Annexe
Probabilités et statistique
Séries ES et S

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,

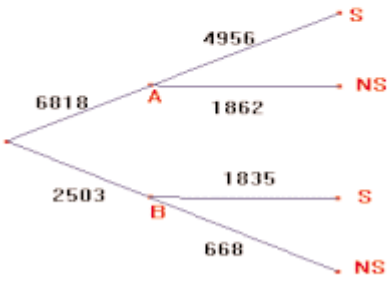
la donnée concernant le statut manquait). On peut ainsi représenter les données par l'un des tableaux (1) ou (2) ou par l'un des arbres (1) ou (2) ci-dessous. La seule information perdue par rapport à un tableau à 2 colonnes et 9 321 lignes est l'ordre des lignes.

	A	B
S	4956	1835
NS	1862	668

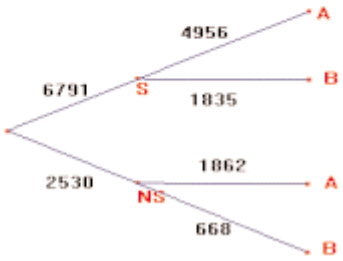
Tableau(1)

	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).

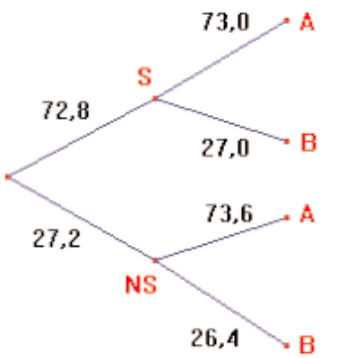
Sur des exemples, les élèves devront savoir passer d'un tableau à un arbre et vice-versa. Aucun formalisme n'est à développer.
À partir du tableau (2), on peut construire les tableaux (3) et (4) ou les arbres (3) et (4) :

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

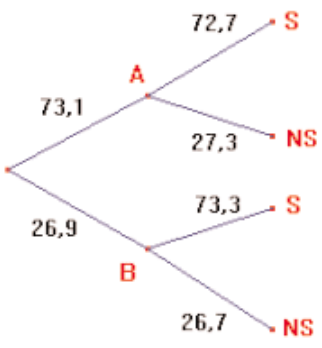
Tableau (3)

	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

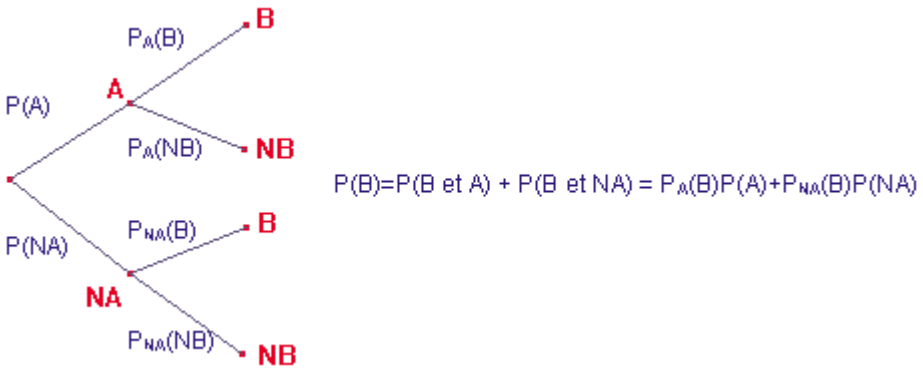
défaillance que B : la cause B est *protectrice*. Par exemple, si D est le décès d'un enfant par accident imputable à un vaccin V contre une maladie M, la probabilité $P_V(D)$ n'est pas nulle (le risque 0 n'existe pas) mais elle est très faible devant la probabilité $P_{\overline{V}}(D)$ de décès par 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 :

