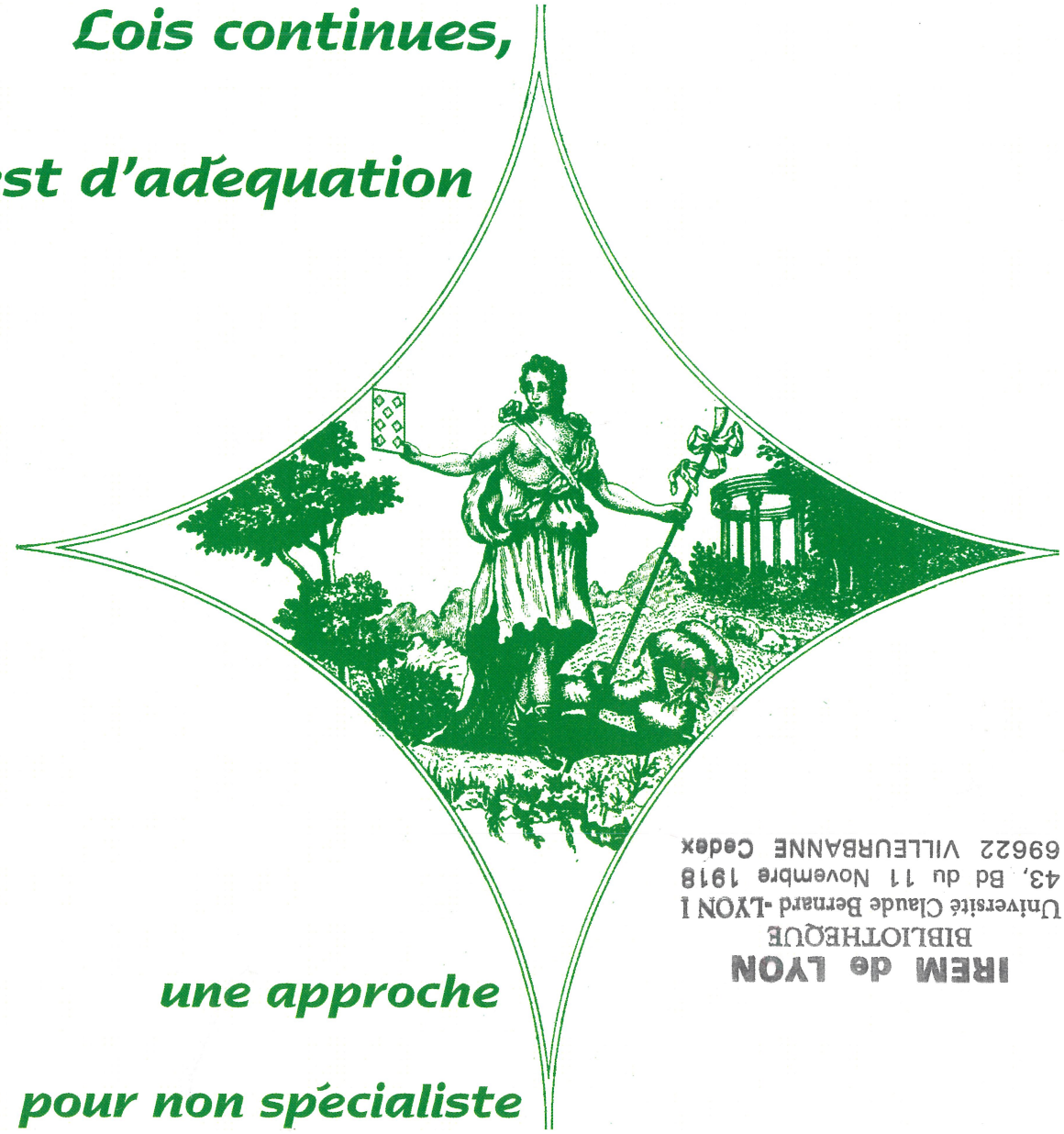
Les Publications de
l'IREM de BESANÇON

Lois continues, test d'adéquation une approche pour non spécialiste



Groupe
PROBABILITÉS & STATISTIQUE

Presses universitaires de Franche-Comté

Lois continues, test d'adéquation

une approche pour non spécialiste

Les Publications de l'IREM de BESANÇON

Directrice de collection HOMBELINE LANGUEREAU

Déjà publié par le groupe PROBABILITÉS & STATISTIQUE

dans la même collection

Arbres et Probabilité, Jean-Pierre Grangé, ISBN 2-913322-96-4, 2000

Probabilité conditionnelle et indépendance, Jean-Pierre Grangé, ISBN 2-913322-59-X, 1999

Les probabilités à l'agrégation externe de mathématiques, Yves Ducel, ISBN 2-913322-57-3, 1999

L'enseignement des statistiques et probabilités, Groupe Techniciens supérieurs, ISBN 2-913322-53-0, 2001

dans la collection *Didactiques*

Autour de la modélisation en probabilité, Michel Henry (ss dir.), ISBN 2-84627-018-X, 2001

Parutions récentes dans la même collection

Le mémoire de Gauss sur les surfaces courbes et la naissance de la géométrie différentielle intrinsèque, Hombeline Langueréau et Claude Merker, ISBN 2-84867-060-6, 2004

De la sphère au plan, Groupes Lycée et Cartographie, ISBN 2-84867-098-3, 2005

Illustration de couverture

Encyclopédie méthodique. Mathématiques, tome troisième.
d'Alembert, Abbé Bossut, Marquis de Condorcet, et alii, Panckoucke, Paris, 1789.

Les Presses universitaires de Franche-Comté bénéficient du soutien financier de la Région Franche-Comté et du Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche dans le cadre du contrat quadriennal.

ISBN 2-84867-101-7

IREM de Franche-Comté

Lois continues, test d'adéquation

une approche pour non spécialiste

Groupe PROBABILITÉS & STATISTIQUE

PRESSES UNIVERSITAIRES DE FRANCHE-COMTÉ, 2005
Diffusé par l'IREM de Franche-Comté

Membres du groupe de travail *Probabilités & statistique* de l'IREM

BARTHÉLÉMY M.-J., professeure certifiée (IREM & Lycée Pergaud, Besançon)

DUCEL Y., maître de conférences (IREM & UFR Sciences et Techniques, Besançon)

GRANGÉ J.-P., professeur agrégé (IREM & Lycée Pergaud, Besançon)

VENDRELY M., professeur certifié (IREM & Lycée Pergaud, Besançon)

Moyens horaires et financiers

Les activités du groupe de travail *Probabilités & statistique* de l'Institut de recherche sur l'enseignement des mathématiques (IREM) de l'Université de Franche-Comté sont financées par l'Université de Franche-Comté dans le cadre du contrat quadriennal d'établissement 2004-2007.

Le groupe de travail *Probabilités & statistique* bénéficie en outre de moyens horaires attribués

- aux personnels du second degré par le Rectorat de l'académie de Besançon ainsi que par la Direction des enseignements scolaires (DESco) du Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche.
- aux personnels de l'enseignement supérieur par le Département de mathématiques de l'UFR des sciences et techniques de l'université.

Table des matières

Introduction

Cette brochure trouve son origine dans deux conférences données en direction de professeurs de lycée, à l'initiative des inspecteurs pédagogiques régionaux de mathématiques de l'académie de Besançon, par les membres du groupe *Probabilités & statistique* de l'IREM de Franche-Comté.

Il s'agissait de proposer à un public non formé à la statistique inférentielle, en un temps relativement court, une réflexion sur la démarche statistique préconisée par les nouveaux programmes de lycée de la seconde à la terminale, notamment dans les activités pédagogiques portant sur le test statistique d'adéquation du modèle d'équirépartition à un phénomène réel.

Le discours proposé est aussi en premier lieu une introduction à la modélisation mathématique des phénomènes aléatoires continus. Cette connaissance est ensuite mise en œuvre pour répondre à la problématique posée par les programmes :

« Comment, à partir de *l'observation* d'un échantillon d'une population, *décider* par une argumentation de nature statistique qu'un *modèle mathématique* (du comportement global de cette population) est en adéquation avec la réalité observée ? »

L'approche adoptée est aussi en filigrane une première réponse à une question générale de nature plus épistémologique sur la modélisation mathématique, qui interroge aussi bien le mathématicien que le non-mathématicien, qu'on pourrait formuler ainsi :

Pourquoi les mathématiques, qui appartiennent au monde du *virtuel*, peuvent-elles prétendre nous dire quelque chose sur le monde *réel* ?

A cet effet, tout au long de ce travail deux axes de lecture doivent être présents à l'esprit du lecteur pour mieux décoder la réponse à la question précédente :

- la relation « discret / continu » dans le traitement statistique
- la relation « réalité de l'observation / virtualité du modèle mathématique ».

Ce document est divisé en quatre chapitres qui sont des réponses à quatre questions posées par la modélisation mathématique des phénomènes continus aux apprentis modélisateurs :

- Pourquoi décrire mathématiquement de façon différente une situation aléatoire discrète et une situation aléatoire continue ?

- Comment modéliser mathématiquement un phénomène aléatoire continu ?

- Y a-t-il des modèles qui jouent un rôle privilégié dans la modélisation statistique ?

- Comment décider de l'adéquation d'un modèle mathématique donné à la réalité étudiée ?

Donnons maintenant quelques précisions sur la méthode utilisée tout au long de ce travail.

Notre démarche se veut avant tout de type heuristique ce qui signifie que l'accent est surtout mis sur le processus d'*invention* des concepts mathématiques. Pour cela, la modélisation est conçue comme une « idéalisation » de l'observation. Le passage du réel au virtuel se fait à ce stade par un choix arbitraire du mathématicien. Ce choix est rendu nécessaire par la rupture existant entre l'objet mathématique (le modèle) et le réel étudié ; il sera validé a posteriori par les interprétations et les prévisions permises par le modèle choisi.

Nous nous intéressons donc plus à la construction du concept qu'à la rigueur du formalisme mathématique. A cet effet, l'utilisation minimale du formalisme est délibérée. En revanche la simulation par ordinateur est très présente pour illustrer les résultats statistiques nécessaires à la construction du test.

Précisons que seules quelques notions de statistiques descriptives sont requises, comme celles d'histogramme, de moyennes et d'écart type, pour aborder la lecture de ce document ainsi accessible à un bachelier scientifique.

Enfin, nous tenons à remercier Julien Michel, Katy Paroux, Bruno Saussereau pour leurs remarques sur le contenu de cette brochure, et Monique Diguglielmo, Christiane Koesler, Hombeline Languereau pour leur aide dans la relecture finale.

Chapitre I

De la modélisation des phénomènes discrets à celle des phénomènes continus

1 - Modélisation d'un phénomène discret

Considérons une variable aléatoire X discrète (i.e. prenant ses valeurs dans $\mathbb{N}$). La connaissance de $P(X = k)$, pour tous les entiers $k \geq 0$, permet d'effectuer tout calcul de probabilités pour n'importe quel événement considéré.

Ainsi le choix d'un modèle mathématique dans l'étude d'un phénomène aléatoire discret peut être considéré comme la donnée de l'application : $k \in \mathbb{N} \to P(X = k) = p_k \in [0,1]$, c'est-à-dire de la suite de réels positifs ou nuls $(p_0, p_1, p_2, \cdots, p_k, \cdots)$ dont la somme vaut 1.

2 - Modélisation d'un phénomène continu

Prenons maintenant une variable aléatoire X continue i.e. susceptible de prendre toutes les valeurs d'un intervalle de $\mathbb{R}$.

A titre d'exemple, considérons le résultat X du choix « au hasard de façon équiprobable » (nous entendons ici cette expression volontairement dans son sens commun) d'un nombre entre 0 et 1.
Cherchons pour $a \in [0, 1]$ fixé, la valeur que l'on peut choisir pour estimer $P(X = a)$.

Proposons trois approches basées sur l'interprétation intuitive de l'expression d'un choix « au hasard de façon équiprobable ».

Première approche : Transposons le raisonnement que nous tiendrions avec une variable discrète. D'après la formule de Laplace, le nombre $P(X = a)$ est le rapport du nombre d'éléments (cas favorables) de l'ensemble $(X = a)$ sur le nombre d'éléments (cas possibles) de l'ensemble [0, 1], ce que nous pouvons écrire abusivement :

$$P(X = a) = \frac{1}{\infty} = 0.$$

Bien sûr, cette approche n'est pas satisfaisante et passera auprès de certains comme une provocation. En effet, nous savons qu'en mathématiques, la manipulation de l'infini doit se faire avec d'infinies précautions et que bien souvent l'intuition nous trompe.

Proposons donc une autre approche.

Deuxième approche : Si nous partageons l'intervalle [0, 1] en deux intervalles d'égales longueurs, l'idée intuitive de l'équiprobabilité nous conduit à considérer que la probabilité que le nombre tiré « au hasard de façon équiprobable » a une chance sur deux de tomber dans l'intervalle [0 ; 0,5].
C'est-à-dire $P(0 \leq X \leq 0{,}5) = 0{,}5$ et de même $P(0{,}5 \leq X \leq 1) = 0{,}5$.

Plus généralement, on peut admettre « intuitivement » que la probabilité que le nombre choisi tombe dans un sous-intervalle I de [0, 1] soit proportionnelle à la longueur de I.
C'est-à-dire $P(X \in I) = \alpha \times longueur(I)$ où α est une constante.

Comme $P(X \in [0, 1]) = \alpha \times longueur([0,1]) = 1$, il vient $\alpha = 1$ et $P(X \in I) = longueur(I)$.
Par suite, pour tout entier $n \geq 1$, en vertu de la monotonie des probabilités,

$$P(X = a) \leq P(a \leq X \leq a + \frac{1}{n}) = longueur([a, a+\frac{1}{n}]) = \frac{1}{n}.$$

Ce qui conduit encore à $P(X = a) = 0$ pour tout nombre $a \in [0, 1]$ fixé avant le tirage.

Proposons enfin une troisième et dernière approche.

Troisième approche : Dire que le nombre est tiré « au hasard de façon équiprobable » signifie intuitivement que tous les nombres de l'intervalle $[0, 1]$ ont la même probabilité de sortir. C'est-à-dire que, pour tout nombre $a \in [0, 1]$ fixé avant le choix, $P(X = a) = p$ où p est une constante indépendante de a. Par suite, si p est non nul, la probabilité que le nombre choisi soit dans l'ensemble $\{1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots\}$ est nécessairement strictement supérieure à 1, ce qui est absurde. Donc $p = 0$.

3 - Conclusion

En résumé de ces trois approches, contrairement à ce que nous avions observé dans le cas des situations discrètes, la modélisation d'un phénomène aléatoire ne peut pas se ramener, dans le cas continu, à la seule connaissance de l'application $k \in \mathbb{R} \to P(X = k)$, car cette application, dans des cas très divers, est tout simplement l'application-nulle. Il faudra associer à la situation réelle à décrire un autre objet mathématique pour résumer l'acte de modélisation.

Chapitre II

De la modélisation mathématique des phénomènes aléatoires continus

1 - Cas du choix « au hasard » d'un nombre de [0 ; 1[

a) Première expérience

Choisissons « au hasard » 10 nombres de $[0 ; 1[$ avec la touche random d'une calculatrice et rangeons-les dans les intervalles $[0 ; ½[$ et $[½ ; 1[$. Cette touche random donne aléatoirement un nombre de 10 chiffres au plus pris uniformément parmi les 10^{10} nombres décimaux d'au plus 10 chiffres de [0 ; 1[[1].

Bien sûr, n'importe quel nombre réel de $[0 ; 1[$ non décimal ou dont l'écriture décimale comporte plus de 10 chiffres ne peut être obtenu par cette méthode. L'expérience décrite ci-dessus consiste à approcher la situation aléatoire idéalisée suivante : « Choisir au hasard 10 nombres de $[0 ; 1[$ ».

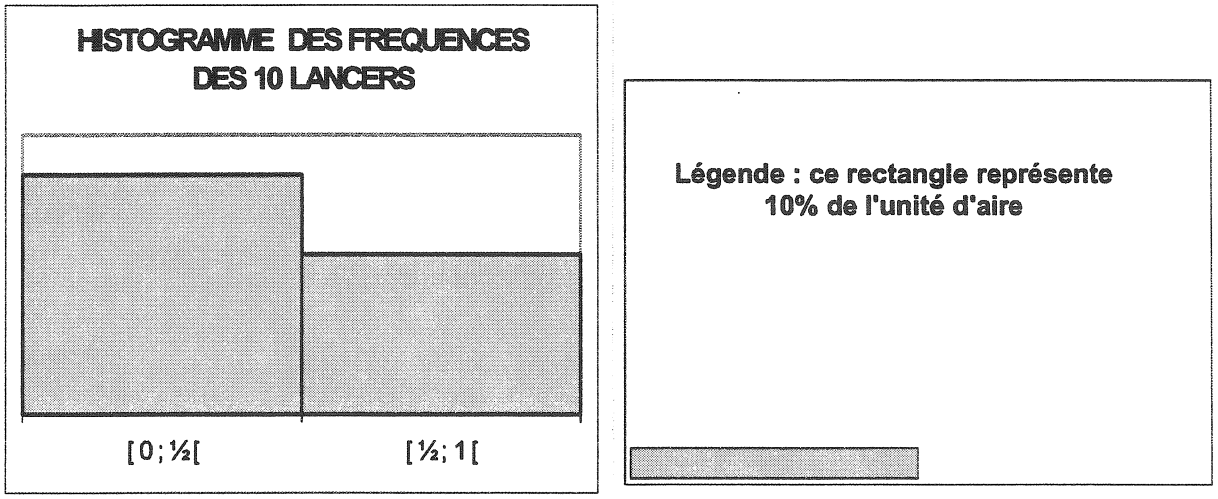
Lors d'expériences concrètes, aléatoires ou non, les valeurs observées sont toujours entachées d'une certaine incertitude due à la précision et à la fiabilité des instruments de mesure et les décimaux sont suffisants pour estimer une observation quantitative. Alors pourquoi choisir l'ensemble des réels de $[0 ; 1[$ pour y définir ici, une nouvelle loi de probabilité ? Une réponse consiste à dire qu'il en est de même pour les fonctions qu'on définit et étudie sur R alors qu'elles servent de modèles à des expériences concrètes ; qu'il en va de même en algèbre pour les résolutions de problèmes etc.. Un des buts essentiels des mathématiques est de proposer des modèles, les modèles numériques en général sont définis sur $\mathbb{R}$.

