LA SIMULATION EN STATISTIQUE

Philippe DUTARTE
CII Lycées Technologiques

Enseignant la statistique depuis déjà quelques années en sections de techniciens supérieurs, il me semblait que le travail des élèves, dans ce domaine, se limitait trop souvent à appliquer des formules et effectuer quelques calculs, dans des situations vite identifiées, correspondant à celles des sujets d'examen.

La «déraisonnable» efficacité des méthodes statistiques, malgré le contexte aléatoire, leur réelle capacité à prévoir, dans certaines limites, le dosage entre prise de décision et risques, tous ces aspects, essentiels, étaient occultés. C'est ainsi que le recours à la simulation est apparu comme un moyen de replacer concrètement l'aléatoire au cœur de la problématique statistique et de montrer l'efficacité réelle des formules exposées en cours. Avec les nouveaux programmes des lycées, cette expérience, acquise en sections de techniciens supérieurs, peut sans doute être utile aux collègues dès la classe de seconde. Après un exposé de ce qu'est la simulation en statistique et comment elle est théoriquement possible, nous en verrons quelques utilisations pédagogiques (les programmes donnés sont ceux utilisés par les élèves, en seconde ou BTS).

1. Qu'est-ce que la simulation ?

Le développement des ***moyens de calcul informatiques*** a modifié bien des pratiques scientifiques. La statistique ne fait pas exception : l'ordinateur ou la calculatrice permettent assez simplement d'expérimenter des situations aléatoires, par simulation à partir d'une loi donnée, en les répétant un grand nombre de fois. On verra que la loi des grands nombres en permet une justification. Il pourra s'agir, soit d'étudier les conséquences d'un modèle, en le faisant «tourner», soit de conjec-

turer certains résultats, là où le calcul est trop compliqué, ou impossible.

Selon *Emile Borel*[1], *« le but principal du calcul des probabilités* [...] *est de calculer les probabilités de phénomènes complexes en fonction des probabilités, supposées connues, de phénomènes plus simples ».* ***En statistique, c'est à partir d'observations que l'on évalue des probabilités***. La simulation permet de fabriquer de telles observations, à partir de probabilités simples, supposées connues, et ainsi de se faire une idée de la qualité des procédures employées.

a — Des exemples

— Dans le ***jeu de pile ou face***, intéressons nous, par exemple, aux séries de lancers consécutifs égaux. La modélisation (admise) consiste à dire qu'à chaque lancer, on a «une chance sur deux» d'avoir pile ou d'avoir face (cette probabilité de 1/2 est admise). Dans le cadre de ce modèle, la simulation permet de conjecturer des résultats non triviaux et non intuitifs, comme par exemple d'évaluer la probabilité d'observer au moins six lancers consécutifs égaux sur 200 lancers.

Le programme sur calculatrice ci-dessous (utilisé en seconde) calcule, pour 200 lancers de pile ou face simulés (on verra comment plus loin), la longueur maximale des lancers consécutifs égaux.

Cinq simulations ont donné, par exemple, comme longueurs maximales de lancers consécutifs égaux sur 200 lancers : 8 ; 7 ; 6 ; 11 ; 7. Sur 200 lancers consécutifs, on a, à chaque fois, observé au moins 6 lancers consécutifs égaux (et parfois beaucoup plus). Ce qui est assez ***contraire à l'intuition***. Bien sûr, cinq simulations, ce n'est pas assez pour faire des statistiques. On verra plus loin comment utiliser le théorème limite central pour déterminer combien effectuer de simulations pour évaluer correctement la probabilité d'avoir au moins 6 lancers consécutifs égaux sur 200 lancers de pile ou face.