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

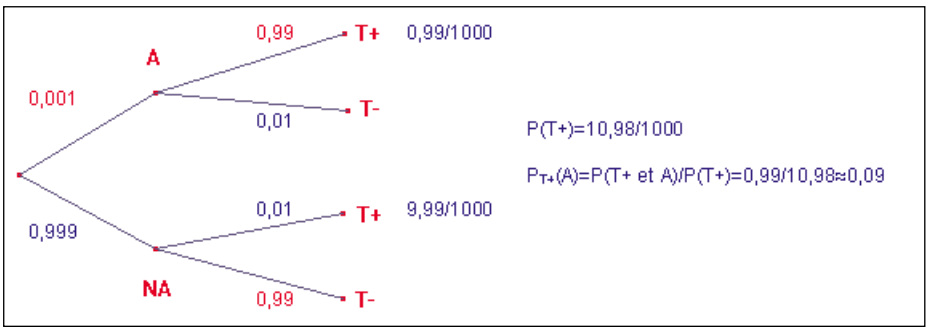
Lorsque des études permettent de déterminer des nombres $P_{C_1}(A)$, $i = 1 \ldots k$, cette formule permet de calculer la probabilité de A dans des populations variées où les C_i se répartissent différemment. Il en est ainsi dans l'application ci-dessous.

> *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 ;

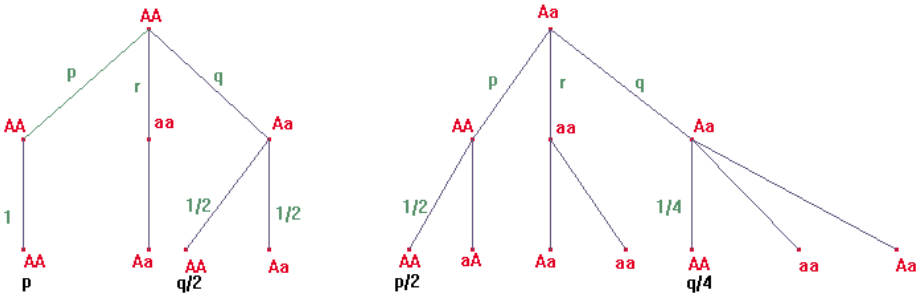
– dans les cas où p est faible, la valeur prédictive du test l'est aussi. Ainsi, si le caractère A révèle la présence d'une maladie rare, un test de dépistage systématique de toute une population aura l'inconvénient majeur de fournir beaucoup de faux positifs (individus non malades dont le test est positif). 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 α petit, $P(\text{Aa})$ est voisin de 2α, *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.

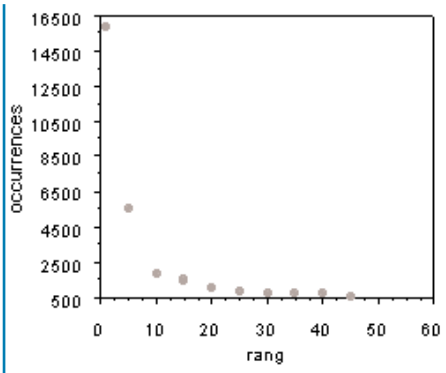
Représentation de données

Exemple

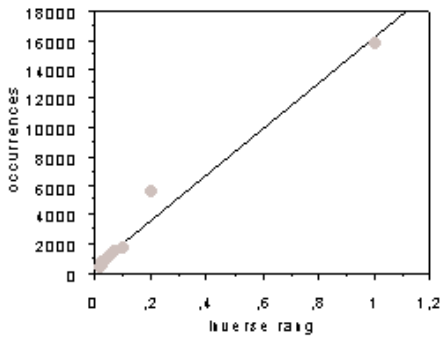
Données du site http://hobart.cs.umass.edu/allan/cs646/char_of_text
Dans un ensemble de 423 courts articles du journal *Time* totalisant 245 412 mots, on a classé les mots du vocabulaire par ordre décroissant de leur nombre d'apparitions dans l'ensemble des articles.
La figure (1), bien que les points 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.