On a choisi « au hasard » 10 nombres de $[0 ; 1[$ avec la touche random d'une calculatrice et on les a rangés dans les classes $[0 ; ½[$ et $[½ ; 1[$ comme l'indique le tableau suivant :

x	Effectifs n_i	Fréquences n_i / N
[0 ; ½[	6	0,6
[½ ; 1[	4	0,4
	$N = 10$	1

[1] Cette touche random active un générateur de nombres pseudo aléatoires considérés comme distribués de façon équiprobable.

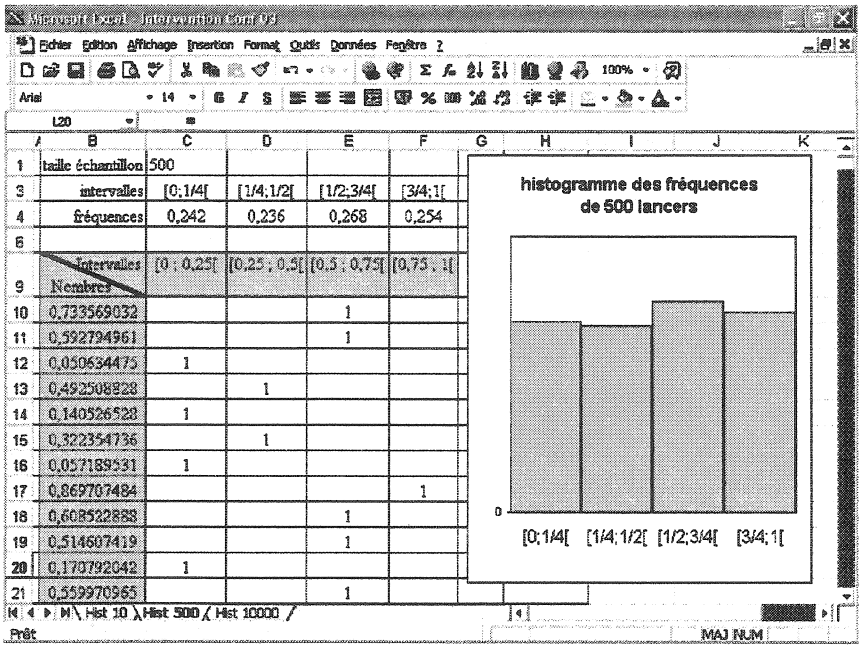
Cette première expérience peut se représenter par l'histogramme[2] suivant dont l'aire est égale à 1.



b) Seconde expérience

Pour avoir une meilleure idée de l'allure des histogrammes des fréquences des nombres choisis au hasard dans *[0 ; 1[* il faut partitionner davantage cet intervalle. Cela nécessite une plus grande quantité de nombres choisis.

Choisissons au hasard 500 nombres de *[0 ; 1[* par un générateur de nombres



	B	C	D	E	F
1	taille échantillon	500			
3	intervalles	[0;1/4[	[1/4;1/2[	[1/2;3/4[	[3/4;1[
4	fréquences	0,242	0,236	0,268	0,254
6					
9	Intervalles / Nombres	[0 ; 0,25[	[0,25 ; 0,5[	[0,5 ; 0,75[	[0,75 ; 1[
10	0,733569032			1	
11	0,592794961			1	
12	0,050634475	1			
13	0,492508828		1		
14	0,140526528	1			
15	0,322354736		1		
16	0,057189531	1			
17	0,869707484				1
18	0,608522883			1	
19	0,514607419			1	
20	0,170792042	1			
21	0,559970965			1	

[2] Le terme histogramme est pris ici dans le sens de surface, c'est à dire réunion de rectangles dont les aires sont proportionnelles au nombre d'observations des classes correspondantes.

pseudo aléatoires et demandons-lui de les ranger dans des classes de longueur 0,25, comme l'indique la copie d'écran ci-dessus.

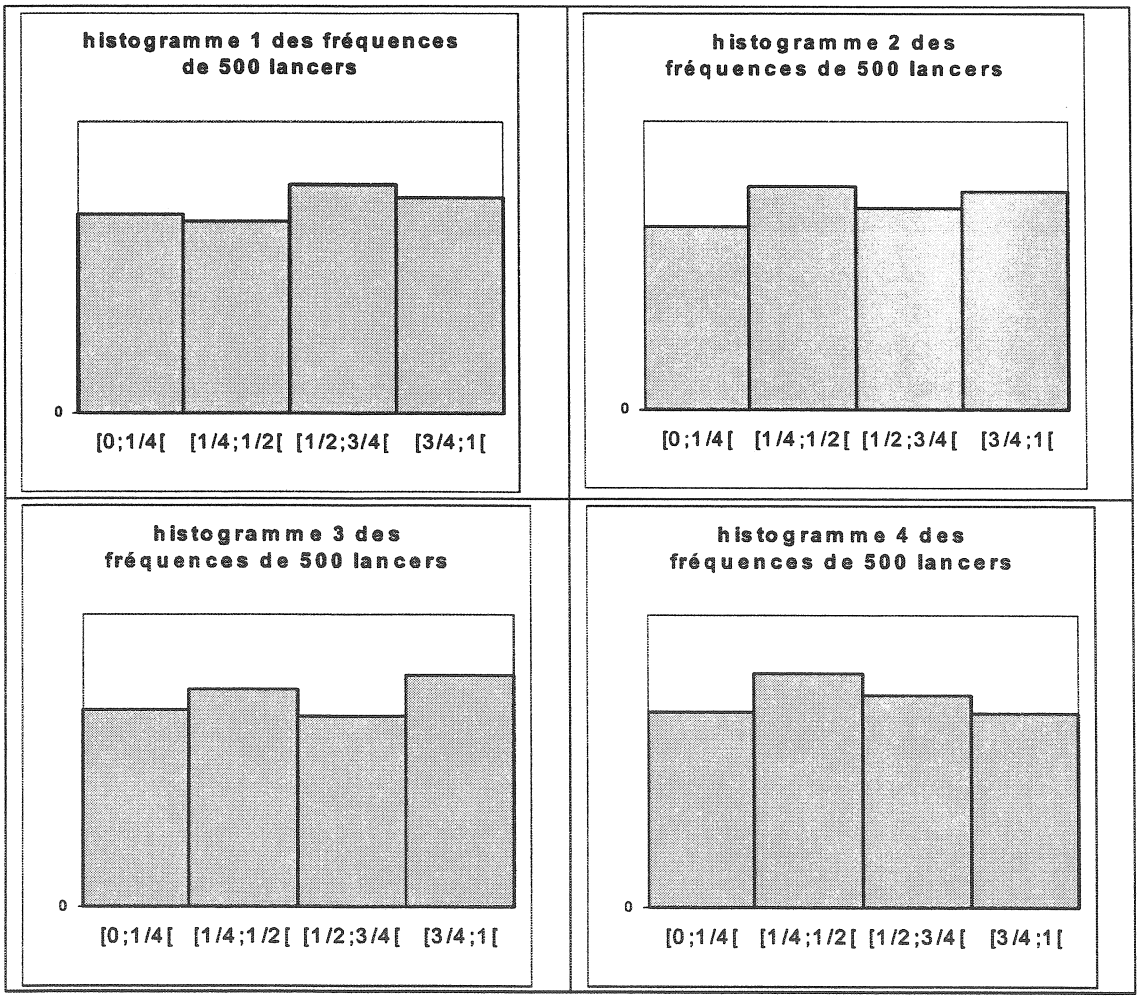
On en déduit que la fréquence des nombres compris entre 1/4 et 3/4 parmi ces 500 nombres est égale à la somme des aires des deux rectangles de bases [¼ ; ½[et [½ ; ¾[.

D'autre part, pour a et b dans $[0\,;\,1[$, en faisant l'hypothèse d'équirépartition des nombres dans les classes, on peut estimer la fréquence des nombres compris entre a et b en calculant l'aire de la partie de l'histogramme limitée à droite et à gauche par les droites d'équation $x = a$ et $x = b$ construites dans le repère associé à l'histogramme.

c) Approche expérimentale du modèle mathématique

Répétition de l'expérience précédente avec un échantillon de taille fixée

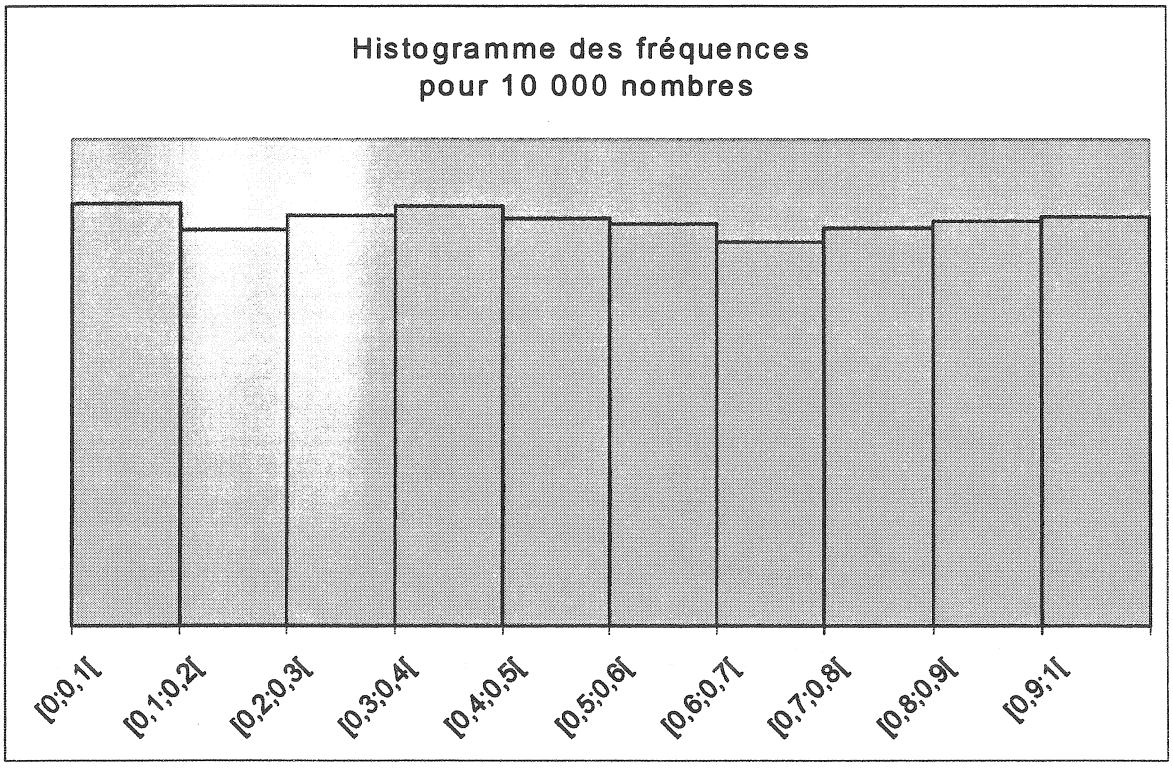
Pour prendre en compte la fluctuation d'échantillonnage et mieux conjecturer le modèle mathématique qui régit la répartition des nombres au hasard dans $[0\,;\,1[$, faisons plusieurs simulations de tels échantillons de taille 500.



Aucune variation significative ne s'observant d'un histogramme à l'autre concernant ces quatre classes, on admet alors que les limites supérieures de ces histogrammes des fréquences se concentrent dans une bande horizontale de faible épaisseur, et on peut conjecturer que la limite supérieure de l'histogramme idéal est un segment horizontal. Celui-ci serait situé à la graduation 1 de l'axe des ordonnées puisque l'aire de l'histogramme ainsi obtenu doit être égale à 1.

Augmentation de la taille n de l'échantillon

L'accroissement de la taille de l'échantillon permet une meilleure stabilisation des fréquences, ce qui infirmera ou confirmera la conjecture précédente. Cet accroissement permet aussi d'augmenter le nombre de classes ou sous intervalles de $[0\,;\,1[$ et ainsi d'observer éventuellement des variations significatives. Par exemple en choisissant n = 10 000, la répartition en dix classes est due à une considération de commodité. Nous avons obtenu l'histogramme ci-dessous :



Là aussi, aucune variation significative de la limite supérieure de cet histogramme n'est observée, la conjecture précédente se confirme.

d) Approche théorique du modèle mathématique

Théoriquement, dans un choix équiprobable un décimal à dix chiffres de $[0\,;\,1[$ a autant de chances d'être dans la classe $[0{,}2\,;\,0{,}3[$ que dans une autre classe de longueur 1/10 ; ou encore, une classe comme $[0{,}6\,;\,0{,}7[$ n'a ni plus ni moins de chances qu'une autre classe de même longueur de contenir un tel nombre choisi au

hasard. Cette équiprobabilité sur chacune des dix classes nous donne une probabilité de 0,1 pour chacune d'elles. L'histogramme théorique est donc formé de dix rectangles ayant pour largeur celle de chaque classe, c'est-à-dire 0,1 et comme hauteur 1. La limite supérieure de cet histogramme théorique est donc un segment horizontal de hauteur 1 sur l'intervalle $[0\,;\,1[$, et la probabilité d'y choisir un décimal compris entre deux nombres a et b est égale à la portion d'aire de cet histogramme limitée par les verticales aux points d'abscisses $x = a$ et $x = b$. Le modèle mathématique choisi consiste à généraliser ce raisonnement pour tout choix "équiprobable" d'un réel de $[0\,;\,1[$.

e) Le modèle probabiliste

Inspiré par les observations et le raisonnement précédents, ce segment horizontal de hauteur 1 est appelée une courbe de densité de probabilité et la loi de probabilité associée à cette courbe de densité est notée U et appelée loi uniforme sur $[0\,;\,1[$ ou $[0\,;\,1]$.

Ainsi, pour deux réels a et b de $[0\,;\,1]$, $U([a\,;\,b])$ = Aire sous cette courbe de densité limitée par les droites d'équations $x = a$ et $x = b$, c'est à dire $U([a\,;\,b]) = b - a$.

La formule de LAPLACE , $P(A) = \dfrac{nombredecasfavorables}{nombredecaspossibles}$ utilisée pour le calcul des probabilités dans le cas de l'équiprobabilité de tous les cas possibles en nombre fini, ne peut pas se généraliser lorsque le nombre de cas possibles devient infini[3]. Le calcul de l'aire sous une courbe de densité est une manière de calculer des probabilités adaptée aux cas continus. Le prix à payer pour utiliser cette manière est qu'on ne peut plus s'intéresser à la probabilité d'obtenir une valeur donnée a priori car elle vaut 0. On ne peut que calculer la probabilité d'obtenir une valeur comprise entre deux nombres a et b.

2 - Densité de probabilité

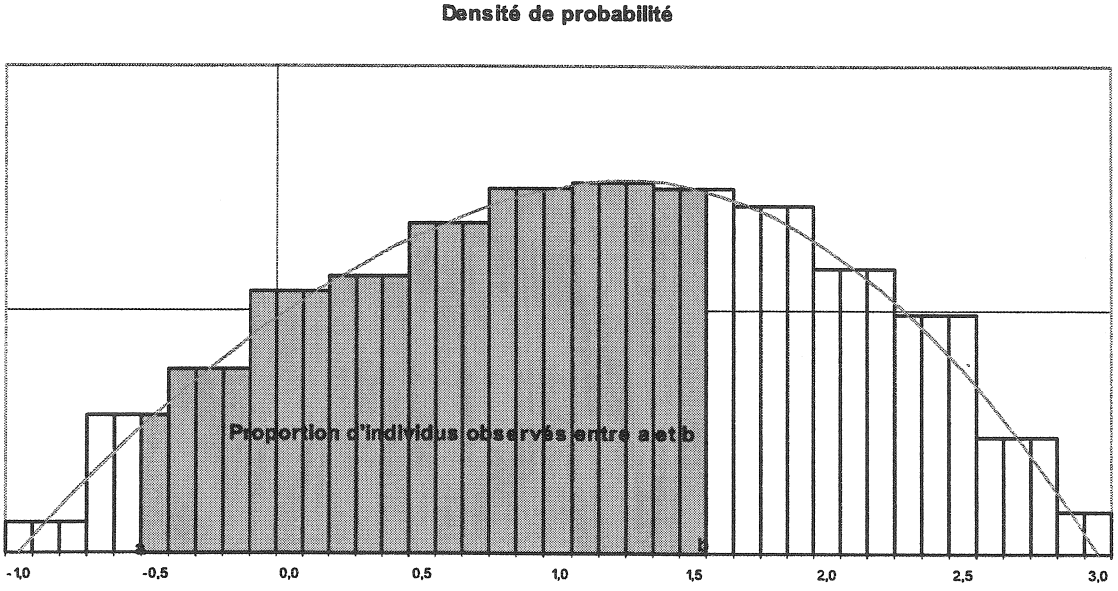
a) Introduction

Si nous augmentons le nombre d'observations d'une variable statistique continue et si on réduit l'amplitude des classes de l'histogramme, les polygones des fréquences qu'on construit peuvent être approchés par une courbe continue située au-dessus de l'axe des abscisses, l'aire sous la courbe étant égale à 1.

Etant donné deux réels a et b quelconques, la proportion d'individus observés donnant des valeurs de la variable comprises entre a et b est égale à la somme des aires des rectangles de base comprise entre a et b et est peu différente de l'aire de la partie du plan comprise entre la courbe, les droites d'équation $x = a$ et $x = b$, et l'axe des abscisses.

