# Statistiques et probabilités dans l'enseignement secondaire

# La statistique au collège

Deux objectifs généraux :

- Le schéma général : s'initier à la lecture et à l'utilisation de représentations, de graphiques.
- Acquérir quelques notions fondamentales de statistique descriptive.

## Classe de sixième

Simple initiation à la lecture, à l'interprétation et à l'utilisation de diagrammes.

## Classe de cinquième

### Programme

| Contenus | Compétences exigibles | Commentaires |
|---|---|---|
| Lecture, interprétation, représentations graphiques de séries statistiques. Classes, effectifs. | Lire et interpréter un tableau, un diagramme à barres, un diagramme circulaire ou semi-circulaire.<br>Regrouper des données statistiques en classes, calculer des effectifs.<br>Présenter une série statistique sous la forme d'un tableau, la représenter sous la forme d'un diagramme ou d'un graphique. | Il importe d'entraîner les élèves à lire et à représenter des données statistiques en utilisant un vocabulaire adéquat.<br>Le calcul d'effectifs cumulés n'est pas une compétence exigible mais il pourra être entrepris, en liaison avec les autres disciplines dans des situations où les résultats auront une interprétation.<br>Le choix de la représentation est lié à la nature de la situation étudiée. |
| Fréquences. | Calculer des fréquences. | La notion de fréquence est notamment utilisée pour comparer des populations d'effectifs différents et faire le lien avec la proportionnalité. Les écritures 4/10, 2/5, 0,4 (ou en notation anglo-saxonne 0.4 ou .4), 40% qui peuvent être utilisées pour désigner une fréquence, permettent d'insister sur les diverses représentations d'un même nombre. |

# Classe de quatrième

## Programme

| Contenus | Compétences exigibles | Commentaires |
|---|---|---|
| Effectifs cumulés, fréquences cumulées.<br>Moyennes pondérées.<br><br><br><br>Initiation à l'utilisation de tableurs-grapheurs. | Calculer des effectifs cumulés, des fréquences cumulées.<br>Calculer la moyenne d'une série statistique.<br>Calculer une valeur approchée de la moyenne d'une série statistique regroupée en classes d'intervalles. | L'élève sera confronté à des situations courantes où la méthode de calcul est à remettre en cause : par exemple, les différences constatées entre la moyenne annuelle des notes d'un élève calculée à partir de l'ensemble des notes de l'année ou à partir de la moyenne des moyennes trimestrielles. |

## Document d'accompagnement des programmes du cycle central

Au collège, l'enseignement de statistique descriptive a pour objet de familiariser progressivement les élèves avec la démarche consistant à synthétiser, sous forme numérique ou graphique, des informations recueillies sur l'ensemble des éléments d'une population. L'essentiel de l'activité des élèves consiste à exploiter, de façon raisonnée, des documents adaptés à chaque classe, afin de développer leur autonomie dans ce domaine ; ces documents gagnent à être choisis en concertation avec d'autres disciplines.

Pour faciliter l'interprétation et l'analyse critique des résultats obtenus, chaque apprentissage est étalé sur deux années de collège. Ainsi, en classe de 5$^{e}$, on poursuit la présentation de relevés statistiques sous forme de tableaux ou de graphiques abordée en classe de 6$^{e}$, en s'intéressant à la pertinence du choix des classes et du mode de représentation graphique retenus. De même, les notions d'effectifs et de fréquences introduites en classe de 5$^{e}$ trouvent un prolongement en classe de 4$^{e}$, avec les effectifs cumulés et les fréquences cumulées.

Avec la moyenne d'une série statistique, qui ne constitue pas une réelle nouveauté pour les élèves, on aborde en classe de 4$^{e}$ une nouvelle phase de la synthèse des informations recueillies. Le programme insiste sur la distinction entre le cas où l'on dispose de données sur l'ensemble des éléments de la population étudiée et celui où les données concernent un regroupement de la population en classes d'intervalles ; dans ce dernier cas, la méthode mise en œuvre ne permet d'obtenir qu'une valeur approchée de la moyenne de la population.

Sans introduire de nouveaux indicateurs de la tendance centrale d'une population, il peut être intéressant de faire observer aux élèves, dès la classe de 4$^{e}$, que la moyenne d'une population dont les éléments sont rangés par ordre croissant ne sépare pas ceux-ci, en général, en deux parties de même effectif.

En 5$^{e}$ et en 4$^{e}$, la partie statistique fait intervenir d'autres rubriques du programme, les activités numériques et graphiques s'appuyant très largement sur la proportionnalité ; elle peut donc contribuer à donner du sens à ce concept dont l'acquisition est un des objectifs de l'enseignement des mathématiques au collège.

L'utilisation de tableurs-grapheurs offre la possibilité de limiter, à propos de quelques exemples nécessaires à une bonne compréhension des règles mises en jeu, le temps consacré à la réalisation manuelle des diagrammes figurant au programme. Avec ces logiciels, il est aussi possible de mener expérimentalement la recherche d'une répartition en classes, adaptée au problème posé, en visualisant rapidement les différentes allures des diagrammes associés.

# Classe de troisième

## Programme

| Contenus | Compétences exigibles | Commentaires |
|---|---|---|
| Caractéristiques de position d'une série statistique. | Une série statistique étant donnée (sous forme de liste ou de tableau, ou par une représentation graphique), proposer une valeur médiane de cette série et en donner la signification. | Il s'agit essentiellement d'une part, de faire acquérir aux élèves les premiers outils de comparaison de séries statistiques, d'autre part de les habituer à avoir une attitude de lecteurs responsables face aux informations de nature statistique.<br>On repère, en utilisant effectifs ou fréquences cumulés, à partir de quelle valeur du caractère on peut être assuré que la moitié de l'effectif est englobée. Les exemples ne devront soulever aucune difficulté au sujet de la détermination de la valeur de la médiane. |
| Approche de caractéristiques de dispersion d'une série statistique.<br>Initiation à l'utilisation de tableurs-grapheurs en statistiques. | Une série statistique étant donnée, déterminer son étendue ou celle d'une partie donnée de cette série. | L'étude de séries statistiques ayant même moyenne permettra l'approche de la notion de dispersion avant toute introduction d'indice de dispersion. On introduira l'étendue de la série ou de la partie de la série obtenue après élimination de valeurs extrêmes.<br>On pourra ainsi aborder la comparaison de deux séries en calculant quelques caractéristiques de position et de dispersion, ou en interprétant des représentations graphiques données.<br>Les tableurs que l'on peut utiliser sur tous les types d'ordinateurs permettent, notamment en liaison avec l'enseignement de la technologie, d'appliquer de manière rapide à des données statistiques les traitements étudiés. |

## Document d'accompagnement des programmes de troisième

Le contenu et les commentaires du programme concernant la statistique constituent un prolongement de ceux des classes antérieures, l'objectif de l'enseignement de statistique descriptive au collège étant indiqué dans le document d'accompagnement des programmes du cycle central.

En classe de 3$^{e}$, il s'agit d'aider les élèves à franchir une nouvelle étape dans le développement de leur autonomie de jugement à propos d'informations qui peuvent être nombreuses. Dans le cas d'un regroupement en classes, les choix effectués peuvent avoir des effets sur les résultats numériques ou les représentations graphiques et leurs interprétations.

En classe de 4$^{e}$, on a pu observer que « la moyenne d'une population dont les éléments sont rangés par ordre croissant ne sépare pas ceux-ci, en général, en deux parties de même effectif », ce qui justifie l'introduction de la médiane en classe de 3$^{e}$. Les élèves disposent alors de deux indicateurs de la tendance centrale d'une population, leur position relative pouvant faire l'objet d'une interprétation dans des situations appropriées.

La nécessité de distinguer deux séries statistiques de même tendance centrale justifie l'intérêt de la notion de dispersion. Dans ce premier contact, le programme se limite à l'étendue d'une série statistique ou à l'étendue d'une partie donnée de celle-ci; cela permet, sans difficulté technique, de familiariser les élèves avec une démarche habituelle en statistique : procéder à une synthèse de l'information sous la forme de nombres mesurant respectivement la position et la dispersion de la série étudiée.

Choix de la représentation d'une série statistique, interprétation des résultats obtenus et comparaison de deux séries statistiques peuvent être conduits, sans répétitions inutiles ni pertes de temps, en utilisant des tableurs-grapheurs ou en répartissant le travail au sein de la classe. De plus, outre son intérêt spécifique, l'enseignement des statistiques contribue au développement des compétences en mathématiques, notamment celles liées au calcul et à la construction, la lecture et l'utilisation de graphiques; toutes les capacités correspondantes peuvent être mises en œuvre au cours d'activités interdisciplinaires.

# Statistiques et probabilités au lycée

## Classe de seconde

### Programme

En seconde le travail sera centré sur :

- la réflexion conduisant au choix de résumés numériques d'une série statistique quantitative ;
- la notion de fluctuation d'échantillonnage vue ici sous l'aspect élémentaire de la variabilité de la distribution des fréquences ;
- la simulation à l'aide du générateur aléatoire d'une calculatrice. La simulation remplaçant l'expérimentation permet, avec une grande économie de moyens, d'observer des résultats associés à la réalisation d'un très grand nombre d'expériences. On verra ici la diversité des situations simulables à partir d'une liste de chiffres.

| Contenus | Capacités attendues | Commentaires |
|---|---|---|
| Résumé numérique par une ou plusieurs mesures de tendance centrale (moyenne, médiane, classe modale, moyenne élaguée) et une mesure de dispersion (on se restreindra en classe de seconde à l'étendue). | Utiliser les propriétés de linéarité de la moyenne d'une série statistique.<br>Calculer la moyenne d'une série à partir des moyennes de sous-groupes.<br>Calcul de la moyenne à partir de la distribution des fréquences. | L'objectif est de faire réfléchir les élèves sur la nature des données traitées, et de s'appuyer sur des représentations graphiques pour justifier un choix de résumé.<br>On peut commencer à utiliser le symbole Σ.<br>On commentera quelques cas où la médiane et la moyenne diffèrent sensiblement.<br>On remarquera que la médiane d'une série ne peut se déduire de la médiane de sous séries. Le calcul de la médiane nécessite de trier les données, ce qui pose de problèmes de nature algorithmique. |
| Définition de la distribution des fréquences d'une série prenant un petit nombre de valeurs et de la fréquence d'un événement.<br>Simulation et fluctuation d'échantillonnage. | Concevoir et mettre en oeuvre des simulations simples à partir d'échantillons de chiffres au hasard. | La touche "random" d'une calculatrice pourra être présentée comme une procédure qui, chaque fois qu'on l'actionne, fournit une liste de n chiffres (composant la partie décimale du nombre affiché). Si on appelle la procédure un très grand nombre de fois, la suite produite sera sans ordre ni périodicité et les fréquences des dix chiffres seront sensiblement égales.<br>Chaque élève produira des simulations de taille n (n allant de 10 à 100 suivant les cas) à partir de sa calculatrice ; ces simulations pourront être regroupées en une simulation ou plusieurs simulations de taille N, après avoir constaté la variabilité des résultats de chacune d'elles. L'enseignant pourra alors éventuellement donner les résultats de simulation de même taille N préparées à l'avance et obtenues à partir de simulations sur ordinateurs. |

L'enseignant traitera des données en nombre suffisant pour que cela justifie une étude statistique ; il proposera des sujets d'étude et des simulations en fonction de l'intérêt des élèves, de l'actualité et de ses goûts.

La notion de fluctuation d'échantillonnage et de simulation ne doit pas faire l'objet d'un cours. L'élève pourra se faire un "cahier de statistique " où il consignera une grande partie des traitements de données et des expériences de simulation qu'il fait, des raisons qui conduisent à faire des simulations ou traiter des données, l'observation et la synthèse de ses propres expériences et de celles de sa classe. Ce cahier sera complété en première et terminale et pourra faire partie des procédures d'évaluation annuelle.

En classe de première et de terminale, dans toutes les filières, on réfléchira sur la synthèse des données à l'aide du couple moyenne, écart-type qui sera vu à propos de phénomènes aléatoires gaussiens et par moyenne ou médiane et intervalle inter-quartile sinon. On amorcera une réflexion sur le problème de recueil des données et sur la notion de preuve statistique ; on fera un lien entre statistique et probabilité.

L'enseignement de la statistique sera présent dans toutes les filières mais sous des formes diverses.

# Ajout au Programme de la classe de seconde (extrait)

BO HS n°6 du 12 août 1999 - Thèmes d'étude

*Pour chacun des chapitres, le professeur choisira, pour l'ensemble des élèves ou pour certains seulement en fonction de leurs centres d'intérêt, un ou plusieurs thèmes d'étude dans la liste ci-dessous.*

## Statistique

- Simulations d'un sondage ; à l'issue de nombreuses simulations, pour des échantillons de taille variable, on pourra introduire la notion de fourchette de sondage, sans justification théorique. La notion de niveau de confiance 0,95 de la fourchette peut être introduite en terme de "chances" (il y a 95 chances sur 100 pour que la fourchette contienne la proportion que l'on cherche à estimer); on pourra utiliser les formules des fourchettes aux niveaux 0,95, 0,90 et 0,99 pour une proportion observée voisine de 0,5 afin de voir qu'on perd en précision ce qu'on gagne en niveau de confiance. On incitera les élèves à connaître l'approximation usuelle de la fourchette au niveau de confiance 0,95, issue d'un sondage sur n individus ($n > 30$) dans le cas où la proportion observée p est comprise entre 0,3 et 0,7, à savoir : $[\, p - 1/\sqrt{n} \,;\, p + 1/\sqrt{n} \,]$.

- Simulations de jeux de pile ou face : distribution de fréquences du nombre maximum de coups consécutifs égaux dans une simulation de 100 ou 200 lancers de pièce équilibrée; distribution de fréquences du gain sur un jeu d'au plus dix parties où on joue en doublant la mise (ou en la triplant) tant qu'on n'a pas gagné. On pourra aussi faire directement l'expérience avec des pièces pour bien faire sentir la notion de simulation..

- Simulations du lancer de deux dés identiques et distribution de la somme des faces. On pourra aussi faire directement l'expérience avec des dés pour bien faire sentir la notion de simulation.... - Simulations de promenades aléatoires sur des solides ou des lignes polygonales, fluctuation du temps et estimation du temps moyen mis pour traverser un cube ou pour aller d'un sommet donné à un autre sommet donné d'une ligne polygonale.

- Simulations de naissances : distribution du nombre d'enfants par famille d'au plus quatre enfants lorsqu'on s'arrête au premier garçon, en admettant que pour chaque naissance, il y a autant de chances que ce soit un garçon ou une fille.

# Document d'accompagnement du programme de la classe de seconde (extrait)

*On trouvera à la suite de ce document d'accompagnement des fiches sur le programme de statistique de la classe de seconde et sur les thèmes associés.*

Les choix, traduits en termes de programme pour la classe de seconde, sont guidés par les perspectives suivantes pour le lycée :

- acquérir une expérience de l'aléatoire et ouvrir le champ du questionnement statistique ;
- voir dans un cas simple ce qu'est un modèle probabiliste et aborder le calcul des probabilités.

Au collège, les élèves se sont familiarisés avec les phénomènes variables et ont appris des éléments du langage graphique (représentations diverses, "camemberts", diagrammes en bâtons) qui permettent de visualiser une série de données expérimentales ; par ailleurs, ils ont travaillé sur la notion de moyenne arithmétique.

En seconde, différents éléments apparaissent au programme :

## La fluctuation d'échantillonnage

Nous appellerons échantillon de taille *n* d'une expérience la série des résultats obtenus en réalisant *n* fois cette expérience ; on dira aussi qu'un échantillon est une liste de résultats de *n* expériences identiques et indépendantes ; on se limite en seconde aux échantillons d'expériences ayant un nombre fini d'issues possibles. La distribution des fréquences associée à un échantillon est le *vecteur* dont les composantes sont les fréquences des issues dans l'échantillon ; on ne donnera pas de définition générale de la notion de distribution des fréquences, on se contentera de la définir comme liste des fréquences dans chacune des situations que l'on traitera. Les distributions des fréquences varient d'un échantillon à l'autre d'une même expérience : c'est ce qu'on appellera en classe de seconde la fluctuation d'échantillonnage.

Aborder la notion de fluctuation d'échantillonnage se fera en premier lieu dans des cas simples (lancers de dés, de pièces), où la notion d'expériences identiques et indépendantes est intuitive et ne pose pas de problème ; l'élève reprendra ainsi contact avec des expériences aléatoires familières (lancer de dés équilibrés) et les enrichira. Historiquement, l'honnête homme du XVII$^{e}$ siècle s'est familiarisé à l'aléatoire en pratiquant les jeux de hasard ; maintenant, les calculatrices et les ordinateurs permettent la production aisée de listes de chiffres au hasard ; la production de telles listes fera partie, à coté des lancers de dés ou de pièces équilibrés, à côté de tirage de boules dans des urnes, du bagage d'expériences de référence de l'élève. L'étude de ces expériences de référence sera ainsi à la base de la formation sur l'aléatoire des élèves.

L'esprit statistique naît lorsqu'on prend conscience de l'existence de fluctuation d'échantillonnage ; en seconde, l'élève constatera expérimentalement qu'entre deux échantillons, de même taille ou non, les distributions des fréquences fluctuent ; la moyenne étant la moyenne pondérée des composantes de la distribution des fréquences est, elle aussi, soumise à fluctuation d'échantillonnage ; il en est de même de la médiane. On observera aussi que l'ampleur des fluctuations des distributions de fréquences calculées sur des échantillons de taille *n* diminue lorsque *n* augmente. Par ailleurs, on n'hésitera pas à parler de la fréquence d'un événement ("le nombre observé est pair", "le nombre est un multiple de trois", etc.) sans pour autant définir formellement ce qu'est un événement, ni donner de formules permettant le calcul automatique de la fréquence de la réunion ou de l'intersection de deux événements.

Le choix pédagogique est ici d'aller de l'observation vers la conceptualisation et non d'introduire d'abord le langage probabiliste pour constater ensuite que tout se passe comme le prévoit cette théorie.

## Simulation

Formellement, simuler une expérience, c'est choisir un modèle de cette expérience puis simuler ce modèle : cet aspect sera introduit ultérieurement en première. Dans le cadre du programme de seconde, simuler une expérience consistera à produire une liste de résultats

que l'on pourra assimiler à un échantillon de cette expérience (voir plus loin la fiche *listes de chiffres au hasard*).

On se contentera de simuler des situations très simples, reposant le plus souvent sur la simulation d'expériences de références où toutes les issues ont des chances égales d'apparaître.

La simulation permettra de disposer d'échantillons de grande taille et d'observer des phénomènes appelant une explication dans le champ des mathématiques. Pour bien comprendre les mathématiques, il est utile d'apprendre quel type de questions sont à adresser à cette discipline et aussi d'apprendre à reformuler ces questions dans le langage propre des mathématiques ; le langage des probabilités présenté en première S, ES et en option de première L, formalisera le langage naïf des *chances* et du *hasard* employé en seconde ; le calcul des probabilités permettra ensuite d'expliquer certains phénomènes observés.

En seconde, on approche dans le cadre d'un langage simple et familier les techniques de simulation ; pour que l'élève ne soit pas écrasé par la puissance des outils modernes de simulation, il convient qu'il ait établi un lien concret entre l'expérience et sa simulation : certaines expériences simples pourront être réalisées par une partie de la classe et simulées par le reste de la classe ; il n'est pas nécessaire, dans un premier temps, de lier les premiers pas vers la simulation de l'aléatoire à l'introduction de concepts théoriques difficiles tel celui de modèle.

## Statistique descriptive

Le programme comporte quelques éléments sur les résumés numériques de séries statistiques, déjà travaillés au collège ; il s'agit essentiellement d'entretenir les acquis, de les réinvestir dans certains thèmes et/ou à l'occasion de certains événements que pourrait offrir l'actualité.

La statistique donne lieu à de nombreuses activités numériques et favorise la maîtrise du calcul ; cependant, de tels calculs ne doivent être demandés que dans la mesure où ils permettent aux élèves de mieux comprendre la spécificité de la série statistique en jeu. Estimer la moyenne de séries de données quantitatives en les regroupant par classe n'est plus une pratique utile en statistique depuis que des ordinateurs calculent la moyenne de milliers de données en une fraction de seconde ; par contre savoir calculer une moyenne à partir de moyennes des sous-groupes ou comprendre la linéarité de la moyenne peut donner lieu à des exercices pertinents au regard de la pratique de la statistique. Calculer simplement, à partir de la moyenne, la moyenne élaguée d'une ou plusieurs valeurs extrêmes montre l'influence d'éventuelles valeurs aberrantes.

## Cahier de statistique

Les élèves pourraient commencer en seconde un cahier de statistique rendant compte des expériences faites ou simulées, en classe ou chez eux, à la demande de l'enseignant ou de leur propre initiative. La rédaction d'un tel document individuel leur permettrait d'organiser et de planifier les expériences et les simulations, de donner forme à la conclusion qu'ils en tirent, aux questions théoriques qui se sont posées et qu'ils pourront reprendre ultérieurement. La tenue de ce cahier pourrait contribuer efficacement à structurer le travail expérimental proposé et aider ultérieurement chaque élève à mieux expliciter le lien entre l'expérience et la théorie ; cela permettrait à l'enseignant de contrôler la qualité des travaux réalisés, de vérifier que ne s'installe pas des perceptions erronées sur les phénomènes aléatoires, de faire des évaluations sur la partie statistique du programme. Ce cahier pourrait être continué en première et terminale : l'enseignant de première pourrait ainsi savoir quels thèmes ont été travaillés par ses élèves en seconde.

La production d'un texte écrit est en soi un élément formateur ; un tel cahier, où se mêlent texte écrit et représentations graphiques, présentant des éléments narratifs et des argumentations, s'inscrit de plus dans le cadre du nouveau programme de français des élèves de seconde.

# Classe de Première ES

## Programme

BO HS n°7 du 31 août 2000

| Contenus | Modalités de mise en œuvre | Commentaires |
|---|---|---|
| **Statistique**<br>Étude de séries de données:<br>- nature des données (effectifs, données, moyennes, indices, pourcentages,…);<br>- lissage par moyennes mobiles; | On s'intéressera en particulier aux séries chronologiques.<br>On effectuera à l'aide d'un tableur le lissage par moyennes mobiles et on observera directement son effet sur la courbe représentant la série. | |
| - histogrammes à pas non constants;<br>- diagrammes en boîte. | Les histogrammes à pas non constants ne seront pas développés pour eux mêmes, mais le regroupement en classes inégales s'imposera lors de l'étude d'exemples comme des pyramides des âges ou de salaires.<br>On apprendra à interpréter diverses formes de diagrammes en boîtes à partir d'exemples.<br>En liaison avec le paragraphe "probabilité", on étudiera plusieurs séries obtenues par simulation d'un modèle; on comparera les diagrammes en boîte. L'utilisation d'un logiciel informatique est indispensable pour accéder à une simulation sur un nombre important d'expériences. | Sans développer de technicité particulière à propos des histogrammes à pas non constants, on montrera l'intérêt d'une représentation pour laquelle l'aire est proportionnelle à l'effectif. |
| Effet de structure lors du calcul de moyennes.<br>Mesures de dispersion: intervalle interquartile, écart-type. | On observera dynamiquement et en temps réel, les effets des modifications des données. | L'objectif est de résumer une série par un couple (mesure de tendance centrale; mesure de dispersion). Deux choix usuels sont couramment proposés: le couple (médiane ; intervalle interquartile), *robuste* par rapport aux valeurs extrêmes de la série, et le couple (moyenne ; écart-type). On démontrera que la moyenne est le réel qui minimise $\Sigma(xi - x)^2$ alors qu'elle ne minimise pas $\Sigma\|xi - x\|$.<br>On notera s l'écart-type d'une série, plutôt que σ, réservé à l'écart-type d'une loi de probabilité. |
| Tableau à double entrée : étude fréquentielle<br>lien entre arbre et tableau à double entrée; notion de fréquence de $A$ sachant $B$. | | La fréquence de $A$ sachant $B$ sera notée $f_B(A)$ ; elle prépare à la notion de probabilité conditionnelle qui sera traitée en terminale. |

| Contenus | Modalités de mise en œuvre | Commentaires |
|---|---|---|
| **Probabilités**<br>Définition d'une loi de probabilité sur un ensemble fini. Probabilité d'un événement, de la réunion et de l'intersection d'événements.<br>Modélisation d'expériences de référence menant à l'équiprobabilité; utilisation de modèles définis à partir de fréquences observées. | Le lien entre loi de probabilité et distribution de fréquences sera éclairé par un énoncé vulgarisé de la loi des grands nombres.<br>On mènera de pair simulation et étude théorique de la somme de deux dés (en liaison avec le paragraphe précédent). | Un énoncé vulgarisé de la loi des grands nombres peut être par exemple:<br>*Pour une expérience donnée, dans le modèle défini par une loi de probabilité P, les distributions des fréquences obtenues sur des séries de taille n se rapprochent de P quand n devient grand.*<br>On indiquera que simuler une expérience consiste à simuler un modèle de cette expérience.<br>On pourra ne pas se limiter à l'étude d'une seule situation et envisager d'autres expériences (produit de deux dés, somme de trois dés…).<br>On pourra repérer les difficultés soulevées par le choix d'un modèle mais sans s'y attarder : on utilisera directement des modèles que la statistique a permis de choisir. |

# Document d'accompagnement programme de la classe de Première ES (extrait)

**A propos du titre "Traitement des données et probabilités"**

Le choix a été fait pour la section ES de donner un rôle important aux séries chronologiques, particulièrement fréquentes dans les cours d'économie de cette section ; mais on veillera à ce que les questions et exemples traités conduisent jusqu'à une réflexion conceptuelle ou axiomatisée, ce qui constitue une bonne préparation à d'éventuelles études ultérieures davantage centrées sur des pratiques professionnelles.

## a) Pourcentages

Comme indiqué dans le programme, il ne s'agit pas d'aborder ici quelque connaissance technique nouvelle, mais d'entretenir un apprentissage de base indispensable pour lire correctement et de façon critique l'information chiffrée. Ce paragraphe recoupe de nombreux autres titres du programme : statistique (fréquences, données en pourcentage, évolution de séries chronologiques,…), suites géométriques (obtenues par augmentations successives), dérivation (approximation affine), … Les notions qu'il décrit devront donc être régulièrement mises en jeu, en particulier à partir de données issues des médias ou de l'environnement scolaire de la classe.

