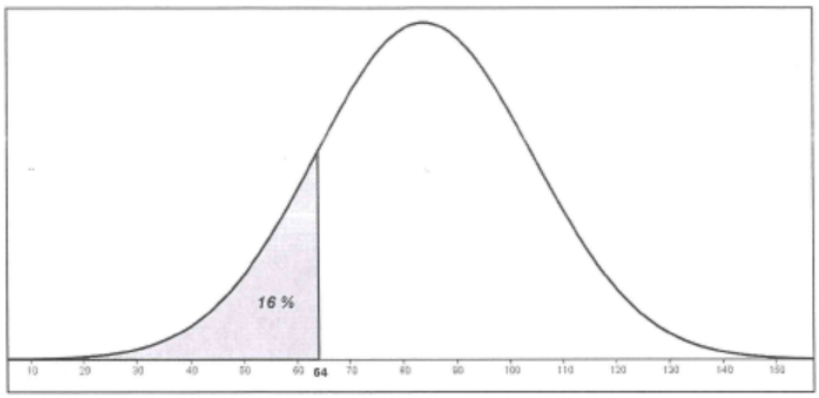
Métropole 2014 était aussi dérangeant quant à la formulation portant sur l'hypothèse à valider ou à rejeter. *« (...) 2. La chaîne de production a été réglée dans le but d'obtenir au moins 97 % de comprimés conformes. Afin d'évaluer l'efficacité des réglages, on effectue un contrôle en prélevant un échantillon de 1 000 comprimés dans la production. La taille de la production est supposée suffisamment grande pour que ce prélèvement puisse*

être assimilé à 1 000 tirages successifs avec remise. Le contrôle effectué a permis de dénombrer 53 comprimés non conformes sur l'échantillon prélevé. Ce contrôle remet-il en question les réglages faits par le laboratoire ? On pourra utiliser un intervalle de fluctuation asymptotique au seuil de 95 %. (...) »

Pour utiliser l'intervalle de fluctuation, on pose ici comme hypothèse que la proportion de comprimés conformes est $p = 0,97$...alors que l'énoncé dit "obtenir au moins 97%". L'intervalle de fluctuation vu en Terminale S permet seulement de "tester" une hypothèse d'égalité.

Année 2015

Pondichéry 2015. Ce 1er sujet sorti contient un exercice de probabilité conséquent puisque sur 6 points. Sa première partie commence par une approximation du calcul d'une probabilité par un calcul d'aire et la détermination du sigma. Par la suite, la détermination de sigma se précise à l'aide de la loi centrée réduite. *« (...) EXERCICE 3 (6 points, commun à tous les candidats)*
Partie A Étude de la durée de vie d'un appareil électroménager. Des études statistiques ont permis de modéliser la durée de vie, en mois, d'un type de lave-vaisselle par une variable aléatoire X suivant une loi normale $\mathcal{N}(\mu, \sigma^2)$ de moyenne $\mu = 84$ et d'écart-type σ. De plus, on a $\mathbb{P}(X \leq 64) = 0,16$. La représentation graphique de la fonction densité de probabilité de X est donnée ci-dessous.



1. a. En exploitant le graphique, déterminer $\mathbb{P}(64 \leq X \leq 104)$.
b. Quelle valeur approchée entière de ? peut-on proposer ?
2. On note Z la variable aléatoire dé ?nie par $Z = \frac{X-84}{\sigma}$. a. Quelle est la loi de probabilité suivie par Z ?
b. Justifier que $\mathbb{P}(X \leq 64) = \mathbb{P}(Z \leq \frac{-20}{\sigma})$. c. En déduire la valeur de σ, arrondie à 10^{-3}.
3. Dans cette question, on considère que $\sigma = 20,1$.
Les probabilités demandées seront arrondies à 10^{-3} près.
a. Calculer la probabilité que la durée de vie du lave-vaisselle soit comprise entre 2 et 5 ans.
b. Calculer la probabilité que le lave-vaisselle ait une durée de vie supérieure à 10 ans... »

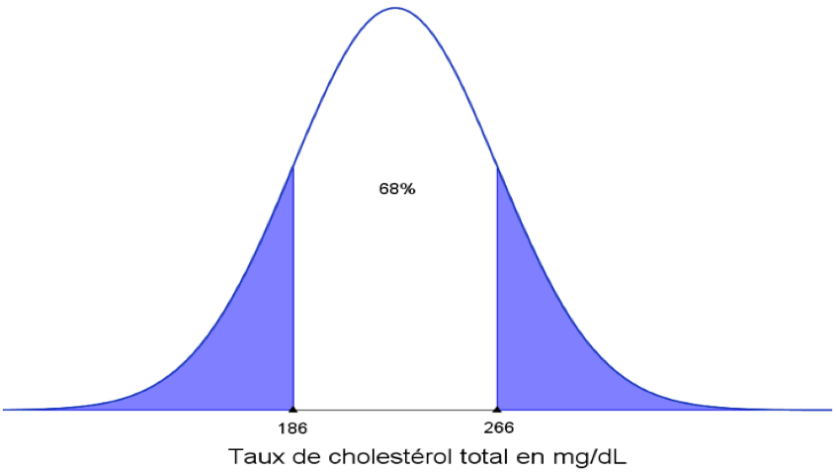
Liban 2015. L'utilisation d'un intervalle de confiance n'est pas précisée. *« (...) 4. L'institut de sondage publie alors les résultats suivants : "52,9% des électeurs voteraient pour le candidat A (estimation fondée sur un sondage d'un échantillon représentatif de 1200 personnes)." Au seuil de confiance de 95 %, le candidat A peut- il croire en sa victoire ? (...) »*

Amérique du Nord 2015. Pas d'intervalle de fluctuation ou d'intervalle de confiance. La loi normale est abordée avec un calcul simple de probabilité à la calculatrice et détermination de sigma avec la touche Normalfrac... ou Invnorm sur calculatrice.

Commentaires : Les exercices de probabilités, statistiques donnés au bac évoluent au fil des années et prennent une place importante dans les sujets. Dommage qu'une part importante des calculs soient à faire à la calculatrice avec certains menus spécifiques des calculatrices (Calculatrices qu'ils n'utiliseront plus dans l'enseignement supérieur). L'élève ou le candidat s'y perd en termes de manipulation de la calculatrice et perd le sens des raisonnements attendus.

Le document ressources de septembre 2014 (Exercices de mathématiques classes de Terminale S, ES, STI2D, STMG) disponible sur Eduscol indique les objectifs recherchés par les exercices proposés. « La version "évaluation avec prise d'initiative", tout en conservant une entrée progressive dans le sujet, permet de mesurer et valoriser la part de créativité et d'autonomie des élèves, compétences indispensables pour une bonne poursuite d'études et une évolution aisée dans la vie professionnelle. (...) Montrer comment il est possible, en respectant le format des sujets de baccalauréat, de concevoir des exercices plus ouverts, nécessitant une prise d'initiative de la part des élèves. »

Un bel exemple est l'exercice proposé p.10 du document ressources. *« Version "évaluation avec prise d'initiative".*
Dans un magazine médical, un article fournit le graphique suivant donnant la répartition du taux de cholestérol total dans une population comportant un grand nombre d'individus. Dans ce graphique, les parties coloriées ont la même aire.



Un médecin lit cet article et souhaite se faire une opinion sur ce graphique. Pour cela, il examine une partie des dossiers médicaux de ses nombreux patients ayant eu une prise de sang dont le résultat faisait apparaître leur taux de cholestérol total. Sur 1 000 dossiers consultés, il constate que 285 patients ont dû subir un examen complémentaire suite à leur dernière prise de sang car ils présentaient un taux de cholestérol total supérieur à 250 mg/dL, considéré comme trop élevé. On modélise le taux de cholestérol total d'une personne, exprimé en mg/dL, par une variable aléatoire X qui suit une loi de probabilité dont la fonction de densité est donnée ci-dessus.
1. Quelle loi est-il raisonnable de faire suivre à X ?
2. Vérifier que la probabilité qu'une personne choisie au hasard ait à subir un examen complémentaire suite à une prise de sang à cause de son taux total de cholestérol trop élevé est égale à 0,27 à 10^{-2}près.
3. Compte tenu des observations que le médecin a réalisées sur son échantillon de 1000 patients, a-t-il des raisons de remettre en cause le graphique publié dans le magazine ? Justifier la réponse... »

L'exercice proposé à Pondichéry cette année en Avril 2015 va bien dans ce sens et est en adéquation avec la volonté d'une entrée progressive dans le sujet.