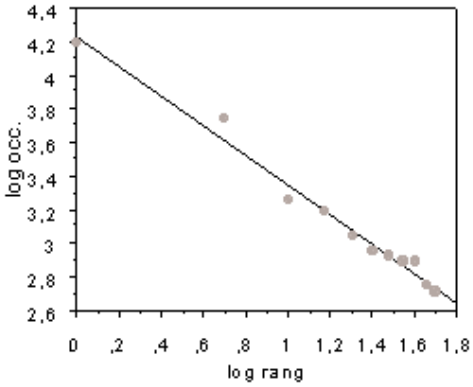
Mot	*the*	*and*	*with*	*on*	*but*	*have*	*so*	*week*	*its*	*new*	*into*
Rang	1	5	10	15	20	25	30	35	40	45	50
Occurrences	15861	5614	1839	1551	1138	914	868	793	793	572	518



(1)



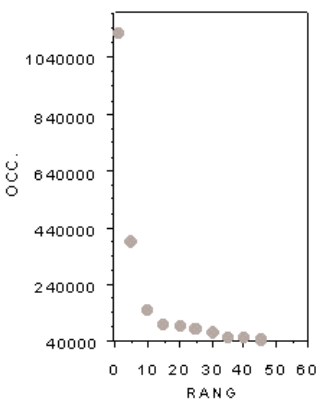
(2)



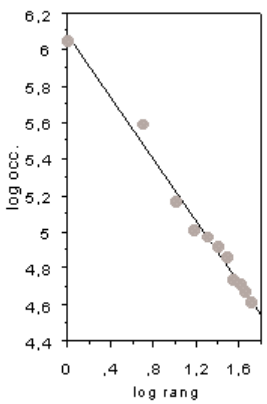
(3)

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

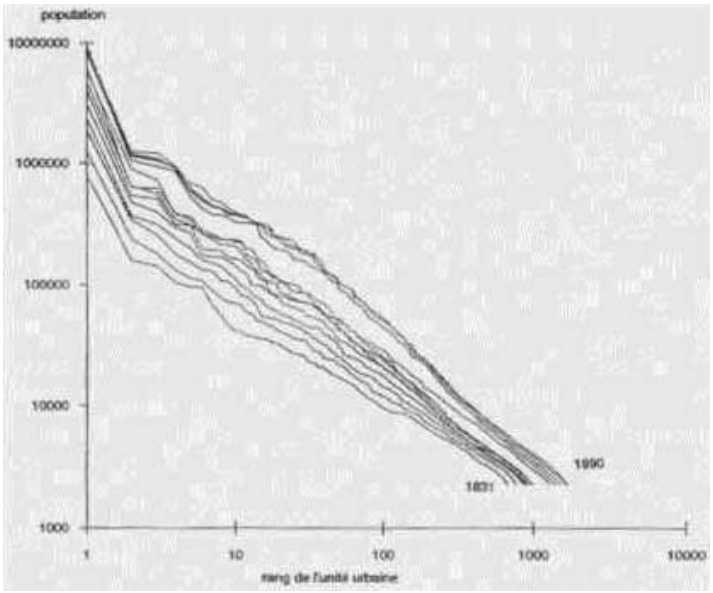


(4)



(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

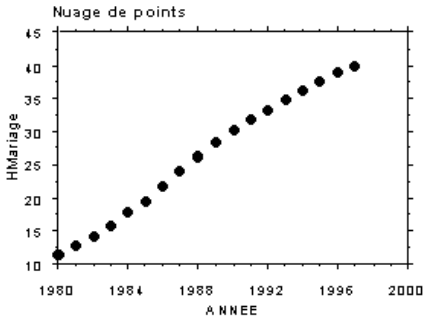
L'objectif du programme est de comprendre, à partir d'exemples simples, ce qu'est une « droite d'ajustement par moindres carrés » et d'illustrer son utilisation pour interpoler ou extrapoler quelques valeurs numériques.