## b) Nature des données

On a parfois prôné, pour l'enseignement de la statistique, le recueil de données par les élèves eux-mêmes, cette pratique étant considérée comme motivante et permettant de percevoir le champ de l'aléatoire. Or la perception de l'aléatoire peut s'acquérir de manière plus profonde par la simulation. Par ailleurs, pour de nombreuses questions que l'on peut se poser dans le champ scolaire, des données existent : elles sont réactualisées chaque année, leur contenu est riche et elles sont accessibles dans des banques de données. Sans exclure complètement un recueil ponctuel de données par les élèves à condition qu'il ne prenne pas beaucoup de temps, on s'appuiera avant tout sur des données existantes. Certaines données sont des données brutes (exemple : série des hauteurs d'un fleuve mesurées en un point géographique précis tous les jours à la même heure) ; d'autres sont obtenues en prenant des moyennes de mesures brutes (séries des températures mensuelles en un point géographique précis, évolution sur une période de 10 ans des dépenses de logement suivant les catégories socioprofessionnelles) ; certaines encore sont des moyennes mobiles : par exemple, dans la présentation des mesures de température quotidienne à une heure donnée en un point

donné, on remplace parfois les températures brutes $x_0,\ldots x_n$ par la série des moyennes mobiles d'ordre k, $y_k,\ldots y_{n-k}$, calculées ainsi : $y_i = \frac{1}{2k+1}\sum_{i-k}^{i+k} x_j$ , pour i=k,…,n-k.

On pourra montrer sur des exemples, avec k=2 ou 3, que la courbe obtenue en remplaçant les données brutes par les moyennes mobiles est plus lisse.

## c) Effets de structure

### Exemple

Le revenu moyen global des individus actifs d'une population (par exemple la population parisienne) peut augmenter avec le temps alors que dans toutes les catégories socioprofessionnelles (CSP) le revenu baisse, l'augmentation globale étant liée à un changement de la répartition en CSP (à Paris, les CSP à faible revenu ont eu tendance à déménager en banlieue). La structure à une date donnée est ici la répartition des CSP à cette date.

## d) Diagrammes en boîtes

Il ne s'agit pas là d'un élément à part ; il sera introduit à l'occasion du traitement de données expérimentales ou d'activités de simulation. On trouvera en annexe de ce document une note sur ce type de diagrammes à l'usage des enseignants.

## e) Étude fréquentielle de tableaux à double entrée

Les commentaires sur les pourcentages des lignes (resp. des colonnes) se feront simplement à partir des distributions de fréquences associées aux marges horizontales (resp. verticales). On ne construira pas les « tableaux théoriques » (on n'introduira pas de façon formelle la notion de sur- et sous-représentation, celles-ci n'ayant vraiment de sens que si elles sont « significatives » au sens statistique).

### Exemple

En 1979, le New-York Times a noté qu'entre 1973 et 1979, en Floride, il a été prononcé 131 peines capitales pour meurtre ; parmi ces condamnés, 55% étaient des Blancs alors que 48 % de ceux qui ont été jugés pour meurtre pendant la même période étaient des Blancs. Qu'en penser ?

Avant d'en penser quoique ce soit, il est opportun de s'intéresser aussi à la couleur de la victime.

1- Entre 1973 et 1979, il y a eu 2433 meurtres en Floride, dont la victime était de couleur blanche. Le tableau ci-dessous donne les détails suivant la couleur de la peau du suspect (blanche, B, ou non blanche, NB), les colonnes indiquant la sentence (PC : peine capitale ; AS : autre sentence)

| | PC | AS | Totaux |
|---|---|---|---|
| NB | 48 | 239 | 287 |
| B | 72 | 2074 | 2146 |
| Totaux | 120 | 2313 | 2433 |

- Dire comment est construit le tableau ci-dessous et ce que signifient les trois lignes :

| | PC | AS | Totaux |
|---|---|---|---|
| NB | 16,7 | 83,3 | 100,0 |
| B | 3,4 | 96,6 | 100,0 |
| Totaux | 4,9 | 95,1 | 100,0 |

Remarque : les différences de pourcentages de condamnation à la peine capitale suivant la couleur du suspect, lorsque la victime est de couleur blanche, est trop grande pour être imputée à la fluctuation d'échantillonnage.

2- Entre 1973 et 1979, il y a eu 4764 meurtres en Floride. Le tableau ci-dessous donne les détails suivant la couleur de peau du suspect (blanche, B, ou non blanche, NB), les colonnes indiquant la sentence (PC : peine capitale ; AS : autre sentence).

| | PC | AS | Totaux |
|---|---|---|---|
| NB | 59 | 2448 | 2507 |
| B | 72 | 2185 | 2257 |
| Totaux | 131 | 4633 | 4764 |

- Construire un tableau analogue au deuxième tableau de la question 1 et le commenter.

- Quel est le pourcentage de suspects de couleur blanche lorsque la victime est de couleur blanche ; quel est le pourcentage de suspects de couleur blanche lorsque la victime n'est pas de couleur blanche ?

- Faire l'étude lorsque la victime n'est pas de couleur blanche.

*Remarque : les différences de pourcentages de condamnation à la peine capitale suivant la couleur du suspect, sans tenir compte de la couleur de la victime, ou lorsque la victime est non blanche, peuvent être imputées à la fluctuation d'échantillonnage, i.e. ne peuvent pas être interprétées comme une tendance des jugements en Floride.*

### Exemple

Que signifient les tableaux ci-dessous et comment sont-ils déduits du premier tableau ?

Les données concernent la répartition suivant le sexe et le poste (AG : agent spécial , P : personnel autre ) des emplois au FBI au 31 janvier 1997.

| | H | F | |
|---|---|---|---|
| AS | 9199 | 1617 | 10816 |
| P | 4535 | 10165 | 14700 |
| | 13734 | 11782 | 25516 |

| | H | F | |
|---|---|---|---|
| AS | 36,1 | 6,3 | 42,4 |
| P | 17,8 | 39,8 | 57,6 |
| | 53,8 | 46,2 | 100,0 |

| | H | F | |
|---|---|---|---|
| AS | 85,0 | 15,0 | 100,0 |
| P | 30,9 | 69,1 | 100,0 |
| | 53,8 | 46,2 | 100,0 |

| | H | F | |
|---|---|---|---|
| AS | 67,0 | 13,7 | 42,4 |
| P | 33,0 | 86,3 | 57,6 |
| | 100,0 | 100,0 | 100,0 |

### Remarque

On distinguera les calculs de l'interprétation, dans le contexte, des données d'observation.

## f) Loi de probabilité

On recensera les propriétés mathématiques élémentaires de l'objet « distributions de fréquences » (cf. tableau ci-dessous) et on définira une loi de probabilité comme un objet mathématique ayant les mêmes propriétés.

| Distribution de fréquences sur $E=\{x_1,\ldots,x_r\}$ | Loi de probabilité sur $E=\{x_1,\ldots,x_r\}$ |
|---|---|
| $(f_1,\ldots,f_r)$<br>$f_i \geq 0$ ; $\sum f_i = 1$<br><br>$A \subset E$ Fréquence de A : $f(A)=\sum f_i$<br>Événement complémentaire : $f(\overline{A})=1-f(A)$<br><br>*Cas numérique :*<br>*Moyenne empirique :* $\overline{x}=\sum f_i x_i$ | $(p_1,\ldots, p_r)$<br>$p_i \geq 0$ ; $\sum p_i = 1$<br><br>$A \subset E$ probabilité de A : $P(A)=\sum p_i$<br>Événement complémentaire : $P(\overline{A})=1-P(A)$<br><br>*Cas numérique :*<br>*Espérance d'une loi P :* $\mu=\sum p_i x_i$ |

Modéliser une expérience aléatoire, c'est lui associer une loi de probabilité (qui est un objet du monde mathématique).

Une fréquence est empirique : elle est calculée à partir de données expérimentales, alors que la probabilité d'un événement est un « nombre théorique » (un objet du monde mathématique). Les distributions de fréquences issues de la répétition d'expériences identiques et indépendantes varient (fluctuent), la loi de probabilité est un invariant associé à l'expérience.

L'objectif est que les élèves comprennent à l'aide d'exemples (cf. paragraphe sur les expériences de référence) que modéliser, c'est ici choisir une loi de probabilité. Il ne s'agit cependant en aucun cas de tenir des discours généraux sur les modèles et la modélisation.

Les élèves devront bien distinguer ce qui est empirique (du domaine de l'expérience) de ce qui est théorique ; en particulier, on réservera la lettre grecque $\sigma$ à l'écart-type d'une loi et on évitera de noter avec cette même lettre un écart-type empirique (il s'agit là de règles de notation internationales) .

En classe de première, une loi de probabilité P sur un ensemble fini est la liste des probabilités des éléments de E ; à partir de cette liste, on définit naturellement la probabilité d'événements, c'est-à-dire implicitement une application de $\mathcal{P}(E)$ dans [0,1], application qui sera encore désignée par P .

Il est inutilement complexe, pour le cas des ensembles finis, de partir d'une application de $\mathcal{P}(E)$ dans [0,1], vérifiant certains axiomes, puis de montrer ensuite que cette application est entièrement caractérisée par $(p_1,\ldots, p_r)$. Le fait de ne pouvoir simplement généraliser cette définition aux ensembles continus, et la nécessité d'une définition ensembliste sera abordée en terminale.

Si tous les éléments d'un ensemble bien défini E ont même probabilité, celle-ci est dite équirépartie ; on parlera aussi d'équiprobabilité et on dira que les éléments de l'ensemble E sont choisis au hasard (i.e. si on ne spécifie rien de plus, le choix au hasard est un choix avec équiprobabilité). Si la loi P n'est pas équirépartie, on parlera de choix d'éléments selon la loi P (ou éventuellement de choix au hasard selon la loi P).

Dans le cas de choix au hasard, la probabilité d'un événement est le quotient de son nombre d'éléments par le nombre d'éléments de l'ensemble.

On évitera tout développement théorique sur le langage des événements et le calcul ensembliste qui en découle : ces notions et la pratique de la logique qu'ils impliquent (étude du complémentaire de l'événement "A ou B", ou de l'événement "A et B") s'acquièrent au fil d'exercices.

## g) Modélisation d'expériences de référence

Modéliser une expérience aléatoire à valeurs dans un espace Ω, c'est choisir une loi de probabilité P définie sur Ω. Ce choix est en général délicat à faire, sauf dans certains cas où des considérations propres au protocole expérimental conduisent à proposer a priori un modèle. Il en est ainsi des lancers de pièces ou de dés pour lesquels des considérations de symétrie conduisent au choix d'un modèle où la loi de probabilité est équirépartie.

Sans faire une liste de conventions terminologiques, on indiquera clairement que les termes *équilibré* et *hasard* indiquent un choix du modèle de l'expérience, modèle où intervient quelque part une probabilité équirépartie.

## h) Modélisation à partir de fréquences.

En dehors des cas évoqués ci-dessus où des considérations quant à la nature des expériences permettent de proposer un modèle, le choix d'un modèle à partir de données expérimentales est beaucoup plus délicat. On se contentera, pour certains exercices, de fournir un modèle en indiquant dans un premier temps que des techniques statistiques ont permis de déterminer et de valider un tel modèle à partir de nombreuses données expérimentales.

Pour déterminer et/ou valider un modèle probabiliste, le premier outil dont on dispose est un théorème de mathématiques appelé loi des grands nombres, dont un énoncé intuitif est :

*Dans le monde théorique défini par une loi P sur un ensemble Ω, les fréquences des éléments de Ω dans une suite de n expériences identiques et indépendantes tendent vers leur probabilité quand n augmente indéfiniment.*

On donnera des exemples où un modèle est déterminé à partir de fréquences ; les exemples les plus compliqués que l'on abordera consistent à associer une loi de probabilité à un tableau à double entrée.

## i) Simulation

L'exemple ci-dessous montre comment mêler diverses composantes d'un travail mathématique : observation, premières conjectures, expérimentation à plus grande échelle, puis obtention et preuve de certains résultats :

### Exemple

**Somme de deux dés.**

En répétant 100 fois de suite le lancer de deux dés, on **observe** que certains résultats s'obtiennent plus souvent que d'autres.

Á l'aide d'un tableur par exemple, il est possible d'**expérimenter** à plus grande échelle : simulation d'un plus grand nombre de lancers de deux dés et construction du tableau des effectifs : l'inégale répartition des fréquences de chaque résultat est flagrante

La recherche d'un modèle théorique adapté avec une loi de probabilité équirépartie permet ensuite calculs et démonstrations : on **prouve** que les résultats sont inégalement probables et on détermine précisément leur probabilité.

Résultats pour 5 séries de 100 lancers

| | | | | | |
|---|---|---|---|---|---|
| 2 | 0,02 | 0,02 | 0,01 | 0,03 | 0,01 |
| 3 | 0,03 | 0,02 | 0,09 | 0,06 | 0,04 |
| 4 | 0,08 | 0,08 | 0,08 | 0,08 | 0,1 |
| 5 | 0,11 | 0,11 | 0,12 | 0,09 | 0,11 |
| 6 | 0,17 | 0,13 | 0,1 | 0,11 | 0,11 |
| 7 | 0,15 | 0,15 | 0,12 | 0,14 | 0,2 |
| 8 | 0,12 | 0,2 | 0,1 | 0,13 | 0,22 |
| 9 | 0,15 | 0,1 | 0,12 | 0,13 | 0,06 |
| 10 | 0,09 | 0,1 | 0,14 | 0,1 | 0,08 |
| 11 | 0,06 | 0,05 | 0,06 | 0,09 | 0,06 |
| 12 | 0,02 | 0,04 | 0,06 | 0,04 | 0,01 |

# Classe de Première L
# Mathématiques – informatique ; enseignement obligatoire

## Programme - Extrait : 2- Statistique

BO HS n°7 du 31 août 2000

En seconde, les élèves ont abordé les notions de fluctuation d'échantillonnage et de simulation. On va maintenant définir de nouveaux paramètres à associer à une série de données numériques ; pour l'interprétation des valeurs de ces paramètres, on gardera à l'esprit qu'ils fluctuent d'une série de données à une autre.

L'objectif de ce chapitre est :

- de familiariser les élèves avec des questions de nature statistique ;
- de montrer, à travers la notion de phénomènes gaussiens, la nature de l'information prévisionnelle apportée par un écart-type ;
- d'étudier des tableaux de pourcentages.

| Contenus | Commentaires |
|---|---|
| **Diagrammes en boîtes**<br>Intervalle inter-quartile<br>Définition de l'intervalle interquartile.<br>Construction de diagrammes en boîtes (aussi appelés *boîtes à moustaches* ou *boîtes à pattes*). | On étudiera des données recueillies par les élèves, tout en choisissant des situations permettant de limiter le temps de recueil de ces données.<br>À cette occasion, on s'attachera à :<br>- définir une problématique ou une question précise motivant un recueil de données expérimentales,<br>- définir les données à recueillir, leur codage et les traitements statistiques qu'on appliquera pour avoir des éléments de réponses à la question posée,<br>- élaborer un protocole de recueil et aborder les problèmes que cela pose.<br>Proposition d'exemples : battements cardiaques, estimation de longueurs, durée des repas du soir, nombre et durée de conversations téléphoniques, temps de passage en caisse dans une grande surface, etc. |
| **Variance, écart-type**<br>Introduction de l'écart-type pour des données gaussiennes. | L'objectif est ici de rendre les élèves capables de comprendre l'information apportée par la valeur de l'écart-type lors de mesures issues de la biologie ou du contrôle industriel.<br>On pourra prendre comme exemple de référence l'étude des courbes de taille et/ou de poids dans les carnets de santé des enfants, en se limitant éventuellement à des âges inférieurs à quatre ou six ans. |
| Définition de la plage de normalité pour un niveau de confiance donné. | On se limitera ici aux exemples de résultats fournis par les laboratoires biologiques lors de certains examens. Pour l'interprétation lorsque le niveau de confiance est 0,95, on notera que le choix de ce dernier résulte d'un consensus pour avoir des formules simples et implique qu'environ une personne sur vingt sorte de cette plage. |
| **Tableaux croisés**<br>Analyse d'un tableau de grands effectifs ; Construction et interprétation :<br>- des marges ;<br>- du tableau des pourcentages en divisant chaque cellule par la somme de toutes les cellules ;<br>- du tableau des pourcentages par ligne en divisant chaque cellule par la somme des cellules de la même ligne ;<br>- du tableau des pourcentages par colonnes en divisant chaque cellule par la somme des cellules de la même colonne. | On ne parlera pas des tableaux théoriques ou dits de proportionnalité ; les commentaires sur les pourcentages des lignes (resp. des colonnes) se feront simplement à partir des distributions de fréquences associées aux marges horizontales (resp. verticales).<br>On pourra prendre comme exemple de référence l'étude de résultats d'élection (classification selon les régions ou les classes d'âge des votes à une élection où plusieurs candidats sont en présence). |

# Document d'accompagnement Programme de la classe de Première L

**Extraits : Le chapitre statistique**

Ce chapitre se divise en trois parties. Dans la première, l'objectif est de voir sur des exemples qu'une question peut trouver une réponse dans le champ de la statistique sous réserve, éventuellement, de transformer la question initiale. À partir de cette nouvelle question, on réfléchira simultanément sur les données à recueillir et sur le traitement statistique que l'on peut envisager pour ces données.

## Exemples

Considérons la question : Quel est le nombre de battements cardiaques à la minute ?

La question est trop imprécise et il convient au moins de spécifier si c'est au repos ou après un effort clairement défini. Il peut se poser alors de nouvelles questions : sur la comparaison des données au repos ou après effort, par exemple. Les élèves peuvent aussi proposer que chacun étudie son propre pouls en faisant plusieurs mesures (il y a alors à la fois la variabilité individuelle de la fréquence cardiaque et les erreurs de mesure qui s'ajoutent) ou faire une étude sur une classe entière.

Considérons une autre question : Sait-on estimer à l'oeil une longueur ? Cette question est, elle aussi, trop imprécise. Au niveau de la population visée, s'adresse-t-on à des gens de tous âges ? De tous métiers ? Par ailleurs, que signifie «estimer une longueur» ? S'agit-il de petites longueurs ou de grandes distances ? Lorsque cette situation a été expérimentée dans des classes, la question initiale s'est transformée pour devenir par exemple : Si on demande à un élève de première de couper 20 cm d'une ficelle sans appareil de mesure, que se passe-t-il ? On peut alors mettre en place un protocole expérimental qui permettra d'observer des données liées à cette question.

L'objectif est donc ici de montrer la diversité des questions qui se posent ainsi que le soin nécessaire à la définition et au recueil des données. Il s'agit aussi de montrer aux élèves qu'une première expérience permet de préciser et de reformuler la question initiale et que, si l'on veut apporter des réponses, généraliser ce qui est fait et interpréter des différences, il faut faire un traitement statistique plus sophistiqué et tenir compte en particulier de la fluctuation d'échantillonnage. Le résumé des données observées pourra se faire à l'aide de diagrammes en boîtes (souvent appelés boîtes à moustaches ou boîtes à pattes), éventuellement accompagnés de la moyenne ou d'une moyenne élaguée.

On trouvera à la fin de ce document une annexe relative aux diagrammes en boîtes donnant toutes les indications nécessaires sur les paramètres utiles (médianes, quartiles), sur les modes de construction de ces diagrammes, ainsi que sur leur utilisation. L'essentiel, dans un tel diagramme, est la construction de la boîte contenant la moitié des valeurs de la série ; pour les « moustaches », on pourra choisir les premier et neuvième déciles ou les valeurs extrêmes, comme l'indique l'annexe citée : en première L, on privilégiera l'utilisation des valeurs extrêmes ; dans tous les cas, les élèves devront légender leur schéma. Le tableur servira avant tout ici à ordonner les valeurs de la série observée et éventuellement à les numéroter : les élèves effectueront ensuite «à la main» le calcul de la médiane et des quartiles ainsi que la construction de la boîte.

Dans la seconde partie, on pourra d'abord définir l'écart type d'une série. On remarquera que l'écart type et la moyenne sont sensibles aux valeurs extrêmes alors que la médiane et l'écart interquartile ne le sont pas. On travaillera ensuite selon l'esprit décrit dans l'annexe ci-dessous relative aux données gaussiennes.

La troisième partie est consacrée à l'étude de tableaux croisés. On trouvera dans le document d'accompagnement de première ES (page 12 et suivantes) quelques exemples d'études de tableaux à double entrée. On s'intéressera aussi à des situations pour lesquelles l'enseignant sait qu'il n'y a pas indépendance entre les deux caractères qualitatifs étudiés sur la population (tableaux présentés lors d'élections, par exemple). Cette partie pourrait

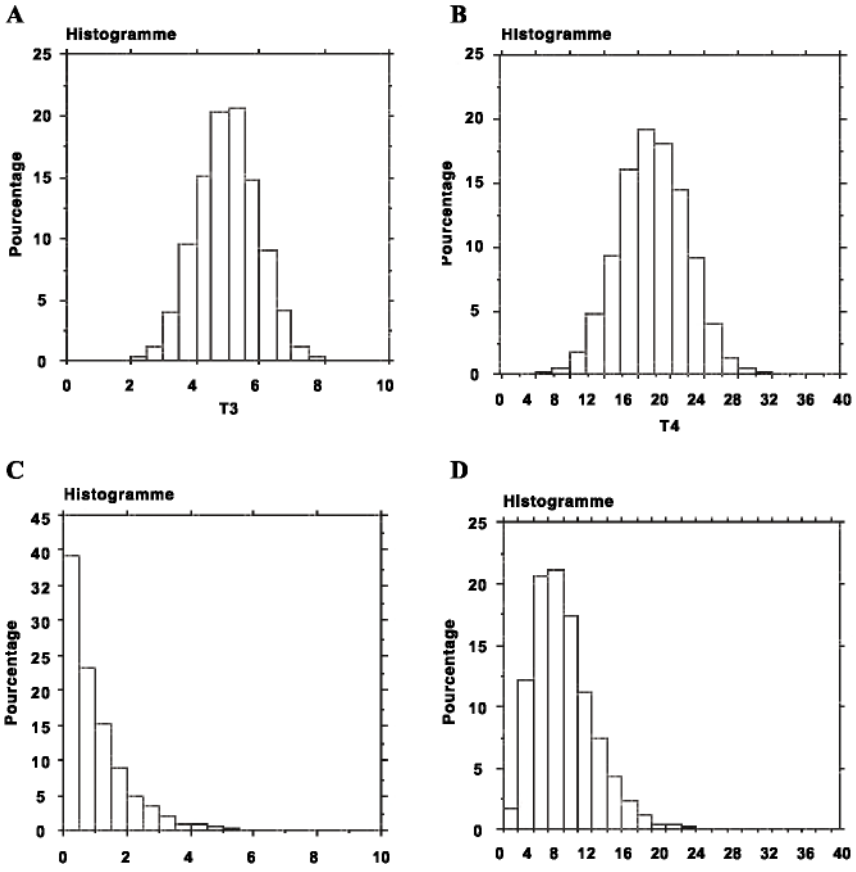
aussi bien figurer dans le chapitre « Information chiffrée» : elle a été mise dans le chapitre « Statistique » afin que l'enseignant puisse indiquer (sans le justifier) que les fluctuations des distributions des fréquences d'une ligne à l'autre (resp. d'une colonne à l'autre) sont éventuellement d'une ampleur que la fluctuation d'échantillonnage ne peut seule expliquer.

Le travail réalisé dans ce chapitre sera rédigé dans le cahier de statistique commencé en seconde.

# Annexe : à propos des données gaussiennes

## Exemple 1 - Bilan de santé

Pour faire le bilan de l'activité thyroïdienne d'un individu, on mesure les quantités de «T3 libre» (tri – iodo thyronine libre) et «T4 libre» (thyroxine libre); ces mesures sont exprimées en pmol/litre (pmol signifie pico-mole, soit $10^{-12}$ mole, une mole étant composée d'environ $6{,}02.10^{23}$ molécules). Les résultats de deux séries de 5 000 mesures chez des individus dont la thyroïde fonctionne normalement sont résumés par les histogrammes A et B; ces histogrammes ont la même allure dite en cloche, nettement différente de celle des histogrammes C et D (ces deux derniers correspondent à des données simulées).




Les deux séries obtenues lors de ces bilans thyroïdiens correspondent à des «données gaussiennes»; pour de telles données on peut déterminer un modèle (modèle gaussien) à partir de deux paramètres $m$ et $\sigma$ calculés sur une série de référence aussi longue que possible :

– la moyenne $m$ de la série de référence : $m = \frac{1}{n}\sum x_i$

– l'écart type de la série de référence : $\sigma = \sqrt{\frac{1}{n}\sum (x_i - m)^2}$

Pour les mesures de T4L, on trouve sur la série de référence (ici de taille 5 000) :

$m = 16{,}7$ et $\sigma = 4{,}0$ (les unités étant des pmol/litre).

Pour les mesures de T3L, on trouve sur la série de référence (ici de taille 5 000) :

$m = 4{,}9$ et $\sigma = 0{,}9$ (les unités étant des pmol/litre).

À partir de ce modèle, on montre que pour des mesures ultérieures de données analogues:

– environ 68 % des mesures sont dans l'intervalle $[\, m - \sigma \,;\, m + \sigma \,]$ ; environ 16 % seront inférieures à $(m - \sigma)$ et environ 16 % seront supérieures à $(m + \sigma)$ ;

– environ 95 % des mesures sont dans l'intervalle $[\, m - 2\sigma \,;\, m + 2\sigma \,]$ ; environ 2,5 % seront inférieures à $(m - 2\sigma)$ et environ 2,5 % seront supérieures à $(m + 2\sigma)$ ;

– environ 99,8 % des mesures sont dans l'intervalle $[\, m - 3\sigma \,;\, m + 3\sigma \,]$; environ 0,1% seront inférieures à $(m - 3\sigma)$ et environ 0,1% seront supérieures à $(m + 3\sigma)$.

Dans l'exemple des analyses biologiques de dosages de T3L et T4L, et dans de nombreux examens, l'intervalle $[\, m - 2\sigma \,;\, m + 2\sigma \,]$ est appelé « plage de normalité » : il contient environ 95 % des valeurs observées chez des individus non-malades. Les plages de normalité sont ici :

– [3,1; 6,7] pour le dosage de T3L;

– [9,7; 25,7] pour le dosage de T4L.