CASIO Graph 25 → 100	**T.I. 80 - 82 - 83**	**T.I. 89 - 92**
Seq(1,I,1,2,1) → List 1↵	:1 → A	:1 → a
Int(Ran# + 0.5) → R↵	:1 → M	:1 → m
For 1 → I To 200↵	:int(rand + 0.5) → R	:int(rand( ) + 0.5) → r
Int(Ran# + 0.5) → S↵	:For(I,1,200)	:For i,1,200
If S = R↵	:int(rand + 0.5) → S	:int(rand( ) + 0.5) → s
Then List 1[1] + 1 → List 1[1]↵	:If S = R	:If s = r
	:Then	:Then
S → R↵	:A + 1 → A	:a + 1 → a
Max(List 1) → List 1[2]↵	:S → R	:s → r
Else 1 → List 1[1]↵	:max(A,M) → M	:max(a,m) → m
IfEnd↵	:Else	:Else
S → R↵	:1 → A	:1 → a
Next↵	:End	:EndIf
List 1[2]	:S → R	:s → r
	:End	:EndFor
	:M	:Disp m

1 Emile Borel – «Les probabilités et la vie» – «Que sais-je» P.U.F. 1947.

Dans cette situation, le calcul est en fait possible (mais pas immédiat). La probabilité qu'une suite de 200 lancers de pile ou face contienne au moins une série de 6 lancers consécutifs égaux est environ 0,965 (voir l'encadré ci-contre).

— Prenons un second exemple dans un cas pratique.

On cherche à étudier ***le temps de rotation moyen d'un bus***, selon la configuration d'une ligne urbaine. On a modélisé la situation. Le temps d'arrêt à chaque station est aléatoire, en fonction du nombre de descentes et de montées (selon un processus de Poisson). De même, le temps entre deux stations dépend des aléas de la circulation.

On peut alors, dans un programme, simuler les différentes variables aléatoires intervenant ici (dont le choix aura été déterminé selon un historique statistique), puis les combiner de façon à simuler une rotation du bus. Ayant simulé ce modèle, on peut alors le faire «tourner» un grand nombre de fois et évaluer, entre autres, le temps moyen de rotation d'un bus (mais aussi observer les phénomènes d'attente).

b — Une définition

Comme on vient de le voir, simuler une expérience aléatoire consiste à produire «virtuellement» des résultats analogues à ceux que l'on aurait obtenus en réalisant «physiquement» l'expérience aléatoire.

Une définition plus précise de la simulation est donnée par *Yadolah Dodge*[2] :

*« La simulation est la méthode statistique permettant la reconstitution **fictive** de l'évolution d'un phénomène. C'est une **expéri-***

La probabilité qu'une suite de 200 lancers de pile ou face contienne au moins une série de six lancers consécutifs égaux est environ 0,965

Notons u_n le nombre de suites de n lancers de pile ou face (x_i) , avec i de 1 à n, ne contenant *aucune* séquence de 6 consécutifs égaux.

Une telle suite peut être de 5 types différents , en considérant les 6 derniers termes, dénombrés ainsi :

- $x_{n-1} \neq x_n$: il y a u_{n-1} telles suites.
- $x_{n-1} = x_n$ et $x_{n-2} \neq x_{n-1}$: il y a u_{n-2} telles suites.
- $x_{n-1} = x_n$; $x_{n-2} = x_{n-1}$ et $x_{n-3} \neq x_{n-2}$: il y a u_{n-3} telles suites.
- $x_{n-1} = x_n$; $x_{n-2} = x_{n-1}$; $x_{n-3} = x_{n-2}$ et $x_{n-4} \neq x_{n-3}$: il y a u_{n-4} telles suites.
- $x_{n-1} = x_n$; $x_{n-2} = x_{n-1}$; $x_{n-3} = x_{n-2}$ et $x_{n-4} = x_{n-3}$ et $x_{n-5} \neq x_{n-4}$: il y a u_{n-5} telles suites.