Exemple 1

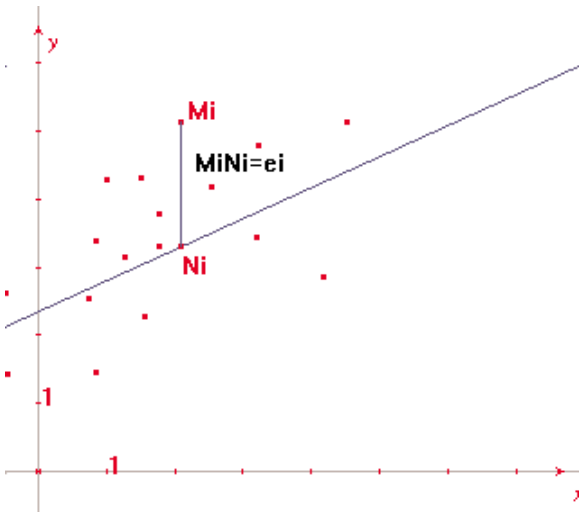
On a représenté graphiquement ci-dessous les valeurs des pourcentages de naissances hors mariages en France, entre 1980 et 1997. Nous nous intéressons ici aux deux questions suivantes :
a) Une description qualitative simple du nuage de points consiste à dire qu'il y a, à peu près, croissance linéaire 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 ε^2 et non $\delta = \sum |y_i - \hat{y}_i| / n$; une des raisons à cela est que la minimisation de ε^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, ε^2 est minimum si la droite passe par l'origine ; à titre d'exercice, on peut alors montrer que ε^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 :

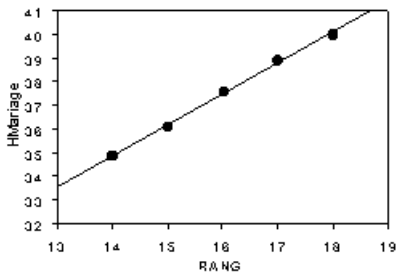
$$\tau(t) = 9{,}47 + 1{,}78(t - 1979).$$

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

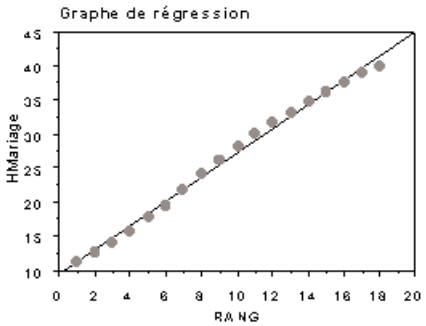
repose sur les 18 années précédentes : est-il vraiment pertinent ici d'utiliser toutes ces données ? Entre 1980 et 1997, les pourcentages observés ont varié de 10 à 40 %, mais un examen un peu plus précis du nuage des 18 points montre un infléchissement pour les dernières années observées ; aussi pour 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 :

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

Δ représente la proportion de la variance des ordonnées expliquée par l'ajustement linéaire. Des calculs simples montrent que Δ 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 Δ (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.

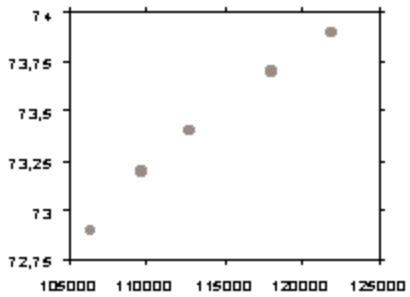
Exemple 2

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 :

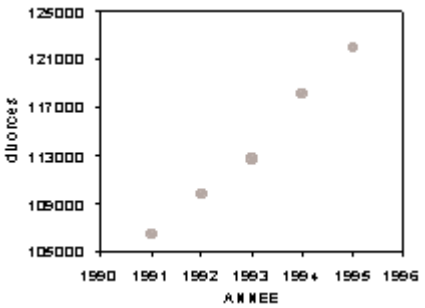
$\Delta y_i \approx a\Delta t$, $\Delta x_i \approx b\Delta t$ soit $\Delta y_i \approx \frac{a}{b}\Delta x_i$, c'est-à-dire que les points (x_i, y_i) sont « presque » alignés.

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.

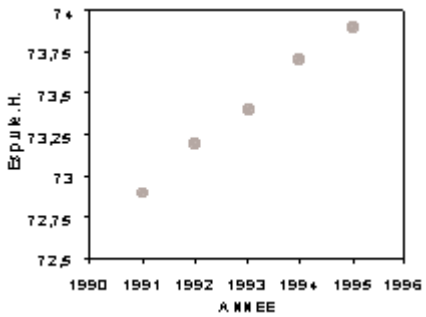


(1)

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 ?