(Les plages de normalité sont indiquées par les laboratoires sur les comptes rendus d'analyse biologique. Ces plages de normalité sont voisines mais pas nécessairement identiques d'un laboratoire à l'autre; elles sont en effet calculées à partir de séries de référence traitées par l'appareil de mesure du laboratoire.)

Si on faisait des dosages de T3L chez des personnes choisies au hasard dans une population donnée, environ une sur vingt aurait une valeur sortant de la plage de normalité. Cela dit, les personnes à qui l'on fait de tels dosages (les individus de la série de référence mis à part) ne sont pas choisies au hasard et présentent en général des symptômes justifiant cet examen; sortir de la plage de normalité constitue un symptôme de plus en faveur d'une maladie de la thyroïde (symptôme d'autant plus marqué que l'on s'éloigne beaucoup de la moyenne : il est classique pour certaines pathologies de dépasser la moyenne de dix écarts types).

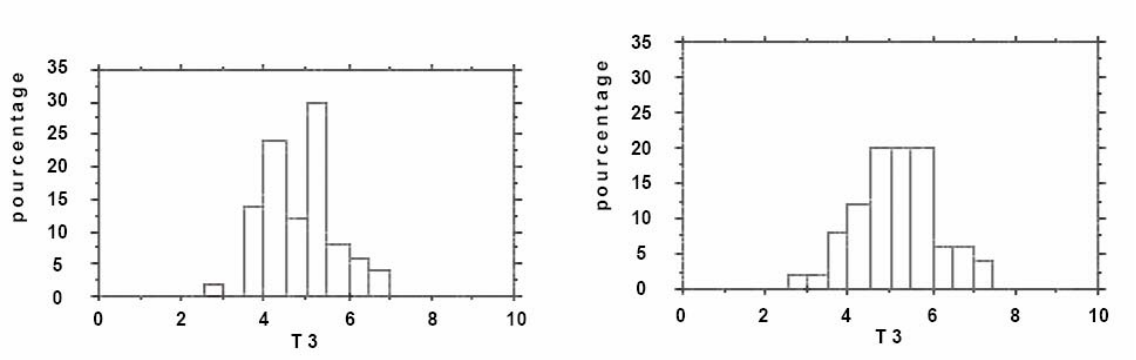
## Exemple 2 - Taille

Les tailles de garçons (resp. de filles) nés la même année constituent des données gaussiennes, et les courbes situées à la fin du carnet de santé des enfants donnent, entre autres, pour chaque âge une plage de normalité égale à $[\, m - 2\sigma \,;\, m + 2\sigma \,]$, où les paramètres $m$ et $\sigma$ pour un âge donné sont calculés sur des séries de référence (malheureusement, ces séries sont anciennes et ne sont plus vraiment des séries de référence pour les enfants qui naissent aujourd'hui). On notera que dans la population concernée par la série de référence, pour chaque âge, environ un individu sur vingt sort de la plage de normalité.

### Commentaires

L'objectif essentiel du paragraphe relatif aux données gaussiennes est de faire comprendre, à partir d'exemples, d'une part, le type d'information qu'apporte l'écart type et, d'autre part, la notion de plage de normalité, en particulier pour une lecture correcte de certains examens biologiques ou des courbes de croissance présentes dans les carnets de santé. On s'est longtemps demandé pourquoi de nombreuses données (mesures biologiques, erreurs de mesure) pouvaient être qualifiées de «gaussiennes»; un théorème de mathématiques appelé «théorème central limite» en propose une explication. Ce théorème est totalement hors

programme. De même, est totalement hors programme la reconnaissance du caractère gaussien de données : on dira aux élèves que des études statistiques ont prouvé qu'il en était ainsi et on leur apprendra simplement à comprendre et utiliser cette information.




On sera attentif à la formulation des conclusions : la normalité évoquée ici est une normalité statistique et il y a une chance sur vingt pour qu'un individu « normal » choisi au hasard soit en dehors de la plage de normalité! De même, un échantillon de petite taille pris au hasard peut s'écarter sensiblement de la forme «en cloche», comme le montrent les deux histogrammes ci-dessous relatifs à deux échantillons de taille cinquante, pris au hasard dans la série de l'exemple 1.

On pourra prolonger la réflexion en simulant, par exemple, la situation suivante où *n* personnes choisies au hasard dans une population de gens en parfaite santé subissent quatre examens médicaux indépendants : on constate alors qu'environ une personne sur cinq a au moins un examen qui sort de la plage de normalité ! Pour faire cette simulation avec une calculatrice, il suffit de disposer de huit chiffres au hasard, de les prendre deux par deux pour fabriquer quatre nombres entre 00 et 99, puis de compter 1 si l'un au moins de ces quatre nombres est supérieur ou égal à 95, 0 sinon; la proportion de 1 est de l'ordre de 1/5 !

# Classe de Première S

## Programme

BO HS n°7 du 31 août 2000

La partie du programme consacrée aux probabilités et à la statistique est centrée :

- sur la mise en place d'éléments de base indispensables pour comprendre ou pratiquer la statistique partout où elle est présente,

- sur l'acquisition de concepts de probabilité permettant de comprendre et d'expliquer certains faits simples observés expérimentalement ou par simulation.

Le programme de la classe de première introduit quelques outils descriptifs nouveaux :

¤ les diagrammes en boîtes qui permettent d'appréhender aisément certaines caractéristiques des répartitions des caractères étudiés et qui complètent la panoplie des outils graphiques les plus classiquement utilisés ;

¤ deux mesures de dispersion : l'écart-type et l'intervalle interquartile.

Ces éléments de statistique pourront notamment être travaillés pour des séries construites à partir de séries simulées; on rencontre ainsi des répartitions variées et on prépare la notion d'estimateur. Cette partie descriptive ne doit pas faire l'objet de longs développements numériques, ni être déconnectée du reste du programme de probabilité et statistique.

| Contenus | Capacités attendues | Commentaires |
|---|---|---|
| **Statistique**<br>Variance et écart-type.<br>Diagramme en boîte; intervalle interquartile.<br>Influence sur l'écart type et l'intervalle interquartile d'une transformation affine des données. | On cherchera des résumés pertinents et on commentera les diagrammes en boîtes de quantités numériques associées à des séries simulées ou non.<br>On observera l'influence des valeurs extrêmes d'une série sur l'écart type ainsi que la fluctuation de l'écart type entre séries de même taille.<br>L'usage d'un tableur ou d'une calculatrice permettent d'observer dynamiquement et en temps réel, les effets des modifications des données. | L'objectif est de résumer une série par un couple (mesure de tendance centrale; mesure de dispersion). Deux choix usuels sont couramment proposés: le couple (médiane; intervalle interquartile), robuste par rapport aux valeurs extrêmes de la série, et le couple (moyenne ; écart-type). On démontrera que la moyenne est le réel qui minimise $\Sigma(xi-x)^2$ alors qu'elle ne minimise pas $\Sigma\|xi-x\|$.<br>On notera s l'écart type d'une série, plutôt que $\sigma$, réservé à l'écart type d'une loi de probabilité. |
| **Probabilités**<br>Définition d'une loi de probabilité sur un ensemble fini. Espérance, variance, écart-type d'une loi de probabilité. Probabilité d'un événement, de la réunion et de l'intersection d'événements. Cas de l'équiprobabilité.<br>Variable aléatoire, loi d'une variable aléatoire, espérance, variance, écart type.<br>Modélisation d'expériences aléatoires de référence (lancers d'un ou plusieurs dés ou pièces discernables ou non, tirage au hasard dans une urne, choix de chiffres au hasard, etc.). | Le lien entre loi de probabilité et distributions de fréquences sera éclairé par un énoncé vulgarisé de la loi des grands nombres. On expliquera ainsi la convergence des moyennes vers l'espérance et des variances empiriques vers les variances théoriques; on illustrera ceci par des simulations dans des cas simples. On pourra aussi illustrer cette loi avec les diagrammes en boîtes obtenus en simulant par exemple 100 sondages de taille n, pour n = 10; 100; 1000.<br>On simulera des lois de probabilités simples obtenues comme images d'une loi équirépartie par une variable aléatoire (sondage, somme des faces de deux dés, etc.). | On pourra par exemple choisir comme énoncé vulgarisé de la loi des grands nombres la proposition suivante :<br>*Pour une expérience donnée, dans le modèle défini par une loi de probabilité P, les distributions des fréquences calculées sur des séries de taille n se rapprochent de P quand n devient grand.*<br>On indiquera que simuler une expérience consiste à simuler un modèle de cette expérience. La modélisation avec des lois ne découlant pas d'une loi équirépartie est hors programme.<br>On évitera le calcul systématique et sans but précis de l'espérance et de la variance de lois de probabilité. |

# Document d'accompagnement 1ère S (GEPS de mathématiques - 08/01/01)

**Annexe 1 : statistique et probabilités**

Le programme de *probabilité-statistique* de la section S introduit les notions de loi de probabilité et de variable aléatoire, notions indispensables pour comprendre l'esprit de la statistique et aborder la problématique propre de cette discipline ; quelques éléments de statistique descriptive sont introduits, mais la statistique descriptive a une part modeste dans cette section.

La terminologie en usage pour la statistique et les probabilités sera réduite au minimum ; on s'efforcera de garder le langage ensembliste et de ne pas développer deux terminologies parallèles : on introduira seulement le terme événement ; l'ensemble des éventualités liées à une expérience gardera le nom d'ensemble, on parlera d'événements complémentaires plutôt que d'événements contraires ou incompatibles ; on pourra , pour éclairer certains exemples mais sans systématiser ou rendre obligatoire un tel langage, parler d'univers, d'événements contraires ou incompatibles ; l'important est de savoir résoudre des problèmes, de relier les résultats de calculs faits sur des observations aux questions posées ; on trouvera en fin de cette annexe un lexique des principaux termes liés à la statistique et que l'élève doit savoir manipuler.

## Statistique descriptive

Schématiquement, la statistique se pratique dans deux contextes :

- pour apporter des éléments de réponse à une question ; cela nécessite le plus souvent de reformuler la question puis de définir un protocole de recueil des données suivies d'une analyse descriptive des données recueillies et une modélisation probabiliste (exemple : les alliages produits par un certain procédé industriel sont-ils conformes aux normes ? téléphoner en voiture est-il un facteur de risque d'accident ? quelle fraction de la population doit-on vacciner pour enrayer une épidémie rapidement ? etc.).

- pour observer ou surveiller régulièrement une population (statistiques de natalité, de mortalité, de la répartition des dépenses des ménages suivant certains secteurs - alimentation, loisir, logement, dépenses de santé, etc- surveillance des taux de chômage, répartition des cultures céréalières dans chaque région, réussite aux bac par filières, etc.).

Dans tous les cas, il y a des données expérimentales et il convient de savoir :

- comment elles sont définies,

- dans quel cadre elles sont recueillies, quelles sont les sources d'erreurs et de variabilité possibles.

Les problèmes afférents au type de données traitées et au mode de recueil dépendent des champs disciplinaires concernés (économie, sociologie, enquêtes de marketing, médecine, finances). Nous avons choisi de ne pas traiter ce qui relève d'une formation professionnelle technique et qui est abordé avec plus d'efficacité dans des études supérieures spécialisées ; certains points pourront par ailleurs être abordés en physique, en biologie, ou en géographie. En première et terminale, on travaillera essentiellement des questions conduisant à une réflexion conceptuelle ou axiomatisée, i.e des questions nécessitant le recours à des calculs probabilistes et on continuera les activités de simulation ; il s'agit là d'éléments qui sont au cœur de tous les champs d'application considérés. Les TPE permettront à certains de faire un réel travail de statistique.

Quelques éléments de statistique descriptive sont présentés pour les séries numériques : écart-type et diagramme en boîte. Il est important ici de démontrer que la moyenne minimise la fonction $x \mapsto \sum_i (x_i - x)^2$ qui mesure la dispersion autour d'un point, et que ce minimum est appelé la variance. Les diagrammes en boîte permettent une comparaison graphique de plusieurs séries de données (voir fiche « sondages » du document

d'accompagnement de seconde) ; on remarquera à l'aide d'exemples que le résumé d'une série par le couple (médiane, écart interquartile) est insensible aux valeurs extrêmes, mais ne peut faire l'objet de calculs par paquets ; résumer une série de données par (moyenne, écart-type) a des propriétés théoriques qui rendent son usage fréquent, notamment en biologie et physique pour ce qui touche aux mesures expérimentales. On regardera comment se transforment les quantités introduites si on change les unités et/ou qu'on décale l'origine ; les résumés numériques pouvant être déterminés à partir de la distribution des fréquences sont aussi affectés par la fluctuation d'échantillonnage : on observera sur des simulations que la variance d'une moyenne est en $n^{-1/2}$, d'où l'intérêt de prendre une moyenne. On notera que la moyenne d'une série de données étant l'isobarycentre des termes et le barycentre des valeurs prises pondérées par les fréquences, elle bénéficie de la propriété d'associativité des barycentres et permet donc les calculs par sous-groupes. La variance permet aussi les calculs par sous groupes ; la formule de décomposition ci-dessous de la variance n'est pas au programme et figure ici à l'usage des enseignants :

Soit une série de n données, de moyenne m et de variance $s^2$, décomposée en k sous-séries de tailles respectives $n_i$, i=1..k. Soit $m_i$ et $s^2_i$ les moyennes et les variances dans chaque sous-série ; alors $s^2$ est la somme de la moyenne des variances $s_i$ pondérées par les $n_i$ /n et de la variance de la série des $m_i$ affectés des poids $n_i$ /n :

$$s^2 = \sum \frac{n_i}{n}(m_i - m)^2 + \sum \frac{n_i}{n} s_i^2$$

Pour de nombreuses questions abordables dans le champ scolaire, des données existent : elles sont réactualisées régulièrement, leur contenu est riche et elles sont accessibles dans des banques de données : on pourra les utiliser. Il est préférable d'aborder un seul exemple motivant, plutôt que des petits exemples dont le traitement se réduit à des calculs numériques sans objet. On pourra aussi traiter de grandes séries de données en fournissant aux élèves des résultats numériques complets ou partiels (par exemple somme des termes et somme des carrés des termes) ; l'important ici n'est pas de faire des calculs soi-même ou d'être virtuose de l'emploi des touches statistiques d'une calculatrice -celles ci ne sont en pratique utilisables que pour des séries de petites tailles- ; quand l'équipement est disponible, le tableur constitue un outil très approprié de traitement des données ; l'objectif reste toujours de comprendre les calculs, de savoir les interpréter.

On a souvent prôné, pour l'enseignement de la statistique, le recueil de données par les élèves eux-mêmes : ce recueil était considéré comme motivant et permettant de percevoir le champ de l'aléatoire. Mais la perception de l'aléatoire peut aussi s'acquérir aujourd'hui par la simulation : les paramètres de dispersion introduits pourront être calculés sur des séries simulées. Un recueil effectif de données par les élèves n'est donc à envisager que s'il ne prend pas beaucoup de temps et traite d'une question que les élèves ont fortement contribué à formuler.

## Probabilité, modélisation, simulation

Le programme de probabilité des classes de première et terminale S concerne la modélisation d'expériences de référence, modélisation définie par la loi de probabilité équirépartie sur un ensemble fini convenablement choisi. Les lois de probabilité non équiréparties rencontrées en première et terminale sont le plus souvent l'image par une variable aléatoire d'une loi équirépartie, et on introduira donc presque simultanément la notion de loi de probabilité et celle de variable aléatoire.

## Loi de probabilité

On recensera les propriétés mathématiques élémentaires de l'objet « distributions de fréquences » (cf. tableau ci-dessous) et on définira une loi de probabilité comme un objet mathématique ayant les mêmes propriétés.

| Distribution de fréquences sur $E=\{x_1,\ldots,x_r\}$ | Loi de probabilité sur $E=\{x_1,\ldots,x_r\}$ |
|---|---|
| $(f_1,\ldots,f_r)$<br>$f_i \geq 0$ ; $\sum f_i = 1$ | $(p_1,\ldots, p_r)$<br>$p_i \geq 0$ ; $\sum p_i = 1$ |
| $A \subset E$ Fréquence de A : $f(A)=\sum f_i$<br>Événement complémentaire : $f(\bar{A})=1-f(A)$ | $A \subset E$ probabilité de A : $P(A)=\sum p_i$<br>Événement complémentaire : $P(\bar{A})=1-P(A)$ |
| Cas numérique :<br>Moyenne empirique : $\bar{x}=\sum f_i x_i$<br>Variance empirique : $s^2=\sum f_i(x_i-\bar{x})^2$<br>Ecart-type empirique : $s=\sqrt{\sum f_i(x_i-\bar{x})^2}$ | Cas numérique :<br>Espérance d'une loi P : $\mu=\sum p_i x_i$<br>Variance d'une loi P : $\sigma^2=\sum p_i(x_i-\mu)^2$<br>Ecart-type d'une loi P : $\sigma=\sqrt{\sum p_i(x_i-\mu)^2}$ |

Modéliser une expérience aléatoire, c'est lui associer une loi de probabilité (qui est un objet du monde mathématique).

Une fréquence est empirique : elle est calculée à partir de données expérimentales alors que la probabilité d'un événement est un « nombre théorique » (on objet du monde mathématique). Les distributions de fréquences issues de la répétition d'expériences identiques et indépendantes varient (fluctuent), la loi de probabilité est un invariant associé à l'expérience.

L'objectif est que les élèves comprennent à l'aide d'exemples (cf. paragraphe sur les expériences de référence) que modéliser, c'est ici choisir une loi de probabilité. Il ne s'agit en aucun cas d'avoir des discours généraux sur les modèles et la modélisation.

Les élèves devront bien distinguer ce qui est empirique (du domaine de l'expérience) de ce qui est théorique ; en particulier, on réservera la lettre grecque $\sigma$ à l'écart-type d'une loi et on évitera de noter avec cette même lettre un écart-type empirique (il s'agit là de règles de notations internationales) .

En classe de première, une loi de probabilité P sur un ensemble fini est la liste des probabilités des éléments de E ; à partir de cette liste, on définit naturellement la probabilité d'événements, c'est à dire implicitement une application de $\mathcal{P}(E)$ dans [0,1], application qui sera encore désignée par P .

Il est inutilement complexe, pour le cas des ensembles finis, de partir d'une application de $\mathcal{P}(E)$ dans [0,1], vérifiant certains axiomes, puis de montrer ensuite que cette application est entièrement caractérisée par $(p_1,\ldots, p_r)$. Le fait de ne pouvoir simplement généraliser aux ensembles continus cette définition, et la nécessité d'une définition ensembliste seront abordés en terminale.

Si tous les éléments d'un ensemble bien défini E ont même probabilité, celle-ci est dite équirépartie ; on parlera aussi d'équiprobabilité et on dira que les éléments de l'ensemble E sont choisis au hasard (i.e. si on ne spécifie rien de plus, le choix au hasard est un choix avec équiprobabilité). Si la loi P n'est pas équirépartie, on parlera de choix d'éléments selon la loi P (ou éventuellement de choix au hasard selon la loi P).

Dans le cas de choix au hasard, la probabilité d'un événement est le quotient de son nombre d'éléments par le nombre d'éléments de l'ensemble.

On évitera tout développement théorique sur le langage des événements et le calcul ensembliste qui en découle : ces notions et la pratique de la logique qu'ils impliquent (étude du complémentaire de l'événement A ou B, ou de l'événement A et B) s'acquièrent au fil d'exercices.

# Variable aléatoire

## Exemples d'activités préparatoires

- On s'intéresse au jeu suivant : on lance deux dés distincts et on gagne 1F si la somme des résultats est paire, on perd 1F sinon.

L'ensemble E des issues possibles est celui des couples (i,j) de nombres entiers entre 1 et 6. On considère l'application gain, notée G, qui à (i,j) fait correspondre 1 si i+j est pair, -1 sinon. On associe à une distribution de fréquences observée lors de n parties la distribution des fréquences des valeurs prises par G : c'est une distribution de fréquences sur {-1,1}.

- n lancers de deux dés discernables (n>30) .

On considère l'ensemble E des couples (i,j) de nombres entiers entre 1 et 6. Soit l'application T qui à (i,j) fait correspondre i+j. On associe, par le calcul, à une distribution de fréquences sur E obtenue en lançant n fois deux dés la distribution des fréquences des valeurs prises par T : c'est une distribution de fréquences sur {2, …,12}.

- n lancers de 5 pièces discernables (n>30) .

L'ensemble E considéré ici sera celui des listes $(x_1, ...,x_5)$ de nombres 0 ou 1.

On considère l'application T qui à $(x_1, ...,x_5)$ fait correspondre $x_1+ ...+x_5$. On associe à une distribution de fréquences sur E, obtenue en lançant n fois cinq pièces, la distribution des fréquences de T : c'est une distribution de fréquences sur {0,...,5}

Une application T de E dans E' (E et E' finis) permet de transporter la loi P définie sur E en une loi P' définie sur E' par le procédé suivant :

$$P'(x') = P(T=x')$$

où P(T=x') désigne la probabilité de l'ensemble des éléments de E dont l'image par T est x' (cet ensemble est souvent noté {T=x'}, on écrit en fait P(T=x') au lieu de P({T=x'})).

On notera sur des exemples que même si la loi P est équirépartie, la loi transportée P' par T, appelée *loi image de T*, n'est en général pas l'équiprobabilité (cf. activités préparatoires).

Dans ce cadre-là, ce qui nous intéresse dans l'application T, c'est ce qui se passe à l'arrivée : partant d'un élément x choisi dans E selon la loi P, on arrive à un élément x' choisi dans E' selon la loi P'. Le terme usuel pour nommer T est de parler d'une variable aléatoire.

Dans le cas numérique, l'espérance de la loi de P' est appelée plus simplement espérance de T et notée E(T).

*Remarque :Les lois de probabilités manipulées dans les exercices pourront bien sûr aussi être définies sur des ensembles discrets non ordonnés et les variables aléatoires ne sont pas nécessairement numériques.*

## Exemples

Que gagne-t-on en moyenne si on lance deux dés distincts en gagnant 1F si la somme des résultats est paire, et en perdant 1F sinon.?

Le modèle associé est l'équiprobabilité sur l'ensemble des couples (i,j) de nombres entiers entre 1 et 6. On considère la variable aléatoire gain G qui à (i,j) fait correspondre 1 si i+j est pair, -1 sinon. On peut alors déterminer la loi de G et son espérance.

Il est important, à propos du concept de variable aléatoire, de comprendre qu'une fois fait le choix d'un modèle pour un expérience, d'autres lois de probabilité en découlent par le calcul probabiliste: on ne fait qu'une seule fois un choix de modèle pour traiter un problème.

## Exercices

- Lancers de deux dés discernables

E= $\{1,\ldots,6\}^2$ et P est la probabilité équirépartie. Un tel choix de modèle pourra être confronté à l'expérience elle-même. On considère l'application T qui à (i,j) fait correspondre i+j. Calcul et simulation de la loi de T.

On considère l'application U de E dans {0,1} qui à (i,j) fait correspondre 0 si i et j ne sont pas premiers entre eux. Calcul et simulation de la loi de U.

- Lancers de 2 pièces de couleurs distinctes dont les faces sont codées 0 et 1.

On considère l'ensemble $\{0,1\}^2$ et la probabilité équirépartie.

Un tel choix de modèle pourra être confronté à l'expérience elle-même dans le cas où les élèves ne l'ont pas déjà fait en seconde. On considère l'application T qui à (i,j) fait correspondre i+j. Calculer et simuler sa loi.

Supposons maintenant qu'un observateur ne voit pas les couleurs des pièces :

Pour lui, les issues observables de cette expérience sont « double 0 », « double 1 », « 1 et 0 ». Pour justifier le choix de la probabilité (0.25,0.25,0.5) pour ces trois issues, on pourra, outre la validation expérimentale en lançant des pièces, se ramener au cas précédent. Le résultat final ne doit pas dépendre du fait que l'observateur voit ou non les couleurs des pièces. On voit ainsi que pour certaines expériences, on est amené à choisir un modèle calculé à partir d'un modèle d'équiprobabilité sur un ensemble qui n'est pas celui des issues observables.

- Lancers de deux dés indiscernables .

Imaginons qu'on ne puisse observer que la somme des faces. L'ensemble des issues observables est donc {2,...,12}. On cherche à modéliser cette expérience. On pourra se faire une idée du résultat par expérimentation ou simulation et déterminer le modèle associé à cette expérience en partant d'un modèle équiréparti sur $\{1,\ldots,6\}^2$ et en introduisant la variable aléatoire somme.

- Urne dont les boules sont numérotées et de couleurs différentes

Lorsqu'on choisit au hasard une boule et qu'on s'intéresse à la loi de probabilité de la variable aléatoire couleur de la boule, la probabilité d'une couleur est égale à la proportion de boules de cette couleur dans l'urne.

- Urne à boules indiscernables

Le plus souvent la seule distinction entre les boules porte sur leur couleur et donc le seul résultat observable est une couleur ; mais la loi sur l'ensemble des couleurs ne doit pas dépendre de notre capacité à distinguer les boules individuellement. On conviendra implicitement (ou en l'explicitant si besoin est)que choisir au hasard des boules coloriées, c'est choisir une couleur selon une loi telle que la probabilité de chaque couleur bleue est égale à la proportion de boules de cette couleur dans l'urne (dans cette phrase le hasard signifie que la probabilité de chaque boule d'être tirée est la même).

Dans un casino, on peut jouer aux jeux suivants ; quel jeu choisir ?

Jeu A : On choisit un chiffre selon une loi de probabilité telle que la probabilité du chiffre i est proportionnelle à i+1. Le gain associé au chiffre i est 2520/(i+1).

Jeu B : On choisit un chiffre selon une loi de probabilité telle que la probabilité du chiffre i est proportionnelle à 1/(i+1). Le gain associé au chiffre i est 1000(i+1).

## Modélisation d'expériences de référence

Modéliser une expérience aléatoire, c'est associer à cette expérience une loi de probabilité sur l'ensemble des issues possibles. Ce choix, c'est à dire la modélisation de l'expérience, est en général délicat à faire, sauf dans certains cas où des considérations propres au protocole expérimental conduisent à proposer a priori un modèle. Il en est ainsi des lancers de pièces ou de dés pour lesquels des considérations de symétrie conduisent au choix d'un modèle où la loi de probabilité est équirépartie. On se restreindra donc aux expériences de référence en évitant tout discours général sur ce qu'est ou n'est pas la modélisation.

La traduction mathématique de « une pièce a autant de chances de tomber sur pile que sur face » est ainsi « la probabilité de pile et de face sont égales ». On indiquera clairement que les termes *équilibré* et *choix au hasard* indiquent par convention un choix du modèle de l'expérience, où la probabilité est équirépartie.