Conclusion : Comme le dit chaque préambule des nouveaux programmes : *« L'enseignement des mathématiques au collège et au lycée a pour but de donner à chaque élève la culture mathématique indispensable pour sa vie de citoyen et les bases nécessaires à son projet de poursuite d'études. (...) L'apprentissage des mathématiques cultive des compétences qui facilitent une formation tout au long de la vie et aident à mieux appréhender une société en évolution. Au-delà du cadre scolaire, il s'inscrit dans une perspective de formation de l'individu. »* Même si pour beaucoup des candidats, le jour de l'épreuve de mathématiques du baccalauréat sera le dernier où ils côtoieront la loi normale, il ne faut pas que son évaluation se limite à l'application de formules (recettes) et à la simple utilisation de la calculatrice, on doit surtout vérifier qu'elle a bien contribué au développement des compétences visées dans notre enseignement : mettre en œuvre une recherche de façon autonome ; mener des raisonnements ; avoir une attitude critique vis-à-vis des résultats obtenus. La parution du document ressource de septembre 2014 va sûrement y contribuer.

2.3.2 Activité sur les populations de communes

On veut évaluer par sondage la taille de la population du Morbihan. Comment faire ? Le sondage 1 est le suivant : on tire 50 communes au hasard, on fait la somme des populations obtenues et on estime la population du Morbihan en multipliant cette somme par 261/50 (261 est le nombre de communes dans le Morbihan). Ceci peut se faire avec un tableur de la façon suivante : on prend le tableau contenant les 261 communes et leurs populations en colonne ; on ajoute une colonne aléatoire fabriquée avec ALEA() puis on trie suivant cette colonne ; les 50 premières communes triées forment notre échantillon. Nous l'avons

fait 12 fois. Nous avons obtenu les résultats suivants :

Essai	1	2	3	4	5	6
Estimation	669580	614023	741804	955823	512917	620648

Essai	7	8	9	10	11	12
Estimation	909746	571621	887984	923037	551639	524808

Les résultats obtenus sont très dispersés.

On propose le sondage 2 : on prend en compte les communes de plus de 10000 habitants, puis on tire au hasard 50 communes parmi les autres. On estime la population en prenant : la somme des communes de plus de 10000 habitants à laquelle on ajoute la somme des populations des 50 sélectionnées multipliée par 252/50. Nous l'avons fait 12 fois. Nous avons obtenu les résultats suivants :

Essai	1	2	3	4	5	6
Estimation	680046	697550	915762	709000	735063	706219

Essai	7	8	9	10	11	12
Estimation	707010	782988	673938	761896	740224	670218

Les résultats obtenus sont nettement moins dispersés. Par ailleurs le nombre d'estimations proches de la population exacte (721657) est bien plus grand. Avec le sondage 1, seule une estimation appartenait à l'intervalle $[680000, 760000]$; avec le sondage 2, 7 estimations y étaient. Pour l'intervalle $[670000, 770000]$ les proportions respectives sont : 1/12 et 10/12. Il semble donc bien que certaines méthodes de sondage soient meilleures que d'autres.

Sondage avec poids

Dans l'exemple précédent on a vu deux types de sondage. Il peuvent être vus comme des cas particuliers de sondages pour lesquels on partage la population à étudier en classes dans lesquelles on tire au hasard avec des probabilités distinctes. Soyons un peu plus formels.

On a une population Ω de taille n réunion de l sous-population Ω_1,..,Ω_l de tailles respectives n_1,..., n_l. À chaque individu ω de Ω est associée une grandeur T_ω dont on cherche à estimer la moyenne. On tire au hasard k_i individus dans chaque Ω_i. On obtient un échantillon de $k = k_1 + \ldots + k_l$ individus. Notons $E = \{e_1, \ldots, e_k\}$ cet échatillon (aléatoire). La somme des grandeurs T associées aux éléments de cet échantillon est

$$n/k \sum_{\omega \in \Omega} \mathbf{1}_{\omega \in E} T_\omega$$

Calculons l'espérance de cette somme :

$$\begin{aligned}\mathbb{E}\left(n/k\sum_{\omega\in\Omega}\mathbf{1}_{\omega\in E}T_\omega\right) &= n/k\sum_{\omega\in\Omega}\mathbb{P}\left(\omega\in E\right)T_\omega\\ &= n/k\sum_{i=1}^{l}\sum_{\omega\in\Omega_i}\mathbb{P}\left(\omega\in E\right)T_\omega\\ &= n/k\sum_{i=1}^{l}\sum_{\omega\in\Omega_i}k_i/n_iT_\omega.\end{aligned}$$

Ce n'est pas la somme des T_ω que l'on cherche à estimer. On dit que l'estimateur est biaisé (cela ne signifie pas qu'il est mauvais, mais ici il est mauvais). Que faut-il faire pour obtenir une estimateur sans biais ? Prendre non la somme mais

$$\sum_{i=1}^{l}\sum_{\omega\in\Omega_i}\mathbf{1}_{\omega\in E}n_i/k_iT_\omega.$$

Pourquoi avons-nous $\mathbb{P}\left(\omega\in E\right)=k_i/n_i$ si $\omega\in\Omega_i$? Les éléments de Ω_i sont sélectionnés en prenant au hasard une partie à k_i éléments de Ω_i. Quelle est la probabilité qu'un élément ω donné de Ω_i appartienne à la partie tirée au hasard ? Elle est égale au rapport du nombre de parties à k_i éléments de Ω_i contenant ω et du nombre de parties à k_i éléments de Ω_i c'est à dire

$$\frac{\binom{n_i-1}{k_i-1}}{\binom{n_i}{k_i}}=\frac{k_i}{n_i}$$

d'après la formule démontrée plus haut.

Réduire la variance

Une information sur la structure de la population peut permettre de réduire la variance d'un estimateur et d'améliorer la précision des sondages obtenus. Prenons un exemple.

Une population est composée de 200 individus ayant pour un caractère mesuré les valeurs 0, 1, 9, 10 avec 50 individus par valeur.

On veut sonder la moyenne du caractère (5).

Sondage 1 : on prend 50 personnes parmi les 200 et on fait la moyenne.

Sondage 2 : on prend 25 personnes dans les 100 dont le caractère vaut 0 ou 1, 25 dans la population dont le caractère vaut 9 ou 10. On fait la moyenne des moyennes.

Calcul de l'espérance.

Sondage 1.

$$\begin{aligned}\mathbb{E}\left(1/50\sum_{\omega\in\Omega}\mathbf{1}_{\omega\in E}T_\omega\right) &= 1/50\sum_{\omega\in\Omega}\mathbb{P}\left(\omega\in E\right)T_\omega\\ &= 1/50\sum_{\omega\in\Omega}50/200T_\omega\\ &= 1/200\sum_{\omega\in\Omega}T_\omega\\ &= 5.\end{aligned}$$

Sondage 2.

$$\begin{aligned}&1/2\left(\mathbb{E}\left(1/25\sum_{\omega\in\Omega_1}\mathbf{1}_{\omega\in E}T_\omega\right)+\mathbb{E}\left(1/25\sum_{\omega\in\Omega_1}\mathbf{1}_{\omega\in E}T_\omega\right)\right)\\ &= 1/50\sum_{\omega\in\Omega_1}25/100T_\omega+ = 1/50\sum_{\omega\in\Omega_2}25/100T_\omega\\ &= 1/200\sum_{\omega\in\Omega}T_\omega\\ &= 5.\end{aligned}$$

Nous allons maintenant calculer les variances. Dans un premier temps étudions ce qu'elles seraient si les sondages se faisaient avec remise (les calculs sont plus simples).

Sondage 1. La variance est le 50ème de la variance d'une variable aléatoire prenant avec probabilité 1/4 les valeurs 0,1,9,10. Une telle variable a pour variance

$$1/4(0+1+81+100)-5^2=20,25.$$

La variance de l'estimateur vaut

$$20,25/50=0,405$$

Sondage 2. La variance est la moitié du 25ème de la variance d'une variable aléatoire prenant avec probabilité 1/2 les valeurs 0,1. Une telle variable a pour variance 1/4. La variance de l'estimateur vaut

$$1/200.$$

Le sondage 2 donne une variance 80 fois plus petite, un écart type 9 fois plus petit.

Considérons maintenant le cas sans remise (plus compliqué).