La probabilité cherchée est donc $\frac{u_{200}}{2^{200}}$ où la suite (u_n) est définie par :

$u_1 = 2$; $u_2 = 4$; $u_3 = 8$; $u_4 = 16$; $u_5 = 32$ puis

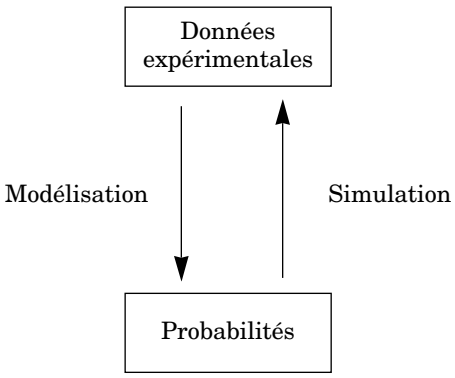
$u_n = u_{n-1} + u_{n-2} + u_{n-3} + u_{n-4} + u_{n-5}$ pour $n \geq 6$.

Ceci permet le calcul de u_{200} de proche en proche. On a ainsi $\frac{u_{200}}{2^{200}} \approx 0{,}965313$.

2 Statistique. Dictionnaire encyclopédique. Dunod 1993

mentation *qui* ***suppose la constitution d'un modèle théorique*** *présentant une similitude de propriétés ou de relations avec le phénomène faisant l'objet de l'étude. »*

Ceci peut être schématisé ainsi :



En produisant des données, sous un certain modèle, la simulation permettra d'examiner les conséquences, souvent non évidentes, de ce modèle, et, éventuellement, son adéquation aux données réelles. De façon générale, la simulation permet d'obtenir (ou de conjecturer) des résultats difficiles, ou impossibles, à calculer (c'est dans ce cadre qu'elle est utile aux statisticiens)[3].

Pour imiter le hasard, les simulations sont basées sur le calcul de nombres pseudo-aléatoires qui, non seulement sont imprévisibles, mais encore ont le «goût» (statistiquement) du hasard.

3 Il existe également des simulations non aléatoires (dans l'étude de systèmes dynamiques, par exemple).

2. Comment peut-on simuler le hasard ?

a — Hasard ou pseudo-hasard ?

Les premières tables de nombres au hasard ont été construites à partir des ***numéros gagnants de la loterie***. Cette pratique a conduit à désigner par *«méthode de Monte-Carlo»* les procédés de calcul d'aire utilisant ces nombres au hasard. Ainsi, alors que le statisticien *Karl Pearson* (1857-1936) eut beaucoup recours à des lancements de pièces ou de dés, embauchant pour ce faire amis et élèves, son fils *Egon Pearson* (1895-1980), à l'origine de la théorie des tests, utilisa ce qu'on appela plus tard la simulation, grâce à des tables de nombres au hasard produites dans les années 1925. En 1955, la *Rand Corporation* édita une table *«A Million Random Digits»* obtenue à partir de ***bruits de fond électroniques*** (fluctuations du débit de tubes électroniques). Il s'agit alors d'un générateur aléatoire physique.

Avec l'apparition des ordinateurs, on chercha à générer des nombres aléatoires, à l'aide d'***algorithmes***. Il ne s'agit plus de hasard physique mais d'un hasard calculé. On comprend bien l'antagonisme entre les deux termes. On ne peut pas calculer des nombres au hasard, puisqu'il sont alors le résultat d'un algorithme déterministe.

Cela nous conduit à nous poser la question : «quand peut-on dire qu'une suite de nombres est une suite au hasard ?» On peut se limiter à une suite de 0 et de 1, et la question devient : «quand peut-on considérer qu'une suite de 0 et de 1 est une suite au hasard ?» C'est à dire résultant d'un tirage à pile ou face, ou encore, de façon plus mathématique, comme étant les résultats successifs

d'une suite de variables aléatoires X_i indépendantes et valant 0 ou 1 avec une probabilité 0,5.