En dehors de tels cas où des considérations quant à la nature des expériences permet de proposer un modèle, le choix d'un modèle à partir de données expérimentales est beaucoup plus délicat et ne sera pas abordé dans l'enseignement secondaire. On se contentera, si nécessaire, de fournir un modèle en indiquant que des techniques statistiques ont permis de déterminer et de valider un tel modèle.

La modélisation ne relève pas d'une logique du vrai et du faux : un modèle n'est ni vrai ni faux : il peut être validé ou rejeté au vu de données expérimentales. Une des premières fonctions de la statistique dite inférentielle est d'associer à une expérience aléatoire un modèle, ou une gamme de modèles compatibles en un certain sens à définir avec les données expérimentales dont on dispose, et aussi de définir des procédures de validation d'un modèle.

Pour valider un modèle probabiliste, le premier outil dont on dispose est un théorème de mathématiques appelé loi (forte) des grands nombres, dont un énoncé intuitif est :

*Dans le monde théorique défini par une loi P sur un ensemble fini, les fréquences des éléments de cet ensemble dans une suite de n expériences identiques et indépendantes « tendent » vers leur probabilité quand n augmente indéfiniment.*

Ou encore

*Si on choisit n éléments selon une loi de probabilité P, indépendamment les uns des autres, alors la distribution des fréquences « tend » vers P lorsque n tend vers l'infini.*

Il s'agit là d'un énoncé vulgarisé. Pour être un peu plus précis, le théorème appelé loi forte des grands nombres dit que dans l'ensemble des suites infinies d'éléments choisis selon P, le sous-ensemble des suites pour lesquelles la distribution des fréquences ne converge pas vers P est « négligeable » . Par exemple, dans une suite finie de n lancers d'une pièce équilibrée, on peut n'obtenir que des faces (code 0) ou que des pile (code 1), ou 001001…001, etc. ; ces trois suites finies ont chacune une probabilité $2^{-n}$ ; si on « prolonge » ces suites, les fréquences de 1 et de 0 ne tendent pas vers 1/2, mais la probabilité de l'ensemble de ces trois suites tend vers 0. Le théorème ci-dessus indique que l'ensemble de toutes les suites imaginables pour lesquelles les fréquences ne tendent pas vers 0,5 est de « probabilité nulle » - pour une loi de probabilité construite sur l'ensemble des suites infinies de 0 et de 1 à partir de l'équiprobabilité sur {0,1} . Le mathématicien dira qu'il y a convergence presque sûre des fréquences des éléments vers leur probabilité. Le statisticien, s'il observe dans une longue série d'expériences des distributions de fréquences qui fluctuent de moins en moins choisira comme modèle, en vertu de cette loi des grands nombres, une loi de probabilité « proche » de la dernière distribution observée : il ne va pas choisir un modèle pour lequel ce qu'il a observé constitue un événement quasi-négligeable !

Une validation du modèle de la loi équirépartie pour le lancer d'un dé consistera à vérifier que la distribution des fréquences est « proche » de (1/6,…,1/6) sur {1,...,6} quand le nombre de lancers est grand (cf. l'annexe du projet de terminale S).

### Exercice

1- Dans un jeu de pile ou face où on gagne le double de la mise sur pile et où on perd la mise sur face, un joueur qui dispose de 1000 F commence par miser un franc, double sa mise tant qu'il perd et ne s'arrête que s'il gagne où s'il ne peut plus miser. Simuler ce jeu, puis calculer l'espérance de gain du joueur (on admettra par généralisation que si on joue n fois le modèle est l'équirépartition sur les n-listes dont les éléments valent 0 ou 1).

2- Dans un jeu de pile ou face où on gagne le double de la mise sur pile et où on perd la mise sur face, un joueur qui dispose de 1000 F commence par miser un franc, triple sa mise tant qu'il perd et ne s'arrête que s'il gagne où s'il ne peut plus miser. Simuler ce jeu, puis calculer l'espérance de gain du joueur.

## Simulation de chiffres au hasard

On clarifiera brièvement les positions respectives de la modélisation et de la simulation : modéliser consiste à associer un modèle à des données expérimentales, alors que simuler consiste à produire des données à partir d'un modèle prédéfini. On parlera de simulation d'une loi de probabilité P ; la simulation d'une telle loi avec des listes de chiffres au hasard ne peut se faire que si P peut être construite comme loi image d'une loi équirépartie. Pour simuler une expérience , on associe d'abord un modèle à l'expérience en cours, puis on simule la loi du modèle ; on pourra détailler ces étapes, sans cependant le faire systématiquement dans les cas simples des expériences de référence .

### Exemple d'activités utilisant des simulations

1- On considère deux stratégies de choix de nombres parmi les nombres 1,3,5.

(i)- Choix au hasard

(ii)- Choix selon la loi ( 0.1, 0.1,0.8)

Calculer l'espérance des deux lois ci-dessus. Simuler n choix selon l'une des deux méthodes ; regarder si au vu de la moyenne des séries observées, quelqu'un ne connaissant pas le modèle choisi pour simuler peut le deviner. On fera varier n.

2- Pour la grille suivante, on considère trois procédures de choix de cases blanches :

| | | |
|---|---|---|
| | N | N |
| | N | |
| N | | |

(i)- On choisit une case blanche au hasard parmi les 5 cases blanches

(ii)- On choisit au hasard une ligne puis dans la ligne une case au hasard parmi les cases blanches de cette ligne

(iii)- On choisit au hasard une colonne puis dans la colonne une case au hasard parmi les cases blanches de cette colonne

Déterminer dans chacun des trois cas la loi de probabilité mise en jeu sur l'ensemble des cases blanches.

Pour une des trois lois de probabilité, un élève simuler n choix de cases blanches ; au vu de la distribution des fréquences obtenu, un autre élève qui ignorerait quelle loi a été simulé peut-il la deviner ?

Dans les deux exemples ci-dessus, on insistera sur le fait que l'observation de résultats simulés ne permet pas de remonter au modèle à coup sûr : une des fonctions de la statistique est de calculer le probabilité que l'on a de se tromper en « remontant d'une distribution de fréquence à une loi de probabilité ».

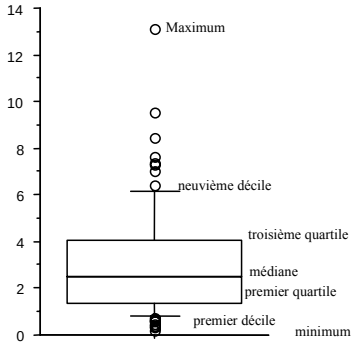
## Exercice

Une expérience consiste à choisir successivement deux chiffres d et u au hasard. On considère l'application T qui à (d,u) fait correspondre 10d + u . Calculer la probabilité que 10d + u $\leq$ k, où k est un entier entre 0 et 9. On justifie ainsi que pour construire une liste de nombres au hasard entre 0 et 99, on lise les chiffres d'une table de chiffres au hasard deux par deux.

# Lexique

**Choix au hasard** dans un ensemble E : la loi de probabilité en jeu sur E est l'équiprobabilité.

**Diagramme en boîte** (ou à pattes ou à moustaches ou diagramme de Tuckey) : on divise l'intervalle des valeurs de la série non plus en intervalles de même longueur comme pour de nombreux histogrammes mais en intervalles qui contiennent des pourcentages des données fixés à l'avance ; par exemple :




**Écart interquartile :** différence entre le troisième et le premier quartile

**Ecart-type** : racine carré de la variance ; l'unité de l'écart-type est celle des données.

**Espérance** d'une loi de probabilité sur un ensemble E de nombres : moyenne des valeurs des éléments de E pondérés par leur probabilité.

**Espérance d'une variable aléatoire** : espérance de la loi de cette variable aléatoire.

**Étendue** : différence entre la plus grande et la plus petite valeur de la série.

**Événement A et B** : si A,B sont deux événements, A et B est l'événement A $\cap$ B

**Événement Aou B** : si A,B sont deux événements, A ou B est l'événement A $\cup$ B.

**Événement** : dans le champ des probabilités, un événement est un sous-ensemble de l'ensemble des issues possibles d'une expérience aléatoire.

**Expériences aléatoires identiques** : signifie qu'on associe à chacune d'elles le même modèle. Les expériences de référence ont d'ailleurs aussi ceci de remarquable que par exemple le même modèle est pertinent pour des pièces (ou des dés) de fabrications différentes lancés par des individus différents : on parlera donc dans ce cas d'expériences identiques.

**Expériences de référence** : lancers de dés, de pièces équilibrées ; choix de nombres au hasard, tirages au hasard de boules coloriées dans une urne, de cartes dans un jeu etc. Considérons une expérience aléatoire modélisée par une loi de probabilité P sur un ensemble fini E. On pourra le plus souvent trouver une expérience de référence qui à un recodage près des issues possibles est régie par le même modèle ; les questions relatives à l'expérience originelle pourront être traduites dans le cadre de l'expérience de référence. Associer l'expérience originelle à une expérience de référence n'est pas une activité à systématiser : l'élève aura recours à cette possibilité selon ses besoins.

**Intervalle interdécile** : intervalle dont les extrémités sont le premier et le neuvième décile.

**Intervalle interquartile** : intervalle dont les extrémités sont premier et le troisième quartile.

**Loi de probabilité** sur $E=\{x_1,\ldots,x_r\}$ : c'est une liste $(p_1,\ldots, p_r)$ de nombres positifs et de somme 1 , associés aux éléments de E.

Loi des grands nombres : en langage imagé :

Si on choisit n éléments d'un ensemble fini E selon une loi de probabilité P, indépendamment les uns des autres, alors la distribution des fréquences « tend » vers la loi de probabilité P lorsque n tend vers l'infini.

**Loi image** : Soit T une variable aléatoire défini sur E à valeurs dans E'. Soit P une loi de probabilité définie sur E. La loi image de T est une loi de probabilité définie sur E' par :

**Médiane empirique** : on ordonne la série des observations par ordre croissant ; si la série est de taille $2n + 1$, la médiane est la valeur du terme de rang n+1 dans cette série ordonnée ; si la série est de taille 2 n, la médiane est la demi-somme des valeurs des termes de rang n et n+1 dans cette série ordonnée. La définition de la médiane n'est pas figée : certains logiciels et certains ouvrages définissent la médiane comme étant le second quartile ou le cinquième décile : dans la pratique de la statistique, les différences entre ces deux définitions sont sans importance ; au lycée, on évitera tout développement la dessus qui ne serait pas une réponse individuelle à une question d'un élève.

**Modèle d'une expérience aléatoire** : c'est une loi de probabilité sur un ensemble, qui est souvent celui des issues observables de l'expérience.

**Neuvième décile (empirique)**: : c'est le plus petit élément *d'* des valeurs des termes de la série, ordonnées par ordre croissant, tel qu'au moins 90% des données soient inférieures ou égales à *d'*.

où P(T=x') désigne la probabilité de l'ensemble des éléments de E dont l'image par T est x' (cet ensemble est souvent noté {T=x'}, on écrit en fait P(T=x') au lieu de P({T=x'})).

$$P'(x') = P(T=x')$$

**Pièce équilibrée** : on choisit pour modéliser le lancer d'une telle pièce l'équiprobabilité de *pile* et de *face*.

**Premier décile (empirique)**: c'est le plus petit élément *d* des valeurs des termes de la série, ordonnées par ordre croissant, tel qu'au moins 10% des données soient inférieures ou égales à *d*.

**Premier quartile (empirique)**: c'est le plus petit élément *q* des valeurs des termes de la série, ordonnées par ordre croissant, tel qu'au moins 25% des données soient inférieures ou égales à *q*.

**Probabilité d'un événement** : c'est la somme des probabilités des éléments qui le composent.

**Troisième quartile (empirique)**: c'est le plus petit élément q' des valeurs des termes de la série, ordonnées par ordre croissant, tel qu'au moins 75% des données soient inférieures ou égales à q'.

**Variable aléatoire** : application définie sur un ensemble muni d'une loi de probabilité P ; son rôle est de transporter P sur un autre ensemble (voir loi image).

**Variance d'une loi de probabilité**: espérance des carrés des écarts à l'espérance ; c'est aussi la différence entre l'espérance des carrés et le carré de l'espérance.

Variance d'une variable aléatoire : c'est celle de sa loi image.

**Variance empirique** : moyenne des carrés des écarts à la moyenne ; c'est aussi la différence entre la moyenne des carrés et le carré de la moyenne.

# Premières STI (toutes spé.), STL (spé. Physique-Chimie

Au collège et en Seconde, les élèves ont étudié la description de séries statistiques à une variable. Le programme de Première comporte un premier contact avec les probabilités. L'objectif est d'entraîner les élèves à décrire quelques *expériences aléatoires* simples, et à *calculer des probabilités*. On évitera tout développement théorique. Pour introduire la notion de probabilité, on s'appuiera sur l'étude de séries statistiques obtenues par répétition d'une expérience aléatoire, en soulignant les propriétés des fréquences et la relative stabilité de la fréquence d'un événement donné lorsque cette expérience est répétée un grand nombre de fois. La description d'expériences aléatoires amène aussi à organiser des données : on se limitera à *quelques exemples* permettant de mettre en valeur les idées, mais ne comportant pas de difficultés combinatoires. Il est important que les élèves puissent se familiariser avec les probabilités pendant une durée suffisante ; l'étude de ce chapitre ne doit pas être bloquée en fin d'année.

| **Programme** | **Commentaires** |
|---|---|
| Événements, événements élémentaires ; la probabilité d'un événement est définie par addition de probabilités d'événements élémentaires.<br>Événements disjoints (ou incompatibles), événement contraire, réunion et intersection de deux événements.<br>Cas où les événements élémentaires sont équiprobables. | Seul est au programme le cas où l'ensemble des événements élémentaires est fini.<br>Les élèves doivent savoir calculer la probabilité de la réunion d'événements disjoints, d'un événement contraire Â, et savoir utiliser la formule reliant les probabilités de $A \cup B$ et de $A \cap B$.<br>Les notions de probabilité conditionnelle, d'indépendance, de probabilité produit et de variable aléatoire ne sont pas au programme. |
| **Travaux pratiques**<br>Exemples simples d'emplois de partitions et de représentations (arbres, tableaux,...) pour organiser et dénombrer des données relatives à la description d'une expérience aléatoire.<br>Exemples simples d'étude de situations de probabilités issues d'expériences aléatoires (modèles d'urnes, jeux...). | L'étude du dénombrement des permutations, arrangements et combinaisons est hors programme.<br>On s'attachera à étudier des situations permettant de bien saisir la démarche du calcul des probabilités, et non des exemples comportant des difficultés techniques de dénombrement. Dans certaines situations, par exemple l'étude de caractères d'une population, les événements élémentaires ne sont pas donnés *a priori* ; on les construit en effectuant une partition de la population. |

# Classe de Terminale ES (extraits)

## Programme

BO HS n°4 du 30 août 2001

À titre indicatif, la répartition horaire entre les différents chapitres peut être : 60 % pour l'analyse (18 semaines); 40% pour la statistique et les probabilités (12 semaines).

| Contenus | Modalités de mise en œuvre | Commentaires |
|---|---|---|
| **Statistique et probabilités** | | |
| Nuage de points associé à une série statistique à deux variables numériques.<br>Point moyen. | On proposera aussi des exemples où la représentation directe en (x ; y) n'est pas possible et où il convient par exemple de représenter (x ; ln y) ou (ln x; y) et on fera le lien avec des repères semi-logarithmiques. | |
| Ajustement affine par moindres carrés. | On fera percevoir le sens de l'expression « moindres carrés » par le calcul sur tableur, pour un exemple simple, de la somme : $\Sigma (yi - axi - b)^2$ .<br>On évoquera sur des exemples l'intérêt éventuel et l'effet d'une transformation affine des données sur les paramètres a et b.<br>On étudiera avec des simulations la sensibilité des paramètres aux valeurs extrêmes.<br>On proposera des exemples où une transformation des données conduit à proposer un ajustement affine sur les données transformées.<br>On proposera un ou deux exemples où les points (xi ; yi) du nuage sont "presque" alignés et où cet alignement peut s'expliquer par la dépendance "presque" affine à une troisième variable. | L'objectif est de faire des interpolations ou des extrapolations.<br>On admettra les formules donnant les paramètres de la droite des moindres carrés : coefficient directeur et ordonnée à l'origine.<br>On traitera essentiellement des cas où, pour une valeur de x, on observe une seule valeur de y (par exemple les séries chronologiques).<br>Le coefficient de corrélation linéaire est hors programme (son interprétation est délicate, notamment pour juger de la qualité d'un ajustement affine).<br>On verra ainsi que pouvoir prédire y à partir de x ne prouve pas qu'il y ait un lien de causalité entre x et y. |
| Simulation. | On étudiera un exemple traitant de l'adéquation de données expérimentales à une loi équirépartie. | L'élève devra être capable de poser le problème de l'adéquation à une loi équirépartie et de se reporter aux résultats de simulation qu'on lui fournira.<br>Le vocabulaire des tests (hypothèse nulle, risque de première espèce) est hors programme. |
| Conditionnement et indépendance. | On justifiera la définition de la probabilité de B sachant A, notée PA(B), par des calculs fréquentiels. | Un arbre de probabilité correctement construit constitue une preuve. |
| Conditionnement par un événement de probabilité non nulle puis indépendance de deux événements.<br>Formule des probabilités totales. | On utilisera à bon escient les représentations telles que tableaux, arbres, diagrammes….efficaces pour résoudre des problèmes de probabilités.<br>On appliquera entre autre cette formule à la problématique des tests de dépistage. | Les élèves doivent savoir appliquer la formule des probabilités totales sans aide dans des cas simples. |

| Contenus | Modalités de mise en œuvre | Commentaires |
|---|---|---|
| Modélisation d'expériences indépendantes.<br>Cas de la répétition d'expériences identiques et indépendantes. | On retravaillera les expériences de références vues en seconde et première (dés, pièces, urnes…). | On conviendra, en conformité avec l'intuition, que pour des expériences indépendantes, la probabilité de la liste des résultats est le produit des probabilités de chaque résultat. |
| Lois de probabilités discrètes. | | Les situations abordées à ce niveau ne nécessitent pas le langage formalisé des variables aléatoires ; ces dernières ne figurent pas au programme. |
| Espérance et variance d'une loi numérique. | À l'aide de simulations et de la loi des grands nombres, on fera le lien avec moyenne et variance d'une série de données. | |
| Expériences et lois de Bernoulli.<br>Lois binomiales. | On se limitera pour les calculs sur ces lois à des petites valeurs de n ($n<5$) : on pourra utiliser des arbres. | On donnera des exemples variés où interviennent des lois de Bernoulli et des lois binomiales. |

# Document d'accompagnement T ES (CNDP - 2002)

Annexe : Probabilités et statistique en terminale ES et S

L'annexe est uniquement disponible sur le cédérom joint au document d'accompagnement

# Classe de Terminale S

## Programme

BO HS n°4 du 30 août 2001

II.3 Probabilités et statistique

Après avoir introduit en classe de seconde la nature du questionnement statistique à partir de travaux sur la fluctuation d'échantillonnage, on poursuit ici la présentation entreprise en première des concepts fondamentaux de probabilité dans le cas fini avec la notion de conditionnement et d'indépendance et l'étude de quelques lois de probabilité.

On vise aussi, en complément à l'usage des simulations introduit dès la seconde, une première sensibilisation à d'autres classes de problèmes, notamment celui de l'adéquation d'une loi de probabilité à des données expérimentales.

| Contenus | Capacités attendues | Commentaires |
|---|---|---|
| **Conditionnement et indépendance** | | |
| Conditionnement par un événement de probabilité non nulle puis indépendance de deux événements. | On justifiera la définition de la probabilité de B sachant A, notée $P_A$ (B), par des calculs fréquentiels. | Un arbre de probabilité correctement construit constitue une preuve. |
| Indépendance de deux variables aléatoires. | On utilisera à bon escient les représentations telles que tableaux, arbres, diagrammes… efficaces pour résoudre des problèmes de probabilités. | |
| Formule des probabilités totales. | Application à la problématique des tests de dépistage en médecine et à la loi de l'équilibre génétique lors d'appariements au hasard. | Les élèves doivent savoir appliquer sans aide la formule des probabilités totales dans des cas simples. |
| Statistique et modélisation<br>Expériences indépendantes.<br>Cas de la répétition d'expériences identiques et indépendantes. | Application aux expériences de références vues en seconde et première (dés, pièces, urnes…). | On conviendra, en conformité avec l'intuition, que pour des expériences indépendantes, la probabilité de la liste des résultats est le produit des probabilités de chaque résultat. |
| **Lois de probabilités** | | |
| Exemples de lois discrètes | On introduira la notation n!. | |
| Introduction des combinaisons, notées $\binom{n}{p}$ | L'élève devra savoir retrouver les formules : $\binom{n}{p}=\binom{n-1}{p-1}+\binom{n-1}{p}$ et $\binom{n}{p}=\binom{n}{n-p}$ | Le symbole $\binom{n}{p}$ peut être désigné par la locution "p parmi n".<br>Pour les dénombrements intervenant dans les problèmes, on en restera à des situations élémentaires résolubles à l'aide d'arbres, de diagrammes ou de combinaisons. |
| Formule du binôme. | | |
| Loi de Bernoulli, loi binomiale ; espérance et variance de ces lois. | On appliquera ces résultats à des situations variées. | La formule donnant l'espérance sera conjecturée puis admise; la formule de la variance sera admise. |
| Exemples de lois continues | | |
| Lois continues à densité : | | |
| - loi uniforme sur [0,1] ; | | |
| - loi de durée de vie sans vieillissement. | Application à la désintégration radioactive :<br>loi exponentielle de désintégration des noyaux. | Ce paragraphe est une application de ce qui aura été fait en début d'année sur l'exponentielle et le calcul intégral. |

| Contenus | Capacités attendues | Commentaires |
|---|---|---|
| Statistique et simulation | Étude d'un exemple traitant de l'adéquation de données expérimentales à une loi équirépartie. | L'élève devra être capable de poser le problème de l'adéquation à une loi équirépartie et de se reporter à des résultats de simulation qu'on lui fournit.<br>Le vocabulaire des tests (test d'hypothèse, hypothèse nulle, risque de première espèce) est hors programme. |

# Document d'accompagnement T S (CNDP - 2002)

Annexe : Probabilités et statistique en terminale ES et S

L'annexe est uniquement disponible sur le cédérom joint au document d'accompagnement

# Terminales STI (toutes spé.) et STL (spé. Physique-Chimie)

Quelques notions de calcul des probabilités ont été introduites en Première ; en Terminale, on poursuit l'étude de phénomènes aléatoires. Le programme comporte une consolidation des acquis de Première et l'introduction, sur des exemples simples, du concept de variable aléatoire. On se limite à des ensembles finis; toute théorie formalisée est exclue et les notions de probabilité conditionnelle, d'indépendance et de probabilité produit ne sont pas au programme.

Pour les variables aléatoires, le programme ne porte que sur l'étude d'exemples.

| **Programme** | **Commentaires** |
|---|---|
| Variable aléatoire (réelle) prenant un nombre fini de valeurs et loi de probabilité associée ; fonction de répartition, espérance mathématique, variance, écart type. | On prendra un point de vue très simple : certaines situations de probabilité s'expriment commodément par l'affectation de probabilités $p_1$ $p_2$, ..., $p_n$, aux valeurs $x_1$, $x_2$, ... , $x_n$. d'une grandeur numérique X associée à une expérience aléatoire ; on dit alors que X est une variable aléatoire. Les événements $(X = x_1)$, $(X = x_2)$, … , $(X = x_n,)$ sont les événements élémentaires de la loi de probabilité de X.<br>Pour la fonction de répartition, on emploiera la convention $F(x) = p(X \leq x)$. |
| **Travaux pratiques**<br>Exemples d'emploi de partitions et de représentations (arbres, tableaux, ...) pour organiser et dénombrer des données relatives à des situations aléatoires.<br>Exemples d'étude de situations de probabilités issues d'expériences aléatoires (modèles d'urnes, jeux, ...).<br>Exemples simples d'étude de situations menant à l'étude d'une variable aléatoire. | L'étude du dénombrement des permutations, arrangements et combinaisons est hors programme.<br>On conserve le même point de vue qu'en Première ; en particulier, on s'attachera à étudier des situations permettant de bien saisir la démarche du calcul des probabilités, et non des exemples comportant des difficultés techniques de dénombrement.<br>Des indications doivent être données sur la méthode à suivre. |

# Annexe probabilités statistiques série ES et S

Rappel

À la rentrée 2002 la brochure « Accompagnement des programmes de Mathématiques rentrée 2002 », éditée par le CNDP, a été envoyée à tous les lycées, accompagnée d'un CDROM.

Le document qui suit est extrait de ce CDROM.

ANNEXE

# Introduction

Le programme de probabilités et de statistique prend la suite des programmes des années précédentes et utilise largement le vocabulaire et les concepts introduits (tirage au hasard, loi de probabilité, variable aléatoire pour la série S). Comme en classe de première, les calculs au niveau des fréquences sont transposés au niveau des probabilités d'événements et des lois de probabilité des variables aléatoires (conditionnement et indépendance). On garde constamment à l'esprit que les distributions de fréquences fluctuent, la loi de probabilité restant fixe, d'où l'émergence de nouvelles questions liées à la reconnaissance d'une loi de probabilité à partir de données fréquentielles.
Les conventions de langage concernant la notion d'expériences identiques et indépendantes sont explicitées.
Le programme revient sur des situations déjà rencontrées dans les années antérieures (calcul de la probabilité d'avoir deux fois *pile* lorsqu'on lance deux pièces équilibrées, tirage au hasard des boules colorées dans une urne – la probabilité d'une couleur est alors égale à la proportion de boules de cette couleur dans l'urne).
On insiste toujours sur le lien entre concepts probabilistes et données empiriques. Des données provenant d'expériences de référence (tirage au hasard de boules ou lancers de pièces ou de dés) permettent de poser des questions sur les liens entre propriétés des distributions des fréquences et propriétés des lois de probabilité. Ainsi, chercher à savoir si un dé est équilibré illustre une problématique classique, même s'il s'agit là d'un cas d'école qui ne reflète pas la pratique professionnelle de la statistique (il peut être bon de le dire aux élèves !). Les problèmes de modélisation pour des données plus complexes ne peuvent pas être traités en terminale.
Si les expériences de référence classiques (le plus souvent simulées en terminale) sont indispensables pour comprendre la théorie des probabilités, elles ne sont cependant pas de nature à convaincre les élèves de l'importance de cette théorie en mathématiques comme dans les autres sciences. Aussi, le programme de probabilité de la série scientifique a partie liée avec un autre chapitre important du programme de mathématique de terminale concernant l'intégration (loi uniforme sur un intervalle borné et loi exponentielle) et une convergence thématique forte apparaît avec le chapitre « Radioactivité » du programme de physique : en physique on étudie la radioactivité au niveau macroscopique et en mathématiques, on l'étudie au niveau microscopique. C'est l'occasion de traduire dans le champ des mathématiques la notion d'absence d'usure (voir l'annexe portant sur l'étude de la radioactivité) ; ce travail de modélisation illustre une pratique que les élèves n'ont en général pas eu l'occasion de rencontrer.