---

[3] cf. première partie



b) Définitions

Si on réalise une expérience aléatoire permettant de définir une variable aléatoire X continue, effectuer des calculs de probabilités relatifs à cette variable aléatoire revient à manipuler une fonction « densité de probabilité » répondant aux propriétés suivantes :

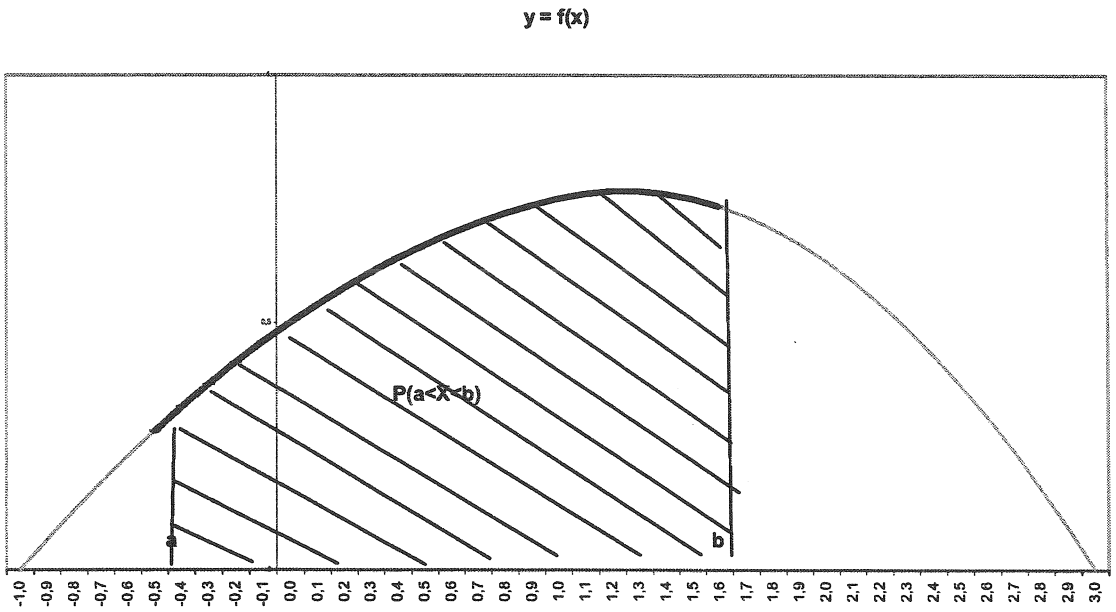
Une fonction f définie et intégrable sur $\mathbb{R}$ est une densité de probabilité si elle possède les propriétés suivantes :

- $\forall x \in \mathbb{R}, f(x) \geq 0.$
- $\int_{-\infty}^{+\infty} f(x)dx = 1.$

Les calculs de probabilité qu'on sera amenés à effectuer avec ce modèle ne feront intervenir que des expressions de la forme :

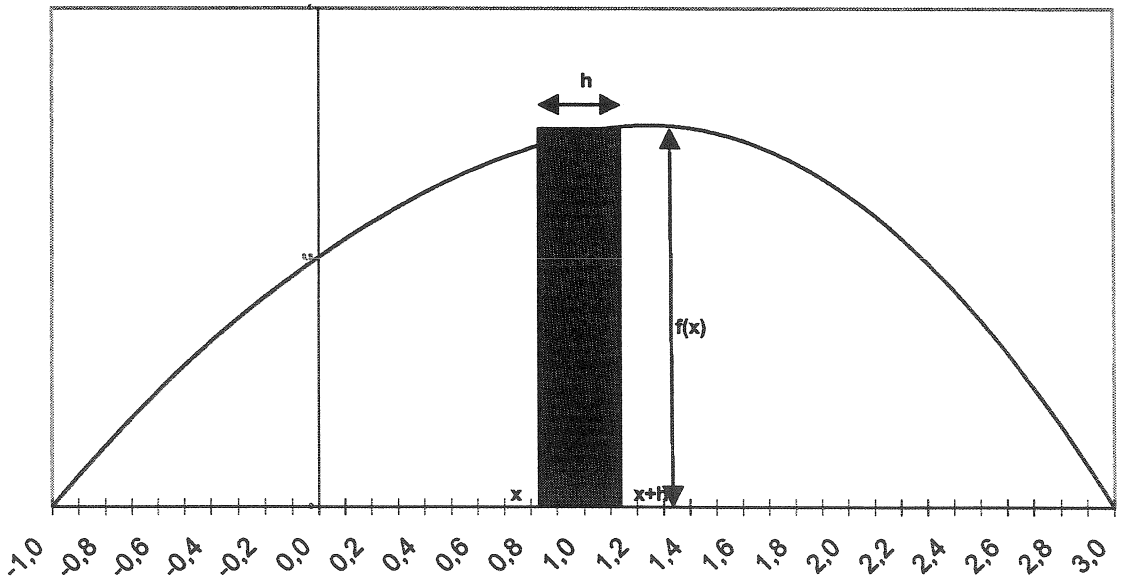
$$P[X \in (a,b)] = \int_a^b f(x)dx$$

où a, b sont des réels, éventuellement infinis, avec $a \leq b$, et (a,b) désigne l'un des intervalles $[a, b]$, $[a, b[$, $]a, b]$, $]a, b[$.



La fonction de répartition d'une variable aléatoire X est la fonction F définie sur $\mathbb{R}$ par $F(x) = P(X \leq x)$.

c) Justification de la terminologie « densité de probabilité »



La mesure de l'aire hachurée représente $P(x \leq X \leq x + h)$ par définition d'une densité de probabilité.

Supposons la densité f continue sur $\mathbb{R}$. Pour h très petit, on peut considérer que la fonction est pratiquement constante dans un voisinage de x, et égale à $f(x)$ sur ce voisinage. Par suite $P(x \leq X \leq x + h) \approx f(x) \times h$.

Le réel $f(x)$ a donc la dimension d'une probabilité par unité de longueur. Par analogie avec le concept de densité linéique de masse en physique on est amené à assimiler f à une densité de probabilité sur la droite réelle $\mathbb{R}$.

3 - Exercices d'application

Les exercices suivants sur les densités de lois continues sont extraits des annexes du programme de terminale S en page 143. La plupart des fonctions étudiées en analyse peuvent servir de modèle de densité de probabilité.

Les exercices proposés sont l'occasion de réinvestir les outils mathématiques pour résoudre des problèmes de probabilités.

Pour les probabilités définies sur un ensemble fini, les dénombrements, les diagrammes, les arbres ont montré leur efficacité.

Pour les probabilités à densité continue, le calcul intégral, les limites, les suites, les calculs d'aires dans le plan, et les calculs de volumes dans l'espace prennent le relais.

On pourra par ailleurs noter le glissement de la notation $\sum_{x\in I} p(x)$ à $\int_I f(x)dx$.

Par exemple pour la médiane M, le passage de $\sum_{x\leq M} p(x) = 0,5$ à $\int_{x\leq M} f(x)dx = 0,5$.

a) Exercice faisant intervenir des calculs d'aires

Rappelons que si $I = [a, b]$ et f est une fonction définie continue positive sur I avec $P(I) = \int_a^b f(x)dx = 1$, alors f est une densité de probabilité.

Exercice 1

Soit $I = [0 ; 1]$ et f la fonction définie sur I par $f(t) = 4\,t^3$

1°) Vérifier que f est une densité de probabilité.

2°) On note P la loi de probabilité associée à f, calculer $P([0,25 ; 0,75])$.

3°) Calculer m pour que $P([0 ; m]) = 0,5$.

Eléments de réponse

On vérifie les résultats suivants :

1°) $P(I) = \int_0^1 4t^3dt = 1$

2°) $P([0,25 ; 0,75]) = \int_{0,25}^{0,75} 4t^3dt = 0,3125$

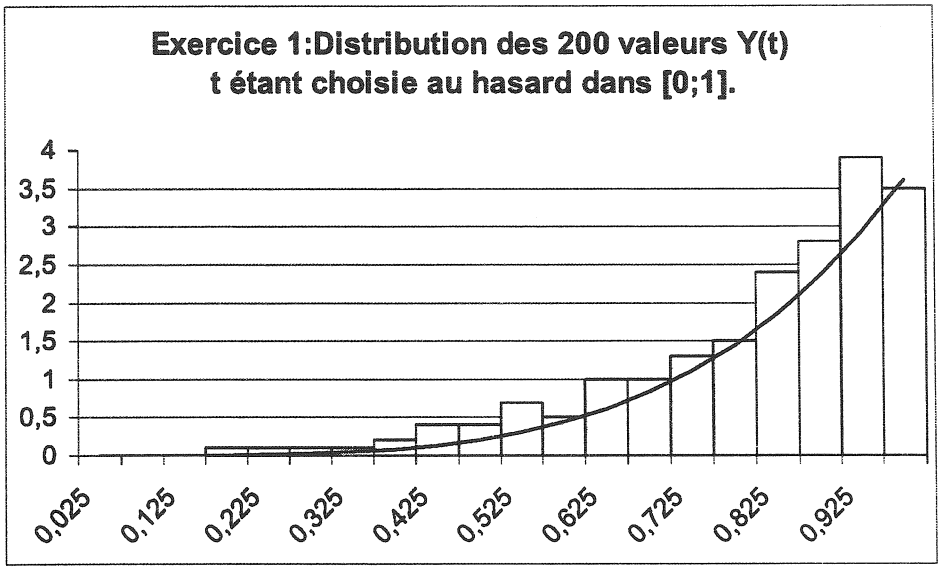
3°) $P([0 ; m]) = \int_0^m 4t^3dt = \left[t^4\right]_0^m = m^4$ · D'où l'équation $m^4 = 0,5$ a pour solution la valeur de la médiane $m = \sqrt[4]{0,5}$.

On proposera aux élèves d'inventer d'autres densités de probabilité.

Simulation

Dans l'annexe II on démontre que si nous choisissons, au hasard 200 valeurs de t, uniformément réparties sur [0 ; 1], les valeurs de $Y(t)=\sqrt[4]{t}$ correspondantes constituent un échantillon de 200 valeurs simulées qui sont distribuées conformément à la loi de probabilité de densité f. L'allure de l'histogramme des fréquences de ces valeurs simulées est proche de la représentation graphique de f.

Graphique



b) Exercice faisant intervenir des limites et des suites

Nous pouvons choisir comme fonctions f celles proposées au programme (voir annexe II). Rappelons que si $I = [a ; +\infty[$ et f est une fonction définie continue positive sur I, vérifiant $\lim_{t\to\infty} \int_a^t f(x)dx = 1$, alors f est une densité de probabilité.

Exercice 2

Soit $I = [0 ; +\infty[$ et f une fonction définie sur I par $f(t) = k\,e^{-2t}$

1°) Déterminer k pour que f soit une densité de probabilité.

2°) On note P la loi de probabilité associée à f, calculer $P([n ; n+1])$.

3°) Calculer m pour que $P([0 ; m])=0,5$.

Eléments de réponse

1°) D'après le rappel, il suffit que $\lim_{t\to\infty} \int_a^t f(x)dx = 1$.

Or $\int_a^t f(x)dx = \int_0^t ke^{-2x}dx = k\left(-\frac{e^{-2t}}{2}+\frac{1}{2}\right)$ et $\lim_{t\to\infty} \int_a^t f(x)dx = \frac{k}{2} = 1$. D'où l'équation $0,5 \times k = 1$ qui donne $k = 2$.

Nous avons affaire à une loi exponentielle (cf. annexe III) d'espérance mathématique $E(X) = 0,5$

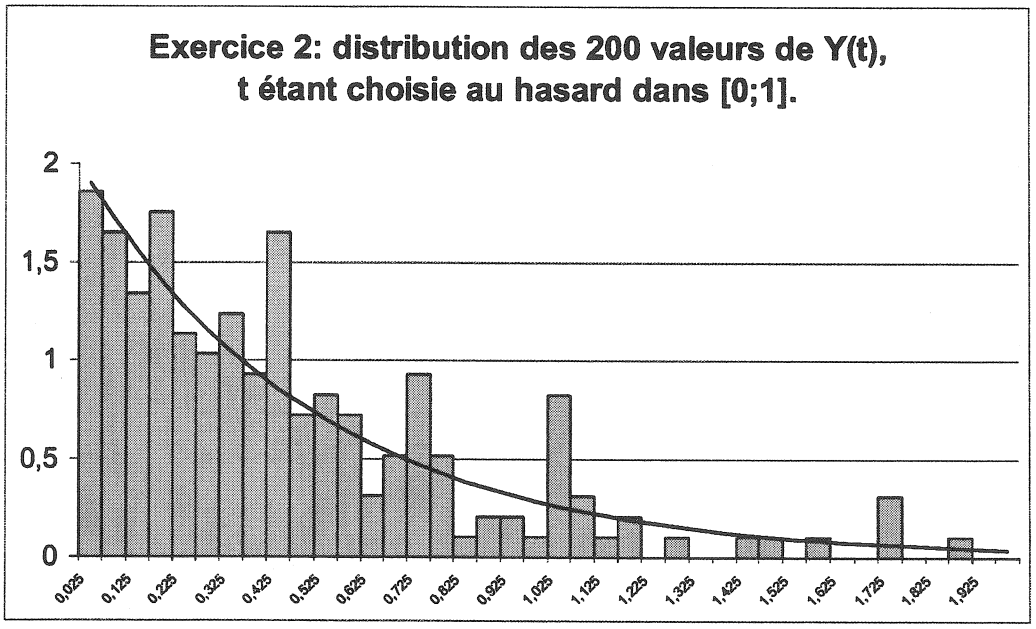
2°) Pour tout entier naturel n, $P([n\ ;\ n+1])=\int_n^{n+1} 2e^{-2x}dx = e^{-2n}(1-e^{-2})$. (Cette question débouche sur l'étude d'une suite dont il est facile de calculer la somme des termes)

3°) $P([0\ ;\ m])=\int_0^m 2e^{-2x}dx = 1-e^{-2m}$. L'équation $1 - e^{-2m} = 0{,}5$ donne pour valeur de la médiane $m = -\dfrac{\ln(0{,}5)}{2}$.

Simulation

Dans l'annexe II on démontre que si nous choisissons, au hasard 200 valeurs de t, uniformément réparties sur $[0\ ;\ 1]$, l'histogramme des fréquences des 200 valeurs de $Y(t)=-\dfrac{\ln(t)}{2}$ peut être comparé à la représentation graphique de f.

Graphique

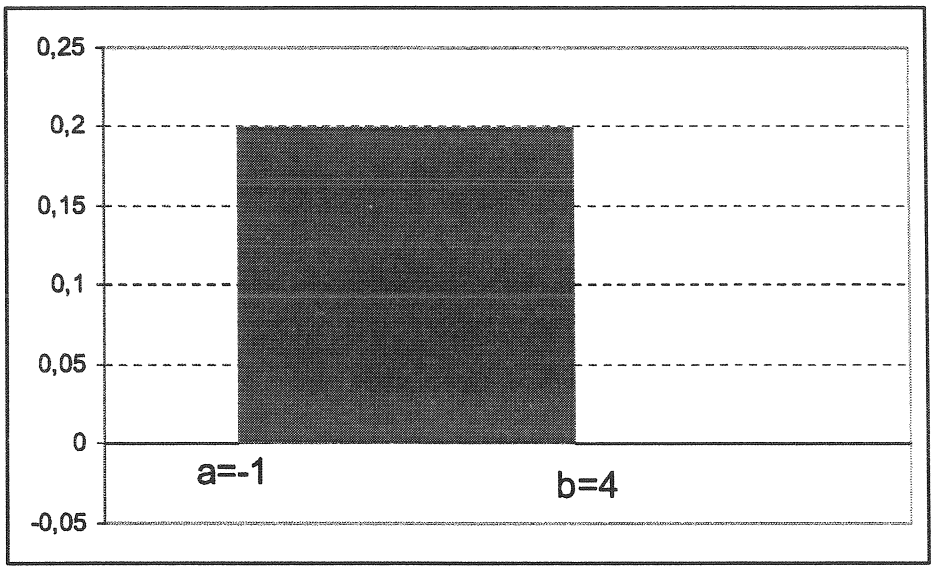


4 - Exemples classiques de densités de probabilité

Nous présentons trois exemples de lois. Les deux premiers font partie du programme de terminale, le troisième est important par ses applications en statistique.

a) Loi uniforme sur $[a, b]$

La fonction f, définie sur $\mathbb{R}$ par $f(x) = \dfrac{1}{b-a}$ si $x \in [a, b]$ et $f(x) = 0$ si $x \notin [a, b]$, est la densité de probabilité de la loi dite uniforme sur $[a, b]$.

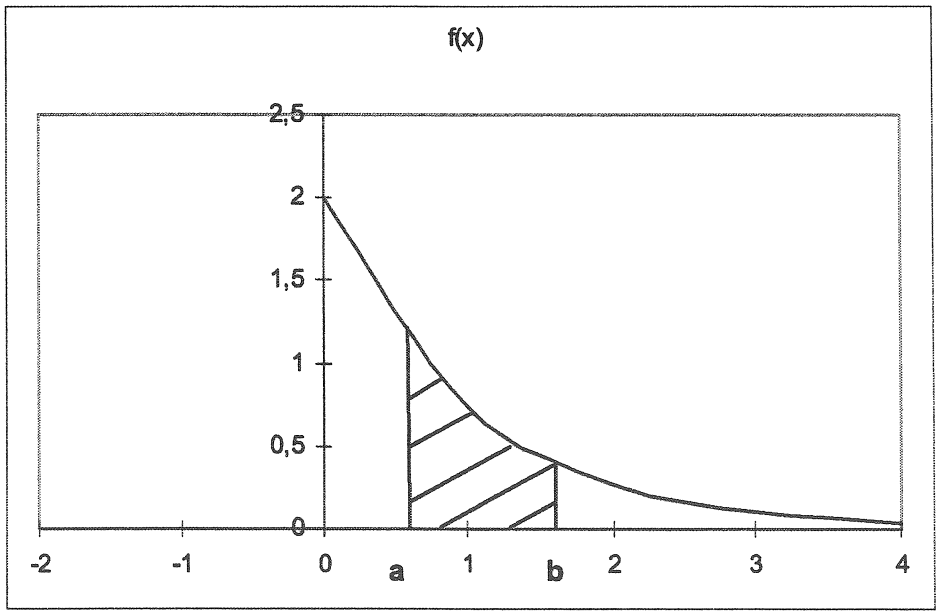


Exemple : Représentation de la densité de la loi uniforme pour $a = -1$ et $b = 4$.