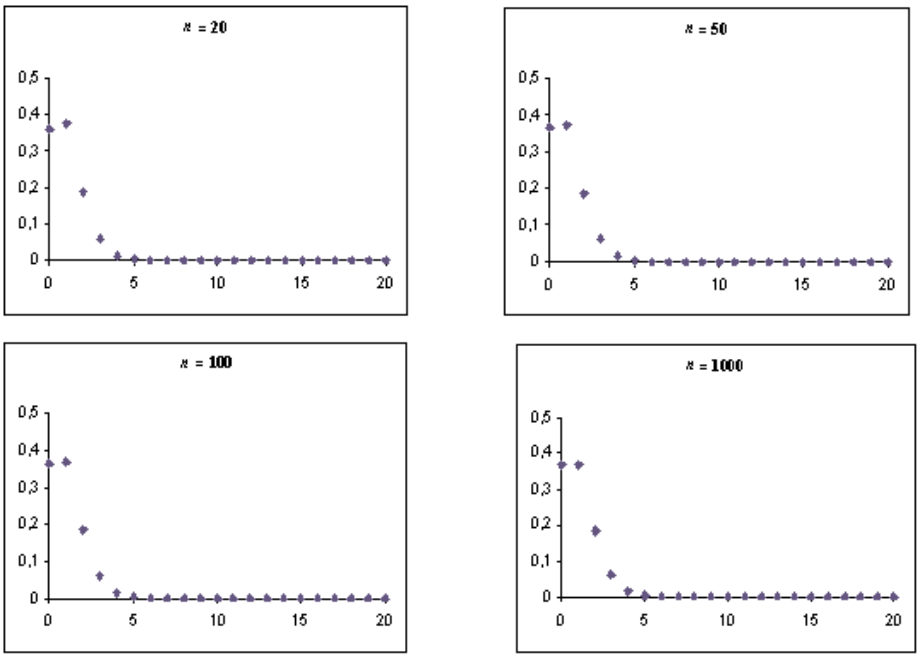
Il semble que cette question n'ait pas beaucoup de sens, si l'on ne possède pas d'autres informations. Et pourtant, il est possible d'émettre quelques hypothèses en utilisant des propositions classiques sur les limites ; ce qu'on explique ici s'appliquera, plus généralement, aux événements rares qui 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.

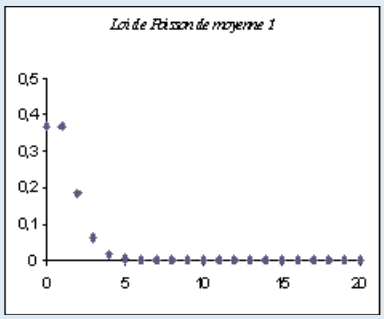
Il est en fait facile de trouver la limite exacte : pour une loi binomiale de paramètres n et $1/n$, la probabilité d'avoir 0 est $\left(1-\frac{1}{n}\right)^n$, et l'on sait (lemme 1) que la suite de terme général $\left(1-\frac{1}{n}\right)^n$ tend vers $1/e$ quand n tend vers $+\infty$: une valeur 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 $L_2 = 1/2e$.

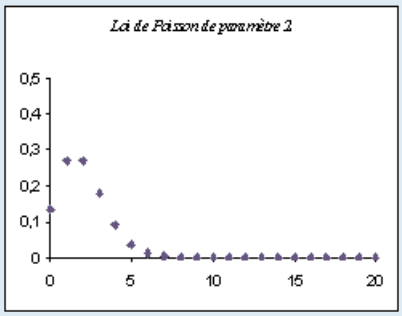
Remarque – Une récurrence montre que $P_n(k)$ tend vers $L_k = e^{-1}/k!$; les lois B(n,1/n) sont définies sur [0,n] ; lorsque n tend vers l'infini, s'il y a une loi de probabilité limite P, celle-ci est définie sur $\mathbb{N}$. On admettra la propriété suivante : $e = \lim\limits_{n\to+\infty} u_n$ 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.