# Étude de deux variables qualitatives. Fréquence conditionnelle

L'esprit humain ne peut appréhender visuellement des listes ou des tableaux de nombres ; aussi doit-on en chercher des modes de représentation éclairants. Une liste de $n$ nombres donnant les valeurs d'un caractère qualitatif à $k$ modalités est le plus souvent représentée par un tableau à $k$ lignes ou $k$ colonnes donnant les effectifs de chaque modalité, ou par un diagramme en bâtons : la seule information perdue entre la liste et le tableau ou le diagramme est l'ordre des termes dans la liste. Pour un tableau de $n$ lignes et deux colonnes donnant les valeurs de deux variables qualitatives valeurs sur $n$ individus (un individu pouvant être un être humain, une ville, un objet manufacturé, etc), deux modes de représentation des données peuvent être trouvés par des élèves : tableau à $k$ lignes et $k$' colonnes, où $k$ et $k$' sont les nombres de modalités des deux variables, ou un arbre.

**Exemple**

Une enquête de marketing portant sur le choix entre deux abonnements A et B lors de l'achat d'un téléphone portable et le statut de l'acheteur (salarié ou non) a conduit au recueil des données sur 9 321 nouveaux acheteurs, enregistrées consécutivement sur un fichier client (l'étude portait sur 10 000 acheteurs, mais pour 679 d'entre eux,

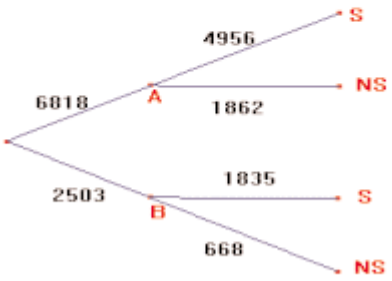
la donnée concernant le statut manquait). On peut ainsi représenter les données par l'un des tableaux (1) ou (2) ou par l'un des arbres (1) ou (2) ci-dessous. La seule information perdue par rapport à un tableau à 2 colonnes et 9 321 lignes est l'ordre des lignes.

| | A | B |
|---|---|---|
| S | 4956 | 1835 |
| NS | 1862 | 668 |

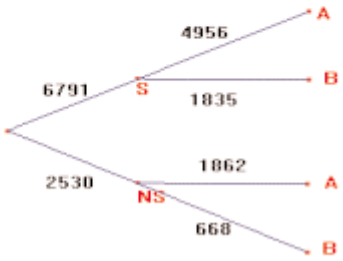
Tableau(1)

| | A | B | Totaux |
|---|---|---|---|
| S | 4956 | 1835 | 6791 |
| NS | 1862 | 668 | 2530 |
| Totaux | 6818 | 2503 | 9321 |

Tableau (2)




Arbre (1)




Arbre(2)

On notera qu'une seule de ces quatre représentations des données permet de reconstituer les trois autres.
Dans certaines études (par exemple si les lignes sont les années des deux dernières élections présidentielles, les colonnes donnant le nombre de votants et le nombre d'abstentions), les totaux par colonnes (dans l'exemple des élections) ou par ligne n'ont pas d'intérêt : dans ce cas un seul des deux arbres ci-dessus est utile (le second dans l'exemple des élections).

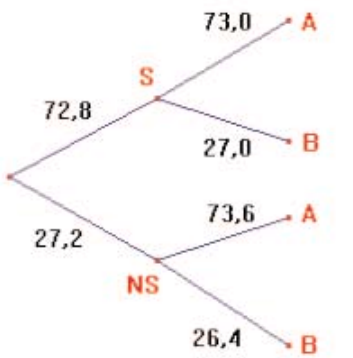
Sur des exemples, les élèves devront savoir passer d'un tableau à un arbre et vice-versa. Aucun formalisme n'est à développer.
À partir du tableau (2), on peut construire les tableaux (3) et (4) ou les arbres (3) et (4) :

| | A | B | |
|---|---|---|---|
| S | 72,7 | 73,3 | 72,8 |
| NS | 27,3 | 26,7 | 27,2 |
| | 100 | 100 | 100 |

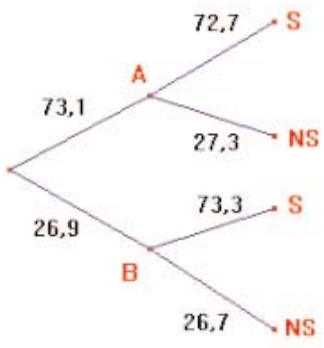
Tableau (3)

| | A | B | |
|---|---|---|---|
| S | 73 | 27 | 100 |
| NS | 73,6 | 26,4 | 100 |
| | 73,1 | 26,9 | 100 |

Tableau (4)




Arbre(3)




Arbre(4)

À titre d'exercice, on pourra dans un exemple analogue à celui-ci, reconstruire les tableaux ou les arbres des effectifs à partir d'un des deux arbres ou d'un des deux tableaux ci-dessus et du nombre $n$ total d'individus. Pour reconstruire le tableau (2) connaissant le tableau (3) et $n = 9321$, on a à résoudre par exemple le système :

$$r + r' = 9321 \text{ et } 72{,}6r + 73{,}3r' = 72{,}8 \times 9321.$$

On peut à ce niveau réfléchir au mode de calcul de la fréquence $f_A(S)$ des salariés parmi les clients choisissant A et arriver à la formule :

$$f_A(S) = \frac{f(A \text{ et } S)}{f(A)} \approx 72{,}7.$$

On notera dans cet exemple que cette fréquence est sensiblement la même que celle des salariés dans l'échantillon considérée (72,8). On verra dans la partie « Test d'indépendance » comment interpréter ces données.
Dans le paragraphe suivant, on donne une sens à l'égalité ci-dessus lorsqu'on remplace les fréquences d'événements par des probabilités.

## Probabilité conditionnelle et indépendance

Étudions une expérience de référence : dans une urne, il y a des pièces indiscernables au toucher, argentées ou dorées (A ou D), certaines en euros, d'autres en francs. Il y a 60 pièces dorées, dont $k$ sont en francs et 40 pièces argentées, dont $30 - k$ sont en francs. À cette situation, on peut associer un tableau ou des arbres donnant les probabilités des événements D et €, D et F, A et €, A et F. Par analogie avec les distributions de fréquences manipulées dans la paragraphe 1, peut alors définir la probabilité de F sachant D par :

$$P_D(F) = \frac{P(D \text{ et } S)}{P(D)}.$$

Si $P_D(\text{F}) = P(\text{F})$, (soit $k/60 = 0{,}3$, soit $k = 18$), c'est-à-dire si le fait de savoir que la pièce tirée est dorée ne change pas sa probabilité d'être en franc, on dit que F est indépendant de D, ce qui s'écrit aussi : $P(\text{D et F}) = P(\text{D}) \times P(\text{F})$ ; on en déduit la notion d'indépendance entre deux événements ; les trois assertions suivantes sont équivalentes pour des événements de probabilités non nulles :

i) $P_D(\text{F}) = P(\text{F})$ ;
ii) $P(\text{D et F}) = P(\text{D}) \times P(\text{F})$ ;
iii) $P_{\text{F}}(\text{D}) = P(\text{D})$.

En utilisant les propriétés des lois de probabilité, on peut démontrer que si D et F sont indépendants, les événements D et € le sont aussi, ainsi que de A et € et A et F ; les variables aléatoires *métal* et *monnaie* sont dites indépendantes.
On peut alors généraliser et définir la probabilité conditionnelle d'un événement B quelconque sachant un événement A de probabilité non nulle, puis l'indépendance de A et B.

Pour deux variables aléatoires, on introduit la définition suivante :
Deux variables aléatoires $X$ et $Y$ définies sur un ensemble E muni d'une loi de probabilité $P$, pouvant prendre les valeurs $(x_1,\ldots, x_k)$ et $(y_1,\ldots, y_r)$, sont indépendantes si pour tout couple $(i,j)$ :

$$P(X = x_i \text{ et } Y = y_j) = P(X = x_i) \times P(Y = y_j).$$

Pour $n$ tirages de pièces avec remise, la fluctuation d'échantillonnage fait que le tableau donnant les fréquences des événements D et €, D et F, A et €, A et F ne sera quasiment jamais identique au tableau donnant les probabilités de ces quatre événements. Il en sera presque sûrement d'autant plus proche que $n$ est grand. On peut alors se poser la question inverse : au seul vu d'un tableau d'effectifs ou de fréquences, comment pourrait-on reconnaître qu'il y a indépendance des variables *métal* et *monnaie* dans l'urne considérée ?

On évitera de masquer la difficulté d'établir un lien entre des propriétés d'un modèle (ici indépendance de deux événements) et la seule connaissance de données empiriques ; si on reprend l'exemple du paragraphe1, on trouve, à partir du tableau (1) :

$$f(\text{A et S}) = 4956/9321 \approx 0{,}5317$$

$$f(\text{A}) \times f(\text{S}) = (6818/9321) \times (6791/9321) \approx 0{,}5329$$

Ces nombres sont *presque* égaux et la question de l'indépendance entre les événements considérés, ou encore ici entre les deux caractères étudiés (abonnement et statut) se pose naturellement. Le sens que donnent les statisticiens à cette question est le suivant : peut-on considérer que ces 9321 résultats pourraient être obtenus par tirage avec remise dans une urne comportant des boules marquées A ou B d'une part, S ou NS d'autre part ? Cette question reste ouverte pour l'élève au niveau de la terminale ; l'enseignant pourra se reporter au paragraphe « Test d'indépendance » pour éclaircir cette question.

*Remarque* – La notion formelle d'indépendance (on dit aussi indépendance stochastique) entre deux événements est une propriété numérique à l'intérieur d'un modèle.
Ainsi, soit un ensemble de 97 pièces telles celles de l'exemple ci-dessus ; comment faire pour que les variables *métal* et *monnaie* soient indépendantes (on suppose qu'aucune de ces variables n'est constante) ? Une telle question provoquera chez un mathématicien une autre question : mais pourquoi 97 pièces et pas 98 ou 99 ou 100 ? Il trouvera alors très vite qu'il n'y a pas de solutions. En effet, $P(\text{D et €}) = P(\text{D}) \times P(€)$ implique :

$$97\text{Card}\,(\text{D} \cap €) = \text{Card}\,(\text{D}) \times \text{Card}\,(€),$$

où Card(D) désigne le nombre d'éléments de l'ensemble D ; comme 97 est un nombre premier, l'égalité ci-dessus est impossible. Il s'agit là de considérations numériques.

Comment alors rattacher la notion formelle d'indépendance à des modes de pensée intuitifs et qualitatifs ? Un premier pas consiste à dire, comme cela est fait ci-dessus, que B est indépendant de A lorsque savoir que A s'est réalisé ou non ne permet pas de changer les prévisions sur la réalisation de B ; la symétrie de la notion d'indépendance doit alors être prouvée.

Dans le langage courant, la notion de dépendance ou d'indépendance a souvent, dans l'histoire des probabilités, été associée à une notion de causalité. On peut se demander si la dépendance stochastique est ou non une traduction formelle de la notion de dépendance causale. La cas simple est celui d'une causalité déterministe correspondant à un événement A inclus dans un événement B : on a alors $P(\text{A et B}) = P(\text{A})$ et il n'y a pas indépendance stochastique (sauf si B est de probabilité 1).
Comme on le voit dans les lignes ci-dessous, des liens entre dépendance causale et stochastique peuvent être explicités : la dépendance causale implique la dépendance stochastique, mais la réciproque est inexacte. Plus précisément :
– *En pratique*, s'il y a dépendance causale avérée entre une cause A et un effet B, cela se traduit à l'intérieur d'un modèle par une dépendance stochastique entre A et B ; et l'indépendance, ou l'indépendance conditionnellement à un événement C (à savoir : $P_C(\text{A et B}) = P_C(\text{A}) \times P_C(\text{B})$) s'interprètent comme une absence de causalité entre A et B.
Le terme *en pratique* signifie ici qu'on peut inventer des exemples fictifs, qu'on ne rencontre pas dans la pratique, mais qui constituent des contre-exemples aux règles énoncées.
Considérons en effet l'exemple fictif suivant : un défaut de fabrication B cause, pour des raisons d'ordre mécanique, la défaillance D d'un moteur avec une probabilité $p$, *i.e.* $P_B(\text{D}) = p$ ; mais on pourrait imaginer que $P_{\overline{B}}(D)$, où $\overline{B}$ est le complémentaire de B, soit aussi égal à $p$ ; par exemple si on se place dans un ensemble E de moteurs ayant tous des défauts et si, lorsque B est absent, d'autres défauts sont présents dont la combinaison provoque D avec la même probabilité $p$. On a alors $P(\text{D et B}) = P(\text{D}) \times P(\text{B})$ et on ne peut pour autant nier la causalité mécanique de B vis-à-vis de D. L'indépendance stochastique résulte ici d'une coïncidence numérique.
Notons cependant que $P_B(\text{D})$ peut être inférieur à $P_{\overline{B}}(D)$ ; il en est ainsi si le défaut B exclut la présence d'autres défauts, ceux-ci provoquant plus facilement la

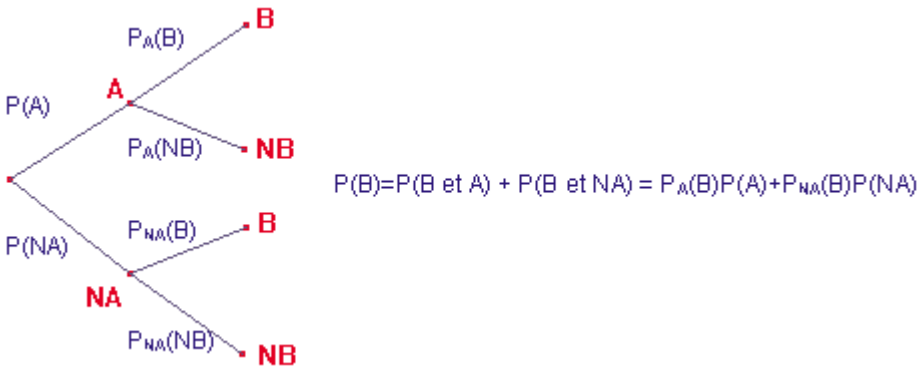
défaillance que B : la cause B est *protectrice*. Par exemple, si D est le décès d'un enfant par accident imputable à un vaccin V contre une maladie M, la probabilité $P_V(D)$ n'est pas nulle (le risque 0 n'existe pas) mais elle est très faible devant la probabilité $P_{\overline{V}}(D)$ de décès par la maladie M.
– La dépendance stochastique n'implique pas la dépendance causale des phénomènes modélisés.
Prenons le contre-exemple fictif suivant : un candidat à la mairie d'un arrondissement R d'une grande ville envoie à 90% des habitants de cet arrondissement une lettre (événement L) exposant sa politique, et indépendamment (au sens stochastique), il envoie à 50 % d'entre eux une boîte de chocolats (événement C) ; il n'envoie lettre et chocolat que dans son arrondissement, lequel regroupe 10 % des habitants de la ville ; la probabilité qu'un habitant de la ville rencontré par hasard ait reçu une lettre (resp. des chocolats) est $P(\text{L}) = 0{,}09$ (resp. $P(\text{C}) = 0{,}05$) ; les événements L et C ne sont pas stochastiquement indépendants car :

$P(\text{L et C}) = 0{,}9 \times 0{,}5 \times 0{,}1 = 0{,}045$ et $P(\text{L}) \times P(\text{C}) = 0{,}09 \times 0{,}05 = 0{,}0045$.

On serait ici peu enclins à dire que la lettre est une cause de la réception d'une boîte de chocolats ou vice-versa. Les événements L et C sont indépendants conditionnellement à la cause « être domicilié dans l'arrondissement R » et on retrouve ainsi l'indépendance causale entre L et C. Ce contre-exemple illustre un phénomène non exceptionnel.
En dehors de quelques situations simples, la causalité est une notion délicate à cerner et à manipuler ; une dépendance de nature causale peut être conjecturée ou validée par des études statistiques ; cependant, ni la causalité ni l'absence de causalité ne peuvent être prouvées avec certitude sans recours à des considérations propres au domaine où l'on se place. Et si dans un modèle, il y a indépendance, ou indépendance conditionnelle de deux événements, c'est le plus souvent parce qu'on a construit le modèle pour qu'il en soit ainsi (voir « Test d'indépendance »).

## Formule des probabilités totales

On explicitera ainsi dans le cas général les éléments des représentations graphiques en arbre (en particulier, on note B plutôt que (B et A) pour alléger l'écriture ; sur les *premières* flèches on indique des probabilités, puis sur les autres des probabilités conditionnelles). L'élève pourra, pour répondre à certaines questions, tracer un arbre et donner un résultat utilisant la formule des probabilités totales en ajoutant directement les produits des probabilités des arcs composant les chemins menant à un sommet terminal.




Arbre (3)

On en déduira la formule des probabilités totales pour une partition à deux éléments, et on pourra alors généraliser.

**Formule des probabilités totales :**
Soit E un ensemble muni d'une loi de probabilité P, et $C_1,\ldots, C_k$ des ensembles de probabilités non nulles formant une partition de E. Pour tout événement A de E, on a :

$$P(A) = P_{C_1}(A) \times P(C_1) + \ldots + P_{C_k}(A) \times P(C_k)$$

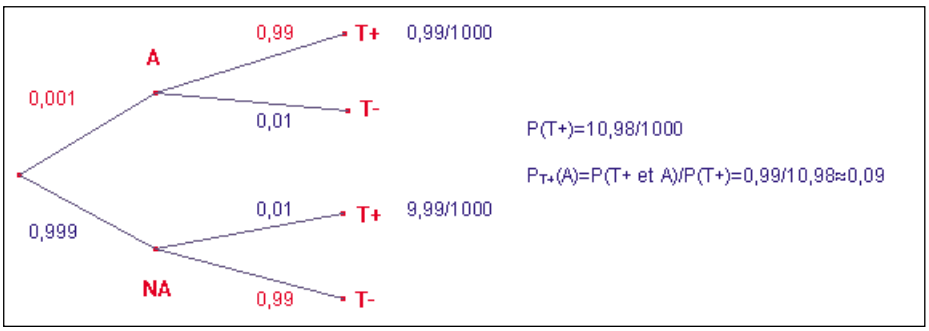
Lorsque des études permettent de déterminer des nombres $P_{C_1}(A)$, $i = 1 \ldots k$, cette formule permet de calculer la probabilité de A dans des populations variées où les $C_i$ se répartissent différemment. Il en est ainsi dans l'application ci-dessous.

> *Remarque* – Certains auteurs appellent formule des probabilités totales l'égalité $P(A) = P(A \text{ et } C_1) + \ldots + P(A \text{ et } C_k)$.
> L'essentiel n'est pas tant le nom que l'écriture d'une formule correcte.
> On admettra en pratique que $P(A) = P(A \text{ et } C_1) + \ldots + P(A \text{ et } C_k)$
> et $P(A) = P_{C_1}(A) \times P(C_1) + \ldots + P_{C_k}(A) \times P(C_k)$
> sont deux écritures de la formule des probabilités totales. L'essentiel n'est en effet pas le nom de la formule employée, mais son exactitude et son efficacité.

On pourra faire un ou deux exercices utilisant la formule des probabilités totales pour calculer la probabilité d'un événement A à partir d'une partition comportant plus de deux éléments.

## Tests de dépistage systématique

Un test de dépistage à la naissance d'un caractère génétique noté A, de probabilité $p = 0{,}001$, est fourni par une firme avec les spécificités suivantes :
– la probabilité qu'un individu ayant le caractère A ait un test positif est 0,99 ;
– la probabilité qu'un individu n'ayant pas le caractère A ait un test négatif est 0,99.
On se demande quelle est la probabilité pour qu'un individu dont le test est positif ait le caractère A.
L'encadré ci-dessous fournit la réponse ; le programme indique qu'un arbre de probabilité bien construit constitue une preuve, de même qu'un tableau ; cela signifie sur cet exemple que l'encadré ci-dessous constitue une réponse parfaitement justifiée à la question posée.
Les données du problème sont représentées sur l'arbre ci-dessous.




Un individu dont le test est positif a une probabilité 0,09 d'avoir le caractère A.

Le test peut aussi être appliqué à diverses populations, la probabilité $p$ qu'un individu ait le caractère A variant d'une population à l'autre. La valeur prédictive du test est la probabilité qu'un individu dont le test est positif soit malade.
Dans le cadre d'un cours de mathématiques, il est intéressant d'établir pour ce test la formule :

$$P_{T+}(A) = \frac{99p}{98p+1}.$$

D'où un tableau donnant quelques valeurs de $P_{T+}(A)$ :

| $p$ | 0,001 | 0,010 | 0,100 | 0,300 | 0,500 | 0,800 |
|---|---|---|---|---|---|---|
| $P_{T+}(A)$ | 0,090 | 0,500 | 0,912 | 0,977 | 0,990 | 0,997 |

On remarque que :
– la valeur prédictive du test n'est pas une notion intrinsèque au test lui-même : elle varie fortement selon la population ciblée. Pour un fabricant, améliorer la qualité du test, c'est faire en sorte d'augmenter $P_A(T+)$ et de diminuer $P_{\bar{A}}(T+)$ ; par contre, le fabricant d'un tel test n'a aucune maîtrise sur la valeur de $p$ ;

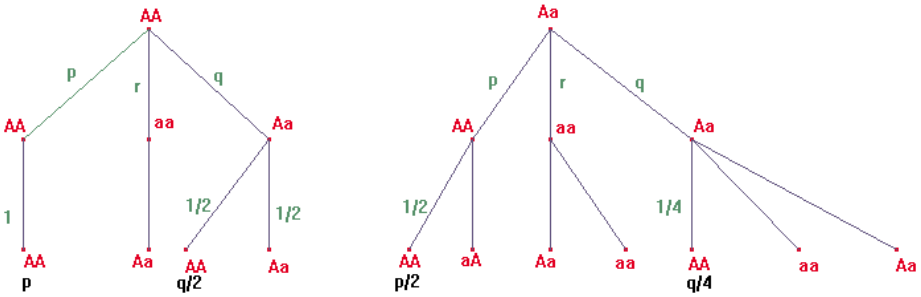
– dans les cas où $p$ est faible, la valeur prédictive du test l'est aussi. Ainsi, si le caractère A révèle la présence d'une maladie rare, un test de dépistage systématique de toute une population aura l'inconvénient majeur de fournir beaucoup de faux positifs (individus non malades dont le test est positif). Pour ces derniers, l'inquiétude liée à la découverte d'un test positif peut-être grande : c'est là un des problèmes éthiques liés à la mise en place des tests de dépistage systématique d'une maladie rare.
On remarquera cependant que, par exemple pour $p = 0{,}01$, la connaissance de la positivité du test multiplie par 50 la probabilité d'être atteint de la maladie : un test positif est toujours un élément à prendre en compte dans un processus de diagnostic ;
– si la population ciblée est celle d'individus présentant des symptômes évocateurs de la présence du caractère A (il ne s'agit plus alors de dépistage systématique) ou une population dite à risque pour la pathologie révélée par A, $p$ n'est pas faible : la positivité du test sera un élément important du diagnostic ;
– en inversant les rôles de $p$ et $1 - p$, pour $p < 0{,}10$, on voit que la probabilité qu'un individu dont le test est négatif ne soit pas atteint de la maladie étudiée est supérieure à 0,999 : le test est utile pour exclure le caractère A.

## Loi de Hardy-Weinberg

Dans les cas simples, un gène peut prendre deux formes (ou allèles) A et a et un individu peut avoir l'un des trois génotypes suivants : AA, Aa, aa. Considérons une population (génération 0) dont les proportions respectives de ces génotypes sont $p$, $q$, $r$. Un enfant hérite d'un gène de chaque parent, chaque choix de gène se faisant au hasard. On admet que les couples se forment au hasard quant aux génotypes considérés (appariement aléatoire).
Comment évoluent les proportions de génotypes dans la population à chaque génération ? On note $p_n$, $q_n$, $r_n$ les proportions des génotypes AA,Aa, aa à la génération $n$.
Commençons par la première génération.
Si on sait calculer $p_1$ en fonction de $p$, $q$, $r$, on déterminera $r_1$ en intervertissant dans la formule donnant $p_1$ les lettres $p$ et $r$ ; on en déduira $q_1$ par la formule $1 = p_1 + q_1 + r_1$.
Pour calculer $p_1$, on peut conditionner par le génotype du père. Comme l'enfant ne peut pas être AA si le père est aa, en tenant compte des deux arbres ci-dessous, on trouve :

$$p_1 = (p + q/2)p + (p/2 + q/4)q = (p + q/2)^2 = (1 + p - r)^2/4.$$

D'où $r_1 = (1 + r - p)^2/4$.




Dans chacun des deux arbres ci-dessus, on a mis en première ligne le génotype du père, en seconde ligne celui de la mère ; en troisième ligne on en déduit les génotypes possibles pour un enfant.
Notons $d = p - r$. On a alors :

$$p_1 = (1 + d)^2/4 \text{ et } r_1 = (1\text{-d})^2/4,\ q_1 = (1 - d^2)/2$$

Mais $p_1 - r_1 = d$, il s'ensuit que :

$$p_2 = (1 + d)^2/4 \text{ et } r_2 = (1 - d)^2/4,\ q_2 = (1 - d^2)/2$$

et plus généralement, pour $n > 0$ : $p_n = (1 + d)^2/4$ et $r_n = (1 - d)^2/4$, $q_n = (1 - d^2)/2$.
Il apparaît ainsi que la répartition des génotypes est stable à partir de la première génération : c'est ce qu'on appelle la loi de Hardy-Weinberg ; cette loi a été établie en 1905 conjointement par le mathématicien anglais G.H. Hardy et un médecin allemand W. Weinberg.