Application : la loi uniforme sur $[a, b]$ est le modèle probabiliste continu usuellement choisi pour représenter l'expérience aléatoire décrite comme le choix « au hasard » d'un nombre dans l'intervalle $[a, b]$.

b) Loi exponentielle de paramètre λ ($\lambda > 0$)

La fonction f, définie sur $\mathbb{R}$ par $f(x) = \lambda e^{-\lambda x}$ si $x \geq 0$, et $f(x) = 0$ si $x < 0$, est la densité de probabilité de la loi dite exponentielle de paramètre λ ($\lambda > 0$).



Exemple : Représentation de la densité de la loi exponentielle pour $\lambda = 2$.

Nous pouvons alors retrouver facilement les résultats suivants :

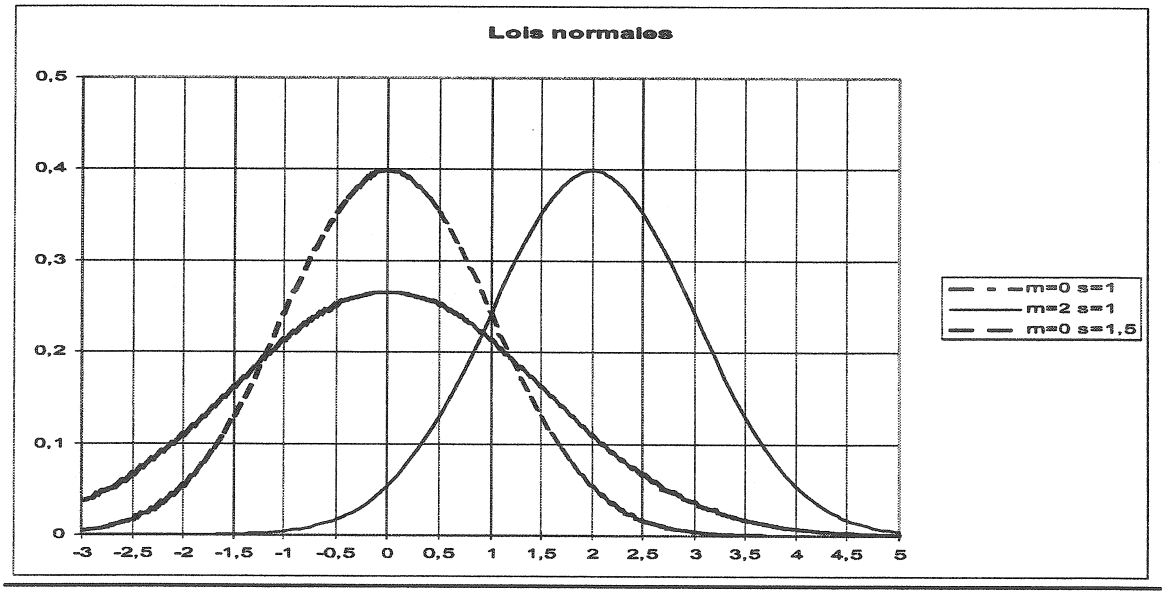
Pour tous réels a et b vérifiant $0 < a < b$,
$P(a \leq X \leq b) = \int_a^b 2e^{-2x}dx = \left[-e^{-2x}\right]_b^a = -e^{-2b} + e^{-2a}$; $P(X \leq a) = 1 - 2e^{-2x}$ et $P(X \geq a) = 2e^{-2x}$.

Application : La durée de vie d'un appareil, c'est-à-dire le temps de fonctionnement avant la première panne à partir de la mise en service (instant $t = 0$), est une variable aléatoire T continue qui suit une loi exponentielle de paramètre λ où λ est appelé taux instantané d'avarie, c'est la limite, quand h tend vers 0, du taux moyen d'avarie par unité de temps entre les instants t et $t + h$. On vérifie que le temps moyen d'attente avant la première panne est égal à $E(T) = \frac{1}{\lambda}$.

c) Loi normale de paramètres m et σ (m réel et $\sigma > 0$)

La fonction f définie sur $\mathbb{R}$ par $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$ est la densité de probabilité du modèle dit normal de paramètres m et σ.

Les paramètres m et σ sont exprimables à l'aide de la densité f par les relations $m = \int_{\mathbb{R}} xf(x)dx$ et $\sigma^2 = \int_{\mathbb{R}} (x-m)^2 f(x)dx$. Ces paramètres sont respectivement appelés espérance et écart type de cette loi.



Exemples : Représentation des densités des trois lois normales de paramètres $m = 0$ et $\sigma = 1$; $m = 2$ et $\sigma = 1$; $m = 0$ et $\sigma = 1,5$.

Chapitre III

Importance du modèle normal

1 - Simulation de deux situations aléatoires

Nous venons de voir que modéliser un phénomène aléatoire continu revient à se donner une fonction f de $\mathbb{R}$ dans $\mathbb{R}$, positive, intégrable dont l'aire totale sous la courbe représentative est égale à 1.

Formellement on peut donc construire toute une famille de modèles potentiels pour décrire de telles situations en construisant mathématiquement de telles fonctions. Dans la pratique certaines de ces fonctions joueront un rôle plus important que d'autres. Nous allons montrer que c'est en particulier le cas de la fonction de Gauss-Laplace qui définit le modèle normal. En effet ce modèle intervient très fréquemment dans la nature.

Nous allons d'abord illustrer cela à partir de la simulation de deux situations simples, puis nous expliquerons ce rôle fondamental du modèle normal en nous appuyant sur un théorème mathématique de la théorie des probabilités, le théorème-limite central, qui est fondamental dans les applications de la statistique.

Situation 1 : Considérons la simulation (avec Excel) de l'expérience aléatoire élémentaire du lancer d'un dé équilibré. Simulons 200 lancers consécutifs et calculons la moyenne des points obtenus au cours de ces 200 lancers. Répétons cette opération 500 fois. Nous obtenons donc 500 moyennes, dont nous visualisons la répartition statistique par un histogramme (cf. histogramme 1). L'histogramme obtenu a vaguement l'allure d'une cloche. Si nous recommençons plusieurs fois, le calcul de 500 autres moyennes, nous obtenons un histogramme différent du premier, mais nous constatons que l'allure générale de tous ces histogrammes est toujours celle d'une cloche. Cela permet de penser que le modèle normal pourrait être un bon candidat pour décrire mathématiquement le comportement aléatoire de la distribution des moyennes des faces obtenues chaque fois qu'on effectue 200 lancers consécutifs.

Situation 2 : Considérons une autre simulation où l'expérience aléatoire de base consiste à simuler un gain à trois valeurs -1, 0 et +9, de probabilités respectives 3/6, 2/6 et 1/6. Cette expérience ne relève plus d'une situation d'équiprobabilité comme dans le cas précédent. Répétons le même scénario que précédemment à partir de

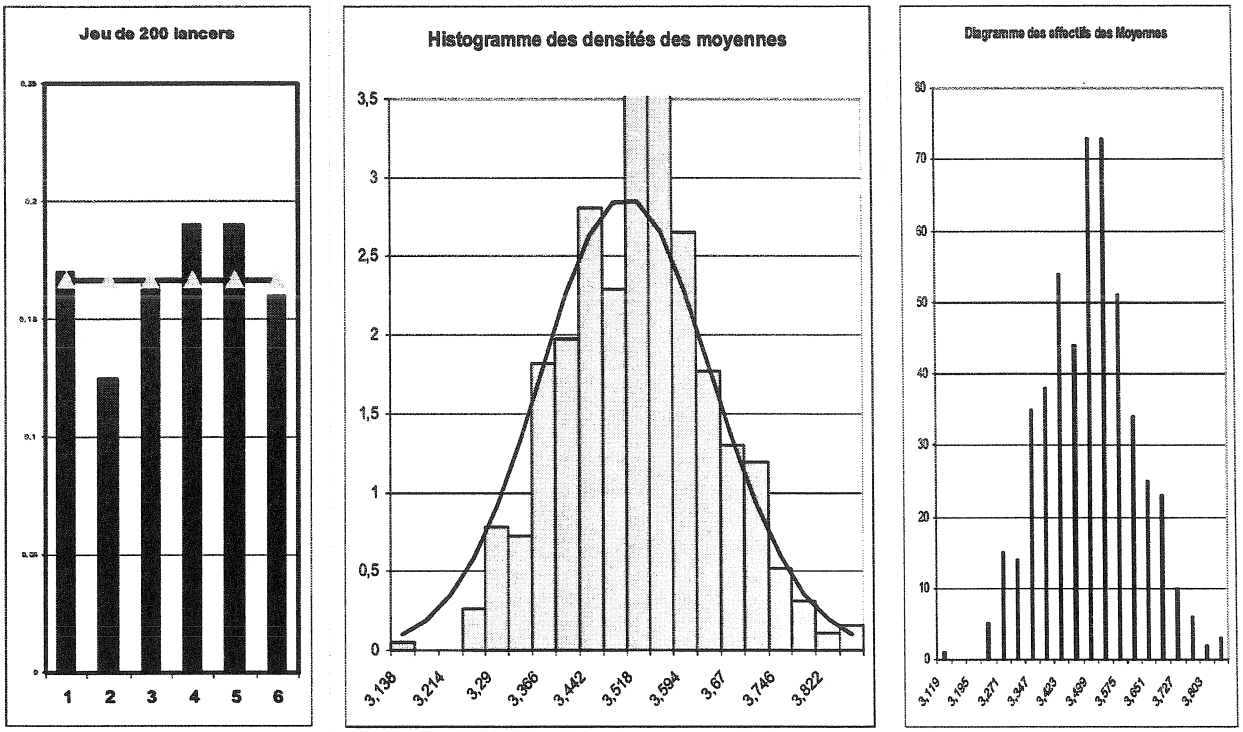
cette expérience élémentaire. Simulons 200 lancers consécutifs et calculons la moyenne des faces obtenues au cours de ces 200 lancers. Répétons cette opération 500 fois. Nous obtenons donc 500 moyennes, dont nous visualisons la répartition statistique par un histogramme (cf. histogramme 2). Si nous recommençons, nous obtenons encore des histogrammes de même allure en cloche. Ce qui nous conduit à la même conclusion, quant au choix du modèle possible, que dans la situation 1.

Ces deux constatations sont corroborées par la théorie mathématique qui établit que si, sous certaines hypothèses de régularité, nous calculons la moyenne empirique observée i.e. la moyenne des valeurs obtenues dans la répétitions de façon indépendante de n expériences aléatoires élémentaires, la distribution aléatoire de ces moyennes empiriques observées obéit à un modèle normal, d'autant mieux que le nombre entier n est grand. La théorie précise alors les valeurs que doivent prendre les paramètres μ et σ du modèle normal pertinent.

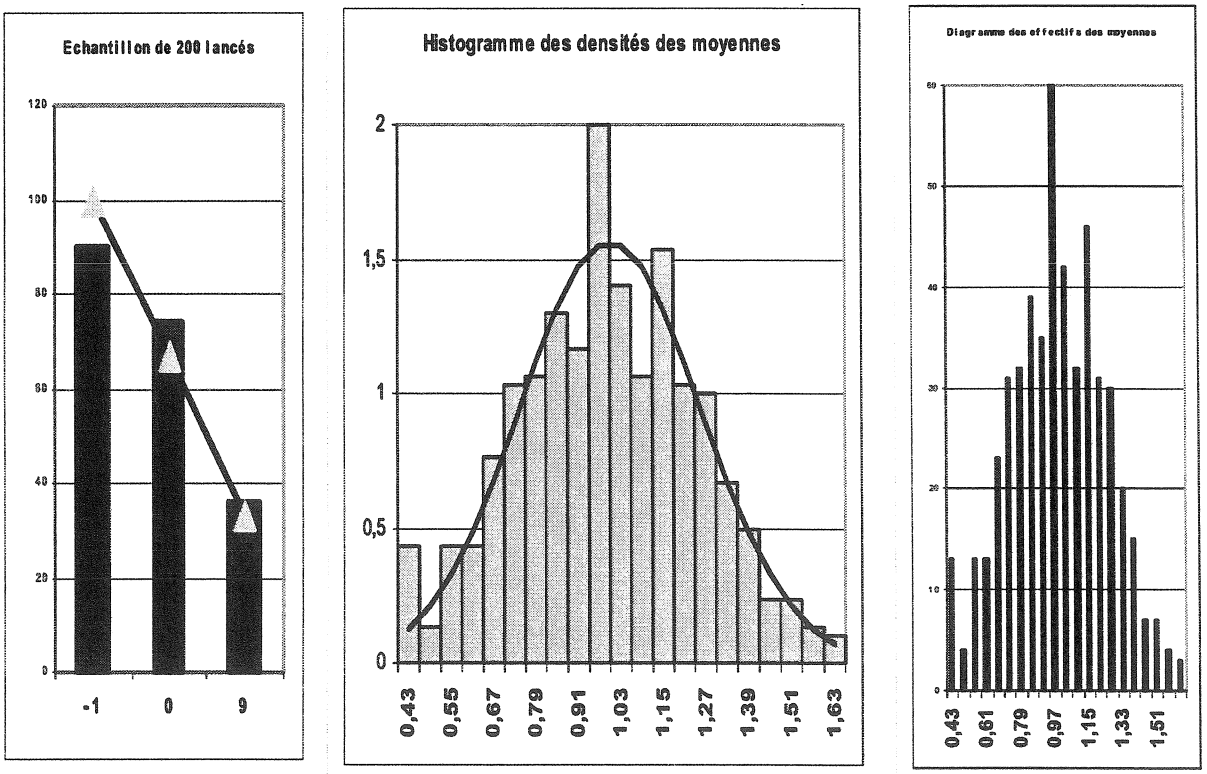
2 - Le théorème-limite central (TLC)

On exprime ce résultat en disant que la moyenne empirique est une variable aléatoire asymptotiquement normale. Un tel théorème de probabilités qui donne un résultat d'autant plus exact que n est grand, s'appelle un théorème-limite. Vu le rôle central que le résultat énoncé plus haut joue dans la théorie statistique, il est connu sous le nom de Théorème-Limite Central (TLC).

Le TLC explique l'importance du modèle normal dans la modélisation de nombreux phénomènes aléatoires. En effet, l'étude de situations physiques ou économiques par exemples, consiste souvent à faire des moyennes sur un grand nombre d'observations. Le modèle normal intervient alors naturellement en vertu du TLC. Le TLC est d'autant plus important que la conclusion de la normalité de la moyenne empirique est valable quel que soit le modèle élémentaire, connu ou inconnu, décrivant le caractère étudié.



Histogramme 1 : Jeu de dé (distribution de 500 moyennes de jeux de 200 lancers de dé)



Histogramme 2 : Jeu dissymétrique (distribution de 500 moyennes de jeux de 200 répétitions)

Soit X une variable aléatoire d'espérance μ et d'écart type σ modélisant le résultat d'une expérience aléatoire quelconque. Alors, la moyenne empirique d'un échantillon de taille n de cette variable X, c'est-à-dire la variable aléatoire

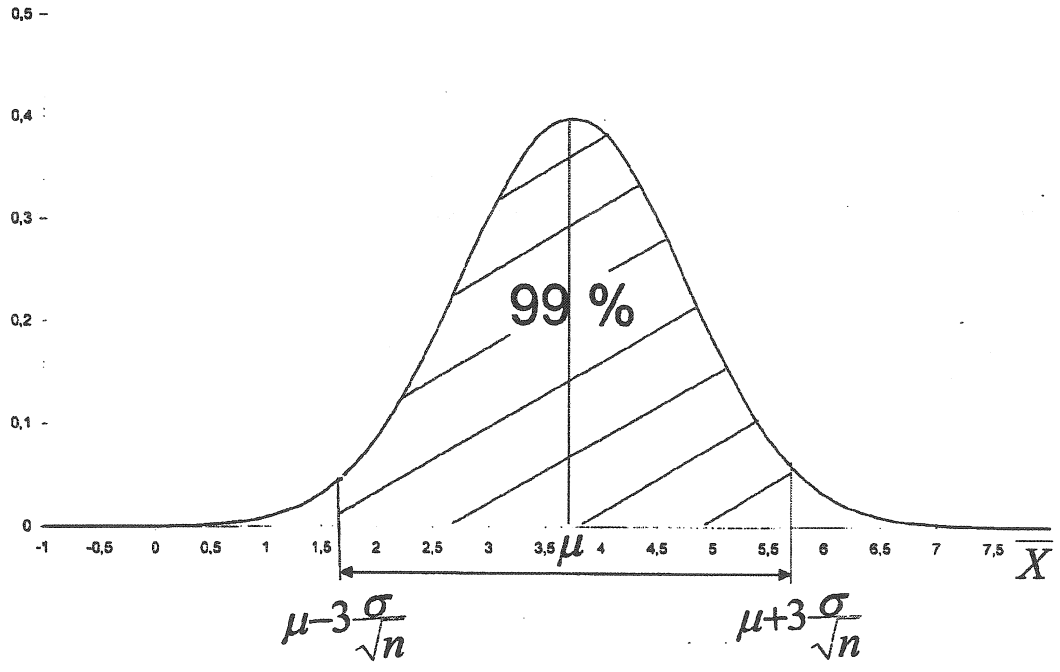
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

est asymptotiquement une variable aléatoire normale.