Cette question est mathématiquement très difficile. Une réponse théorique à été apportée en 1966 par *Martin-Löf* : *« Une suite de chiffres est aléatoire quand le plus petit algorithme nécessaire pour l'introduire dans l'ordinateur contient à peu près le même nombre de bits que la suite »*. Cette définition, exclut donc toute possibilité d'une règle effective.

Un objectif plus raisonnable est de trouver un algorithme produisant une suite de nombres, telle qu'un statisticien en l'analysant, ne soit pas capable de détecter si elle a été produite par un procédé mathématique ou un réel phénomène aléatoire physique : qu'il lui soit impossible, par exemple, sur une suite assez grande de 0 et de 1 (disons 200) de savoir s'ils ont été générés par un ordinateur, ou en lançant une pièce de monnaie bien équilibrée. Une telle suite est ***pseudo-aléatoire***. Ces suites, construites sur des procédés récurrents, sont nécessairement périodiques, puisque l'on travaille avec un nombre fini de décimales. On cherche donc à ce que la période soit très grande et il faut être sûr de son générateur lorsque l'on a besoin d'une très grande quantité de nombres au hasard.

Pour simuler une variable aléatoire de loi donnée, le principe consiste à «déformer» un générateur de nombres pseudo-aléatoires correspondant à une distribution uniforme sur l'intervalle [0, 1].

b - Simuler une distribution uniforme sur [0 , 1]

La plupart des générateurs de nombres (pseudo) aléatoires, simulent le tirage au hasard d'un nombre réel (ou plutôt décimal) entre 0 et 1. De façon plus précise, on simule les réalisations d'une suite de variables aléatoires indépendantes X_i de même loi U ([0 , 1]).

Les procédés les plus courants consistent en des suites récurrentes, de grande période, et dont le comportement chaotique ***satisfait à divers tests statistiques***, permettant de valider l'hypothèse qu'il s'agit de réalisations de X (un premier test peut se construire sur l'observation des fréquences d'apparition des différents chiffres).

Un premier exemple de générateur de nombres aléatoires dans [0 , 1]

Si les constructeurs de calculatrices actuels ne donnent pas de renseignement quant à leur générateur de nombres aléatoires, ce n'était pas le cas dans les années soixante dix, où le générateur suivant correspond à un ancien modèle de calculatrice *Hewlett-Packard*.

En mode habituel de calcul, effectuer :

— Sur CASIO,

0.5 → X EXE puis
Frac (9821X + 0.211327) → X EXE
et appuyer plusieurs fois sur EXE.

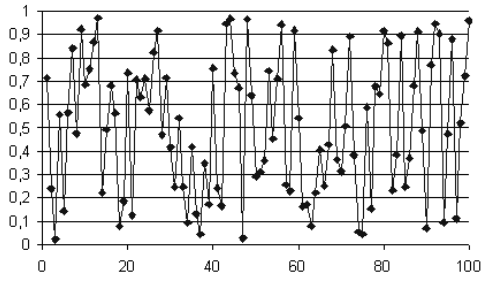
— Sur T.I. ,

0.5 → X ENTER puis
fPart(9821X + 0.211327) → X ENTER
et appuyer plusieurs fois sur ENTER.

Il s'agit donc d'une suite (x_n) définie par récurrence par $x_0 = 0{,}5$ et x_{n+1} = partie fractionnaire de $9821x_n + 0{,}211327$.

En voici les dix premiers termes :

0,71327
0,237940002
0,022028641
0,55655523
0,142180072
0,563753843
0,839757939
0,475991181
0,92266285
0,685117385



Le graphique des 100 premiers termes, calculés sur Excel, montre son caractère chaotique. (Ces suites, chaotiques, diffèrent souvent selon le type de calculateur). Reste à vérifier qu'elle possède les propriétés statistiques de la loi uniforme sur [0, 1]…

Un second générateur obtenu selon la méthode de *Lucas-Lehmer*

De nombreux générateurs sont obtenus à partir de propriétés arithmétiques, en particulier suite aux travaux de *Lehmer*, dans les années soixante dix. Certaines suites congruentes possèdent, en effet, des propriétés structurelles démontrées, comme la grande longueur de leur période (pour les propriétés arithmétiques, voir l'encadré ci-contre), qui en font, *a priori*, de bons candidats pour servir de générateur aléatoire. On leur fait subir ensuite toutes sortes de tests statistiques (on en verra des exemples plus loin) pour sélectionner le plus satisfaisant. Mais, cette fois, il n'y aura pas de certitude. La méthode statistique ne démontre pas qu'un générateur donné est toujours satisfaisant pour une simulation donnée.