*Remarque* – Ce résultat pourra faire l'objet d'un problème ; mais, de même que le professeur peut parfois exposer aux élèves une *belle* démonstration que ceux-ci n'auraient pu faire eux-mêmes, il peut aussi développer devant eux le calcul ci-dessus, pour illustrer l'intérêt de combiner réflexion et calcul.

#### Application

Une maladie M est causée par la présence d'un allèle récessif ; soit, si on note A cet allèle :
– un individu AA est malade ;
– un individu Aa n'est pas malade mais peut transmettre la maladie (porteur sain) ;
– un individu aa n'est pas malade et ne peut pas transmettre la maladie.
Sachant qu'en Europe, la répartition de la maladie est stabilisée avec un enfant atteint sur 2 500, comment estimer la proportion de porteurs sains ?
On suppose que la loi de Hardy-Weinberg s'applique. La répartition stable vérifie : $P(\text{AA}) = \alpha^2$, $P(\text{aa}) = (1-\alpha)^2$ et $P(\text{Aa}) = 2\alpha(1-\alpha)$. Soit $\alpha = (1/2500)^{1/2} \simeq 0{,}02$.
La probabilité d'être porteur sain dans ce modèle est $2\alpha(1-\alpha) = 0{,}0392$ ; on remarque, dans les formules ci-dessus, que pour $\alpha$ petit, $P(\text{Aa})$ est voisin de $2\alpha$, *i.e.* la probabilité d'être porteur sain est voisine du double de la racine de la probabilité d'être malade.

## Expériences indépendantes ; expériences indépendantes et identiques

Choisir un chiffre au hasard signifie qu'on adopte le modèle défini par l'équiprobabilité sur l'ensemble des chiffres. Dans le même ordre d'idée, dire que $k$ expériences sont indépendantes, c'est se placer dans un modèle pour lequel la probabilité d'une liste de $k$ résultats est le produit des probabilités de chacun d'entre eux.
Dire que des expériences sont *identiques* signifie que le modèle adopté pour chacune d'elles est le même : on pourra dire ainsi que lancer deux pièces équilibrées, c'est faire deux expériences identiques.

#### Exercice : Anonymat

On fait une enquête sur le tabac dans un lycée. On fabrique pour cela le questionnaire suivant :
Lancer une pièce à *pile* ou *face*. Si elle tombe sur *pile*, répondez à la question :
*Est-ce que vous fumez plus d'un paquet de cigarettes par semaine ?*
La réponse est donnée en cochant l'une des deux cases oui ou non en bas du questionnaire.
Si elle tombe sur *face*, relancer la pièce une deuxième fois et répondez par oui ou non à la question :
*Est-ce que vous êtes tombé sur pile au deuxième lancer ?*
La réponse est donnée en cochant l'une des deux cases oui ou non en bas du questionnaire.
Lorsqu'un questionnaire porte la réponse oui (resp. non), il est impossible de savoir s'il s'agit d'une réponse à la question 1 ou à la question 2. On suppose que grâce à ce procédé les élèves donnent des réponses sans mentir.
On recueille une proportion $p$ de oui. Modéliser la situation et estimer en fonction de $p$ la proportion de fumeurs dans ce lycée.

La répétition de $k$, $k < 4$, expériences de Bernoulli (expériences ayant deux issues possibles) indépendantes peut donner lieu à des représentations en arbre, l'indépendance permettant de remplacer des probabilités conditionnelles par des probabilités : le professeur pourra ou non appeler de tels arbres des schémas de Bernoulli. L'étude de tels arbres n'est pas un objectif du programme.
Enfin, l'étude de l'évolution temporelle de systèmes pouvant prendre deux états peut conduire, après utilisation de la formule des probabilités totales, à l'étude de suites $(p_n)$ du type $p_{n+1} = ap_n + b$.

*Remarques*
– Parler de $k$ expériences identiques et indépendantes, c'est considérer la loi de probabilité qui à tout élément $(r_1,\ldots, r_k)$ associe $P(r_1) \times \ldots \times P(r_k)$, où $P$ est la loi de probabilité qui modélise chacune des expériences. Ce modèle est construit de telle sorte que les variables aléatoires $X_1,\ldots, X_k$ sont indépendantes, où $X_i(r_1,\ldots, r_k) = r_i$, $i = 1\ldots k$. En effet, par définition, l'indépendance de $k$ variables aléatoires signifie que pour tout $(r_1,\ldots, r_k)$ :

$$P(X_1 = r_1 \text{ et} \ldots \text{ et } X_k = r_k) = P(X_1 = r_1) \times \ldots \times \mathrm{P}(X_k = r_k\ ).$$

– On évitera de dire que deux expériences relevant du même modèle mais qui n'ont rien à voir entre elles sont identiques (par exemple : opérer un malade avec une probabilité $10^{-5}$ de complications graves et jouer à un jeu de hasard avec une probabilité $10^{-5}$ de gagner).
– Il a été vu en première qu'un modèle d'une expérience aléatoire est une loi de probabilité $P$ sur l'ensemble des résultats E. On associe parfois mentalement à l'expérience réelle une expérience de référence relevant du même modèle (faire de telles associations n'est pas un objectif du programme). Il doit être cependant clair que le modèle est la loi de probabilité $P$ sur l'ensemble E et non un tirage de boules dans une urne (on évitera pour cela de parler de *modèle d'urne*).
– L'indépendance est liée à l'absence de mémoire ; on dit ainsi souvent que les résultats à la roulette sont indépendants car *la roulette est sans mémoire.*
La convergence, sur un grand nombre d'expériences, des fréquences vers leurs probabilités est empiriquement remarquablement vérifiée lors de la réalisation d'un processus expérimental *sans mémoire*. Au niveau de l'intuition, il y a là un paradoxe : comment un processus sans mémoire peut-il conduire à une régularité prévisible ? Le paradoxe disparaît si on acquiert l'intuition que le changement d'échelle fait passer d'un modèle aléatoire à un modèle déterministe : étudier les expériences une par une nécessite un modèle aléatoire et les étudier par paquet de $n$, $n$ très grand, conduit à un modèle déterministe ; l'aléatoire a partie liée à l'échelle où se situe l'observateur.

# Études de deux variables quantitatives

Le programme de la classe terminale ES fait une place importante à la réflexion autour du traitement de l'information chiffrée. Il s'agit, pour ce chapitre, de trouver un équilibre entre le traitement mathématique des données et la prise en compte du contexte dont elles sont issues.
Il n'est pas utile de multiplier les exercices courts et techniques, mais on veillera à motiver les activités par un questionnement.

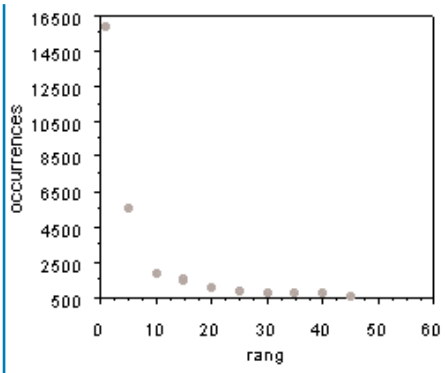
## Représentation de données

**Exemple**

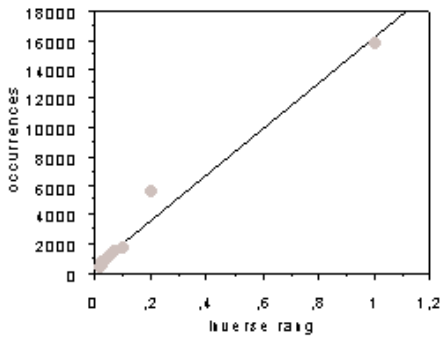
*Données du site http://hobart.cs.umass.edu/allan/cs646/char_of_text*
Dans un ensemble de 423 courts articles du journal *Time* totalisant 245 412 mots, on a classé les mots du vocabulaire par ordre décroissant de leur nombre d'apparitions dans l'ensemble des articles.
La figure (1), bien que les points d'ordonnée inférieurs à 2 000 soient mal représentés, incite à regarder si la décroissance de $y$ est simplement en $1/x$ et pour cela à représenter $y$ en fonction de $1/x$ ; la figure (2) représentant $y$ en fonction de $1/x$ fait apparaître des points presque alignés. Mais le problème de la qualité de la représentation se pose toujours. Pour y remédier, on peut prendre le logarithme des abscisses et des ordonnées. En effet, si les produits sont à peu près constants, alors $\log(y_i)$ sera presque une fonction affine de $\log(x_i)$. La figure (3) semble aller dans ce sens ; il conviendrait de continuer avec les autres mots de cet ensemble de textes pour confirmer ce phénomène ; nous ne disposons pas de ces données. En revanche, nous disposons des données résultant d'une autre expérience, faite cette fois-ci en dénombrant les mots distincts d'un corpus de 46 500 articles de journaux, totalisant 19 millions de mots.

On a refait la même étude (figures (4) et (5)). On constate à peu près le même phénomène.

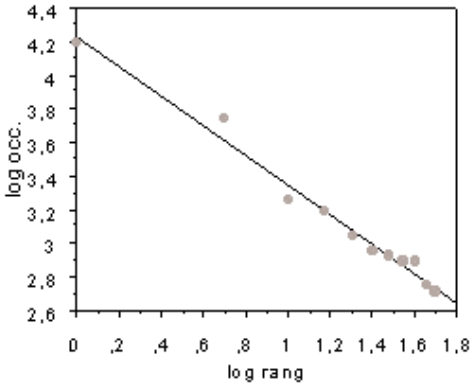
| Mot | *the* | *and* | *with* | *on* | *but* | *have* | *so* | *week* | *its* | *new* | *into* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Rang | 1 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
| Occurrences | 15861 | 5614 | 1839 | 1551 | 1138 | 914 | 868 | 793 | 793 | 572 | 518 |




(1)



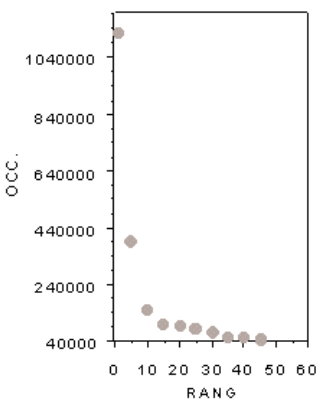

(2)



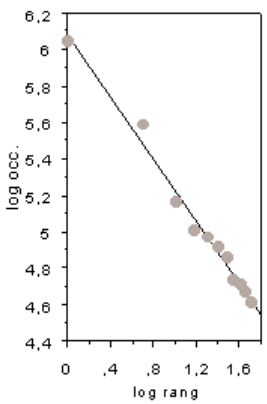

(3)

| Mot | *the* | *in* | *said* | *at* | *million* |
|---|---|---|---|---|---|
| Rang | 1 | 5 | 10 | 15 | 20 |
| Occurrences | 1130021 | 390819 | 148302 | 101779 | 93515 |

| Mot | *company* | *but* | *or* | *would* | *trade* | *their* |
|---|---|---|---|---|---|---|
| Rang | 25 | 30 | 35 | 40 | 45 | 50 |
| Occurrences | 83070 | 71887 | 54958 | 50828 | 47310 | 40910 |

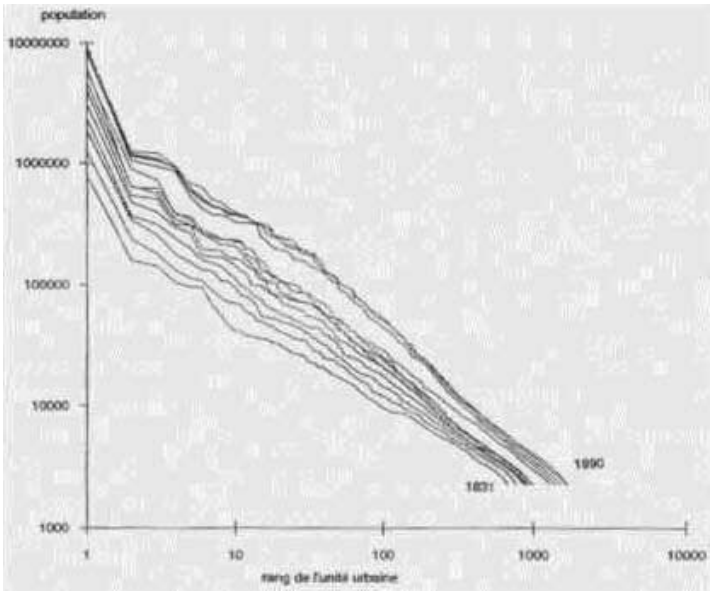



(4)




(5)

Ces deux exemples illustrent effectivement une **loi empirique**, vérifiée pour de nombreuses langues, appelée loi de Zipf ; cette loi énonce que le produit du rang du mot par sa fréquence reste à peu près constant.
Cette loi empirique s'applique aussi pour les données suivantes dans de nombreux pays : on classe les villes par ordre décroissant de leur nombre d'habitants, et en excluant les quelques plus grandes villes et les plus petites, le produit du rang par la taille garde le même ordre de grandeur. Ainsi, sur la figure (6), on trouve, pour quelques années entre 1831 et 1990, des représentations des points de coordonnées $(i, n_i)$, où $n_i$ est le nombre d'habitants de la ville de rang $i$ : on observe un assez bon alignement des points pour les rangs compris entre 10 et 1000, sur une droite de pente environ –1. Les géographes appellent souvent cette loi empirique la loi rang-taille des villes et en font un outil de référence pour lui comparer la répartition rang-taille effectivement observée.
On pourra commenter les échelles dans le graphique ci-dessous.




(6) Évolution de la distribution rang-taille des unités urbaines françaises entre 1831 et 1990.
*Source : Deux siècles de croissance urbaine*, coll. « Villes », Économica, 1993.

## Ajustement par moindres carrés

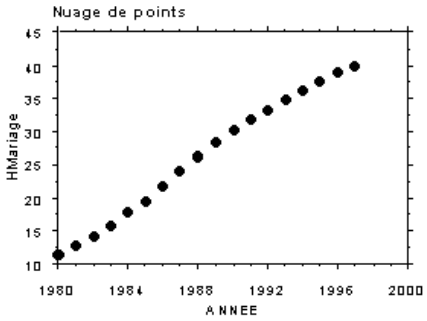
L'objectif du programme est de comprendre, à partir d'exemples simples, ce qu'est une « droite d'ajustement par moindres carrés » et d'illustrer son utilisation pour interpoler ou extrapoler quelques valeurs numériques.

**Exemple 1**

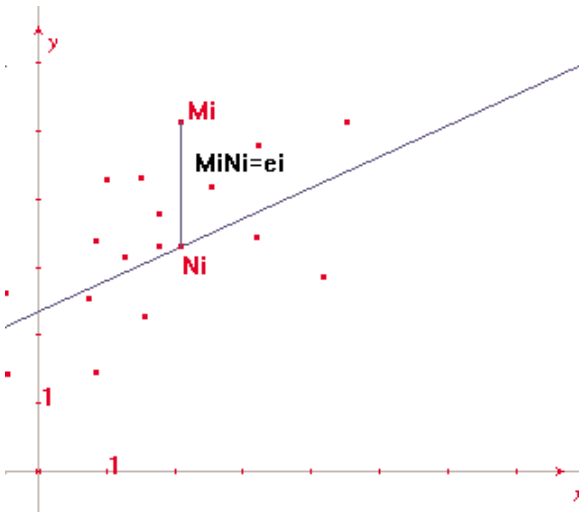
On a représenté graphiquement ci-dessous les valeurs des pourcentages de naissances hors mariages en France, entre 1980 et 1997. Nous nous intéressons ici aux deux questions suivantes :
a) Une description qualitative simple du nuage de points consiste à dire qu'il y a, à peu près, croissance linéaire pendant cette période. Comment rendre compte quantitativement de cette observation ?
b) Comment estimer le taux de naissances hors mariage en 1998 ?

| Année | 1980 | 1981 | 1982 | 1983 | 1984 | 1985 | 1986 | 1987 | 1988 | 1989 |
|---|---|---|---|---|---|---|---|---|---|---|
| Pourcentage | 11,4 | 12,7 | 14,2 | 15,9 | 17,8 | 19,6 | 21,9 | 24,1 | 26,3 | 28,2 |

| Année | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 |
|---|---|---|---|---|---|---|---|---|
| Pourcentage | 30,1 | 31,8 | 33,2 | 34,9 | 36,1 | 37,6 | 38,9 | 40,0 |




On peut envisager de résumer quantitativement l'observation faite en donnant l'équation d'une droite qui *ajuste au mieux* les *n* points du nuage. Pour tenir compte de la deuxième question, on cherche une droite D telle que les erreurs $e_i = y_i - \hat{y}_i$ (où $\hat{y}_i$ est le point de D d'abscisse $x_i$), soient *petites*.




On choisira plus précisément de minimiser $\varepsilon^2 = \sum (y_i - \hat{y}_i)^2 / n$ ; on peut se demander pourquoi considérer $\varepsilon^2$ et non $\delta = \sum |y_i - \hat{y}_i| / n$ ; une des raisons à cela est que la minimisation de $\varepsilon^2$ conduit à une solution unique et à une formule simple, à savoir $\hat{y} = a(x - \bar{x}) + \bar{y}$, où $\bar{x}$ et $\bar{y}$ désignent les moyennes des abscisses et des ordonnées et $a = \dfrac{\sum (y_i - \bar{y}) \times (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = r \dfrac{s_y}{s_x}$.

La droite d'ajustement linéaire par moindre carrés est aussi appelée droite de régression linéaire par moindres carrés, ou plus simplement droite de régression. Pour illustrer ce résultat, on pourra montrer que si on a 3 ou 4 points tels que $\bar{x} = \bar{y} = 0$, alors pour une pente de droite donnée, $\varepsilon^2$ est minimum si la droite passe par l'origine ; à titre d'exercice, on peut alors montrer que $\varepsilon^2$ est minimum pour $a = \sum x_i y_i / \sum x_i^2$.

On indiquera que si les abscisses et les ordonnées ont des dimensions, la dimension de *a* doit être celle des ordonnées divisée par celle des abscisses. On pourra enfin regarder à partir des formules comment se transforme l'équation de la droite d'ajustement linéaire par moindres carrés si on change d'unités sur les abscisses par exemple (*i.e.* par transformation affine des abscisses).

Dans l'exemple considéré, l'équation de la droite d'ajustement est :

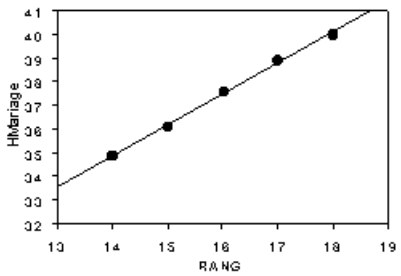
$$\tau(t) = 9{,}47 + 1{,}78(t - 1979).$$

Sous l'hypothèse d'un accroissement annuel du pourcentage de naissances hors mariage égal à 1,78, on trouve $\tau'(1998) = 43{,}3$. Cette extrapolation des données

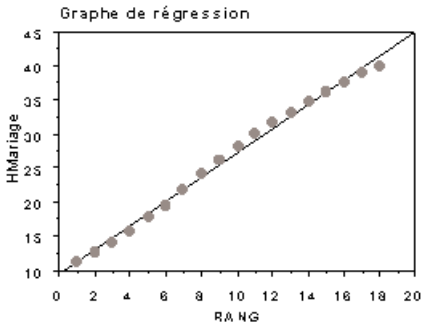
repose sur les 18 années précédentes : est-il vraiment pertinent ici d'utiliser toutes ces données ? Entre 1980 et 1997, les pourcentages observés ont varié de 10 à 40 %, mais un examen un peu plus précis du nuage des 18 points montre un infléchissement pour les dernières années observées ; aussi pour l'extrapolation demandée, on peut limiter aux quelques années précédentes. Avec les cinq années de 1993 à 1997, l'équation de la droite d'ajustement par moindres carrés est :

$$\tau'(t) = 16{,}7 + 1{,}30(t - 1979).$$

Sous l'hypothèse que cette tendance linéaire (accroissement du pourcentage 1,30 par an) se maintienne, on trouve $\tau'(1998) = 41{,}4$. Si on fait les calculs avec les trois années 1995-1996-1997, on trouve 41,2 : une prévision raisonnable est l'intervalle [41,2 ; 41,4]




Droite d'ajustement sur les années 1993-1997 ; taux de naissances hors mariages en fonction de $x$=$t$-1979.




Droite d'ajustement ; taux de naissances hors mariages en fonction de $x$=$t$–1979

*Remarque* – On démontre que $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$ et en divisant par $n$, $s_y^2 = s_{\hat{y}}^2 + \varepsilon^2$.
On notera que si on change d'unité pour les $y_i$, ce qui revient à les multiplier par un nombre $k$, alors la pente et l'ordonnée à l'origine de la droite d'ajustement linéaire par moindre carrés est multipliée par $k$ et la somme des carrés des erreurs est multipliée par $k^2$. Dire que la somme des carrés des erreurs est *petite* n'a donc pas de sens : il convient de regarder si elle est *petite par rapport à la variance* des ordonnées, où, ce qui revient au même, à regarder si D est proche de 1, avec :

$$\Delta = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{s_{\hat{y}}^2}{s_y^2}.$$

$\Delta$ représente la proportion de la variance des ordonnées expliquée par l'ajustement linéaire. Des calculs simples montrent que $\Delta$ est le carré du coefficient de corrélation linéaire $r$, soit :

$\Delta = r^2$, avec $r = \dfrac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\left(\sum(x_i - \bar{x})^2\right)^{1/2}\left(\sum(y_i - \bar{y})^2\right)^{1/2}}$.

Si $\Delta = 1$, les points du nuage sont alignés et plus $\Delta$ (ou $r$) est proche de 1, meilleur est l'ajustement : quantifier ce propos est délicat et nécessite un modèle probabiliste ; la quantification de la qualité de l'ajustement n'est pas un objectif du programme.

On pourra consulter, sur le logiciel *SEL* présent sur le cédérom, le lexique correspondant au terme « régression linéaire simple » et regarder comment varie la somme des carrés des erreurs lorsqu'on ajuste « à la main » un nuage de points par une droite. En consultant le lexique au terme « donnée aberrante », on pourra observer la sensibilité à une valeur aberrante de la droite d'ajustement par moindres carrés.

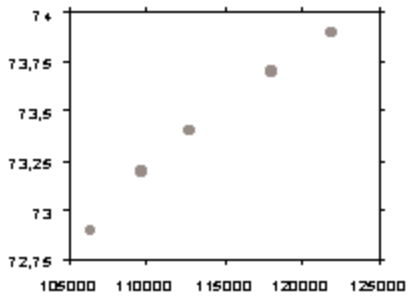
**Exemple 2**

Le graphique (1) page suivante représente, pour les années 1991 à 1995, le nombre des divorces en France entre 1991 et 1995 (en abscisse) et l'espérance de vie à la naissance des hommes (en ordonnée). Les cinq points sont quasiment alignés. On peut

chercher une explication à cet alignement ; ici, les graphiques (2) et (3) indiquent que l'accroissement annuel pour les deux quantités considérées est à peu près constant d'où :

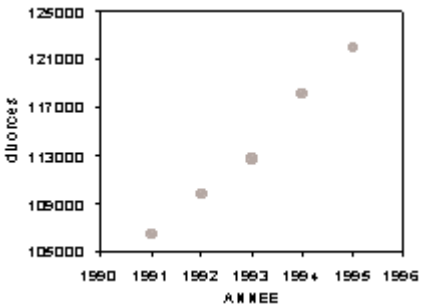
$\Delta y_i \approx a\Delta t$, $\Delta x_i \approx b\Delta t$ soit $\Delta y_i \approx \frac{a}{b}\Delta x_i$, c'est-à-dire que les points $(x_i, y_i)$ sont « presque » alignés.

On notera en conséquence que si un lien de causalité est *a priori* suspecté entre deux quantités $x$ et $y$ (ce n'était pas le cas ici !), le quasi alignement des points du nuage est éventuellement un indice en faveur de ce lien, mais n'en constitue absolument pas une preuve.

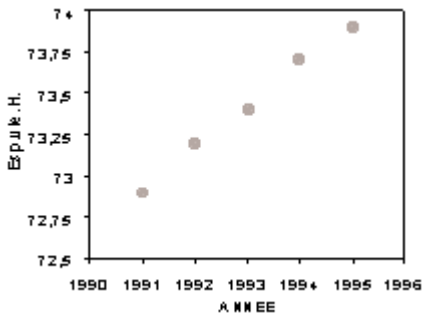



(1)

*De gauche à droite, les points représentent, dans l'ordre les années 1991 à 1995.*




(2)




(3)

# Lois de probabilités

## Lois de probabilités discrètes : loi de Bernoulli, loi binomiale

L'étude de la loi binomiale sera l'occasion d'introduire la notation $n!$. On peut se reporter au document d'accompagnement du programme de l'option facultative de terminale L (présent sur le cédérom joint) pour une première approche.

L'application ci-dessous est un thème d'étude possible, à la croisée de plusieurs chapitres : calculs de limites de suites, exponentielle, radioactivité, sensibilisation à des lois définies sur $\mathbb{N}$.

### Les événements rares suivent-ils une loi ?

À un certain carrefour très fréquenté, il y a en moyenne, depuis 10 ans, un accident par an (c'est-à-dire qu'il y a eu 10 accidents en 10 ans) ; peut-on évaluer la probabilité que, les choses étant inchangées, il n'y ait aucun accident l'an prochain? Ou qu'il y en ait exactement 1 ou 2 ?