L'intérêt de ce résultat réside dans le fait que, pour l'obtenir, il n'est même pas nécessaire de connaître la distribution aléatoire de la variable X d'origine, c'est-à-dire qu'il n'est même pas nécessaire de connaître le modèle probabiliste initial décrivant cette variable.

La variable $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ est donc distribuée suivant une courbe de Gauss. Les paramètres du modèle sont aussi fixés par la théorie. Ils sont uniquement fonctions des paramètres μ et σ du modèle initial décrivant l'expérience aléatoire élémentaire et de la taille n de l'échantillon considéré.

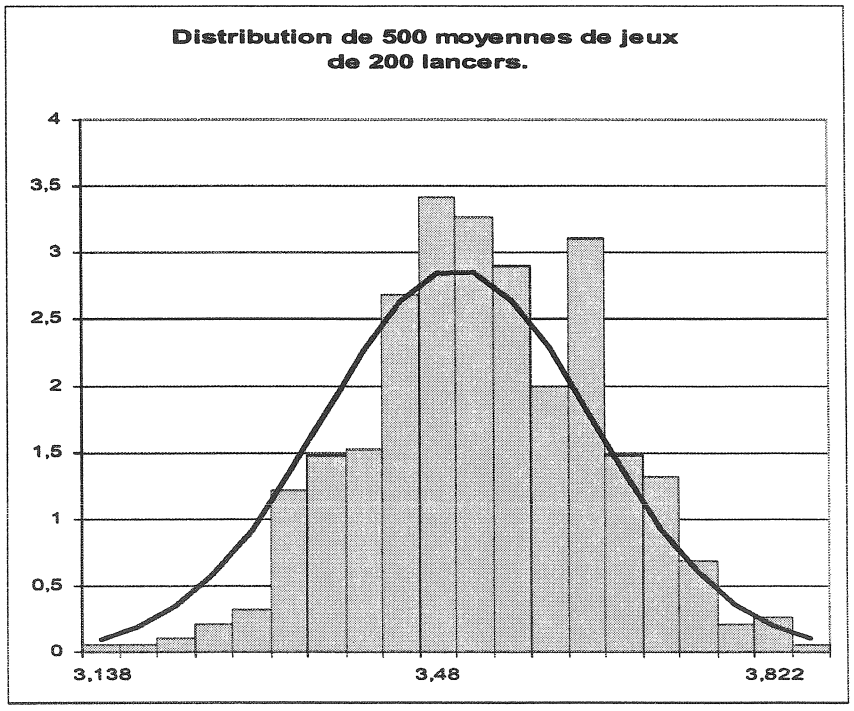
Comme conséquence intéressante du TLC, on montre qu'il y a une probabilité de 99% que le tirage d'un échantillon de taille n conduise à une moyenne empirique observée comprise entre $\mu - \frac{3\sigma}{\sqrt{n}}$ et $\mu + \frac{3\sigma}{\sqrt{n}}$. Ce qui peut être schématisé par le dessin ci-dessous :



Appliquons cette remarque au cas de la situation 1 précédente, où la variable aléatoire X modélise le résultat d'un lancer de dé équilibré. Dans ce cas $\mu = 3,5$ et $\sigma \approx 1,7$. Puisque nous avons observé des échantillons de taille $n = 200$, nous obtenons $\mu - \frac{3\sigma}{\sqrt{n}} \approx 3,14$ et $\mu + \frac{3\sigma}{\sqrt{n}} \approx 3,86$.

Cela signifie que sur une observation de 500 moyennes d'échantillons de taille 200, nous devons nous attendre à en avoir environ 99%, soit environ 495 moyennes, qui auront une valeur comprise entre 3,14 et 3,86.

Nous pouvons vérifier cette affirmation sur l'histogramme ci-dessous obtenu par simulation.



Variable X : résultat d'un lancer de dé équilibré.

Espérance : $\mu = 3,5$

Ecart-type : $\sigma \approx 1,7$

Taille : $n = 200$

et $\mu + 3\frac{\sigma}{\sqrt{n}} \approx 3,86$

$\mu - 3\frac{\sigma}{\sqrt{n}} \approx 3,14$

Chapitre IV

Adéquation d'un modèle probabiliste à la réalité

1 - Problématique des tests d'adéquation

Le choix d'un modèle probabiliste d'une situation aléatoire se ramène, comme nous venons de le voir, soit à la donnée d'une application positive définie sur $\mathbb{R}$, telle que l'aire totale située entre la courbe et l'axe des abscisses est égale à 1 (cas continu), soit à la donnée d'une suite de réels positifs de somme 1 (cas discret).

Une question se pose alors : *Comment choisir cette densité de probabilité ou cette suite de réels ?*

Nous avons vu, dans la partie précédente, que la description statistique d'une situation continue conduisait à des histogrammes dont l'allure générale pouvait nous guider pour sélectionner les fonctions candidates à être cette densité. Cependant si nous ne disposons pas de théorèmes, comme le TLC, pour confirmer ou départager ces candidats possibles, nous en sommes réduits à n'avoir qu'une approche qualitative pour choisir cette densité.

Il nous faut donc mettre en place un procédé mathématique qui permettra de « tester » si un candidat potentiel est digne d'être retenu au regard des observations réelles effectuées. Un tel procédé, appelé test d'adéquation, devra s'appuyer sur une analyse statistique de la situation permettant de proposer une décision à partir de l'observation d'un seul échantillon. Ce test revient donc à confronter à la réalité, via des observations de la situation étudiée, le modèle probabiliste théorique retenu a priori.

Dans cette partie, nous allons construire ce test en justifiant cette construction dans le cas d'une situation discrète, mais le raisonnement s'applique aussi au cas des situations continues. Nous prendrons appui, pour expliciter et illustrer la démarche de construction de ce test, sur l'exemple développé au Chapitre 8 (page 236) du manuel de Terminale S de la collection *Indice* édité chez Bordas[4] dont nous reproduisons ci-dessous un extrait :

---

[4] *Manuel de Terminale S programme 2002*, collection Indice, éditeur Bordas, auteurs : G. Mison, R-L. Gauthier, etc © Bordas/VUEF, 2002, ISBN 204-729597-1 (tous droits réservés)

7. Adéquation de données à une loi équirépartie

Un joueur veut vérifier si le dé qu'il possède est « normal », c'est-à-dire bien équilibré. On sait que, dans ce cas-là, la loi de probabilité associée est la loi uniforme : $P\{1\} = P\{2\} = P\{3\} = P\{4\} = P\{5\} = P\{6\} = \frac{1}{6}$.

Pour cela, le joueur lance 200 fois le dé et note les résultats obtenus :

x_i	1	2	3	4	5	6
n_i	31	38	40	32	28	31
f_i	0,155	0,190	0,200	0,160	0,140	0,155

Pour savoir si la distribution de fréquences obtenue est « proche » de la loi uniforme, on calcule la quantité suivante, qui prend en compte l'écart existant entre chaque fréquence trouvée et la probabilité théorique attendue :

$$d^2 = \left(0{,}155 - \frac{1}{6}\right)^2 + \left(0{,}190 - \frac{1}{6}\right)^2 + \ldots + \left(0{,}155 - \frac{1}{6}\right)^2 \approx 0{,}00268 .$$

▶ Notation

On note cette quantité d^2, car son calcul est celui du carré d'une distance.

Mais rien ne permet de dire pour l'instant si cette quantité trouvée est « petite » ou « grande ». En effet, elle est soumise à la fluctuation d'échantillonnage, puisque sa valeur varie d'une série de lancers à l'autre. On va donc étudier cette fluctuation d'échantillonnage pour convenir d'un seuil entre « petite » et « grande » valeur de d^2 lorsqu'on lance 200 fois un dé. Pour cela, on génère des séries de 200 chiffres au hasard pris dans {1 ; 2 ; 3 ; 4 ; 5 ; 6}. Les résultats trouvés pour le nombre d^2 à partir de 1 000 simulations sont résumés par le tableau suivant :

Minimum	D_1	Q_1	Médiane	Q_3	D_9	Maximum
0,00363	0,00138	0,00233	0,00363	0,00555	0,00789	0,01658

▶ Technique

Q_1 et Q_3 sont le premier et le troisième quartile et D_1 et D_9 sont le premier et le neuvième décile de la série.

Le **neuvième décile** de la série des valeurs simulées de d^2 est 0,00789.
Cela signifie que 90 % des valeurs de d^2 obtenues au cours de ces 1 000 simulations sont dans l'intervalle [0 ; 0.00789].
Comme la valeur observée de d^2 est inférieure à cette valeur seuil de 0,00789, on peut convenir que le dé est équilibré avec un risque de 10 %.
En effet, en utilisant cette méthode sur les données simulées, on se serait trompé dans 10 % des cas. On dit que l'on a un **seuil de confiance** de 90 %.

PROPRIÉTÉ

Soit une épreuve conduisant aux issues $a_1, a_2, \ldots, a_q$.
Expérimentalement, si on répète n fois cette épreuve ($n \geqslant 100$), on obtient les fréquences $f_1, f_2, \ldots, f_q$ pour chacune des issues. Pour vérifier **l'adéquation** de ces données à la **loi équirépartie** sur $\{a_1, a_2, \ldots, a_q\}$, on calcule le nombre $d^2 = \sum_{i=1}^{q} \left(f_i - \frac{1}{q}\right)^2$.

La réalisation d'un grand nombre de simulations de cette épreuve conduit pour la variable d^2 à une série statistique de neuvième décile D_9.

- Si $d^2 \leqslant D_9$, alors on dira que les données sont compatibles avec le modèle de la loi uniforme au seuil de risque 10 %.
- Si $d^2 > D_9$, on dira que les données ne sont pas compatibles avec ce modèle au seuil de risque 10 %.

▶ Technique

Le processus décrit ici est un cas particulier simplifié d'un processus beaucoup plus général et très utilisé en statistiques : le test du χ^2 (khi-deux).

La problématique développée dans l'activité présentée dans ce manuel peut se résumer d'une façon plus générale à la question suivante :

Etant donné une situation aléatoire réelle dont on a observé un échantillon de taille n, doit-on rejeter l'hypothèse, traditionnellement appelée hypothèse-nulle et notée H_0, suivant laquelle un certain modèle M serait un « bon » modèle pour modéliser la situation ?

2 - La règle de décision

Une telle question nécessite de se donner une règle statistique de décision, basée sur l'observation d'un échantillon. Mais l'utilisation de cette règle comportera un risque d'erreur α exprimé en pourcentage appelé le seuil de signification du test. Le principe de base d'une telle règle est un raisonnement classique par l'absurde légèrement accommodé à la sauce probabiliste avec un zeste d'optimisme.

De façon plus précise, une règle (classique) de décision basée sur le raisonnement classique par l'absurde s'énonce : *Si une hypothèse H_0 est vraie et si cette hypothèse implique d'observer un événement A de façon certaine quand on prélève un échantillon, alors, si on n'observe pas A, on rejette l'hypothèse H_0.*

La règle statistique de décision au seuil de signification α que nous adopterons sera la suivante : *Si une hypothèse H_0 est vraie et si cette hypothèse implique d'observer un événement A avec $(100-\alpha)\%$ de chances quand on prélève un échantillon, alors, si on n'observe pas A, on rejette l'hypothèse H_0.*

Remarquons que la règle classique de décision peut être assimilée à une règle statistique de décision de seuil $\alpha = 0\%$.

On comprend bien le risque encouru dans l'application d'une telle règle statistique de décision. Par exemple si α = 5%, l'hypothèse H_0 peut très bien être vraie sans pour autant observer l'événement A, puisque celui-ci a 5% de chances de ne pas se réaliser bien que H_0 soit vraie. Le réel α représente bien un risque, c'est le risque de se tromper dans l'application de la règle de décision. α est la probabilité de rejeter à tort l'hypothèse H_0, appelée aussi risque de première espèce. Le zeste d'optimisme du statisticien est de considérer qu'un risque de 5% est suffisamment faible pour raisonner comme s'il était nul, c'est-à-dire pratiquement suivant la règle classique.

Par exemple pour le lancer du dé de notre exemple dont nous souhaitons tester s'il est équilibré, l'hypothèse H_0 s'énoncera « Le dé est équilibré », ce qui se traduit en termes de modèle par : le modèle d'équiprobabilité des faces est le bon modèle pour décrire le comportement du dé en question.

a) L'événement A de la règle de décision

Nous avons dit que nous souhaitions pouvoir appliquer la règle de décision à partir de l'observation d'un seul échantillon de taille n. Une observation d'un tel échantillon se traduit mathématiquement par la donnée d'une suite de n nombres réels (n-uplet), c'est-à-dire d'un élément de $\mathbb{R}^n$. Pour plus de commodité, introduisons une application T, dite statistique de décision, de $\mathbb{R}^n$ dans $\mathbb{R}$. Une observation de l'échantillon conduit alors par calcul à une valeur de la statistique de décision $T_{observé}$. Le fait que l'échantillon soit prélevé au hasard permet de considérer la statistique de décision comme une variable aléatoire, dont nous pourrons étudier la distribution aléatoire des valeurs par une approche statistique.

L'ensemble des valeurs possibles pour T est a priori l'ensemble des nombres réels $\mathbb{R}$. Comme la décision à prendre est du type « rejet » ou « non rejet », établir la règle revient à partager $\mathbb{R}$ en deux parties disjointes et complémentaires, W dite zone de rejet (ou zone critique) et $\overline{W}$ dite zone de non-rejet. L'événement A sera réalisé si l'observation de l'échantillon conduit à une valeur $T_{observé}$ appartenant à W, il ne le sera pas si elle conduit à une valeur $T_{observé}$ appartenant à $\overline{W}$.

Le problème revient à choisir la statistique de décision. Ce choix dépendra du type de problème envisagé. Nous allons maintenant expliquer quelle statistique de décision nous considèrerons dans la règle de décision d'un test d'adéquation à un modèle discret.

b) Choix de la statistique de décision

Généralisons la situation de notre exemple. Considérons une expérience aléatoire discrète dont les issues possibles sont numérotées de 1 jusqu'à N (dans notre exemple du lancer d'un dé, $N = 6$) qu'on propose de modéliser par la suite des nombres $(p_1, p_2, \cdots, p_N)$ dont la somme vaut 1. Par exemple pour le lancer d'un dé dont nous souhaitons tester s'il est équilibré, nous prendrons $p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$. Si nous répétons de façon indépendante un très grand nombre n de fois l'expérience aléatoire, nous obtenons la suite des fréquences observées d'apparition des issues 1 à N pendant les n répétitions.

Nous avons donc deux suites de N nombres dont l'une $(p_1, p_2, \cdots, p_N)$ traduit le modèle proposé et l'autre $(f_1, f_2, \cdots, f_N)$ la réalité observée, résumées dans le tableau suivant :

Issues	*Réalité observée*	*Modèle proposé*
1	f_1	p_1
2	f_2	p_2
...	...	...
N	f_N	p_N

Pour choisir la statistique de décision, nous allons en quelque sorte « mesurer » l'écart qu'il y a entre la réalité observée et le modèle proposé puis prendre la décision en fonction de cette mesure. Une façon naturelle en mathématiques pour mesurer l'écart entre deux suites de n nombres est de considérer la distance usuelle dans IR^N, c'est-à-dire que, pour calculer $T_{observé}$ à partir des données de l'observation de l'échantillon, nous prendrons l'expression $d^2_{observé} = \sum_{i=1}^{N}(f_i - p_i)^2$. Cette quantité est notée sous forme d'un carré car elle est toujours positive.

Il est naturel de penser que si le modèle proposé est un « bon » modèle (hypothèse H_0), il faut s'attendre à ce que la valeur $T_{observé}$ soit proche de 0. En revanche, plus la valeur $T_{observé}$ sera éloignée de 0 plus nous serons incités à rejeter le modèle proposé. Il nous faut donc dire à partir de quelle valeur t_α, dite <u>valeur critique</u>, nous décrèterons que $T_{observé}$ est trop grand. Cette valeur critique t_α sera bien sûr fonction du seuil de signification α qu'on se sera fixé auparavant.

Une fois cette valeur critique t_α fixée, nous aurons bien partagé $\mathbb{R}$ en deux zones : la zone de non-rejet $\overline{W}_\alpha = [0, t_\alpha[$ et la zone de rejet $W_\alpha = [t_\alpha, \infty[$. Nous prendrons alors la décision en conséquence, suivant que la valeur calculée $T_{observé}$ de T à partir de l'observation de l'échantillon tombe ou non dans la zone de rejet W.