Sondage 1. On a

$$Var\left(1/50\sum_{\omega\in\Omega}\mathbf{1}_{\omega\in E}T_\omega\right) = 1/50^2\left(\sum_{\omega\in\Omega}Var(\mathbf{1}_{\omega\in E})T_\omega^2+\sum_{\omega\neq\omega'}T_\omega T_{\omega'}Cov(\mathbf{1}_{\omega\in E},\mathbf{1}_{\omega'\in E})\right).$$

Par ailleurs

$$Var(\mathbf{1}_{\omega\in E}) = \mathbb{P}(\omega \in E)(1 - \mathbb{P}(\omega \in E)) = 3/16$$

et, pour $\omega \neq \omega'$,

$$\begin{aligned} Cov(\mathbf{1}_{\omega\in E}, \mathbf{1}_{\omega'\in E}) &= \mathbb{P}(\omega \in E, \omega' \in E) - \mathbb{P}(\omega \in E)\mathbb{P}(\omega' \in E) \\ &= \frac{50.49}{200.199} - 1/16 = -9,42.10^{-4} \end{aligned}$$

On obtient

$$\begin{aligned} Var\left(1/50 \sum_{\omega\in\Omega} \mathbf{1}_{\omega\in E} T_\omega\right) &= 1/50.3/16.(1 + 81 + 100) \\ &-1/50^2.9,42.10^{-4}\,(50.49(1 + 81 + 100) + 50.50.(9 + 10 + 90)) \end{aligned}$$

$$Var\left(1/50 \sum_{\omega\in\Omega} \mathbf{1}_{\omega\in E} T_\omega\right) = 0,6825 - 0,168 - 0,102 = 0,41$$

Sondage 2. On mène le calcul par sous-population. Un raisonnement similaire donne :

$$Var\left(1/25 \sum_{\omega\in\Omega_1} \mathbf{1}_{\omega\in E} T_\omega\right) = 1/25^2 \left(\sum_{\omega\in\Omega_1} Var(\mathbf{1}_{\omega\in E}) T_\omega^2 + \sum_{\omega\neq\omega'} T_\omega T_{\omega'} Cov(\mathbf{1}_{\omega\in E}, \mathbf{1}_{\omega'\in E})\right).$$

$$Var(\mathbf{1}_{\omega\in E}) = \mathbb{P}(\omega \in E)(1 - \mathbb{P}(\omega \in E)) = 3/16$$

$$Cov(\mathbf{1}_{\omega\in E}, \mathbf{1}_{\omega'\in E}) = \mathbb{P}(\omega \in E, \omega' \in E) - \mathbb{P}(\omega \in E)\mathbb{P}(\omega' \in E) = \frac{25.24}{100.99} - 1/4 = -1,9.10^{-3}$$

$$Var\left(1/25 \sum_{\omega\in\Omega_1} \mathbf{1}_{\omega\in E} T_\omega\right) = 50/25^2.3/16 - 1,9.10^{-3}.50.49/25^2 = 0,0075$$

Dans le sondage 2 la variance est donc à peu près 0,0038. L'écart type du deuxième sondage est dix fois inférieur à celui du premier !

L'intérêt de la variance petite se voit bien quand on énonce l'inégalité de Bienaymé-Tchebychev :

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}.$$

Par exemple dans le sondage 1 dont on vient de parler, cette inégalité donne (pour $\epsilon = 1$)

$$\mathbb{P}(|Estimation1 - 5| \geq 1) \leq \frac{0,41}{1} = 0,41$$

Pour le sondage 2, l'inégalité est (toujours pour $\epsilon = 1$)

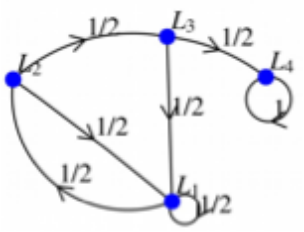
$$\mathbb{P}(|Estimation2 - 5| \geq 1) \leq \frac{0,0038}{1} = 0,0038$$

L'inégalité de Bienaymé-Tchebychev assure donc que l'estimation 1 (donnée par le sondage 1) tombe dans l'intervalle $[4, 6]$ dans plus de 4 cas sur 10, l'estimation 2 dans plus de 996 cas sur 1000. Bien entendu l'inégalité de Bienaymé-Thebychev n'est qu'une majoration ; mais c'est un outil intéressant qui fait bien sentir l'intérêt de disposer de petites variances.

2.4 Chaînes de Markov

2.4.1 Loi des séries ?

Le but de cette situation est de calculer la probabilité d'obtenir au moins une série de n issues identiques lors de la répétition de N expériences aléatoires indépendantes à issues équiprobables. Modéliser cette situation par le graphe ci-dessous relève de l'astuce, on le donne donc aux élèves.



1) Sur 100 lancers d'une pièce équilibrée, on souhaite calculer la probabilité d'obtenir au moins une série d'au moins 4 faces ou d'au moins 4 piles (P P F P F F F F P F... par exemple). Expliquer comment ce graphe probabiliste peut permettre de répondre à la question. Calculer cette probabilité. Commenter.

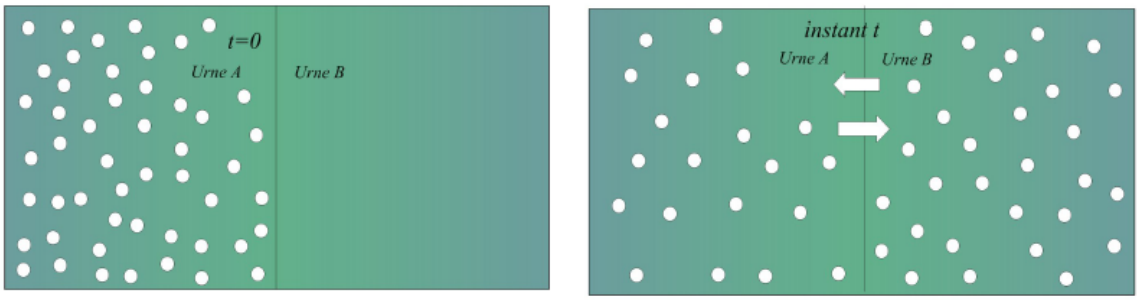
2) Proposer un graphe permettant de calculer la probabilité d'obtenir au moins une série d'au moins 6 nombres identiques sur 200 lancers d'un dé à 6 faces équilibré. Calculer cette probabilité.

2.4.2 Ehrenfest

Une première présentation du modèle

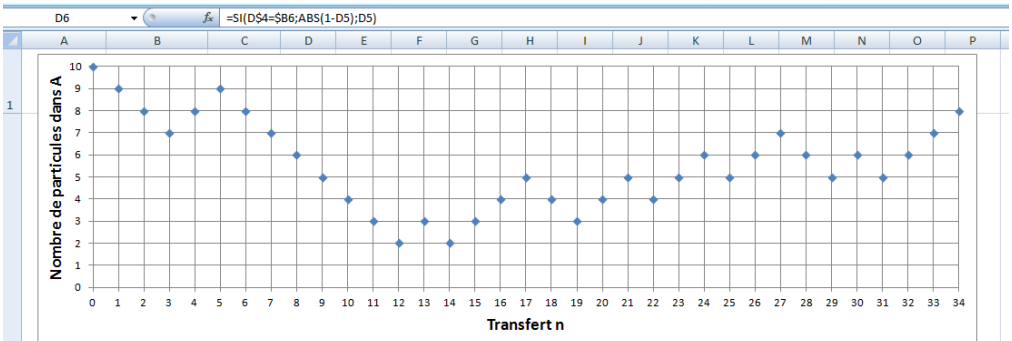
Ce qui suit peut être le sujet d'un devoir à la maison.

Ce modèle doit son nom au couple de physiciens Tatiana et Paul Ehrenfest et à leurs travaux réalisés au début du XXème siècle. Il fut proposé en 1907 afin de décrire en terme de physique statistique les échanges de chaleur entre deux systèmes portés initialement à une température différente. L'étude de ce modèle nous permettra de voir que l'on peut naturellement s'attendre à atteindre un état d'équilibre alors que le comportement du modèle est réversible dans le temps. On considère une enceinte séparée par une paroi en deux urnes A et B et contenant au total N boules numérotées de 1 à N. A l'instant initial, toutes les boules sont dans l'urne A. À chaque instant, on choisit au hasard et de façon équiprobable un nombre compris entre 1 et N. Ainsi, à chaque instant, une boule choisie au hasard change d'urne.



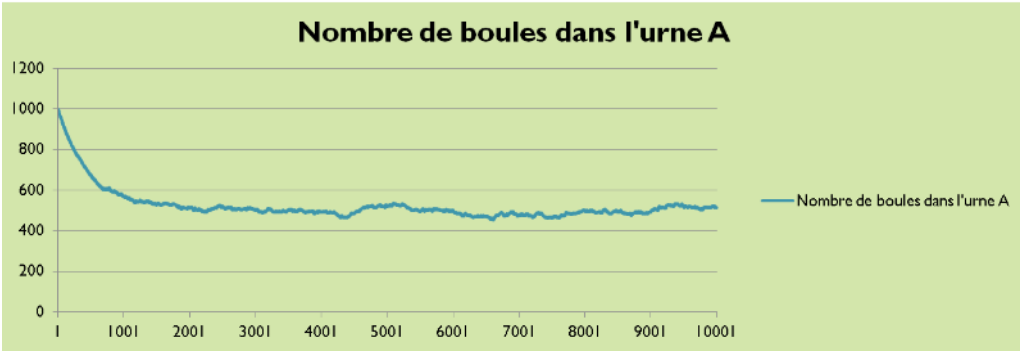
Partie A : Simulations. Effectuer une simulation, avec Excel, pour le cas $N = 10$ et observer l'évolution du nombre de particules dans A.