On choisit des entiers a et m premiers entre eux (m grand, souvent un nombre premier), puis on construit la suite (r_n) d'entiers positifs de [0 , m - 1], définie à partir d'une valeur r_0 , non nulle et première avec m, et de la relation de récurrence :

$$r_{n+1} = a\ r_n \bmod(m)\ ,$$

c'est à dire que r_{n+1} est le reste de la division euclidienne de $a\ r_n$ par m.

La suite (x_n) définie par $x_n = \frac{r_n}{m}$ fournit, pour certains choix de a et m, un générateur de nombres aléatoires dans [0 , 1].

Le choix de a et m dépendent de la configuration de l'ordinateur. Pour un modèle *IBM* des années 80, on avait choisi :

$a = 7^5$; $m = 2^{31} - 1$; $r_{n+1} = a\ r_n \bmod(m)$

puis $x_n = r_n / m$.

On donne page suivante les premières valeurs de (x_n), obtenue sur Excel, avec :

$$a = 7^5$$

$$m = 2^{31} - 1$$

et en débutant avec la valeur $r_0 = 5$.

Propriétés arithmétiques de la suite $r_{n+1} = ar_n \bmod(m)$
$0 < r_0 < m$, r_0 et a premiers avec m

• Tout d'abord, pour tout n dans **N** on a r_n non nul et premier avec m.

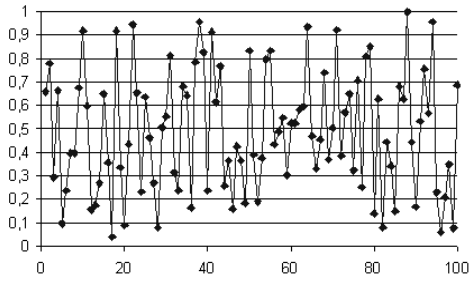
En effet, $r_n = 0$ impliquerait l'existence d'un entier k tel que $ar_{n-1} = km$ mais, puisque m est premier avec a, le lemme de *Gauss* donnerait que m divise r_{n-1}, c'est- à-dire : $r_{n-1} = 0$ (r_{n-1} est un reste modulo m). Par récurrence, on remonterait à $r_0 = 0$, ce qui est exclu.
De même, s'il existait d diviseur commun à r_n et m, alors d diviserait ar_{n-1} et, puisque m est premier avec a, le lemme de *Gauss* donnerait que d divise r_{n-1}. On remonterait à d divise r_0 qui est exclu.

• Dans ces conditions, la suite (r_n) a pour période l'ordre multiplicatif de a modulo m, c'est-à-dire le plus petit entier k tel que $a^k = 1 \bmod m$. En effet :
$r_{n+t} = r_n \Leftrightarrow a^t r_n = r_n \bmod(m) \Leftrightarrow (a^t - 1)r_n = km$ avec k entier, c'est à dire, puisque r_n est premier avec m, $a^t = 1 \bmod m$.