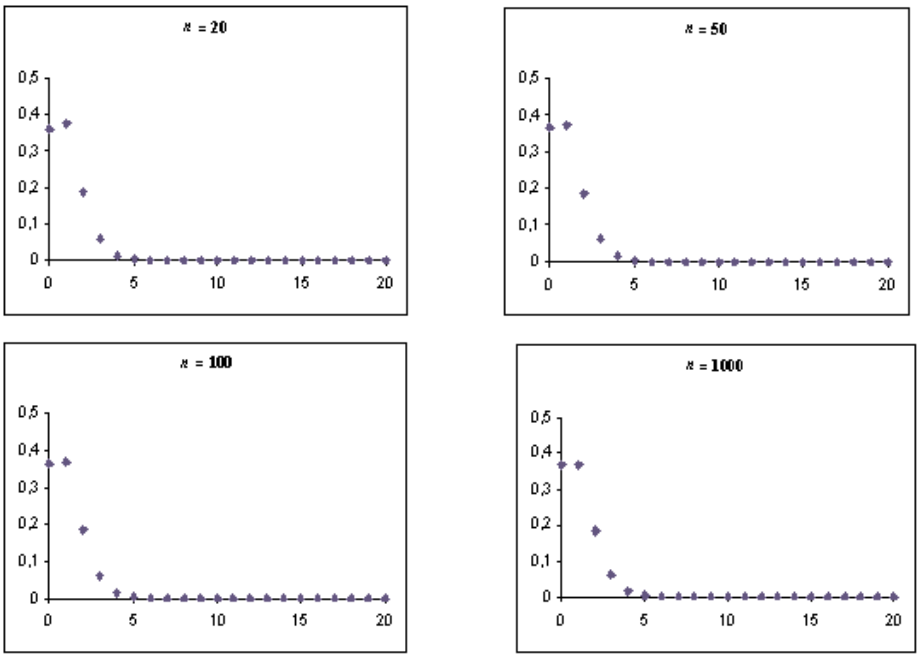
Il semble que cette question n'ait pas beaucoup de sens, si l'on ne possède pas d'autres informations. Et pourtant, il est possible d'émettre quelques hypothèses en utilisant des propositions classiques sur les limites ; ce qu'on explique ici s'appliquera, plus généralement, aux événements rares qui surviennent au cours d'une activité fréquente : par exemple, pannes de matériel, accidents d'avion, randonneurs frappés par un éclair, désintégration des noyaux d'une substance radioactive, etc.

On utilisera uniquement la formule pour la loi binomiale de paramètres $n$ et $p$, et les deux lemmes d'analyse suivants, qui sont au programme de terminale :

**Lemme 1.** On a $\lim\limits_{n\to+\infty}(1-1/n)^n = e^{-1}$ et plus généralement, $\lim\limits_{n\to+\infty}(1-a/n)^n = e^{-a}$.

**Lemme 2.** Si la suite $(u_n)_{n\in\mathbb{N}}$ tend vers $a$ et si la suite $(v_n)_{n\in\mathbb{N}}$ tend vers $b$, alors la suite produit $(u_n.v_n)_{n\in\mathbb{N}}$ tend vers $ab$.

Une modélisation simple, voire simpliste, pour la situation ci-dessus est la suivante : on suppose que chaque année, un million de véhicules passent par le carrefour. Le modèle envisagé est une loi binomiale de paramètres $10^6$ et $10^{-6}$ (on lance $10^6$ fois une pièce qui tombe sur *face* avec probabilité $10^{-6}$ et on compte le nombre de *face*).
Il a pourtant un défaut : en général, on connaît la moyenne, mais on ne sait pas vraiment le nombre de véhicules qui passent ; peut-être y en a-t-il $10^5$ ? ou $10^7$ ? On aurait alors une modélisation avec $10^5$ tirages, avec probabilité $10^{-5}$ à chaque tirage, ou bien $10^7$ tirages, avec probabilité $10^{-7}$ à chaque tirage. Si ces modèles donnaient des résultats très différents, notre modélisation serait inutilisable.
Autrement dit, nous voulons comparer les lois binomiales obtenues par $n$ tirages indépendants avec probabilité $1/n$. Voici le dessin pour des valeurs de $n$ = 20 ; 50 ; 100 ; 1000 ; on n'a représenté que les probabilités d'avoir des valeurs entre 0 et 20. Au-delà de 10, les probabilités sont trop petites pour être correctement représentées avec l'échelle choisie :




En particulier, on voit que la probabilité d'avoir 0 se stabilise aux environs de 0,37.

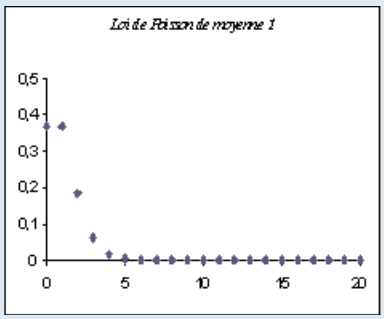
Il est en fait facile de trouver la limite exacte : pour une loi binomiale de paramètres $n$ et $1/n$, la probabilité d'avoir 0 est $\left(1-\frac{1}{n}\right)^n$, et l'on sait (lemme 1) que la suite de terme général $\left(1-\frac{1}{n}\right)^n$ tend vers $1/e$ quand $n$ tend vers $+\infty$ : une valeur approchée de la limite est 0,368. Cherchons maintenant un résultat analogue pour la probabilité de $k$.

Notons $P_n(k)$ la probabilité d'avoir $k$ avec une loi binomiale de paramètres $n$ et $1/n$.
On sait que : $P_n(k) = \binom{n}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{n-k}$.

On vérifie facilement que $P_n(k+1) = \frac{n-k}{(k+1)(n-1)}P_n(k)$.

Or, $\lim\limits_{n\to+\infty} \dfrac{n-k}{(k+1)(n-1)} = \dfrac{1}{k+1}$. Donc, si pour $k$ fixé, $P_n(k)$ tend vers une limite $L_k$, alors d'après le lemme 2, $P_n(k+1)$ tend vers $L_{k+1} = L_k/(k+1)$ quand $n$ tend vers l'infini.
En particulier, puisque $P_n(0)$ tend vers $L_0 = 1/e$, $P_n(1)$ tend vers $L_1 = 1/e$ et $P_n(2)$ vers $L_2 = 1/2e$.

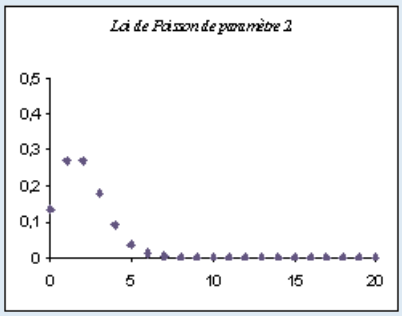
*Remarque* – Une récurrence montre que $P_n(k)$ tend vers $L_k = e^{-1}/k!$ ; les lois B($n$,1/$n$) sont définies sur [0,$n$] ; lorsque $n$ tend vers l'infini, s'il y a une loi de probabilité limite $P$, celle-ci est définie sur $\mathbb{N}$. On admettra la propriété suivante : $e = \lim\limits_{n\to+\infty} u_n$ avec $u_n = 1 + 1/2 + \ldots + 1/n!$ (on pourra d'abord observer ce résultat sur tableur avant de l'admettre, en notant qu'il s'agit là d'un résultat important qui conduit à de nombreux théorèmes et applications). D'où $\lim\limits_{n\to+\infty}\left(L_0 + L_1 + \ldots + L_n\right) = 1$. En posant alors $P(k) = L_k$, on voit que les lois binomiales B($n$,1/$n$) *convergent* vers la loi $P$.
Cette loi limite est une *loi de Poisson* de moyenne 1. Voici le graphe de cette loi de Poisson, très peu différent bien sûr de ceux qui précèdent.

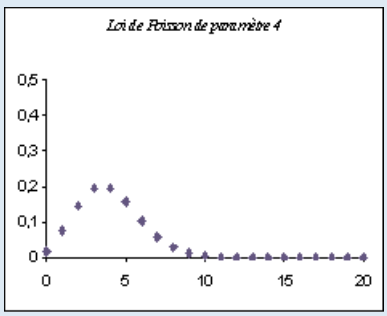



Dans ce modèle, pour des accidents rares qui arrivent en moyenne une fois par an, on a environ 36 % de chances qu'aucun accident ne se produise une année donnée, la même chance qu'il y en ait un, environ 18 % de chances qu'il y en ait 2, 6 % de chances qu'il y en ait 3, et 2 % qu'il y en ait 4 ; au-delà, la probabilité devient très faible.

Et pour les moyennes différentes de 1 ?
On peut reprendre un raisonnement analogue. On considère une loi binomiale de moyenne $a$, donc de paramètres $n$ et $a/n$. La probabilité d'avoir 0 est alors $(1 - a/n)^n$ ; d'après le lemme 1, elle tend vers $e^{-a}$ quand $n$ tend vers l'infini.
On montre comme précédemment que $P_n(0)$ tend vers $L_0 = e^{-a}$, $P_n(1)$ tend vers $L_1 = ae^{-a}$ et $P_n(2)$ vers $L_2 = e^{-a}a^2/2$.

*Remarque* – En reprenant les mêmes notations que ci-dessus, on obtient $P_{n+1}(k) = \dfrac{a(n-k)}{(k+1)(n-a)} P_n(k)$.
On montre que $\dfrac{a(n-k)}{(k+1)(n-a)}$ tend vers $\dfrac{a}{k+1}$ quand $n$ tend vers l'infini ; comme ci-dessus, ceci permet de calculer par récurrence sur $k$ la limite de $P_n(k)$ pour $k$ fixé et $n$ tendant vers l'infini.
On vérifie que cette limite vaut $e^{-a}\dfrac{d^k}{k!}$. On admet la propriété suivante : $e^a = \lim\limits_{n\to+\infty} v_n$ avec $v_n = 1 + a^2/2 + \ldots + a^n/n!$. On retrouve bien le cas précédent en prenant $a = 1$. La loi limite est appelée loi de Poisson de paramètre $a$.
Voici quelques exemples de lois de Poisson (paramètres respectifs 2 et 4)



Il s'avère que les observations que l'on peut faire *suivent* remarquablement de telles lois de Poisson, par exemple pour des accidents à des carrefours ou pour la désintégration des noyaux d'une substance radioactive.

## Lois continues

Nous proposons ci-dessous une introduction aux lois de probabilité à densité continue, qui fait naturellement suite au cours sur l'intégration et l'enrichit d'applications importantes, telles la modélisation de la durée de vie d'un noyau d'une substance radioactive (voir le document à ce sujet). Si quelques exercices faciles peuvent être proposés pour faire fonctionner ce concept de loi de probabilité sur un sous-ensemble de $\mathbb{R}$, il convient de ne pas oublier qu'il s'agit d'une toute première approche.
Aucune difficulté technique ne sera soulevée ; en particulier on ne traitera que des cas menant à des calculs d'intégrales s'exprimant aisément à l'aide des fonctions étudiées en terminale. Pour une loi sur $\mathbb{R}^+$, aucune notion d'intégrale généralisée n'est abordée formellement : l'outil *limite à l'infini* d'une fonction est suffisant.

### Que signifie choisir au hasard un nombre dans (0,1) ?

*Remarque* – (0,1) désignera l'un quelconque des intervalles [0,1], [0,1[, ]0,1] ou]0,1[.
On a fait la convention terminologique que choisir au hasard un élément d'un ensemble E fini, c'est considérer sur E la loi équirépartie, pour laquelle les probabilités des éléments de E sont égales.
Soit $E_2$ l'ensemble des nombres de [0,1[ dont l'écriture décimale comporte au plus 2 chiffres après la virgule ; il y a $10^2$ éléments dans $E_2$ et la loi uniforme sur $E_2$ attribue à chacun de ces nombres la probabilité $10^{-2}$. La probabilité de l'ensemble des éléments de $E_2$ qui sont dans $]a,b]$ (ou $[a,b[$), où $a$ et $b$ sont dans $E_2$, vaut $b - a$. Plus généralement, soit $E_k$ l'ensemble des nombres de [0,1[ dont l'écriture décimale comporte au plus $k$ chiffres après la virgule ; il y a $10^k$ éléments dans $E_k$ et la loi uniforme sur $E_k$ attribue à chacun de ces nombres la probabilité $p = 10^{-k}$. La probabilité de l'ensemble des éléments de $E_k$ qui sont dans $]a,b]$ (ou $[a,b[$), où $a$ et $b$ sont dans $E_k$, vaut $b - a$. Ces calculs montrent que pour définir le choix au hasard d'un nombre réel dans [0,1[, on ne peut plus passer par la probabilité $p$ de chaque élément, puisqu'on devrait alors avoir $p = 0$ : cette difficulté a été un véritable défi pour les mathématiciens et a conduit à repenser la notion de loi de probabilité.
Un autre argument, conduisant à la même impossibilité de passer par la probabilité des éléments de [0,1[ pour définir la notion de choix au hasard, consiste à couper [0,1[ en $n$ ou en $2^n$ intervalles égaux ; si on admet que dans un modèle de choix au hasard, les intervalles de même longueur ont même probabilité, on trouve que la probabilité d'un point $x$ est inférieure à $1/n$ ou $1/2^n$, pour tout $n$, elle est donc nulle.

Par ailleurs, s'il est facile, dans le cas fini, d'imaginer des protocoles expérimentaux (tels des tirages de boules dans une urne) *réalisant* un choix au hasard dont le résultat est connu avec exactitude, la situation est différente pour [0,1[ : le résultat d'une mesure ne fournit qu'un intervalle où le résultat se situe. De même, si on veut donner le résultat d'un tel choix au hasard sous la forme de son écriture décimale, on ne pourra écrire qu'un nombre fini de décimales, ce qui en fait revient à définir un intervalle auquel il appartient.
Ces considérations conduisent à changer de point de vue : pour des ensembles tels que (0,1), une loi de probabilité sera caractérisée non plus par la probabilité des éléments mais par celle de ses intervalles.

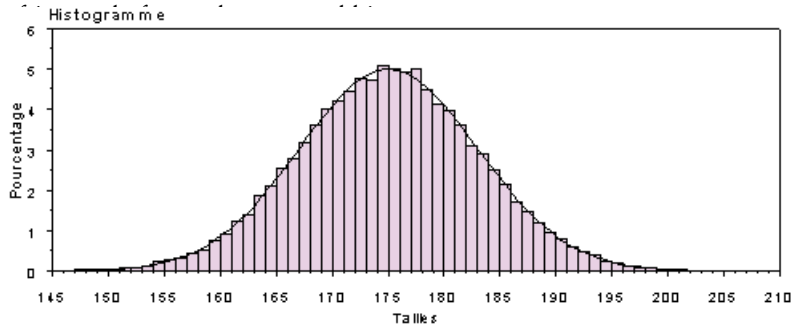
### Histogrammes et aire sous une courbe. Comment définir les probabilités d'intervalles ?

Imaginons la situation suivante :
On dispose d'un échantillon 50 000 tailles d'hommes adultes ; un résumé numérique de cet échantillon de 50 000 données est fourni dans le tableau ci-dessous.

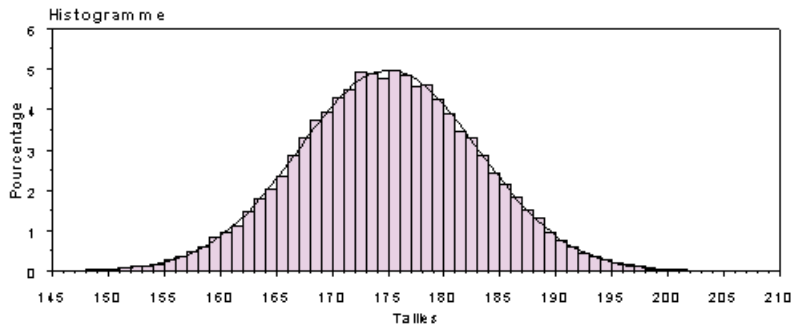
| | Moy. | Ecart typ | Nombre | Minimum | Maximum | Médiane | Interquartile |
|---|---|---|---|---|---|---|---|
| Tailles | 175,0 | 8,0 | 50000 | 145,1 | 209,5 | 175,0 | 10,8 |

Traçons maintenant un histogramme de ces données, avec un pas de 1 cm. Si l'unité d'aire est celle d'un rectangle correspondant à 1 cm en abscisse et un pourcentage de 0,01 en ordonnée, l'aire grisée ci-après est la somme des fréquences pour chaque intervalle et vaut 1. Si on cherche la fréquence des données entre 173 et 177 cm, ce sera la somme des aires des cinq rectangles de bases respectives [173,174[, …,[176,177[. En première approximation, ces rectangles ont tous une hauteur voisine de 0,05, la fréquence est donc voisine de 0,20 : la série comporte 50 000 données, et le modèle doit donc être tel que la probabilité d'avoir une taille entre 173 et 177 cm soit voisine de 0,20. L'idée est ici de trouver une fonction $f$ dont la courbe représentative *épouse* l'histogramme, l'aire sous cette courbe devant être égale à 1 : la probabilité d'un intervalle $(a,b)$ sera alors l'aire sous la courbe délimitée par les droites d'équations $x = a$ et $x = b$, c'est-à-dire le nombre $\int_a^b f(x)dx$. La courbe représentative d'une telle fonction $f$ a été tracée sur l'histogramme ; déterminer une telle fonction est un problème délicat, mais pour de nombreuses situations, dont celle qui est traitée ici, on cherche la fonction $f$ parmi une famille paramétrée de fonctions : il suffit alors d'ajuster les paramètres.

De même que dans un modèle défini par une loi de probabilité $P$ sur un ensemble fini E, les fréquences fluctuent autour de la loi $P$, de même ici, l'invariant est la fonction $f$ et pour des grandes séries de données, les histogrammes *fluctuent autour* du graphe de $f$. On peut voir ci-dessous une seconde série de 50 000 données ; on y a représenté la même fonction $f$ : l'histogramme n'a pas beaucoup bougé et la courbe représenta-




| | Moyenne | Ecart type | Nombre | Minimum | Maximum | Médiane | Interquartile |
|---|---|---|---|---|---|---|---|
| Taille | 175 | 8 | 50000 | 145,4 | 208,5 | 175 | 10,8 |




### Lois de probabilité à densité continue sur un intervalle

Pour la classe terminale, on se limite à des lois de probabilités définies sur un intervalle I borné ou borné à gauche (*i.e.* I = [a,b], ou I = [a, + ∞)) et dites *à densité continue*.

| I = $(a,b)$, loi $P$ de densité $f$. | I = $[a, +\infty)$, loi $P$ de densité $f$. |
|---|---|
| f est une fonction définie sur I, continue positive.<br>$P(I) = P\big((a,b)\big) = \int_a^b f(x)dx = 1$ | $f$ est une fonction définie sur I, continue positive.<br>$\lim\limits_{t\to+\infty} F(t) = 1$ où $F(t) = \int_a^t f(x)dx$. |
| Pour tout intervalle borné $(c,d)$ (ouvert, semi ouvert ou fermé) de I :<br>$P\big((c,d)\big) = \int_c^d f(x)dx$ et<br>– pour J⊂ J', $P$(J) ⩽ $P$(J') ;<br>– la probabilité de la réunion finie d'intervalles deux à deux disjoints est la somme des probabilités de chaque intervalle ;<br>– si J et J' sont deux intervalles complémentaires dans I, $P$(J') = 1 – $P$(J). | Pour tout intervalle borné $(c,d)$ (ouvert, semi ouvert ou fermé) de I :<br>$P\big((c,d)\big) = \int_c^d f(x)dx$ et $P\big((c,+\infty)\big) = 1 - F(c)$.<br>– pour J⊂ J', P(J) ⩽ $P$(J') ;<br>– la probabilité de la réunion finie d'intervalles deux à deux disjoints est la somme des probabilités de chaque intervalle ;<br>– si J et J' sont deux intervalles complémentaires, dans I, $P$(J') = 1 – $P$(J). |

On remarquera que pour de telles lois, la probabilité d'un intervalle réduit à un élément est nulle.
On conviendra alors que choisir un nombre au hasard dans I = $(a, b)$, c'est le choisir selon la loi $P$ dont la densité vaut $1/(b - a)$. La probabilité d'un intervalle inclus dans I est égale au quotient de sa longueur par celle de I.
On pourra faire l'analogie avec les densités de masse (la masse d'un point est nulle, celle d'un segment est proportionnel à sa longueur dans le cas d'une tige de densité constante).
Dans le cas I = $(a, +\infty)$, l'étude de la fonction $F$ n'est pas un objectif du programme.

**Exemples d'exercices**

• Soit I = [0,1] et une loi de probabilité de densité $f$ avec $f(t) = 4t^3$.
Calculer $P$([0,25 ; 0,75]). Calculer $m$ tel que si on choisit un nombre dans I suivant cette loi de probabilité, la probabilité qu'il soit inférieur à $m$ soit 0,5.
• Soit I = $[0, +\infty)$ et une loi de probabilité de densité $f$ avec $f(t) = 2e^{-2t}$.
Calculer P($[n,n + 1]$). Calculer $m$ tel que si on choisit un nombre dans I suivant cette loi de probabilité, la probabilité qu'il soit inférieur à $m$ soit 0,5.
• Soit I = [1,10] et une loi de probabilité de densité $f$ avec $f(t) = \lambda t^{-2}$. Déterminer λ.
• Soit I = $[1, +\infty)$ et une loi de probabilité de densité $f$ avec $f(t) = \lambda t^{-2}$. Déterminer λ.

On définit des probabilités conditionnelles en étendant la définition donnée dans le cas des ensembles finis. Soit I' un intervalle de I, de probabilité non nulle et J un autre intervalle dans I ; la probabilité $P_{I'}$(J) de J sachant que I' est par définition égal à :

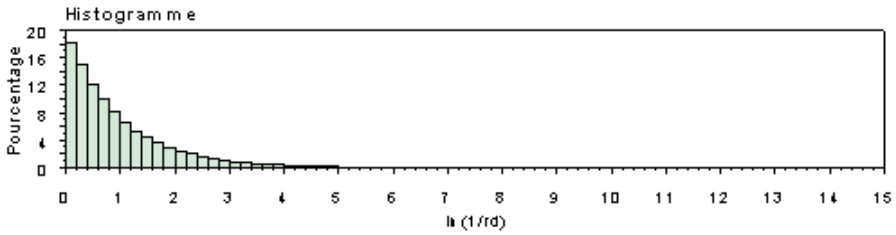
$$P_{I'}(J) = P(I' \cap J) / P(I').$$

En pratique, on se limitera pour les lois continues aux cas où J ⊂ I', pour lesquels $P_{I'}(J) = P(J) / P(I')$.

**Exemple d'exercice**

On choisit un nombre au hasard dans ]0,1[.
1) Sachant qu'il est inférieur à 0,3, quelle est la probabilité que le second chiffre après la virgule soit 1 ?
2) À l'aide d'un tableur, choisir 4 nombres au hasard $x_1,\ldots,x_4$ dans ]0,1[. À cette série de nombres, on associe la série de leurs logarithmes : $\ln(x_1),\ldots,\ln(x_4)$. La moyenne de cette seconde série est-elle égale au logarithme de la moyenne des quatre nombres ?
On considère la variable aléatoire $X$ qui à $x$ fait correspondre $-\ln(x)$.
Calculer $H(t) = P(X \leqslant t)$ ; déterminer la dérivée de la fonction $H$.

Voici par ailleurs l'histogramme correspondant à la série des opposés des logarithmes de 50 000 nombres choisis au hasard dans ]0,1[. Donner une fonction dont la courbe représentative *colle* à cet histogramme.




# Deux problèmes

## Adéquation à une loi

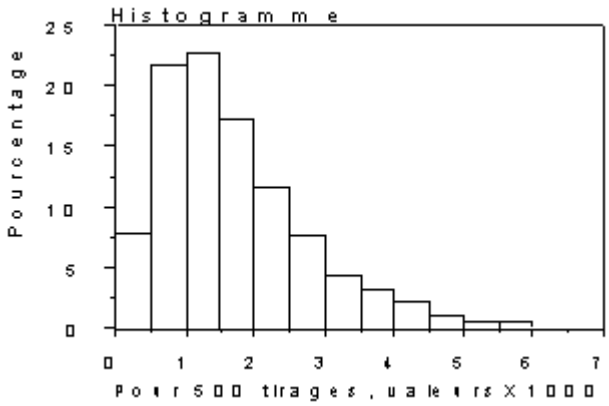
Les problèmes de validation de modèles sont complexes. Pour y sensibiliser les élèves, on peut commencer par des expériences de référence (tirages de boules dans des urnes, de choix au hasard, etc.) pour comprendre de quoi il s'agit, puis traiter des exemples.

*Un joueur veut vérifier si le dé qu'il utilise est équilibré. Comment peut-il faire ?*

Il pourrait étudier la symétrie du dé, mais ce n'est pas exactement de cela qu'il s'agit. Il convient plutôt de vérifier que dans des conditions normales d'utilisation du dé, les résultats sont *compatibles* avec un modèle d'équiprobabilité sur {1,2,3,4,5,6}.
Essayons donc de définir sur cet exemple un critère de compatibilité de données expérimentales avec un modèle.
On peut par exemple s'intéresser à la distance entre la distribution des fréquences $(f_1, \ldots, f_6)$ obtenues en lançant $n$ fois un dé et la loi de probabilité {1/6, …,1/6} et regarder si cette distance est *petite*. En prenant la définition classique de la distance, on peut fonder la notion de compatibilité sur l'étude du carré de cette distance, à savoir :

$$d^2 = (f_1 - 1/6)^2 + (f_2 - 1/6)^2 + \ldots + (f_6 - 1/6)^2.$$

La quantité $d^2$ est soumise à la fluctuation d'échantillonnage, *i.e.* sa valeur varie d'une série de lancers à l'autre. C'est précisément l'étude de la fluctuation d'échantillonnage qui va permettre de convenir d'un seuil entre valeur *petite* et valeur *non petite* de $d^2$.
Imaginons que le joueur ait lancé 500 fois le dé et ait obtenu une distribution de fréquence $(f_1, \ldots, f_6)$, d'où une valeur observée $d^2_{obs}$ qu'on va comparer à d'autres. Pour cela, on simule des séries de $n$ = 500 chiffres au hasard dans {1, …,6}. Ci-dessous, on voit un histogramme de 2 000 valeurs de $d^2$ obtenues par des simulations de séries de 500 chiffres au hasard dans {1,…,6}.




Le 9e décile de la série des valeurs simulées de $d^2$ est 0,003 (soit 90 % des valeurs simulées de $d^2$ sont dans l'intervalle [0 ; 0,003]).

Convenons de la décision suivante :

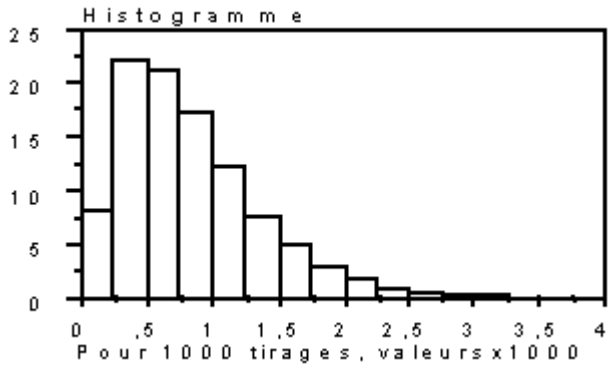
– si $d_{obs}^2 \leqslant 0{,}003$, alors le dé sera déclaré équilibré ;

– si $d_{obs}^2 > 0{,}003$, alors le dé sera déclaré non équilibré.