On pourra créer un nombre entier aléatoire entre 1 et 10 dans la colonne "particule transférée" avec =ALEA.ENTRE.BORNES(1 ;10)

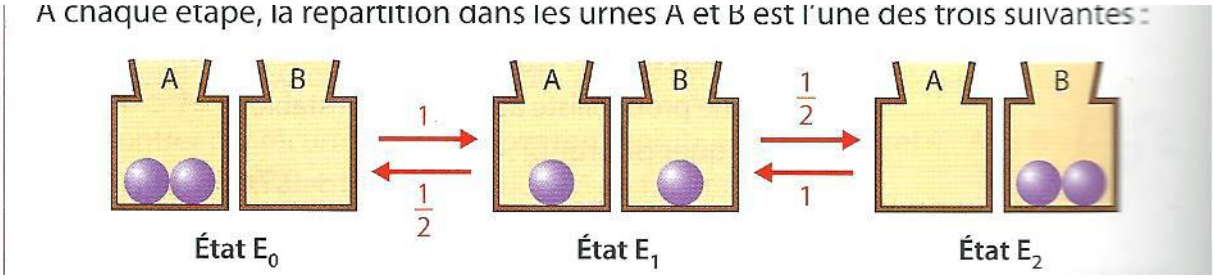


Transfert n	Particule transférée	Position de chaque particule (code 1 pour A, 0 pour B). Particule numéro :										Effectif	
		1	2	3	4	5	6	7	8	9	10	A	B
0		1	1	1	1	1	1	1	1	1	1	10	0
1	3	1	1	0	1	1	1	1	1	1	1	9	1
2	4	1	1	0	0	1	1	1	1	1	1	8	2
3	8	1	1	0	0	1	1	1	0	1	1	7	3
4	4	1	1	0	1	1	1	1	0	1	1	8	2
5	8	1	1	0	1	1	1	1	1	1	1	9	1
6	7	1	1	0	1	1	1	0	1	1	1	8	2

Simulation sur tableur avec $N = 1000$: on observe dans ce cas la stabilisation de la répartition des boules entre les deux urnes. Illustration dans le cas de 10 000 tirages.



Partie B : Modélisation dans le cas N=2.



On modélise la situation par des urnes contenant des boules identiques. On suppose qu'on a placé initialement deux boules dans l'urne A. Les urnes peuvent ensuite se trouver dans trois situations possibles :
1 les deux boules sont dans l'urne A ;
2 chaque urne contient une boule ;
3 les deux boules sont dans l'urne B.
Pour tout entier naturel n, on note X_n la variable aléatoire donnant le nombre de boules dans l'urne A au n-ème échange, et la matrice des probabilités :

$$U_n = (\mathbb{P}(X_n = 2), \mathbb{P}(X_n = 1), \mathbb{P}(X_n = 0)).$$

Ainsi, $U_0 = (1, 0, 0)$.
1. Montrer que, pour tout $n \geq 0$, $U_{n+1} = U_n A$, avec une matrice A à préciser.
2. En déduire que, pour tout $n \geq 0$, $U_n = U_0 A^n$.
3. À l'aide de la calculatrice, conjecturer la valeur de A^n en fonction de n, puis démontrer la conjecture.
4. La répartition des boules se stabilise-t-elle lorsque le nombre d'échanges devient grand ?

Une deuxième présentation du modèle

Modèle d'Ehrenfest : pourquoi ne verra-t-on presque jamais un ballon percé se regonfler ? On dispose de 2 urnes : la première vide et la seconde remplie de 3 boules numérotées de 1 à 3. On répète l'expérience aléatoire suivante : "choisir au hasard un numéro puis changer d'urne la boule correspondante".

1) Modéliser cette situation à l'aide d'un graphe probabiliste où les états correspondront au nombre de boules dans la première urne. Calculer la probabilité d'être revenu à l'état initial après 10 étapes, 11 étapes, commenter en regardant la parité du nombre de boules dans la première urne.

2) Expliquer pourquoi, en remplaçant l'état 0 par un état absorbant et en commençant à l'état 1, on peut calculer la probabilité d'être revenu à l'état initial au cours de n étapes. Calculer cette probabilité pour 10 et 11 étapes.

3) Calculer la probabilité d'être revenu à l'état initial au cours de 10 étapes, de 100 étapes, si la deuxième urne contient au départ 9 boules (c'est un peu long mais avec Xcas cela s'écrit bien). Commenter la phrase "On ne verra presque jamais un ballon percé se regonfler" en considérant le nombre d'Avogadro.

Mesure stationnaire

Une mesure stationnaire pour une chaîne de Markov est une mesure telle que si on choisit au hasard la position initiale suivant cette mesure, alors à tous les instants la probabilité d'être en telle ou telle position est donnée par cette probabilité.

Au modèle de l'urne d'Ehrenfest on associe facilement une chaîne de Markov. Appelons N le nombre total de boules et X_n le nombre de boules contenues dans le compartiment gauche. L'état du système à l'instant n est complètement décrit par X_n (le nombre de boules dans le compartiment droit s'en déduit). Les règles de passage d'un état à un autre sont données par les probabilités conditionnelles

$$\mathbb{P}_{X_n=k}(X_{n+1}=k-1)=\frac{k}{N},\quad \mathbb{P}_{X_n=k}(X_{n+1}=k+1)=\frac{N-k}{N}.$$

Nous allons vérifier que la loi binomiale de paramètre N, $1/2$ est stationnaire. Le problème est le suivant : montrer que si X_n suit cette loi alors X_{n+1} aussi. La loi de X_{n+1} est décrite par les nombres $\mathbb{P}(X_{n+1}=k)$, k variant de 0 à N. Calculons un tel nombre en utilisant la formule des probabilités totales :

$$\begin{aligned}\mathbb{P}(X_{n+1}=k) &= \mathbb{P}_{X_n=k+1}(X_{n+1}=k)\mathbb{P}(X_n=k+1)+\mathbb{P}_{X_n=k-1}(X_{n+1}=k)\mathbb{P}(X_n=k-1)\\ &= \frac{k+1}{N}\mathbb{P}(X_n=k+1)+\frac{N-k+1}{N}\mathbb{P}(X_n=k-1)\\ &= \frac{k+1}{N}\binom{N}{k+1}\frac{1}{2^N}+\frac{N-k+1}{N}\binom{N}{k-1}\frac{1}{2^N}.\end{aligned}$$

Or nous avons vu la relation

$$N\binom{N}{j-1}=j\binom{N}{j}.$$

En l'appliquant avec $k+1$ à la place de j on obtient

$$\frac{k+1}{N}\binom{N}{k+1}=\binom{N-1}{k}.$$

D'autre part on a

$$\binom{N}{k-1}=\binom{N}{N-k+1}$$

la même formule (en remplaçant j par $N-k+1$ cette fois) donne

$$\frac{N-k+1}{N}\binom{N}{k-1}=\frac{N-k+1}{N}\binom{N}{N-k+1}=\binom{N-1}{N-k}=\binom{N-1}{k-1}.$$

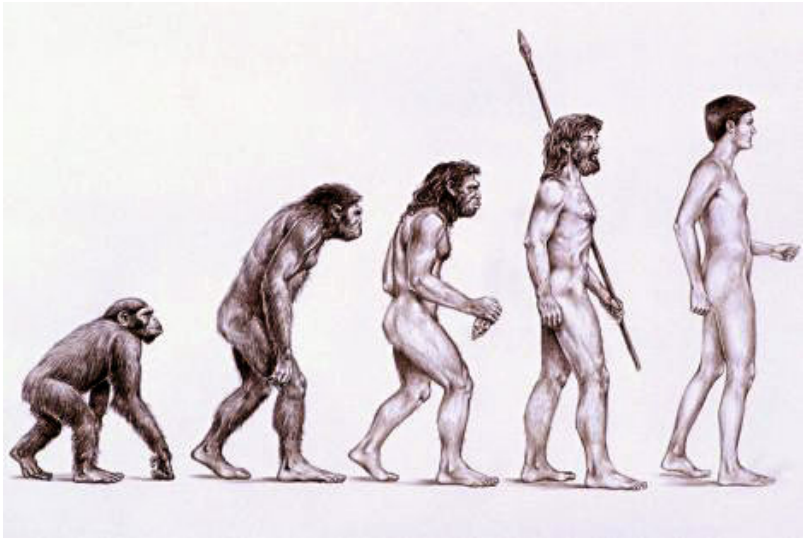
On en déduit

$$\mathbb{P}(X_{n+1}=k)=\frac{1}{2^N}\left(\binom{N-1}{k}+\binom{N-1}{k-1}\right)=\frac{1}{2^N}\binom{N}{k},$$

ce que nous voulions montrer. On peut être gêné par les cas particuliers $k=0$ et $k=N$ (certaines des probabilités écrites plus haut sont nulles). On pourra les traiter à part si l'on préfère.