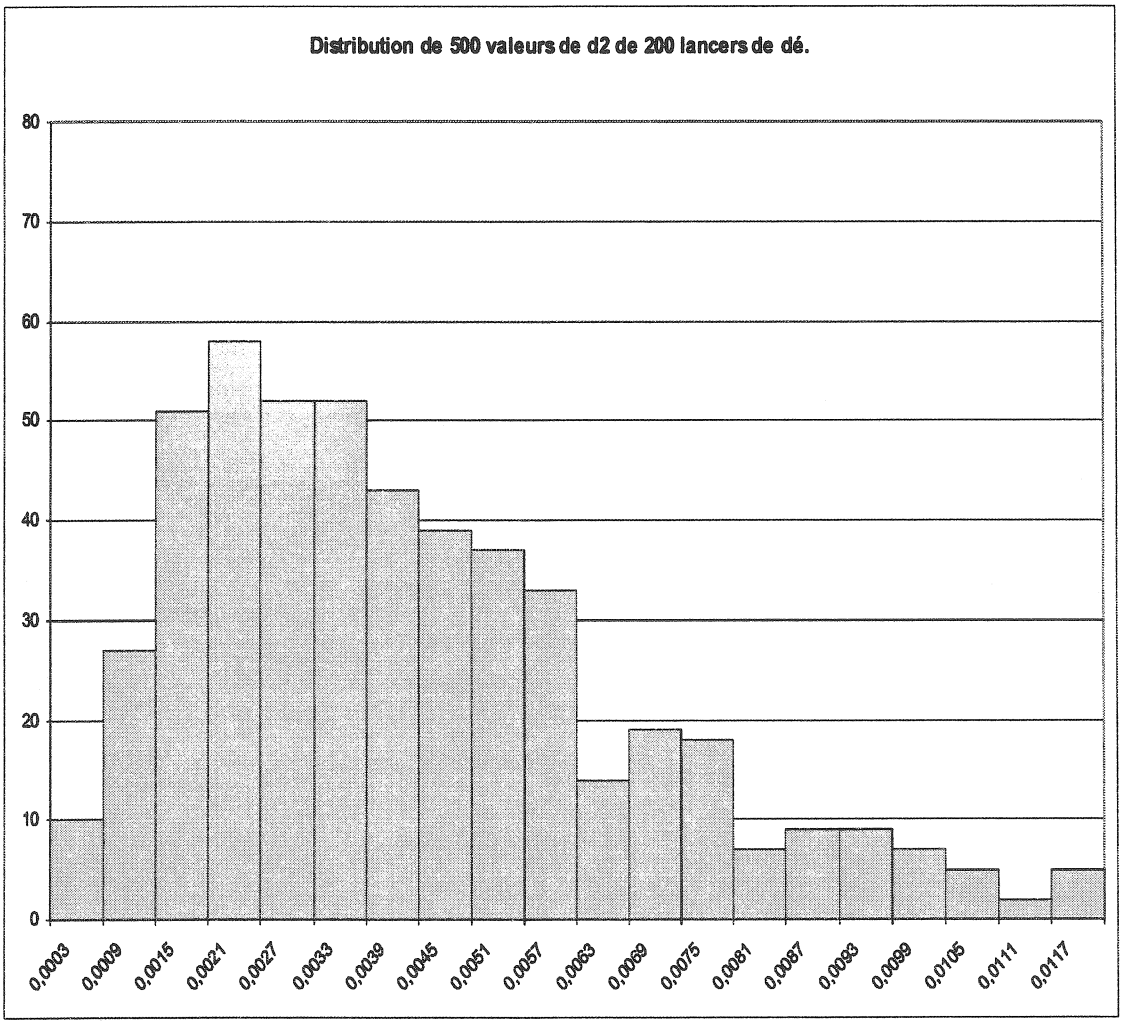
c) Choix de la valeur critique pour un seuil de signification $\alpha = 10\%$

La valeur critique t_α est choisie de sorte que la probabilité empirique de tomber dans la zone de rejet $W_\alpha = [t_\alpha, +\infty[$ soit égale à α. En général on choisit $\alpha = 1\%$, $\alpha = 5\%$ ou $\alpha = 10\%$; ici nous prendrons $\alpha = 10\%$.

Pour cela nous allons simuler par ordinateur une étude statistique des valeurs de $d^2_{observé} = \sum_{i=1}^{N}(f_i - p_i)^2$ en nous plaçant dans le modèle théorique attendu. Dans le cas qui nous intéresse, le modèle théorique est celui d'un dé parfaitement équilibré.

Nous allons donc simuler 200 lancers du dé équilibré c'est-à-dire observer un échantillon de taille $n = 200$ du lancer. Pour cette observation de 200 lancers, nous relèverons la fréquence observée de chaque face et calculerons la valeur de $d^2_{observé} = \sum_{i=1}^{N}(f_i - p_i)^2$. Répétons 500 fois cette simulation, ce qui nous donne une série statistique de 500 valeurs de $d^2_{observé}$, et traçons à l'ordinateur l'histogramme des fréquences des 500 valeurs de $d^2_{observé}$.

Nous obtenons un histogramme de la forme ci-dessous. La valeur déterminée empiriquement pour t_α, à partir de ces 500 observations d'échantillons de taille 200, correspond pour $\alpha = 10\%$, au neuvième décile empirique δ_9 de la série statistique c'est-à-dire à la valeur déterminée graphiquement sur l'histogramme telle que 90% des 500 valeurs observées soient inférieures à δ_9 et 10% des 500 valeurs observées soient supérieures à δ_9. Dans le cas de notre exemple nous avons trouvé $\delta_9 \approx 0{,}0078$ valeur comparable à celle mentionnée dans le manuel.

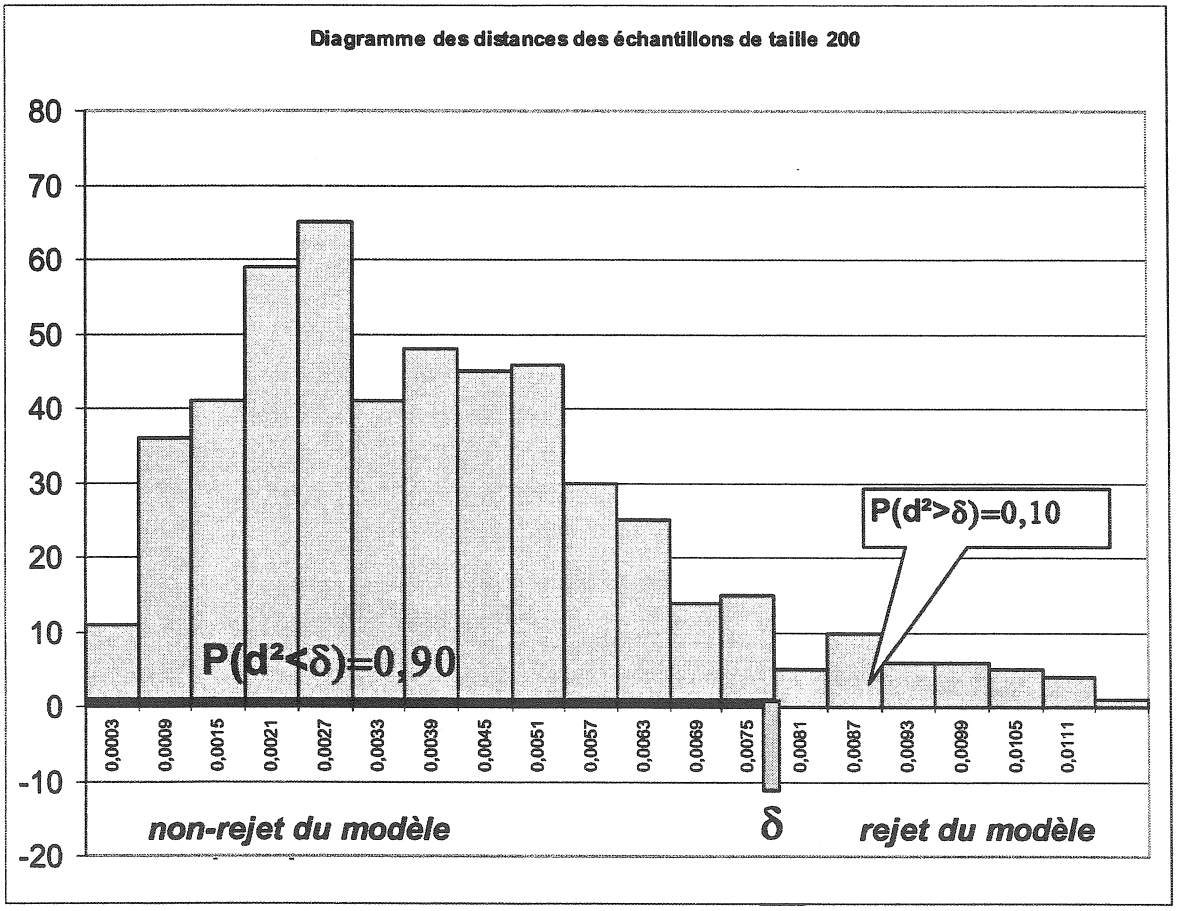


En terme de probabilités, cela signifie que, si le modèle théorique est le bon modèle, et si on convient de prendre sa décision sur l'observation d'un seul échantillon de taille 200, statistiquement il y a :

- 90% de chances d'obtenir un $d^2_{observé}$ inférieur à δ_9 auquel cas nous considérons que l'écart entre 0 et $d^2_{observé}$ est simplement dû au hasard et n'est pas significatif d'une inadéquation du modèle théorique. Il n'y a donc pas de raison statistique de rejeter le modèle théorique : on ne rejette pas l'hypothèse H_0.

- 10% de chances d'obtenir $d^2_{observé}$ supérieur à δ_9, auquel cas l'écart entre 0 et $d^2_{observé}$ est considéré trop grand pour être simplement dû au hasard et il est jugé plutôt significatif d'une inadéquation du modèle théorique : on rejette l'hypothèse H_0.

Toutes ces remarques et cette règle de décision peuvent se résumer dans le graphique ci-dessous.



Annexe I

Un peu de théorie : distance du Khi-Deux

Dans ce qui précède, nous avons mis en œuvre une démarche empirique basée sur la simulation avec un ordinateur du modèle théorique proposé. Cette démarche nous a permis d'étudier l'adéquation d'un modèle à un phénomène aléatoire réel présentant N issues possibles.

Issues	*Réalité observée*	*Modèle proposé*
1	f_1	p_1
2	f_2	p_2
...	...	...
N	f_N	p_N

Pour cela nous avons introduit une statistique à partir de la distance euclidienne $d^2 = \sum_{i=1}^{N}(f_i - p_i)^2$. Nous avons choisi cette distance car c'est la plus naturelle et c'est celle que les élèves ont déjà rencontrée.

Mais certains manuels adoptent d'autres définition de la distance pour cette même activité. Comme par exemple dans le manuel de terminale S de la collection *Transmath* édité chez Nathan[5], nous trouvons au chapitre 11 (pages 305) dont nous reproduisons ci-dessous un extrait avec l'aimable autorisation de l'éditeur :

[5] *Manuel de Terminale S programme 2002*, collection Transmath, éditeur Nathan, auteurs : A. Antibi, R. Bara, J. Morin etc ©Nathan/VUEF 2002, ISBN 209-172420-3.

TD 5 Adéquation avec la loi équirépartie

Dans ce TD, on veut comparer les résultats observés à partir d'expériences avec les valeurs théoriques attendues données par une loi de probabilité, ici la loi équirépartie.

1 Exemple 1

1. Les données

On lance 100 fois une pièce de monnaie, on obtient 55 fois « face » et 45 fois « pile ». Ce sont les résultats observés. Si l'on considère la pièce équilibrée, 50 fois « face » et 50 fois « pile » sont les valeurs attendues.

2. La mesure de l'écart entre les deux distributions

Appelons f_i les fréquences observées et p_i les probabilités des différents événements élémentaires. Pour « mesurer » l'écart entre la distribution observée des fréquences f_i avec celle des p_i liés à la loi équirépartie, on calcule le nombre obtenu en ajoutant toutes les valeurs $\frac{(f_i - p_i)^2}{p_i}$. Notons d^2 cette somme.

• Remplissez la dernière colonne du tableau suivant et calculez alors le nombre d^2.

	fréquence observée (f_i)	probabilité (p_i)	$\frac{(f_i - p_i)^2}{p_i}$
pile	0,45	0,5	
face	0,55	0,5	

Si les valeurs théoriques et observées sont proches les unes des autres, d^2 sera petit et on considérera qu'il y a adéquation entre l'expérience et la loi équirépartie.

L'utilisation de cette nouvelle distance peut s'expliquer par le fait que la théorie des probabilités montre que la distribution théorique de $d^2 = \sum_{i=1}^{N}(f_i - p_i)^2$ fait intervenir, même pour n grand, à la fois la taille n de l'échantillon et le nombre N d'issues de l'expérience aléatoire. Pour éviter cette double dépendance, et surtout la dépendance explicite vis-à-vis de la taille n de l'échantillon, les statisticiens préfèrent utiliser une autre statistique moins naturelle que le d^2, définie par $D^2 = n\sum_{i=1}^{N}\frac{(f_i - p_i)^2}{p_i}$ et appelée distance du Khi-Deux.

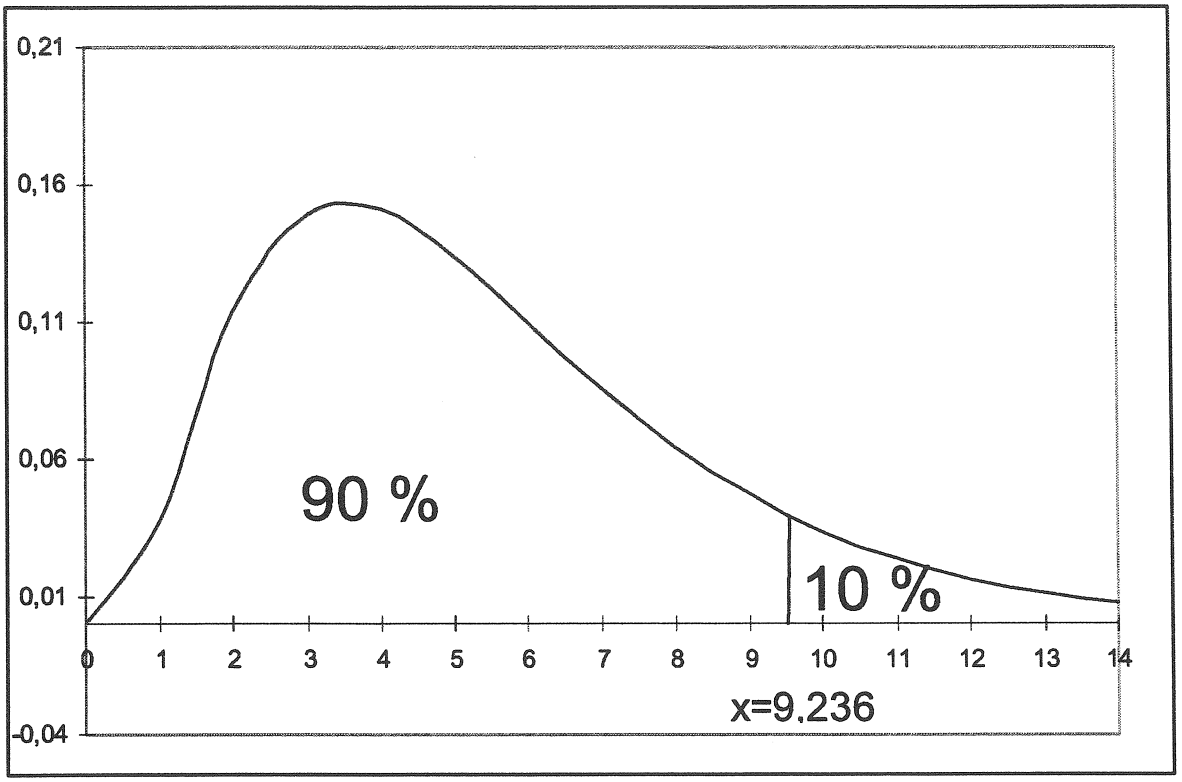
L'intérêt de cette nouvelle statistique est justement que la distribution utilisée pour D^2 ne fait plus intervenir que le nombre N d'issues de l'expérience aléatoire. Elle ne s'exprime donc plus avec la taille n de l'échantillon. On montre que cette statistique, sous l'hypothèse que le modèle théorique est adapté au phénomène étudié (i.e. hypothèse H_0), suit approximativement une loi, dite du Khi-Deux à $N - 1$ degrés de liberté, dont la densité a l'allure de la courbe de la figure de la page suivante.

Appliquons cette remarque au cas du dé dont on souhaite tester l'équilibre par l'étude d'un échantillon de taille 200.

Nous avons $p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$, $N = 6$, $n = 200$

Par un calcul élémentaire, nous obtenons la relation liant d^2 et D^2 :

$$D^2 = 6\, n\, d^2 = 1200\, d^2$$



Si le dé testé est équilibré, la variable aléatoire $D^2 = 1200\, d^2$ aura donc pour densité une courbe de la forme ci-dessus (densité du Khi-deux à 5 degrés de liberté).

Pour déterminer le neuvième décile théorique t_9 (dont δ_9 est une estimation empirique par simulation dans la démarche précédente) défini par $P(d^2 \leq t_9) = 0{,}90$, il suffit de se ramener à l'utilisation de la variable D^2 en remarquant que $P(d^2 \leq t_9) = P(D^2 \leq 1200\, t_9) = 0{,}90$

Ceci nous conduit à déterminer le réel x défini par $P(D^2 \leq x) = 0{,}90$. Il suffit pour cela d'utiliser une table du Khi-deux à 5 degrés de liberté qui nous donne 9,236 pour valeur de x.

Par suite, $1200\, t_9 \approx 9{,}236$. Ce qui donne $t_9 \approx 0{,}007696$, valeur théorique à comparer à la valeur estimée empiriquement par simulation $\delta_9 \approx 0.0078$.

Annexe II

Simulation informatique d'une variable aléatoire

Commençons par prouver, dans un cas simple, une propriété très générale des variables aléatoires.

1 - Propriété de simulation

Si X est une variable aléatoire dont la fonction de répartition F est continue et strictement croissante de $\mathbb{R}$ *sur* $[0;1[$, *alors* $Y = F(X)$ *est une variable aléatoire de loi uniforme sur* $[0;1[$.

2 - Démonstration

Montrons que la variable aléatoire définie par $Y = F(X)$ a pour fonction de répartition G la fonction de répartition d'une variable aléatoire uniforme sur]0, 1[.

- Soit y un réel tel que $0 < y < 1$, par définition de Y et des fonctions de répartition, $G(y) = P(Y \leq y) = P(F(X) \leq y)$. Par suite $G(y) = P(X \leq F^{-1}(y))$, car F est une bijection strictement croissante sur $\mathbb{R}$. D'où $G(y) = F(F^{-1}(y))$, par définition de la fonction de répartition F de X. Ce qui conduit pour ce cas à $G(y) = y$.