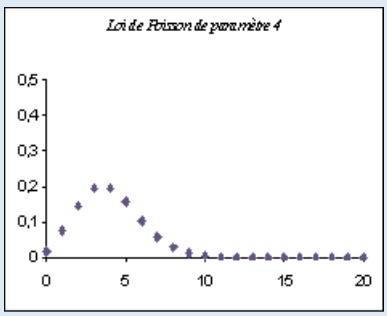


Dans ce modèle, pour des accidents rares qui arrivent en moyenne une fois par an, on a environ 36 % de chances qu'aucun accident ne se produise une année donnée, la même chance qu'il y en ait un, environ 18 % de chances qu'il y en ait 2, 6 % de chances qu'il y en ait 3, et 2 % qu'il y en ait 4 ; au-delà,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

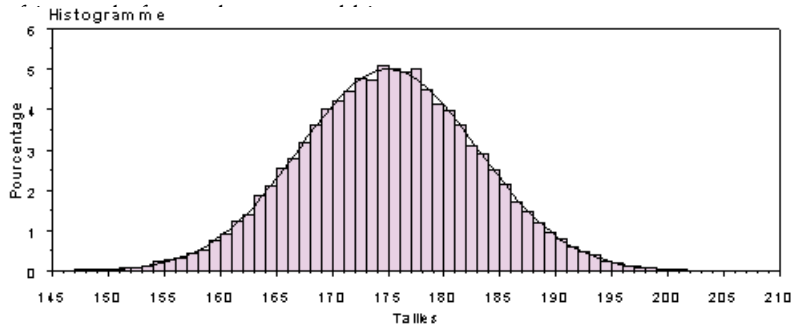
Histogrammes et aire sous une courbe. Comment définir les probabilités d'intervalles ?

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.

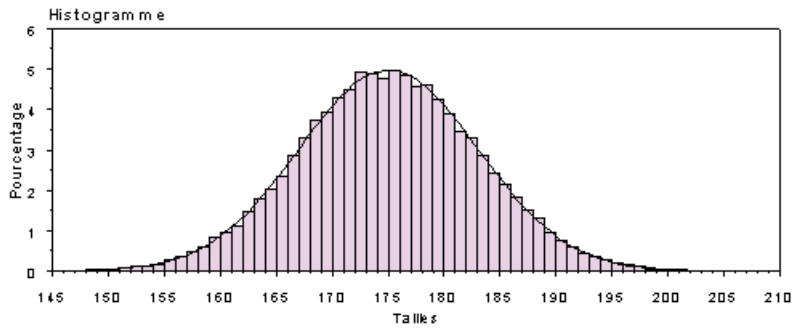
	Moy.	Ecart typ	Nombre	Minimum	Maximum	Médiane	Interquartile
Tailles	175,0	8,0	50000	145,1	209,5	175,0	10,8

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

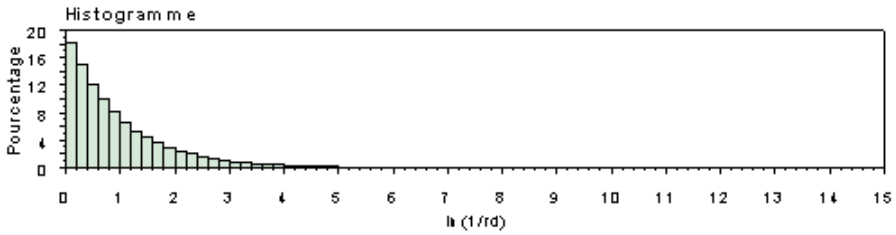
$$P_{I'}(J) = P(I' \cap J) / P(I').$$

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

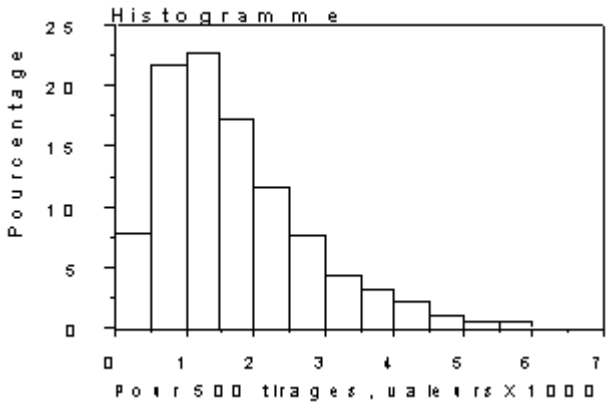
Les problèmes de validation de modèles sont complexes. Pour y sensibiliser les élèves, on peut commencer par des expériences de référence (tirages de boules dans des urnes, de choix au hasard, etc.) pour comprendre de quoi il s'agit, puis traiter des exemples.