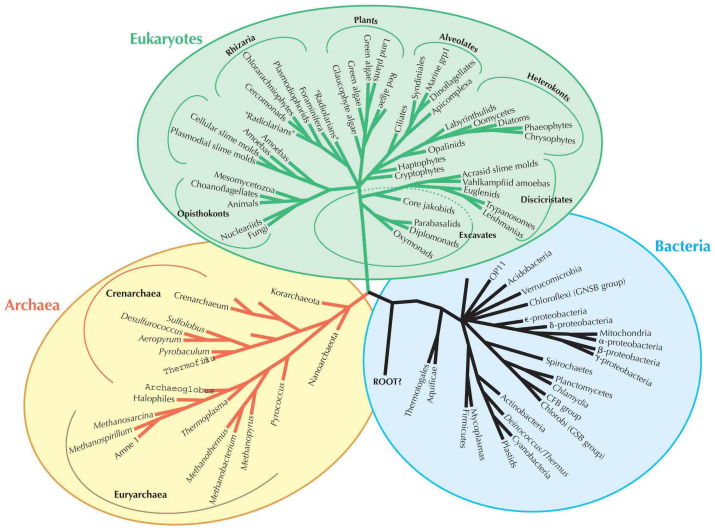
2.4.3 La marche de l'ivrogne

Dans son livre l'éventail du vivant Stephen Jay Gould (1941-2002) évolutionniste américain présente certains raisonnements statistiques et les utilise pour réfléchir à différents sujets (évolution, base-ball, médecine...). Il explique pourquoi selon lui l'idée (répadue) d'une marche vers la complexité dans l'évolution est fausse. Cette idée erronée est bien illustrée par des images comme la suivante :



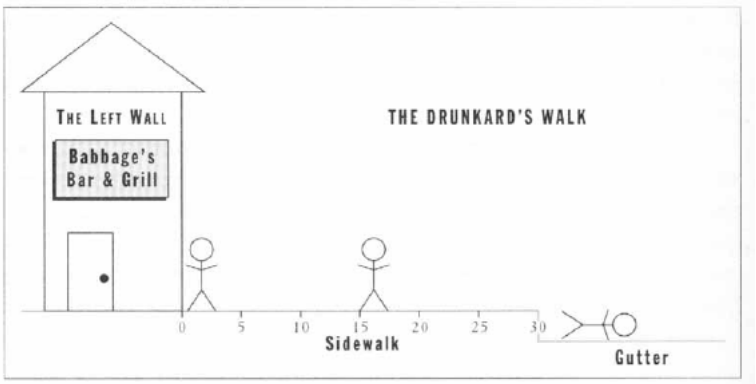
Il n'est pas rare de présenter l'évolution comme l'histoire d'un progrès que serait l'engendrement d'êtres de plus en plus complexes : on partirait de créatures unicellulaires pour progressivement aboutir au summum, le cerveau humain. Il est implicitement admis que les êtres complexes ont un avantage dans la lutte pour la vie qui expliquerait pourquoi ils sont sélectionnés. Gould conteste ce point : les êtres unicellulaires sont toujours massivement présents et remarquablement adaptés à leur milieu ; de nombreuses espèces complexes ont disparues.

La source de cette erreur est peut-être en rapport avec une vision linéaire de l'évolution du vivant : on ne suit qu'une lignée d'espèces (en particulier une lignée aboutissant à l'homme). Or il est bien préférable de l'imaginer sous la forme d'un arbre foisonnant.



Gould prétend que la bactérie est la forme de vie la plus répandue et la plus capable d'adaptation à différents milieux (l'idée d'une adaptabilité supérieure des créatures complexes serait fausse). Pour battre en brèche l'idée d'une tendance à la complexité Gould prend l'exemple de la marche de l'ivrogne :

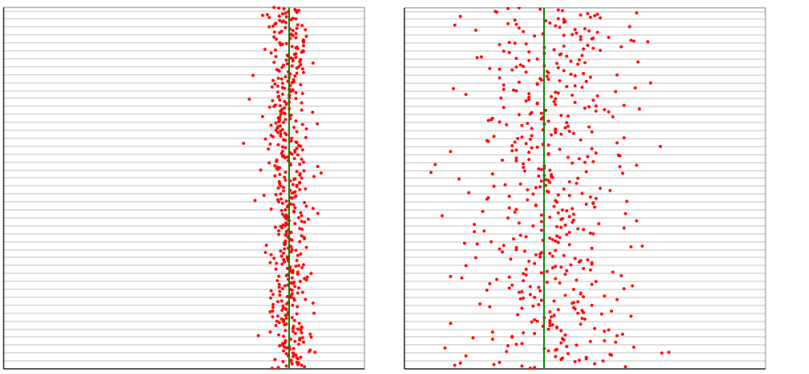
Un homme sort d'un bar complètement saoul et titube sur le trottoir, entre le mur du bar et le caniveau. S'il atteint le caniveau, il s'écroule ivre mort et le jeu s'arrête. Supposons que le trottoir ait trois mètres de large et que notre ivrogne marche au hasard, avec un pas moyen de cinquante centimètres, en avant ou en arrière. [...] Où va-t-il aboutir si on le laisse tituber suffisamment longtemps ? Dans le caniveau, inéluctablement, et pour la raison suivante : chaque pas, en avant ou en arrière, a une probabilité égale de 1/2. Le mur du bar constitue une "barrière infranchissable". [...] Autrement dit, son mouvement ne peut se développer que dans une seule direction : vers le caniveau. [...] J'ai exhumé ce vieil exemple pour illustrer un point capital : [...] L'ivrogne tombe à chaque fois dans le caniveau, mais son mouvement ne témoigne d'aucune tendance à cette forme de perdition. D'une manière similaire, une mesure moyenne ou extrême de la vie peut progresser dans une direction donnée même si aucun avantage évolutif ou aucune tendance intrinsèque ne favorise ce mouvement.



On peut simuler grâce à un ordinateur cette marche de l'ivrogne. Le programme suivant écrit avec le logiciel Algobox simule 500 fois une marche pour une probabilité d'aller à droite égale à $1-p$ où p est demandé à l'utilisateur. Le nombre de pas effectués n est aussi demandé. Le programme affiche en rouge (sur la version électronique ; peut-être en gris sinon) les positions atteintes par l'ivrogne au bout de n pas. Dans les images obtenues recopiées plus bas le nombre de pas demandé est 1000.

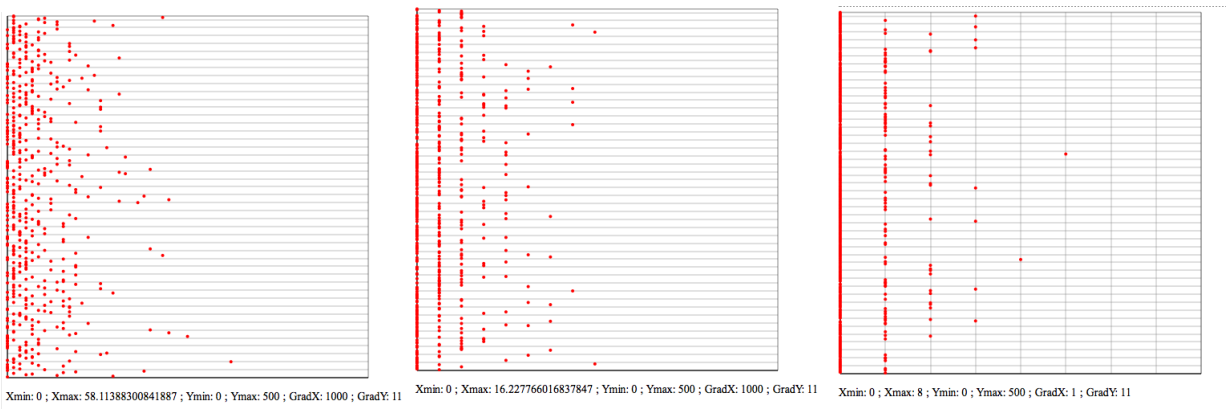
```
VARIABLES
    n EST_DU_TYPE NOMBRE
    k EST_DU_TYPE NOMBRE
    j EST_DU_TYPE NOMBRE
    position EST_DU_TYPE NOMBRE
    alea EST_DU_TYPE NOMBRE
    p EST_DU_TYPE NOMBRE
DEBUT_ALGORITHME
    LIRE n
    LIRE p
    POUR k ALLANT_DE 1 A 500
        DEBUT_POUR
        position PREND_LA_VALEUR 0
        POUR j ALLANT_DE 1 A n
            DEBUT_POUR
            alea PREND_LA_VALEUR random()
            SI (alea>p) ALORS
                DEBUT_SI
                position PREND_LA_VALEUR position+1
                FIN_SI
            SI (alea<p) ALORS
                DEBUT_SI
                SI (position>0) ALORS
                    DEBUT_SI
                    position PREND_LA_VALEUR position-1
                    FIN_SI
                FIN_SI
            FIN_POUR
        TRACER_POINT (position,k)
        FIN_POUR
    TRACER_SEGMENT (n*(1-2*p),0)->(n*(1-2*p),500)
FIN_ALGORITHME
```