• Lorsque m est premier, le petit théorème de *Fermat* affirme que si m est premier et ne divise pas a, alors $a^{m-1} = 1 \bmod(m)$, donc, pour a premier avec m, l'ordre multiplicatif de a divise $m - 1$.
Un théorème de *Legendre* assure alors, pour m premier, l'existence de nombres d'ordre multiplicatif maximum $m - 1$ modulo m.
Par exemple, modulo 5, le nombre 2 est d'ordre multiplicatif maximum 4 car $2^2 = 4 \bmod(5)$ puis $2^3 = 3 \bmod(5)$ et $2^4 = 1 \bmod(5)$.
Il n'existe malheureusement pas d'algorithme permettant de trouver ces nombres d'ordre maximum. Le théorème de *Legendre* précise cependant qu'entre 1 et $m - 1$, il en existe $\varphi(m - 1)$ correspondant au nombre d'entiers premiers avec $m - 1$ dans $\{1 , 2 , \ldots , m - 2\}$. Les nombres d'ordre maximum ne sont donc pas rares, et, avec l'aide de l'ordinateur, on peut trouver un tel a, qui nous assurera une suite dont la plus petite période est $m - 1$, m étant un très grand nombre premier.

• Lorsque m n'est pas premier, un théorème d'*Euler* donne que, si a et m sont premiers entre eux, alors $a^{\varphi(m)} = 1 \bmod(m)$ où φ est l'indicateur d'*Euler* correspondant au nombre d'entiers premiers avec m dans $\{1 , 2 , \ldots , m - 1\}$. Ainsi, l'ordre multiplicatif de a est alors un diviseur de $\varphi(m)$. Mais on n'est pas assuré qu'il existe un tel nombre d'ordre multiplicatif maximum $\varphi(m)$, modulo m.
Par exemple, $\varphi(21) = 12$ et, modulo 21, l'ordre multiplicatif de 5, par exemple, est 6 ($5^6 = 1 \bmod(21)$) qui est un diviseur de 12 et l'ordre multiplicatif maximum, modulo 21.

Les 100 premiers termes du second générateur sont :



(n=2) 0,657688941
0,778026611
0,29325066
0,663836187
0,094795932
0,235223081
0,394323584
0,396482029
0,67346448
0,917510387
0,59708186
0,154826731
0,172860553
0,267308175
0,648500967
0,35574692
0,038490931
0,917078254
0,334211188
0,087429872

Remarque : Le nombre $2^{31} - 1$ est un nombre de *Mersenne* premier.

Les nombres de *Mersenne* ne sont pas tous premiers (la calculatrice TI 89 donne, par exemple, en faisant « factor(2^30 –1) » :
$2^{30} - 1 = 3^2 \times 7 \times 11 \times 31 \times 151 \times 331$) ,
mais leur intérêt réside dans le fait qu'il existe un test (découvert par *Lucas* en 1878) permettant de savoir s'ils sont premiers, et que leur manipulation est très commode dans le système binaire des ordinateurs, puisque $2^{31} - 1$ s'écrit avec 31 chiffres 1 consécutifs.

Edouard Lucas (1842-1891), professeur au lycée St-Louis, puis Charlemagne, à Paris, est célèbre pour ses résultats en théorie des nombres, et ses «Récréations mathématiques» (il est l'inventeur du problème des «Tours de Hanoï»).

Tester un générateur de nombres aléatoires

On construit, à partir de la suite (x_n) fournie par le générateur, une suite de chiffres aléatoires parmi 0, 1, 2, ..., 9, en faisant Ent(10 x_n) où Ent désigne la partie entière, instruction qui ne retient que la première décimale.

Le premier test à effectuer est celui des ***fréquences d'apparition des différents chiffres***. Chaque chiffre doit avoir une probabilité 1/10 de sortir, et sur n chiffres consécutifs fournis par le générateur, la fréquence observée d'un chiffre doit se répartir, suivant les échantillons, approximativement selon la loi normale $N\left(\frac{1}{10}, \sqrt{\frac{\frac{1}{10} \times \frac{9}{10}}{n}}\right)$ (d'après le théorème limite central).