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 :

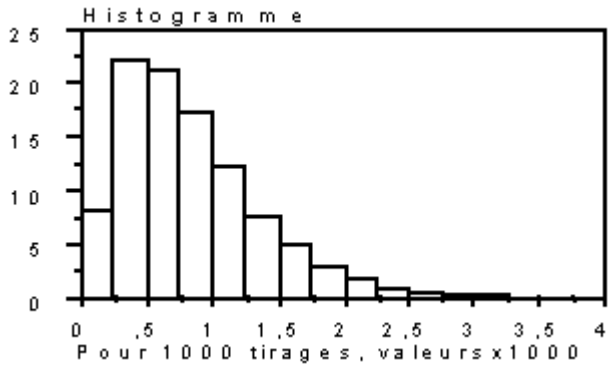
– si $d_{obs}^2 \leqslant 0{,}003$, alors le dé sera déclaré équilibré ;

– 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 α 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 ?

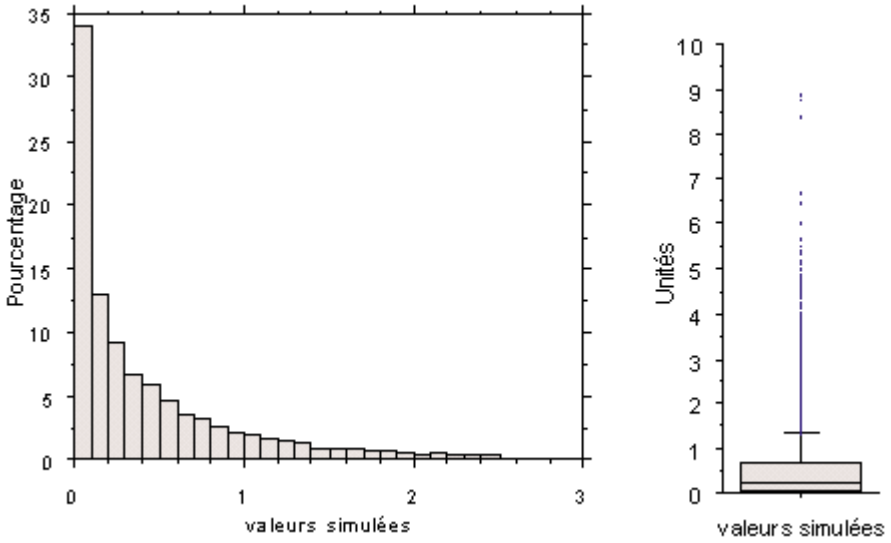
La traduction mathématique de cette question est ici : la loi de probabilité (0.5,0.5) est-elle pertinente pour modéliser la répartition des sexes à la naissance ?

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')

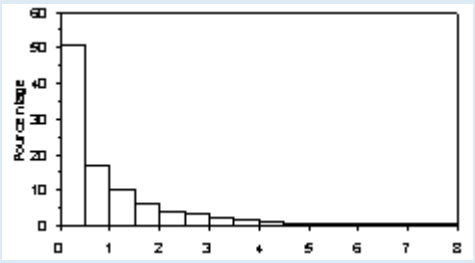
$$z = N\left(\frac{(a' - \hat{p}\hat{r})}{\hat{p}\hat{r}} + \frac{(b' - \hat{p}(1-\hat{r}))^2}{\hat{p}(1-\hat{r})} + \frac{(c' - (1-\hat{p})\hat{r})^2}{(1-\hat{p})\hat{r}} + \frac{(d' - (1-\hat{p})(1-\hat{r}))^2}{(1-\hat{p})(1-\hat{r})}\right)$$

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 χ^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 ∞.
– 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.