Voici les positions finales que l'on obtient pour des probabilités d'aller vers la droite égales à 0,8 puis 0,55 :



Xmin: 0 ; Xmax: 758.113883008419 ; Ymin: 0 ; Ymax: 500 ; GradX: 1000 ; GradY: 11 Xmin: 0 ; Xmax: 258.11388300841895 ; Ymin: 0 ; Ymax: 500 ; GradX: 1000 ; GradY: 11

Pour des probabilités d'aller vers la droite égales à 0,45, 0,35, 0,2 :



Que constatons-nous ? Si la probabilité d'aller à droite est supérieure à 0,5 alors les positions au bout de 1000 pas se trouvent presque toutes éloignées à droite. Lorsque la probabilité d'aller à droite est inférieure à 0,5 alors les positions finales sont concentrées près du bar (d'autant plus que p est grande). Si p est proche de 1 il semble très difficile de s'éloigner beaucoup du bar.

L'analogie est la suivante : la position 0 est celle d'un être simple unicellulaire, les positions éloignées du bar celles d'êtres complexes. Dire qu'il existe une tendance à la complexité serait dire que la probabilité d'aller vers la droite est supérieure à 0,5 (il n'y aurait pas de tendance à la complexité si cette probabilité valait 0,5 ; il y aurait une tendance à la simplification si cette probabilité était inférieure à 0,5). Quand existe une tendance à droite alors les positions près du bar disparaissent. Ce n'est pas ce qu'on observe car les bactéries n'ont pas disparu (et, s'il faut parier, semblent bien moins menacées que beaucoup d'autres). Conclusion : il n'y a pas de tendance à la complexité en évolution.

Je ne nie pas que les créatures les plus complexes soient devenues plus complexes au cours du temps. Mais ce n'est pas une indication que le système s'est éloigné d'une marche au hasard. Cela se serait produit de toute manière dans n'importe quel système dirigé par le hasard et débutant à proximité d'une limite infranchissable à gauche de la courbe de distribution. Je propose une analogie avec la marche de l'ivrogne qui sort d'un bar. Il se retrouve avec un mur à gauche et le trottoir à droite. A gauche il va heurter le mur, à droite il va finir par tomber dans le caniveau.

L.R. : L'analogie de l'ivrogne ne nous dit rien sur l'émergence de la complexité.

S.J.G. : Non, mais c'est une bonne analogie. L'ivrogne marche au hasard, et en raison du mur à gauche le hasard le conduit immanquablement dans le caniveau. La seule raison de l'existence d'une directionalité est ce mur à gauche et la marche au hasard. L'histoire de la vie montre exactement la même chose. Elle avance au hasard, avec ce mur à gauche qui interdit à un organisme vivant d'être plus simple qu'un certain degré minimal de

complexité, plus simple que des cellules sans noyau. Je ne dis pas qu'il ne se produit pas des événements de complexification, je dis que si l'on regarde l'ensemble de l'histoire, l'ensemble des variations effectives, la maison au complet avec tous ses habitants, on ne décèle pas de préférence pour la complexité. Le fait que l'homme soit plus complexe que les trilobites, qui sont plus complexes que les algues, qui sont plus complexes que les bactéries, ce fait-là, que je ne nie pas, est mineur au regard de l'histoire du vivant prise dans sa totalité.

L.R. : Vous montrez d'ailleurs que l'évolution se dirige souvent dans le sens d'une simplification des organismes...

S.J.G. : Sans doute même aussi souvent que dans l'autre sens. Et peut-être même plus souvent... Si l'on considère l'histoire d'organismes nés plus ou moins récemment dans l'histoire de la vie, donc qui n'étaient pas limités par un mur à gauche, on observe aussi bien des évolutions vers moins de complexité que le contraire. C'est manifestement le cas de nombreux parasites, ceux qui vivent profondément installés dans le corps de leur hôte : ils n'ont besoin ni d'organes de locomotion ni d'organes de digestion.

Ajoutons une nouvelle remarque. Même si la probabilité p est proche de 1, si elle ne vaut pas 1, alors l'ivrogne atteindra tôt ou tard n'importe quelle position $c > 0$. Pour le voir il suffit de remarquer que si l'ivrogne fait c pas vers la droite alors il se trouve en une position plus à droite que c et donc est passé par c. Quelle est la probabilité que l'ivrogne ne fasse pas c pas vers la droite ? Réponse : $1-(1-p)^c$ (probabilité du complémentaire de "faire c pas vers la droite"). Quelle est la probabilité que lors d'une série de n fois c pas, il ne fasse jamais c pas vers la droite ? Réponse : $(1-(1-p)^c)^n$. Lorsque n tend vers l'infini, cette quantité tend vers 0 (sauf si $p = 1$). Autrement dit, lorsque le temps passe la probabilité que l'ivrogne ne soit pas passé par la position c tend vers 0. Dans le langage de l'analogie précédente cela signifie que ce n'est pas parce qu'existent des êtres complexes qu'existe une tendance à la complexité.

Cherchons une suite $(p_n)_{n\geq 0}$ définissant une probabilité stationnaire pour la marche de l'ivrogne. Une telle suite, si elle existe, vérifie :

$$p_n = \mathbb{P}(X_{k+1} = n) = \mathbb{P}(X_{k+1} = n | X_k = n-1)p_{n-1} + \mathbb{P}(X_{k+1} = n | X_k = n+1)p_{n+1},$$

pour $n \geq 1$, et

$$p_0 = \mathbb{P}(X_{k+1} = 0) = \mathbb{P}(X_{k+1} = 0 | X_k = 0)p_0 + \mathbb{P}(X_{k+1} = 0 | X_k = 1)p_1.$$

En remplaçant les probabilités conditionnelles par leurs valeurs, on obtient :

$$p_0 = p_0 p + p_1 p, \quad \forall n \geq 1 \; p_n = p_{n+1}p + p_{n-1}(1-p).$$

On en déduit (posons $q = 1-p$) $p_1 = \frac{q}{p}$, puis $p_2 p = p_1 - p_0 q = p_0(\frac{q}{p} - q) = p_0 \frac{q^2}{p}$, soit $p_2 = \frac{q^2}{p^2} p_0$. On montre ensuite par récurrence que, si la suite $(p_n)_{n\geq 0}$ existe, elle satisfait

$p_n = \frac{q^n}{p^n} p_0$. Pour que $(p_n)_{n\geq 0}$ définisse une probabilité il faut que les p_n soient des nombres positifs ou nuls de somme égale à 1. Ils ne peuvent donc être tous nuls. Si la formule $p_n = \frac{q^n}{p^n} p_0$ est vérifiée pour tout n, aucun des nombres p_n n'est nul. Si $p \leq 1/2$, alors $\frac{q}{p} \geq 1$ et la suite (p_n) est croissante. La somme des p_n est alors infinie : il n'existe pas dans ce cas de probabilité stationnaire pour la marche de l'ivrogne. Si $p > 1/2$, alors $\frac{q}{p} < 1$ et on a

$$\sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} \frac{q^n}{p^n} p_0 = p_0 \frac{1}{1-\frac{q}{p}} = p_0 \frac{p}{p-q} = p_0 \frac{p}{2p-1}.$$