On associera à cette conclusion le risque $\alpha = 0{,}1$ correspondant au fait suivant : en utilisant cette règle de décision sur les données simulées, on se serait « trompé » dans 10 % des cas.

Deux éléments semblent arbitraires dans cette façon de conclure ou non à l'équilibre du dé : que se passerait-il pour une autre simulation, de taille égale ou non, et que se passerait-il si au lieu de lancer 500 fois le dé, on le lançait par exemple 1 000 fois ?

Qu'à cela ne tienne : étudions les résultats de 5 000 simulations de séries de 1 000 tirages de chiffres au hasard dans {1,…,6} ; le neuvième décile des valeurs $d^2$ est 0,0015, *i.e.* 90 % des valeurs de $d^2$ simulées sont dans l'intervalle [0 ; 0,0015]. Comparons les résumés graphiques dans les deux séries simulées : même allure, à l'échelle près. L'existence d'un changement d'échelle est conforme au théorème des grands nombres vu en première : plus le nombre de tirages est grand, plus la distribution des fréquences est proche de la loi de probabilité, donc plus les valeurs de $d^2$ sont petites.




Des résultats théoriques expliquent la ressemblance de forme entre les histogrammes : on sait démontrer que la loi de probabilité de $nd^2$, où $n$ est le nombre de tirages ($n$ vaut 500 puis 1 000 dans notre étude), ne bouge sensiblement plus avec $n$ lorsque $n$ est suffisamment grand ; c'est là un des nombreux résultats de la théorie des probabilités. En pratique, on pourra considérer que la loi de probabilité de $nd^2$ dépend peu de $n$, pour tout $n > 100$.

On peut adapter ce qui vient d'être fait en changeant le risque a, ou en testant l'adéquation à une loi équirépartie sur un ensemble à $k$ éléments pour diverses valeurs de $k$. La répartition des valeurs de loi de probabilité de $nd_k^2$ où :

$$d_k^2 = \left(f_1 - 1/k\right)^2 + \left(f_2 - 1/k\right)^2 + \ldots + \left(f_k - 1/k\right)^2$$

ne dépend quasiment plus de $n$ pour $n$ grand mais dépend cependant de $k$.

Le choix du risque $\alpha$ dépend du contexte : il est fonction de l'enjeu lié à la question et parfois on conclut non par rapport à un risque, mais par rapport à une plage de risques ; ce risque n'est en général pas choisi par le statisticien. On prend souvent par défaut $\alpha = 0{,}05$ : il est important que les élèves sachent qu'il s'agit d'un consensus et non d'une constante immuable. L'enjeu est de comprendre sur quoi porte le risque (refuser à tort le modèle) et que plus le risque est petit, plus on aura tendance à accepter le modèle de l'équiprobabilité.

**Exemple**

La répartition des sexes à la naissance est-elle équilibrée ?

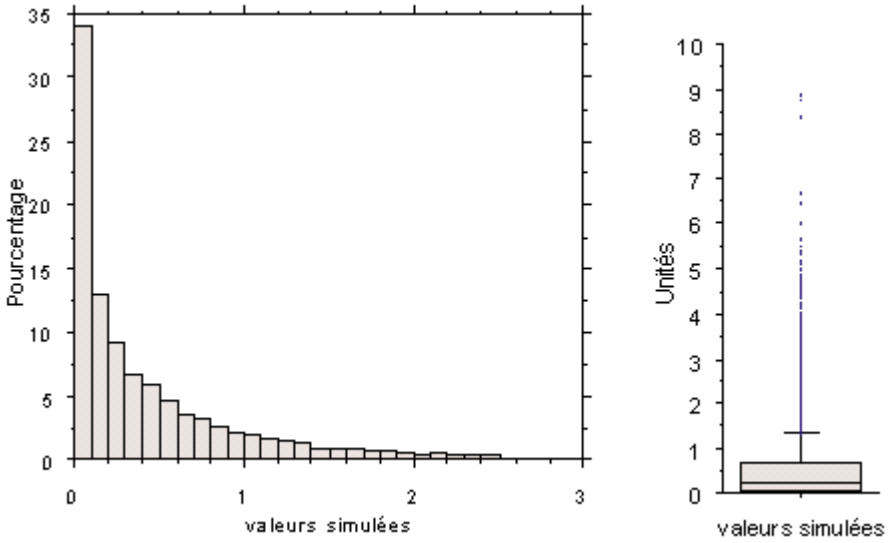
La traduction mathématique de cette question est ici : la loi de probabilité (0.5,0.5) est-elle pertinente pour modéliser la répartition des sexes à la naissance ?

On dispose pour répondre à cette question des données suivantes : sur $n = 53041$ naissances consécutives dans la ville de Grenoble, on observe 25 946 naissances de

filles. Comme on l'a vu ci-dessus, les variations en $n$ de la répartition des valeurs de $nd_2^2$ pouvaient être négligées dès que $n$ est grand, où :

$$d_2^2 = (f_1 - 1/2)^2 + (f_2 - 1/2)^2 = 2(f_1 - 1/2)^2.$$

On trouvera ci-dessous des résumés numériques et graphiques de résultats correspondants à 10 000 valeurs simulées de $nd_2^2$.




L'histogramme ne contient pas toutes les valeurs simulées de $n$ et est complété à sa droite par un diagramme en boîte.

| moyenne | écart-type | nombre | minimum | maximum | médiane | interquartile |
|---|---|---|---|---|---|---|
| 0,50 | 0,71 | 10000 | 0 | 8,84 | 0,23 | 0,59 |

Aux arrondis près, l'intervalle [0;1,4] contient 90 % des valeurs simulées et l'intervalle [0;2] en contient 95%.
Ici, la valeur observée de $nd_2^2$ est environ 12,5 : au vu des données et au risque $\alpha = 0{,}05$ (donc aussi au risque $\alpha = 0{,}1$), on rejette le modèle d'équiprobabilité et on conclut qu'il n'y a pas égalité des sexes à la naissance. La question se pose de généraliser ce résultat à d'autres villes, d'autres pays, et aussi de regarder si la proportion de filles est stable géographiquement, et au cours du temps. On pourra consulter à ce sujet le hors série n° 6 de la revue *La Recherche* paru en 2001.

*Remarques*

1) Une idée naturelle serait ici d'étudier les fluctuations de $\delta = |f_1 - 1/2|$ et de les comparer aux fluctuations de la même quantité lorsqu'on simule un grand nombre de séries de taille 53 041 de nombres (0,1) tirés au hasard. Mais on peut écrire : $d_2^2 = 2\delta^2$ et les deux études sont donc identiques. En particulier, comme on a vu que les variations en $n$ de la répartition des valeurs de $nd_2^2$ pouvaient être négligées dès que $n$ est grand, il en est de même pour $\sqrt{n\delta}$. La règle de décision adoptée ici est d'accepter le modèle équiréparti au niveau 0,95 si $\sqrt{n\delta} \leqslant 1$, soit si 1/2 est dans l'intervalle de confiance au niveau 0,95 de la fréquence $f_1$ observée (voir sur le cédérom le complément théorique de la fiche sur les sondages du document d'accompagnement de seconde).

2) Il existe en fait un théorème plus général, à savoir : dans le monde théorique défini par une loi $P = (p_1,\ldots, p_k)$, alors la loi de probabilité de la quantité :

$$\chi^2 = n\sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i}$$

est, pour $n$ grand, distribuée selon une loi qui ne dépend que de $k$. Cette loi s'appelle loi du khi deux à $k-1$ degrés de liberté et on trouve dans des tables numériques la liste des 9e déciles de ces lois (voir tableau ci-dessous). On peut ainsi généraliser la méthode ci-dessus et définir un critère de compatibilité d'une série de données avec une loi quelconque sur un ensemble fini.

| $k-1$ | 1 | 2 | 3 | 4 | 5 | 6 | 10 | 20 | 30 |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha = 0{,}1$ | 2,71 | 4,61 | 6,25 | 7,78 | 9,24 | 10,64 | 15,99 | 28,41 | 40,26 |
| $\alpha = 0{,}05$ | 3,84 | 5,99 | 7,81 | 9,49 | 11,07 | 12,59 | 18,31 | 31,41 | 43,77 |

Pour $k = 2$, on part d'une loi binomiale et les calculs peuvent se retrouver autrement (voir sur le cédérom, « Compléments aux documents d'accompagnement »).
L'objectif ici n'est pas que les élèves fassent eux-mêmes la simulation, mais qu'ils soient capables de définir une règle de décision et d'exploiter les résultats de simulations.

## Test d'indépendance

Dans le paragraphe « Probabilités conditionnelles et indépendance », on s'est posé la question de l'indépendance des variables abonnement et statut pour lesquelles on dispose du tableau suivant, donnant les résultats de ce couple de variables sur $N = 9321$ individus (voir tableau (1) ci-dessous).

| | A | B |
|---|---|---|
| S | 4956 | 1835 |
| NS | 1862 | 668 |

Tableau (1)

La traduction dans le champ de la statistique de cette question est : peut-on trouver un modèle compatible avec les données, défini par deux nombres $p$ et $r$ tels que les probabilités des 4 événements en jeu soient celles qui sont données dans le tableau ci-dessous :

| | A | B |
|---|---|---|
| S | $pr$ | $p(1-r)$ |
| NS | $(1-p)r$ | $(1-p)(1-r)$ |

Si tel est le cas, la quantité $d^2$ suivante doit être *petite* :

$$d^2 = \left(\frac{4956}{9321} - pr\right)^2 + \left(\frac{1835}{9321} - p(1-r)\right)^2 + \left(\frac{1862}{9321} - (1-p)r\right)^2 + \left(\frac{668}{9321} - (1-p)(1-r)\right)^2.$$

Mais on ne connaît ni $p$ ni $r$. Un objectif en statistique est ici de trouver une fonction des données d'un tableau tel le (1) dont la répartition se stabilise, lorsque le nombre de données devient grand, vers une répartition qui ne dépend ni de $p$ ni de $r$ ; c'est le cas pour la fonction définie ci-dessous, les données du tableau (2) étant remplacées par des lettres (tableau (2')) :

| | A | B | Totaux |
|---|---|---|---|
| S | 4956 | 1835 | 6791 |
| NS | 1862 | 668 | 2530 |
| Totaux | 6818 | 2503 | 9321 |

Tableau (2)

| | A | B | Totaux |
|---|---|---|---|
| S | a | b | n |
| NS | c | d | N-n |
| Totaux | m | N-m | N |

Tableau(2')

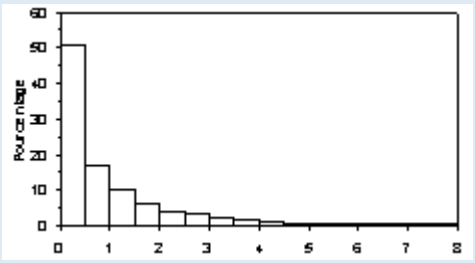
$$z = N\left(\frac{(a' - \hat{p}\hat{r})}{\hat{p}\hat{r}} + \frac{(b' - \hat{p}(1-\hat{r}))^2}{\hat{p}(1-\hat{r})} + \frac{(c' - (1-\hat{p})\hat{r})^2}{(1-\hat{p})\hat{r}} + \frac{(d' - (1-\hat{p})(1-\hat{r}))^2}{(1-\hat{p})(1-\hat{r})}\right)$$

où : $\hat{p} = \frac{n}{N}, \hat{r} = \frac{m}{N}, a' = \frac{a}{N}, \ldots, d' = \frac{d}{N}$

(on remarque que s'il existe un modèle compatible avec les données, défini par $p$ et $r$, où les événements A et S sont indépendants, alors $\hat{p} = \frac{n}{N}$ et $\hat{r} = \frac{m}{N}$ seront voisins de $p$ et $r$).

On peut écrire $z$ plus simplement sous la forme suivante : $z = \dfrac{N(ad-bc)^2}{nm(N-n)(N-m)}$.

On démontre en théorie des probabilités que la répartition asymptotique (*i.e.* pour N tendant vers $+\infty$) des valeurs de $z$ ne dépend ni de $p$ et $r$ ; en pratique, pourvu que $N$ soit suffisamment grand et $p$ et $r$ pas trop voisins de 0 ou 1, on approchera la loi de probabilité de $z$ par la loi limite, à savoir la loi du $\chi^2$ à 1 degré de liberté. Pour avoir l'allure de cette répartition simulons des tableaux avec $N = 1000$, $p = r = 1/2$ et calculons pour chaque tableau la valeur de $z$.




Histogramme de 2000 valeurs de z

On voit que 90 % des valeurs de la série simulée sont inférieures à 2,7 ; on peut convenir de la règle de décision suivante :
– si la valeur observée de $z$ est $\leqslant 2{,}7$, alors les variables en jeu seront dites indépendantes ;
– si la valeur observée de $z$ est $> 2{,}7$, alors les variables en jeu seront dites non indépendantes.
On associera à cette conclusion le risque $\alpha = 0{,}1$ correspondant au fait suivant : en utilisant cette règle de décision sur les données simulées, on se serait trompé dans 10 % des cas.
Si la valeur observée de $z$ est inférieure ou égale à 2,7, on pourra choisir le modèle défini par :

$$P(A \text{ et } S) = \hat{p}\hat{r},\ P(B \text{ et } S) = \hat{p}(1-\hat{r}),\ P(A \text{ et } NS) = (1-\hat{p})\hat{r},\ P(B \text{ et } NS) = (1-\hat{p})(1-\hat{r}).$$

Pour le tableau (1) la valeur observée de $z$ est 0,36 : les variables en jeu sont dites indépendantes au risque 0,1, ou au niveau de confiance 1-0,1 = 0,9. On pourra prendre le modèle suivant :

$$P(A \text{ et } S) = \hat{p}\hat{r} = 0{,}53,\ P(B \text{ et } S) = \hat{p}(1-\hat{r}) = 0{,}20,\ P(A \text{ et } NS) = (1-\hat{p})\hat{r} = 0{,}20,\ P(B \text{ et } NS) = (1-\hat{p})(1-\hat{r}) = 0{,}07\ .$$

La démarche suivie est donc de tester l'hypothèse qu'il existe un modèle où A et S sont indépendants et qui est compatible avec les données. Si cette hypothèse est acceptable, on construit alors une loi P qui vérifie $P(\text{A et S}) = P(\text{A})P(\text{S})$.

Cette conclusion suppose que les données manquantes ne masquent pas un phénomène spécifique ; ainsi, si les 679 cas exclus au départ sont tous des non salariés qui prennent l'abonnement B, alors la valeur observée de $z$ est 225 et la conclusion change !

## Statistique et TICE

Le développement rapide de l'usage de la statistique est lié à celui de l'informatique. Pour une sensibilisation à la statistique, dans le cadre d'un enseignement de mathématiques, il convient cependant de cerner en quoi les outils logiciels sont indispensables. Il ne s'agit pas d'initier les élèves à un logiciel spécialisé de statistique, ni même de les entraîner à utiliser systématiquement les possibilités de logiciels comme les tableurs ou les logiciels de géométrie (il serait utile que les enseignants acquièrent une bonne maîtrise de tels outils). On pourra se limiter à quelques possibilités indispensables à une mise en œuvre efficace des programmes de seconde, première et terminale.
On distinguera notamment les usages suivants :
– calculs simples, tels ceux de moyenne, d'écart type qui peuvent être faits sur calculatrice par un ordinateur quasi-instantanément même sur des séries de grandes tailles ;

– représentations graphiques diverses (l'usage d'un ordinateur permet de réfléchir sur le choix d'un pas convenable pour un histogramme), etc. ;
– calculs nécessitant le tri d'une série : médianes, quartiles, déciles. Le principal apport d'un logiciel est ici la possibilité de trier très rapidement un grand nombre de données. L'élève, pour tracer un diagramme en boîte associé à une longue série, pourra le faire « à la main », à partir de la simple observation de la série triée.
– la simulation : le programme de seconde insiste sur la nécessité de construire à partir de situations vécues le lien entre expériences réelles et simulations (à l'occasion de lancers de deux dés par exemple). Apprendre à simuler une expérience est un exercice formateur, tant au plan de la connaissance des phénomènes aléatoires qu'à celui du raisonnement. Le paragraphe « Deux problèmes » propose une approche de la notion de test, à l'aide de simulations.
Comme application de la notion de choix au hasard dans un intervalle $[a,b]$ et de celle d'expériences indépendantes, on peut estimer des aires à l'aide de simulations : si on choisit au hasard des nombres dans des intervalles I et I' bornés, la probabilité d'un rectangle de $I \times I'$ est le quotient de son aire par celle de $I \times I'$ ; on admet alors que la probabilité d'un sous-ensemble de $I \times I'$ est son aire.
On trouvera sur le cédérom des « appliquettes » à ce sujet, dans la section consacrée aux « Compléments aux documents d'accompagnement ».

## Sondages

On pourra reprendre la fiche « sondages » du document d'accompagnement de la classe de seconde et adapter, en tenant compte des connaissances de la classe terminale, l'aperçu théorique complétant cette fiche.
L'appliquette sur les fourchette de sondage peut permettre aux élèves de se familiariser avec la notion de fourchette de sondage. On pourra aussi consulter le site www.eduscol.education.fr/culturemath.
Il conviendra, pour les sondages effectivement réalisés par des instituts et relatifs par exemple à des élections, de bien séparer les situations suivantes :
– les sondages préélectoraux : cette situation est analogue au tirage de boules coloriées dans une urne ; les boules peuvent changer de couleur au cours du temps et le sondage reflète au mieux la répartition des couleurs à la date où il est pratiqué. De plus, dans cette situation, certaines boules, en sortant de l'urne, changent de couleur – mais reprennent leur couleur originelle si on les remet dans l'urne (les personnes sondées ne disent pas toujours à l'enquêteur leur choix réel). Certaines études faites sur des élections antérieures, comparées à la réalité des votes après dépouillement, permet d'établir un modèle dans lequel on connaît la loi du changement de couleur lors du tirage. Cela permet alors de « redresser » les calculs et d'estimer, dans le cadre de ce modèle, les proportions de chaque couleur dans l'urne. Les calculs faits sont « justes » à l'intérieur de ces modèles, mais la loi de changement de couleur peut évoluer d'une élection à l'autre et dans ce cas, les estimations faites ne sont plus pertinentes ;
– les estimations faites « à 20 heures » à partir du dépouillement d'échantillons de bulletins de vote. La situation est comparable ici au tirage au hasard d'un grand nombre de boules dans une urne (les boules ne changent plus de couleurs au cours du temps ou en sortant de l'urne). Ces estimations sont très précises et assorties d'un niveau de confiance élevé : on a extrêmement peu de chances de donner des chiffres éloignés de ceux qui tomberont après le dépouillement de la totalité des urnes.
Il convient enfin de distinguer d'une part la question de la fiabilité des sondages (fourchette de sondage et estimation qualitative de la fiabilité des lois de fausses réponses lors de l'enquête) des usages et interprétations des résultats qu'ils apportent.

## Liens avec les autres disciplines

On trouvera sur le cédérom :
– un extrait du document d'accompagnement de physique pour la terminale de la série S, où une expérience de lancers de dés et des simulations sont proposées pour éclairer le processus de désintégration radioactive ;

– un extrait du document d'accompagnement de chimie de la même classe. On simule des courbes d'évolution de réactions chimiques en comparant ce processus à des tirages de boules dans des urnes sous différentes conditions.

## Problèmes divers

Les possibilités de calculs sur tableur peuvent motiver la recherche de formules exploitables au plan numérique pour résoudre un problème, indépendamment de l'intérêt théorique d'une telle formule.

**Exemple**

Dans la fiche statistique « Faites vos jeux » du document d'accompagnement de la classe de seconde (disponible sur le cédérom joint), on s'intéresse à la probabilité qu'il y ait au moins 6 résultats consécutifs égaux dans une série de $n$ lancers d'une pièce équilibrée. On peut simuler cette situation, comme cela est proposée dans la fiche (ou faire des calculs matriciels tout à fait hors de portée d'un élève de terminale).
On peut aussi se demander si le résultat est calculable à partir d'une formule simple et exploitable sur tableur pour les valeurs de $n$ susceptibles de nous intéresser : établir une telle formule est l'objet du texte ci-dessous.
Les lancers d'une pièce de monnaie équilibrée sont associés comme on l'a vu précédemment à un modèle bien déterminé. Si on note $X_n$ le résultat du $n$-ème lancer :

$$P(X_n = 0) = \frac{1}{2} \text{ et } P(X_n = 1) = \frac{1}{2}.$$

On construit un compteur pouvant prendre les valeurs 1, …,6, la valeur 6 indiquant la présence d'au moins une séquence de 6 résultats consécutifs égaux. Un exemple est donné ci-dessous, où les résultats des lancers sont en première ligne et la valeur du compteur en deuxième ligne.

| $x(n)$ | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y(n)$ | 1 | 2 | 3 | 1 | 2 | 1 | 1 | 1 | 2 | 3 | 1 | 1 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 6 | 6 | 1 | 2 | 3 | 1 | 1 |

Cela revient à considérer les variables aléatoires $(Y_n)_{n>0}$, définies par :

$$Y_1 = 1 \text{ et } Y_n = \begin{cases} 6 \text{ si } Y_{n-1} = 6 \\ Y_{n-1} + 1 \text{ si } Y_{n-1} < 6 \text{ et } X_n = X_{n-1} \\ 1 \text{ sinon} \end{cases}.$$

Alors $P(Y_n = 6)$ est la probabilité pour qu'il y ait au moins 6 résultats consécutifs égaux dans une série de $n$ lancers (on peut généraliser les résultats qui suivent à des valeurs différentes de 6, voir sur le cédérom).
Soit $p_n = P(Y_n = 6)$. On a $p_1 = p_2 = p_3 = p_4 = p_5 = 0$ et $p_6 = \frac{2}{2^6}$.

Si on lance $n$ fois la pièce avec $n > 6$, alors l'événement « $Y_n = 6$ » se produit dans les cas suivants :
– lorsque $Y_{n-1} = 6$ ;
– ou dans l'un des deux cas suivants :
- on vient d'obtenir 6 fois 1 (événement $A_n$), $X_{n-6} = 0$ et $Y_{n-6} < 6$,
- on vient d'obtenir 6 fois 0 (événement $B_n$), $X_{n-6} = 1$ et $Y_{n-6} < 6$.

Il s'ensuit que, pour tout $n > 6$ :
$p_n = P(Y_{n-1} = 6) + P(A_n \text{ et } X_{n-6} = 0 \text{ et } Y_{n-6} < 6) + P(B_n \text{ et } X_{n-6} = 1 \text{ et } Y_{n-6} < 6)$,
Comme $A_n$ (resp. $B_n$) est indépendant de l'événement « $X_{n-6} = 0$ et $Y_{n-6} < 6$ » (resp. « $X_{n-6} = 1$ et $Y_{n-6} < 6$ »), on obtient :
$p_n = p_{n-1} + 1 / 2^6 \times [P(X_{n-6} = 0 \text{ et } Y_{n-6} < 6) + P(X_{n-6} = 1 \text{ et } Y_{n-6} < 6)]$.
Or, $P(X_{n-6} = 0 \text{ et } Y_{n-6} < 6) + P(X_{n-6} = 1 \text{ et } Y_{n-6} < 6)] = P(Y_{n-6} < 6) = 1 - p_{n-6}$.
D'où, pour tout $n > 6$ : $p_n = p_{n-1} + \frac{1}{2^6}(1 - p_{n-6})$. (1)

On peut ainsi calculer de proche en proche $p_n$ pour $n > 6$ sur tableur ou calculatrice. On trouve les valeurs suivantes :

| $n$ | 10 | 20 | 50 | 150 | 200 |
|---|---|---|---|---|---|
| $p_n$ | 0,094 | 0,237 | 0,544 | 0,918 | 0,965 |

*Remarques*
– Dans le cadre de cette activité, les élèves pourraient eux-mêmes définir le compteur sur des exemples de suites de lancers. L'enseignant pourrait établir la formule (1) (formule assez exotique pour définir une suite), les élèves ayant ensuite à faire les calculs numériques : ils sont en général surpris du résultat pour $n = 150$ ou $n = 200$, même si des simulations ont montré que la probabilité était forte pour de telles valeurs de $n$. On peut remplacer 6 par 3 ou 4, mais les élèves sont alors moins surpris et intéressés par le résultat final.
– Les élèves motivés peuvent faire des expérimentations numériques (pas tous les mêmes), en admettant la formule correspondant à « au moins $k$ coups consécutifs égaux », à savoir :

$$p_n = p_{n-1} + \frac{1}{2^k}\left(1 - p_{n-k}\right)$$

et en regardant soit à partir de quelle valeur de $n$ la probabilité devient supérieure à 0,5, soit la valeur de la probabilité pour $n = 500$ fixé par exemple.
– La suite $(p_n)$ est croissante et majorée par 1. Elle converge donc vers une limite $l$ qui vérifie $l = l + (1 - l)/2^6$ soit $l = 1$. En d'autres termes, la probabilité d'obtenir au moins une fois six résultats consécutifs égaux tend en croissant vers 1 lorsque $n$ tend vers $\infty$.
– Une autre formule permettant de calculer $p_n$ de proche en proche est donnée dans « Enseigner la statistique au lycée : des enjeux aux méthodes », par P. Dutarte et J.-L. Piednoir, brochure n° 112, commission inter IREM, *Lycées technologiques*, p. 96. Si on note $u_n$ le nombre de suites de taille $n$ de chiffres dont les termes sont 0 ou 1, ne contenant aucune séquence de 6 termes consécutifs égaux, on a $u_n = u_{n-1} + u_{n-2} + u_{n-3} + u_{n-4} + u_{n-5}$ et $p_n = 1 - u_n/2^n$. Des élèves pourraient faire les calculs avec cette deuxième formule, et regarder si on obtient les mêmes résultats qu'avec la formule (1).

## Cahier de statistique

On pourra inciter les élèves à faire dans ce cahier un bilan de ce qu'ils ont acquis depuis la seconde en probabilités et statistique, des questions résolues et de celles qui sont restées ouvertes.
L'évaluation du chapitre « probabilités et statistique » pourra tenir compte de la rédaction de ce cahier.