- Soit y un réel strictement négatif, alors $G(y) = P(Y \leq y) = 0$ car Y est une variable positive.

- Soit y un réel supérieur ou égal à 1, alors $G(y) = P(Y \leq y) = 1$ car $0 \leq Y \leq 1$.

En conclusion, la fonction de répartition $G : \mathbb{R} \to [0,1]$ de la variable Y est définie, pour tout $y \in \mathbb{R}$, par :

$$\begin{cases} G(y) = 0 & si\ y \leq 0 \\ G(y) = y & si\ 0 < y < 1 \\ G(y) = 1 & si\ y \geq 1. \end{cases}$$

Ce qui prouve que Y est bien une variable aléatoire uniforme sur]0 ; 1[.

3 - Principe de simulation

Une variable uniforme sur [0, 1] peut être simulée par les fonctions *RAND* ou *ALEA()* des calculatrices ou des ordinateurs. De $Y = F(X) = RAND$, nous déduisons

$$X = F^{-1}(Y) = F^{-1}(RAND).$$

Cette remarque élémentaire conduit au principe de simulation suivant :

Pour simuler une variable aléatoire X dont on connaît la fonction de répartition F, admettant une fonction réciproque F^{-1}, Il suffit de simuler avec une machine la variable $F^{-1}(RAND)$.

Dans le cadre plus général où la fonction F n'est pas inversible, on montre qu'on peut remplacer F^{-1} par une fonction $F,^{*}$ construite à partir de F et ayant à peu près les mêmes propriétés que F^{-1}, à laquelle on fait jouer le même rôle que F^{-1} dans le principe de simulation.

4 - Exemples

a) Dans l'exercice 1 de la page 18

D'après la réponse à la question 3 de cet exercice, la fonction de répartition F de la variable X est définie sur $\mathbb{R}$ par $F(t) = P([0\,;\,t]) = t^4$. Par suite $F^{-1}(t) = \sqrt[4]{t}$.

Le principe de simulation entraîne que la simulation d'une variable X de loi définie par la fonction de densité $f(t) = 4\,t^3$ est obtenue en programmant la fonction $\sqrt[4]{RAND}$.

b) Dans l'exercice 2 de la page 19

La fonction de répartition F d'une variable aléatoire de loi exponentielle de paramètre λ est définie sur $\mathbb{R}$ par $F(t) = 1 - e^{-\lambda t}$, pour tout réel t positif.

Par suite $F^{-1}(t) = -\dfrac{\ln(1-t)}{\lambda}$.

Le principe de simulation entraîne que la simulation d'une variable X de loi exponentielle de paramètre λ est obtenue en programmant la fonction $-\dfrac{\ln(1-RAND)}{\lambda}$. Dans le cas particulier de l'exercice 2, il faut donc programmer la fonction $-\dfrac{\ln(1-RAND)}{2}$.

Annexe III

Deux variables aléatoires continues

au programme de terminale S

1 - Variable aléatoire de loi uniforme sur [0 ; 1]

Sa fonction de densité f est définie par $f(x) = 1$ si $x \in [0 ; 1]$ et $f(x) = 0$ sinon.

Sa fonction de répartition est définie par $F(x) = x$ si $x \in [0 ; 1]$, $F(x) = 0$ si x est négatif et $F(x) = 1$ si x est supérieur à 1.

Son espérance mathématique est $E(X) = \int_0^1 x\, f(x)dx = \int_0^1 x\, dx = 0,5$.

Sa variance est $Var(X) = E(X^2) - [E(X)]^2 = \int_0^1 x^2\, f(x)dx - \frac{1}{4} = \int_0^1 x^2\, dx - \frac{1}{4} = \frac{1}{12}$.

Son écart type est $\sigma = 1/\sqrt{12}$.

2 - Variable aléatoire de loi exponentielle de paramètre $\lambda > 0$

Sa fonction densité f est définie par $f(x) = \lambda e^{-\lambda x}$ si x est positif et $f(x) = 0$ sinon.

Sa fonction de répartition est définie par $F(x) = 1 - e^{-\lambda x}$ si x est positif et $F(x) = 0$ sinon.

Son espérance mathématique est $E[X] = \lim_{t\to\infty} \int_0^t xf(x)dx = \frac{1}{\lambda}$, car par une intégration par parties,

$$\int_0^t x\, f(x)dx = \int_0^t \frac{1}{\lambda} x\, e^{-\lambda x}dx = \left[-xe^{-\lambda x}\right]_0^t + \int_0^t e^{-\lambda x}dx = \left[-xe^{-\lambda x} - \frac{1}{\lambda}e^{-\lambda x}\right]_0^t = -te^{-\lambda t} - \frac{1}{\lambda}e^{-\lambda t} + \frac{1}{\lambda}.$$

Sa variance est donnée par

$$Var(x) = E(X^2) - [E(X)]^2 = \lim_{t\to\infty} \int_0^t x^2\, f(x)dx - \frac{1}{\lambda^2} = \lim_{t\to\infty} \int_0^t \frac{1}{\lambda} x^2 e^{-\lambda x}\, dx - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Son écart type est $\sigma = 1/\lambda$.

Annexe IV

Protocole pour l'utilisation d'un tableur en classe

1 - Pour fabriquer une présentation

Utiliser le camescope gratuit de Microsoft.

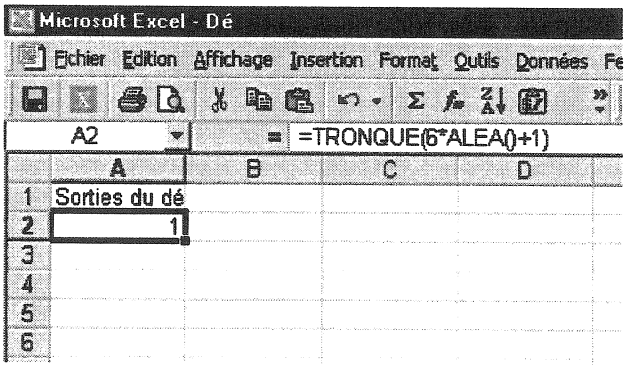
Exécuter le programme Camcordr.exe qui se trouve dans le répertoire valupack\Mscam du CD-Rom d'installation d'office 97. Une fois installé, ce petit programme vous permet d'enregistrer dans un fichier vidéo au format Avi (ou dans un programme exécutable) tout ce qui se déroule à l'écran : les activations de fenêtres, de menus, les déplacements de souris, etc. Ce fichier se visionne sur n'importe quel autre micro-ordinateur. Un moyen relativement simple de présenter sous forme de vidéo un document de formation à l'utilisateur d'un logiciel quelconque.

2 - Simulation du jeu de dé

But : Faire apparaître des nombres entiers « au hasard », compris entre 1 et 6.
Outils utilisés : Un tableur et sa fonction *ALEA()*, générateur de nombres « au hasard », et sa fonction *TRONQUE* qui donne la partie entière d'un nombre.
Fonction utilisée : *TRONQUE(6*ALEA()+1)*, car $0 < ALEA() < 1$ donne $1 < 6*ALEA() + 1 < 7$ dont la partie entière est un nombre entier compris entre 1 et 6.

Procédure : Prendre une feuille de calcul et la nommer « de.XLS »

Dans la cellule A1 mettre le titre : Sorties du dé
Dans la cellule A2 écrire la fonction : *=TRONQUE(6*ALEA()+1)*.
La validation de cette fonction donne un nombre entier compris entre 1 et 6.
On rejoue en utilisant la touche F9 .



3 - Simulation de l'exercice 1 page 18

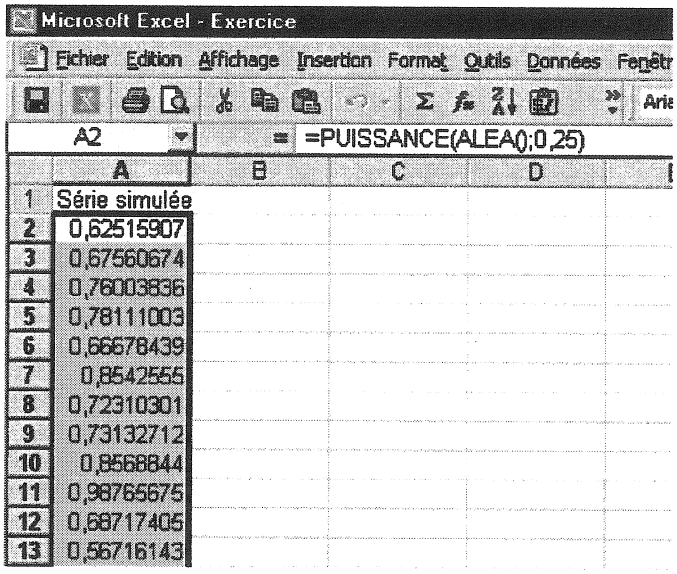
But : Fabriquer un échantillon de 200 valeurs simulées d'une variable aléatoire dont la fonction de densité est définie par $f(t) = 4\,t^3$.
Outils utilisés : Un tableur et sa fonction *ALEA()*, générateur de nombres « au hasard »,

- sa fonction *PUISSANCE(t ; n)* qui donne la valeur de t^n.
- la fonction réciproque de la fonction de répartition F définie par $F(t) = t^4$, ce qui donne $F^{-1}(t) = t^{0,25}$ (voir annexe II)

Fonction utilisée : $\sqrt[4]{ALEA()}$, voir annexe II pour la justification.

Procédure : Prendre une feuille de calcul, la nommer « exercice »

- Dans la cellule A1 mettre le titre : Série simulée
- Dans la cellule A2 écrire la fonction : *=PUISSANCE(ALEA();0,25)*
- La validation de cette fonction donne un nombre compris entre 0 et 1.
- Recopier 200 fois vers le bas cette fonction dans la plage A2:A201 en tirant la poignée de la cellule A2 vers le bas.
- Un autre échantillon apparaît grâce à la touche F9.



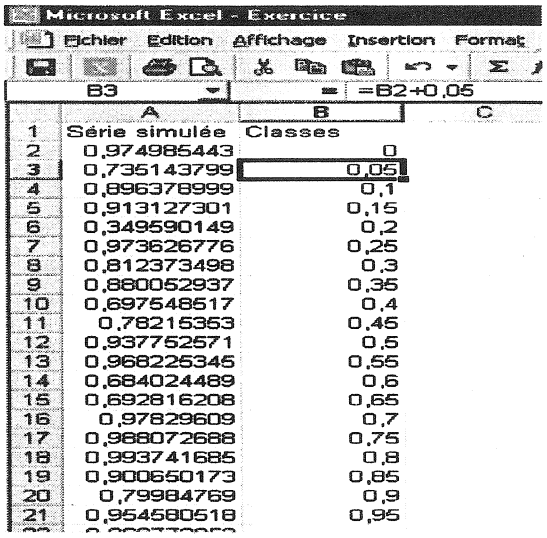
4 - Regroupement des données statistiques

But : Regrouper dans des classes d'amplitude 0,05 les 200 valeurs simulées .

Outils utilisés : La fonction *FREQUENCE(Plage des données, plage des bornes sup des classes)*. Cette fonction matricielle est délicate à manipuler et doit être validée par les trois touches **<Ctrl>**, ⇧, ↵ enfoncées ensemble.

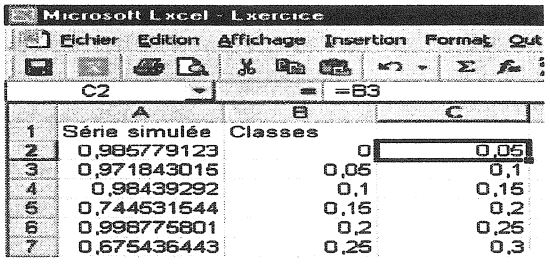
Procédure : Prendre la feuille de calcul « exercice »

Construction de la liste des classes
D'abord les bornes inférieures :
Dans la cellule B1 placer le titre : Classes
Dans la cellule B2 placer le nombre 0.
Dans la cellule B3 écrire la formule =B2+0,05 qui signifie « ajouter 0,05 au contenu de la cellule du dessus », et recopier cette formule jusqu'à avoir 0,95 dans la cellule B21.



	A	B	C
1	Série simulée	Classes	
2	0,974985443	0	
3	0,735143799	0,05	
4	0,896378999	0,1	
5	0,913127301	0,15	
6	0,349590149	0,2	
7	0,973626776	0,25	
8	0,812373498	0,3	
9	0,880052937	0,35	
10	0,697548517	0,4	
11	0,78215363	0,45	
12	0,937752571	0,5	
13	0,968225345	0,55	
14	0,684024489	0,6	
15	0,692816208	0,65	
16	0,97829609	0,7	
17	0,988072688	0,75	
18	0,993741685	0,8	
19	0,900650173	0,85	
20	0,79984769	0,9	
21	0,954580518	0,95	

Construction des bornes supérieures
Dans la cellule C2 écrire la formule *=B3* qui reporte le contenu de B3 en C2, et recopier cette formule jusqu'à avoir *0,95* dans la cellule C20.
Placer *1* dans la cellule C21.

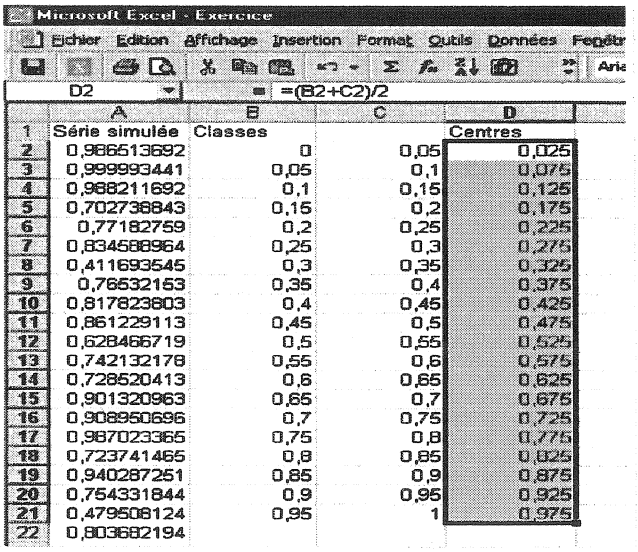


	A	B	C
1	Série simulée	Classes	
2	0,985779123	0	0,05
3	0,971843015	0,05	0,1
4	0,98439292	0,1	0,15
5	0,744531544	0,15	0,2
6	0,998775801	0,2	0,25
7	0,675436443	0,25	0,3

Construction de la liste des centres des classes

Dans la cellule D1 placer le titre : Centres

Dans la cellule D2 écrire la formule *=(B2+C2)/2* et recopier cette formule jusqu'à avoir *0,975* dans la cellule D21.



	A	B	C	D
1	Série simulée	Classes		Centres
2	0,986513692	0	0,05	0,025
3	0,999993441	0,05	0,1	0,075
4	0,988211692	0,1	0,15	0,125
5	0,702738843	0,15	0,2	0,175
6	0,77182759	0,2	0,25	0,225
7	0,834588964	0,25	0,3	0,275
8	0,411693545	0,3	0,35	0,325
9	0,76532153	0,35	0,4	0,375
10	0,817823803	0,4	0,45	0,425
11	0,861229113	0,45	0,5	0,475
12	0,628466719	0,5	0,55	0,525
13	0,742132178	0,55	0,6	0,575
14	0,728520413	0,6	0,65	0,625
15	0,901320963	0,65	0,7	0,675
16	0,908950696	0,7	0,75	0,725
17	0,987023365	0,75	0,8	0,775
18	0,723741465	0,8	0,85	0,825
19	0,940287251	0,85	0,9	0,875
20	0,754331844	0,9	0,95	0,925
21	0,479508124	0,95	1	0,975
22	0,803682194			

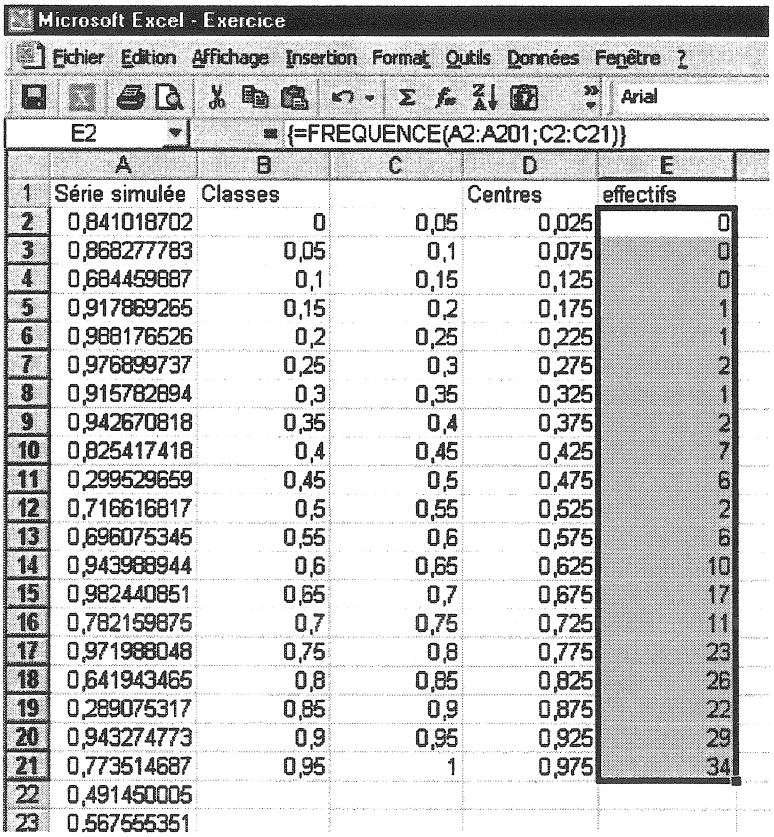
Construction de la liste des effectifs des classes

Dans la cellule E1 écrire : effectifs

Sélectionner la plage E2:E21 et taper =FREQUENCE(A2:A201;C2:C21)

Valider avec les 3 touches <Ctrl>, ⇑, ↵

La validation de cette fonction donne les effectifs de chaque classe.