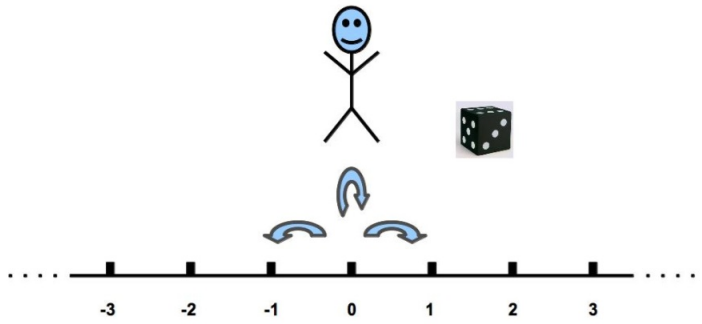
Si $p > 1/2$, il existe une seule probabilité stationnaire définie par :

$$p_n = \frac{2p-1}{p} \frac{q^n}{p^n}.$$

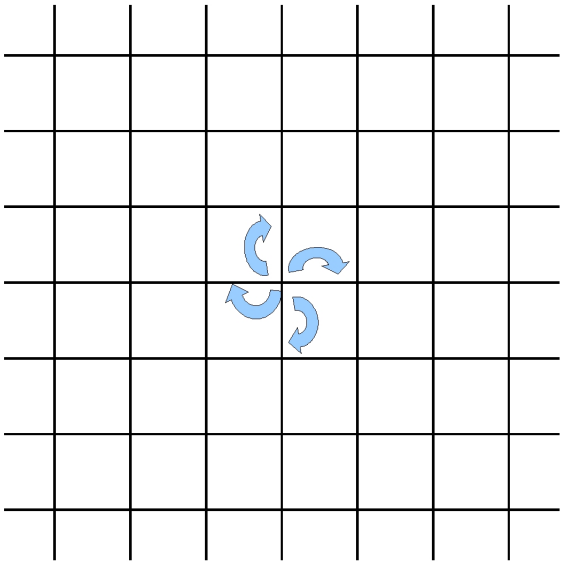
2.4.4 Les marches au hasard sur $\mathbb{Z}^d$

Cette section contient des mathématiques qui ne sont pas accessibles aux lycéens. Mais là encore les idées présentées peuvent être facilement présentées et des algorithmes les illustrant facilement écrits.

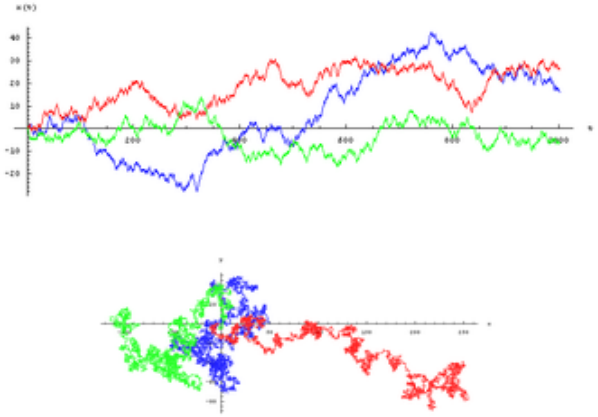
Considérons un piéton qui se déplace sur une ligne droite de mètre en mètre. Il choisit au hasard avec probabilité 1/2 d'avancer ou de reculer.

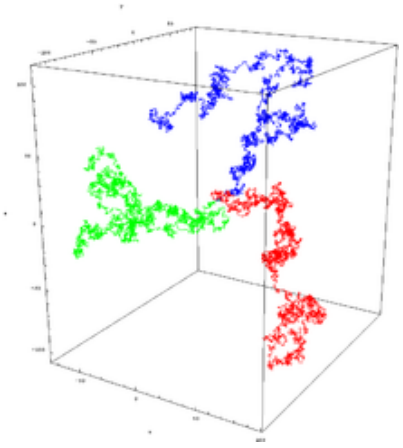


Il pourrait aussi se déplacer sur un plan en choisissant avec probabilité 1/4 l'une des quatre possibilités avant, arrière, droite, gauche :



Il pourrait aussi marcher dans l'espace et choisir avec probabilité 1/6 à chaque instant, chacune des directions avant, arrière, droite, gauche, haut, bas. Quelle allure aura sa trajectoire ? Voici dans chacun des trois cas évoqué plus haut, trois trajectoires (en trois couleurs) vues de loin.





Voyez-vous des différences ? Les trajectoires de la marche sur $\mathbb{Z}^3$ ont l'air de se séparer plus nettement, peut être.

Appelons S_n la position à l'instant n, X_i le pas fait entre l'instant $i-1$ et i. Supposons que nous partions de l'origine à l'instant 0 : $S_0 = 0$. Les variables X_i sont supposées indépendantes à valeurs dans $\mathbb{Z}$, $\mathbb{Z}^2$ ou $\mathbb{Z}^3$.

Nous nous intéressons à la probabilité de revenir en 0 à un instant strictement positif :

$$\mathbb{P}(\exists n > 0 \quad S_n = 0).$$

Cette probabilité est égale à

$$\mathbb{P}(\bigcup_{n>0}(S_n = 0)).$$

Comme les événement $S_n = 0$ ne sont pas disjoints la probabilité de leur réunion n'est pas la somme de leurs probabilités. On peut introduire des événements qui eux seront disjoints et auront la même réunion : E_n, être en 0 à l'instant n pour la première fois :

$$E_n = (S_1 \neq 0, \ldots, S_{n-1} \neq 0, S_n = 0).$$

Par définition les événements E_n sont disjoints et ont même réunion que les $S_n = 0$ (car pour tout $n \geq 1$ on a $(S_n = 0) = \cup_{k=1}^{n} E_k$; réfléchir...). On a

$$\mathbb{P}(\bigcup_{n>0}(S_n = 0)) = \sum_{n=1}^{\infty} \mathbb{P}(E_n).$$

Pour alléger l'écriture plus bas introduisons les notations

$$a_n = \mathbb{P}(S_n = 0), \quad b_n = \mathbb{P}(E_n).$$

$$\begin{aligned} \mathbb{P}(S_n = 0) &= \sum_{k=1}^{n} \mathbb{P}(E_k \cup (S_n = 0)) \\ &= \sum_{k=1}^{n} \mathbb{P}(S_1 \neq 0, \ldots, S_{k-1} \neq 0, S_k = 0, S_n = 0) \end{aligned}$$

Comme les pas X_i sont indépendants, la probabilité

$$\mathbb{P}(S_1 \neq 0, \ldots, S_{k-1} \neq 0, S_k = 0, S_n = 0)$$

est égale à

$$\mathbb{P}(S_1 \neq 0, \ldots, S_{k-1} \neq 0, S_k = 0)\mathbb{P}(S_{n-k} = 0).$$

On obtient la relation

$$a_n = \sum_{k=1}^{n} a_{n-k} b_k,$$

et en posant b_0 par convention,

$$a_n = \sum_{k=0}^{n} a_{n-k} b_k.$$

On reconnaît la forme du coefficient du produit de deux polynômes ou séries entières. Posons

$$F(z) = \sum_{n=0}^{\infty} a_n z^n, \quad G(z) = \sum_{n=0}^{\infty} b_n z^n$$

Comme a_n et b_n sont des probabilités, ces séries convergent pour $|z| < 1$ et on a

$$F(z)G(z) = \sum_{n=0}^{\infty} c_n z^n,$$

où $c_n = \sum_{k=0}^{n} a_{n-k} b_k$, autrement dit $a_n = c_n$ sauf pour $n = 0$, $a_0 = 1$ et $c_0 = 0$. On a donc

$$F(z)G(z) = F(z) - 1,$$

soit

$$F(z)[1 - G(z)] = 1.$$

En prenant pour z des valeurs réelles positives et en faisant tendre z vers 1, on voit que

$$\sum_{k=1}^{\infty} a_k = +\infty \text{ si et seulement si } \sum_{k=1}^{\infty} b_k = 1.$$

Dire que $\sum_{k=1}^{\infty} b_k = 1$, c'est dire que l'on est sûr de revenir en 0. Dans ce cas on est sûr d'y revenir une infinité de fois. On dit alors que la marche est récurrente. Dire que $\sum_{k=1}^{\infty} b_k < 1$, c'est dire qu'on ne revient pas en 0 avec probabilité positive. Dans ce cas, à coup sûr, on n'y reviendra un nombre fini de fois avant de partir vers l'infini. On dit alors que la marche est transiente.

Les probabilités a_n sont plus faciles à calculer que les b_n et il est plus facile de dire si une série converge ou non que de calculer sa somme.

Dans le cas de la marche aléatoire simple sur $\mathbb{Z}$ on ne peut être en 0 qu'aux instants pairs. La probabilité d'être en 0 à l'instant $2n$ est

$$\frac{\binom{2n}{n}}{2^{2n}}.$$

Pour avoir une idée de la taille de ce nombre nous allons utiliser la formule de Stirling

$$m! \sim \sqrt{2\pi m}\left(\frac{m}{e}\right)^m.$$

En utilisant cette formule pour $n!$ et $(2n)!$ apparaissant dans $\binom{2n}{n}$ on obtient

$$\frac{\binom{2n}{n}}{2^{2n}} \sim \frac{\sqrt{2\pi 2n}\left(\frac{2n}{e}\right)^{2n}}{2^{2n}2\pi n\left(\frac{n}{e}\right)^{2n}},$$

et après simplification

$$\frac{\binom{2n}{n}}{2^{2n}} \sim \frac{1}{\sqrt{\pi n}}.$$

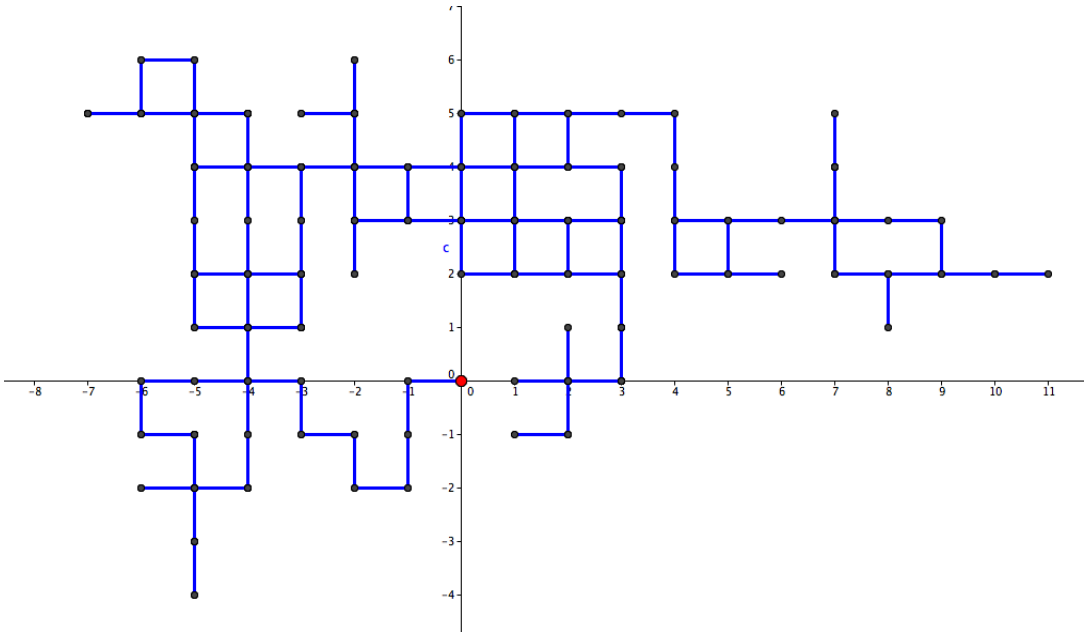
La série de terme général $1/\sqrt{n}$ étant divergente on en déduit que dans le cas de la marche aléatoire simple, la somme des a_n est infinie. La marche est récurrente.

On peut montrer que dans le cas de la marche aléatoire simple sur $\mathbb{Z}^2$ a_{2n} est de l'ordre de c/n pour une certaine constante $c > 0$. La série de terme général $1/n$ étant divergente, on est encore dans le cas récurrent.

Pour la marche aléatoire simple sur $\mathbb{Z}^3$ a_{2n} est de l'ordre $cn^{-3/2}$. Cette fois la série est convergente. La marche est transiente.

Conclusion : en marchant au hasard sur une droite ou sur le plan on finit toujours par retrouver son point de départ (à condition de ne pas favoriser de direction), en volant au hasard dans l'espace on finit par se perdre à jamais.

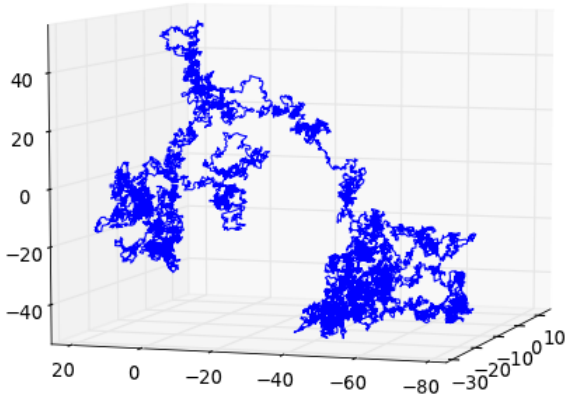
On peut utiliser Geogebra pour visualiser de telles trajectoires :



On peut aussi varier les façons de marcher au hasard. Voici un programme écrit en Python d'une marche aléatoire dans l'espace pour laquelle on choisit une direction (latitude, longitude) au hasard à chaque instant :

```
#-------------------------------------------------------------
# Name:        module1
# Purpose:
#
# Author:      Papa
#
# Created:     26/06/2015
# Copyright:   (c) Papa 2015
# Licence:     <your licence>
#-------------------------------------------------------------
#!/usr/bin/env python

def main():
    pass

if __name__ == '__main__':
    main()
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
import matplotlib.pyplot as plt

fig = plt.figure()
ax = fig.gca(projection='3d')

N=10000

x,y,z=0,0,0
X,Y,Z=[0],[0],[0]

for I in range(N):
    theta = np.random.rand()*2*3.14159265
    phi   = np.random.rand()*2*3.14159265
    x = x + np.sin(phi) * np.cos(theta)
    y = y + np.sin(phi) * np.sin(theta)
    z = z + np.cos(phi)

    X.append(x)
    Y.append(y)
    Z.append(z)

plt.figure(1)


ax.plot(X,Y,Z)#, zdir='z', label='marche aleatoire')



#ax.legend()
ax.set_xlim3d(min(X),max(X))
ax.set_ylim3d(min(Y),max(Y))
ax.set_zlim3d(min(Z),max(Z))

```

Et voici le type de trajectoire qu'on obtient :



Références

[1] J.-P. Benzecri, *Histoire et préhistoire de l'analyse des données*, Dunod, 1982.

[2] D. Perrin, *Géométrie projective, Postface*, http://www.math.u-psud.fr/~perrin/Livre_de_geometrie_projective.html.

[3] S. J. Gould, *L'éventail du vivant*, Seuil, Points Sciences, 1997.

[4] M. Gromov, *Introduction aux mystères*, Actes Sud, 2012.

[5] B. Pascal, *Les pensées*, Le livre de poche, 1962.

[6] *Images des mathématiques*, http://images.math.cnrs.fr/.

[7] *Enseigner les probabilités au lycée*, publié par le réseau des IREM.

[8] Y. Chevallard, http://yves.chevallard.free.fr.

[9] C. Schwartz, *La preuve par les chiffres : de quoi s'agit-il ?*, http://publications-sfds.math.cnrs.fr/index.php/StatEns/article/view/125/115.

FICHE SIGNALÉTIQUE

Titre : Statistique et probabilités au lycée.

Auteurs : Guy CASALE - Jean-Baptiste FAURE - Laurence GAUDE - Gilbert GARNIER - Claude GUIBERT - Stéphane LE BORGNE - Isabelle LE NAOUR - Valérie MONBET

Editeur : IREM de Rennes.

Date : OCTOBRE 2015

Niveau : Lycée

Mots clés : statistique, probabilités, nouveaux programmes des lycées

Résumé : Cette brochure est composée de deux parties.
La première fait le point sur les notions de statistique descriptive, statistique inférentielle et chaînes de Markov, présentes aux programmes des lycées. Nous donnons en particulier une présentation synthétique des principes probabilistes en jeu dans la statistique inférentielle qui pourra être utile aux professeurs ayant à les enseigner (en ES et S, mais aussi en STMG, STI2D, STL). Nous donnons des éléments (parfois hors programmes) permettant aux professeurs de mieux comprendre les différents aspects des programmes : loi des grands nombres et théorème limite central, ou bien comportements observés dans les problèmes faisant intervenir les produits de matrices (en particulier les chaînes de Markov). Cette première partie contient aussi quelques réflexions critiques sur les programmes et des suggestions de progression.
La deuxième partie est un recueil d'activités ou exercices. Plusieurs de ces activités ont des liens avec d'autres disciplines (sciences naturelles, physique, géographie, démographie).

Format	Nombre de pages	Prix	Tirage
21 x 29,7	**79**	15 euros	**100 ex**

ISBN 2-85728-078-5

I.R.E.M de RENNES – Université de RENNES 1