	A	B	C	D	E
1	Série simulée	Classes		Centres	effectifs
2	0,841018702	0	0,05	0,025	0
3	0,868277783	0,05	0,1	0,075	0
4	0,684459887	0,1	0,15	0,125	0
5	0,917869265	0,15	0,2	0,175	1
6	0,988176526	0,2	0,25	0,225	1
7	0,976899737	0,25	0,3	0,275	2
8	0,915782894	0,3	0,35	0,325	1
9	0,942670818	0,35	0,4	0,375	2
10	0,825417418	0,4	0,45	0,425	7
11	0,299529659	0,45	0,5	0,475	6
12	0,716616817	0,5	0,55	0,525	2
13	0,696075345	0,55	0,6	0,575	6
14	0,943988944	0,6	0,65	0,625	10
15	0,982440851	0,65	0,7	0,675	17
16	0,782159875	0,7	0,75	0,725	11
17	0,971988048	0,75	0,8	0,775	23
18	0,641943465	0,8	0,85	0,825	26
19	0,289075317	0,85	0,9	0,875	22
20	0,943274773	0,9	0,95	0,925	29
21	0,773514687	0,95	1	0,975	34
22	0,491450005				
23	0,567555351				

Construction de la liste de densités de fréquence

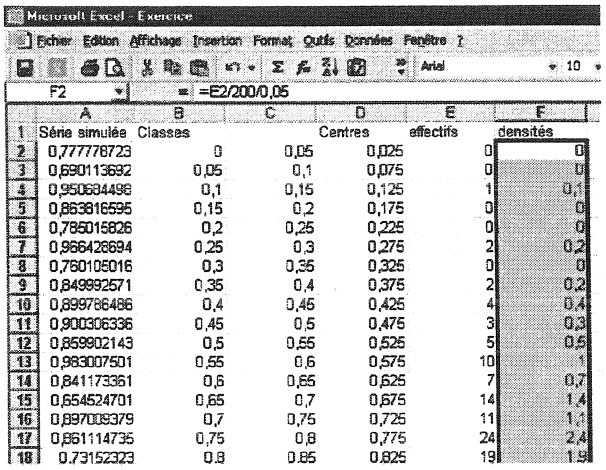
Formule : rapports des fréquences sur les amplitudes.

But : ce sont les ordonnées de l'histogramme.

Dans la cellule F1 écrire : densités

Dans la cellule F2 taper *=E2/200/0,05*

Recopier cette formule jusqu'à la cellule F21.



	A	B	C	D	E	F
1	Série simulée	Classes		Centres	effectifs	densités
2	0,777778723	0	0,05	0,025	0	0
3	0,690113692	0,05	0,1	0,075	0	0
4	0,950684498	0,1	0,15	0,125	1	0,1
5	0,863816595	0,15	0,2	0,175	0	0
6	0,785015826	0,2	0,25	0,225	0	0
7	0,966426694	0,25	0,3	0,275	2	0,2
8	0,760105016	0,3	0,35	0,325	0	0
9	0,849992571	0,35	0,4	0,375	2	0,2
10	0,899786486	0,4	0,45	0,425	4	0,4
11	0,900306336	0,45	0,5	0,475	3	0,3
12	0,859902143	0,5	0,55	0,525	5	0,5
13	0,983007501	0,55	0,6	0,575	10	1
14	0,841173361	0,6	0,65	0,625	7	0,7
15	0,654524701	0,65	0,7	0,675	14	1,4
16	0,897008379	0,7	0,75	0,725	11	1,1
17	0,861114736	0,75	0,8	0,775	24	2,4
18	0,73152323	0,8	0,85	0,825	19	1,9

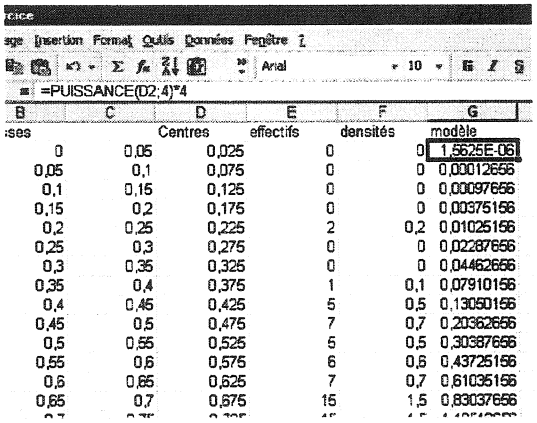
Construction de la liste des valeurs du modèle simulé : $f(t) = 4\,t^4$

Dans la cellule G1 écrire : modèle

Dans la cellule G2 taper :

=PUISSANCE(D2;4)*4

Recopier cette formule jusqu'à la cellule F21.



B	C	D	E	F	G
ses		Centres	effectifs	densités	modèle
0	0,05	0,025	0	0	1,5625E-06
0,05	0,1	0,075	0	0	0,00012656
0,1	0,15	0,125	0	0	0,00097656
0,15	0,2	0,175	0	0	0,00375156
0,2	0,25	0,225	2	0,2	0,01025156
0,25	0,3	0,275	0	0	0,02287656
0,3	0,35	0,325	0	0	0,04462656
0,35	0,4	0,375	1	0,1	0,07910156
0,4	0,45	0,425	5	0,5	0,13050156
0,45	0,5	0,475	7	0,7	0,20362656
0,5	0,55	0,525	5	0,5	0,30387656
0,55	0,6	0,575	6	0,6	0,43725156
0,6	0,65	0,625	7	0,7	0,61035156
0,65	0,7	0,675	15	1,5	0,83037656

5 - Les graphiques

Construction du graphique qui nous permettra d'évaluer l'adéquation de la simulation et du modèle.

Sélectionner la plage F1:G21.

Dans le menu insertion choisir « graphique »

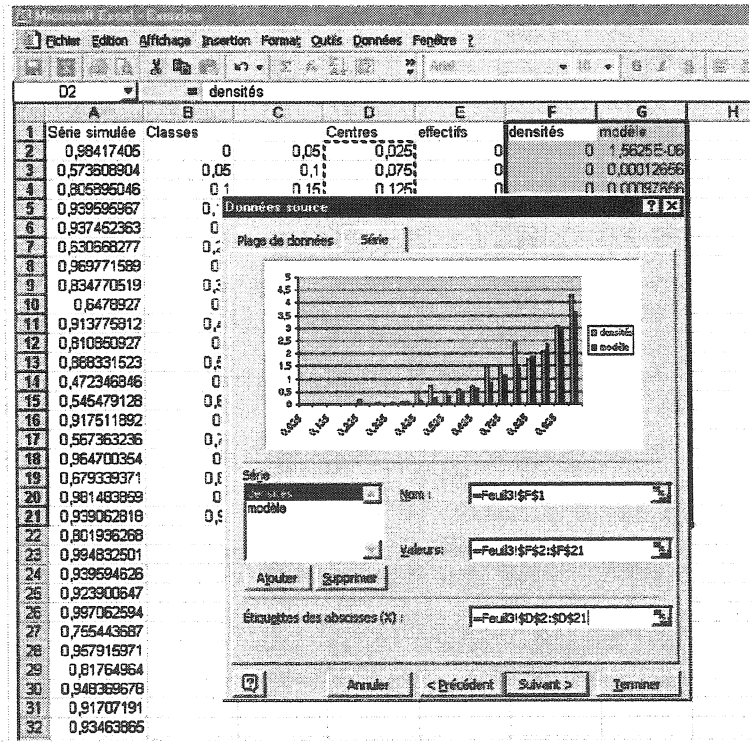
puis « histogramme »

puis « suivant »

puis cliquer sur l'onglet « série ».

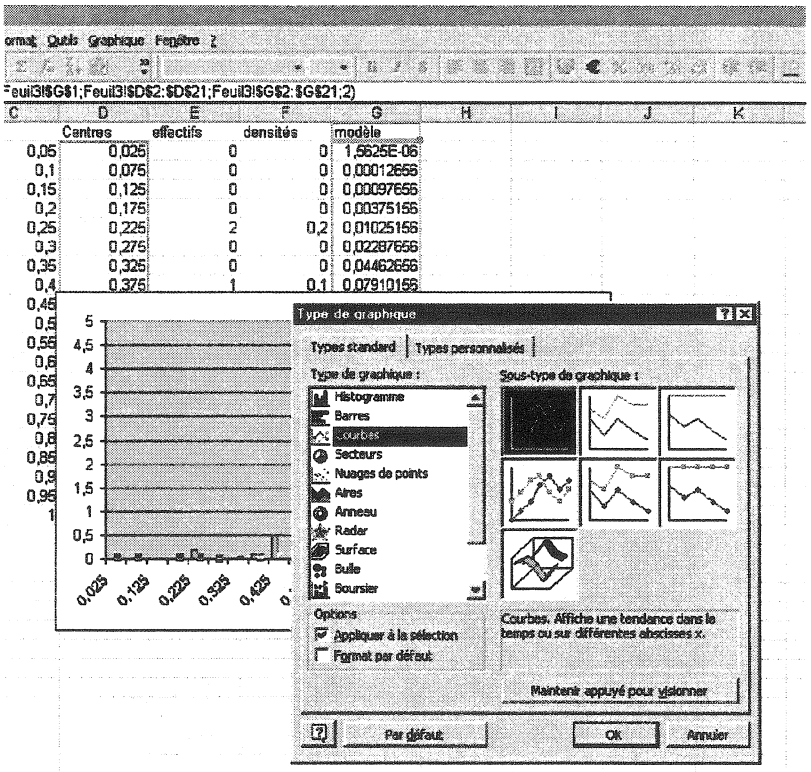
Dans la fenêtre « Etiquettes des abscisses » cliquer,

sélectionner la plage D2:D21 des centres des classes, puis cliquer sur « Terminer »

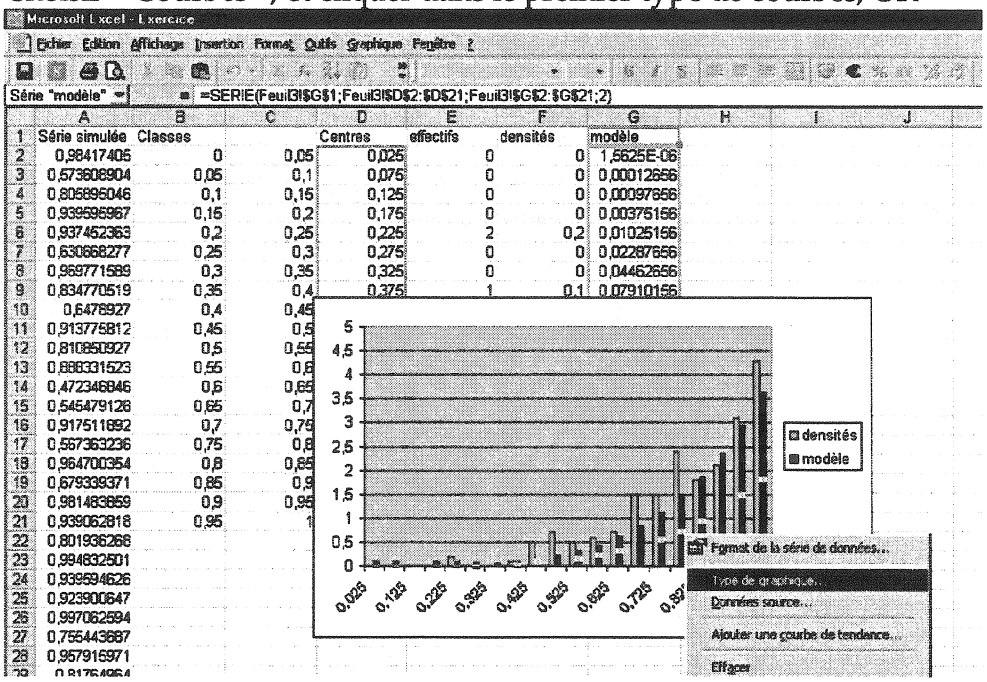


	A	B	C	D	E	F	G	H
1	Série simulée	Classes		Centres	effectifs	densités	modèle	
2	0,98417405	0	0,05	0,025	0	0	1,5625E-06	
3	0,573608904	0,05	0,1	0,075	0	0	0,00012656	
4	0,805895046	0,1	0,15	0,125	0	0	0,00097656	
5	0,939595967	0,15						
6	0,937452363	0						
7	0,630668277	0,2						
8	0,969771589	0						
9	0,834770519	0,3						
10	0,6478927	0						
11	0,913775912	0,4						
12	0,810860927	0						
13	0,868331523	0,5						
14	0,472346846	0						
15	0,545479128	0,6						
16	0,917511892	0						
17	0,567363236	0,7						
18	0,964700354	0						
19	0,679339371	0,8						
20	0,981483959	0						
21	0,939062818	0,9						
22	0,801936268							
23	0,994832501							
24	0,939594526							
25	0,923900647							
26	0,997062594							
27	0,755443587							
28	0,957915971							
29	0,81764964							
30	0,948369678							
31	0,91707191							
32	0,93463865							

Le graphique s'insère dans la feuille de calcul.
Cliquer avec le bouton droit sur l'histogramme du « modèle ».
Une fenêtre avec un menu où « type de graphique » apparaît. Valider.



Choisir « Courbes », et cliquer dans le premier type de courbes, OK



C'est fini. Faire F9 pour animer le tout.

Bibliographie

Antibi A., Bara R., Morin J., et al. : *Manuel de Terminale S, programme 2002*, collection Transmath, éditeur Nathan, ISBN 209-172420-3

Baillargeon G. : *Techniques statistiques*, Les Editions SMG, 1984

Ouvrage d'accès très simple avec beaucoup d'exemples développés et des explications très détaillées et illustrées. Convient à une première approche de la statistique sans formalisme mathématique exagéré.

Bonneval L.-M. : *Test d'équirépartition : qui a dit khi-deux ?*, Bulletin de l'APMEP, N°441, pp. 512-521

Commission Inter-IREM « Statistique & Probabilités » : *Autour de la modélisation en probabilités*, ouvrage coordonné par M. Henry, P.U.F.C., Besançon, 2001

On trouvera dans ce livre un ensemble d'articles sur les concepts fondateurs du calcul des probabilités, leurs origines historiques et leur prise en compte dans l'enseignement secondaire. Des exemples typiques de modèles sont proposés pour des activités en classe.

Droesbeke J.J. & Tassi P. : *Histoire de la statistique*, coll. « Que sais-je ? », PUF, Paris, 1990

Mison G., Gauthier R-L., et al. : *Manuel de Terminale S, programme 2002*, collection Indice, éditeur Bordas, 2002, ISBN 204-729597-1

Vessereau A. : *La statistique*, coll. « Que sais-je ? », PUF, Paris, 1992

Source des reproductions

Page 30 : tous droits réservés

Manuel de Terminale S programme 2002, collection Indice, éditeur Bordas, auteurs : G. Mison, R-L. Gauthier, et al. © Bordas/VUEF, 2002, ISBN 204-729597-1, chapitre 8, page 236.

Page 38 : autorisation de l'éditeur

Manuel de Terminale S programme 2002, collection Transmath, éditeur Nathan, auteurs : A. Antibi, R. Bara, J. Morin, et al. ©Nathan/VUEF 2002, ISBN 209-172420-3, chapitre 11, page 305.

Presses universitaires de Franche-Comté
P.U.F.C. - Université de Franche-Comté
2, place Saint-Jacques - 25030 Besançon Cedex

Imprimerie : **Burs Besançon**
9, rue Lecourbe - 25000 Besançon

Dépôt légal 3e trimestre 2005

Auteur Groupe *Probabilités & statistique*

Titre Lois continues, test d'adéquation. Une approche pour non spécialiste.

Langage Français.

Caractéristiques de l'édition

Édition	Première édition
Éditeur	Presses universitaires de Franche-Comté
Diffuseur	IREM de Franche-Comté
Année	2005
Format	21 x 29,7 cm (A4)
	53 pages
	Support papier
Dépôt légal	3e trimestre 2005
ISBN	2-84867-101-7

Public Professeurs de lycées et lycées professionnels, formateurs et étudiants IUFM, public motivé.

Résumé Cette brochure IREM est issue d'une conférence en direction de professeurs de lycée prononcée par les membres du groupe *Probabilités & statistique* de l'IREM à l'initiative des inspecteurs pédagogiques régionaux de mathématiques de l'académie de Besançon.

Enseignants de lycée et d'université, les auteurs proposent à un public non formé à la statistique inférentielle une réflexion sur la démarche statistique préconisée par les derniers programmes de lycée, notamment dans les activités pédagogiques portant sur les tests statistiques d'adéquation d'un modèle d'équirépartition à un phénomène réel.

Les notions de modélisation d'un phénomène aléatoire, de loi continue et de test d'adéquation sont présentées dans une approche qui privilégie l'explication des concepts en évitant de les noyer dans la rigueur du formalisme mathématique et en prenant appui, pour illustrer le discours, sur la simulation informatique des phénomènes aléatoires

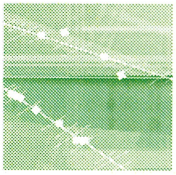
Mots clés Statistique inférentielle, test d'adéquation, théorème-limite central, lois continues, professeurs de lycée, lycées professionnels, formateurs IUFM, simulation informatique, modélisation, test statistique.

UNIVERSITÉ DE FRANCHE-COMTÉ

Institut de Recherche sur l'Enseignement des Mathématiques de l'Université de Franche-Comté

Département de Mathématiques - UFR Sciences et Techniques
16 route de Gray - 25030 BESANÇON Cedex - France

Tél. : 03 81 66 62 25 - Fax : 03 81 66 62 34
Courriel : iremfc@math.univ-fcomte.fr – http : //www-irem.univ-fcomte.fr/



Presses universitaires de Franche-Comté