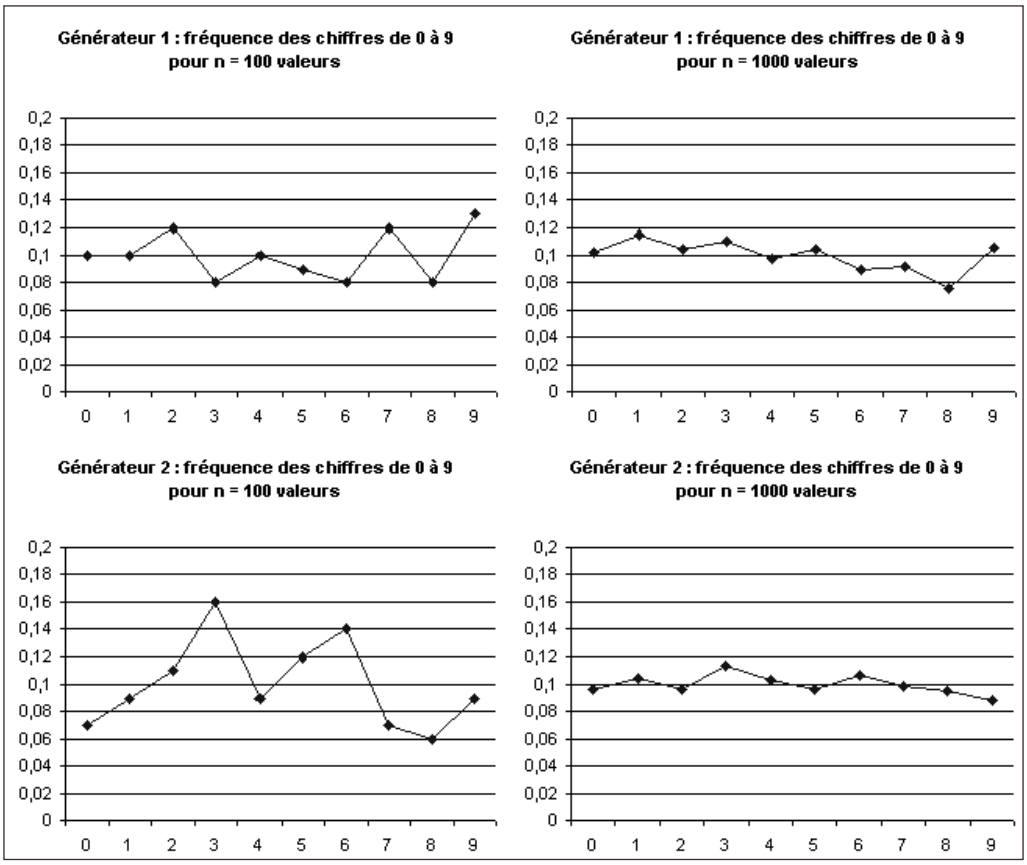
La dispersion «normale» des fréquences observées, sur des échantillons de taille n, doit donc se faire, si le générateur est bon, avec un écart type $\sigma = \frac{0{,}3}{\sqrt{n}}$, c'est-à-dire 0,03 si n = 100 et 0,0095 si n = 1000 (il est à noter qu'une dis-

persion trop faible est aussi suspecte que le contraire !). D'après les propriétés de la loi normale, on devrait donc avoir 95 chances sur 100 d'observer les fréquences d'apparition d'un chiffre à moins de 2σ de 1/10, alors qu'un écart de 3σ est peu probable.

Ainsi, sur des échantillons de taille $n = 100$, on a 95% de chances d'observer la fréquence de sortie d'un chiffre dans la bande [0,04 ; 0,16], alors que pour des échantillons de taille $n = 1000$, les fréquences doivent, à 95%, se situer dans la bande [0,08 ; 0,12]. On peut donc construire un premier test statistique, en écartant un générateur fournissant trop fréquemment une fréquence en dehors de ces intervalles.

En conservant la première décimale des résultats des deux générateurs précédents, on obtient les graphiques ci-dessous.

On observe que le premier générateur a fourni, pour $n = 1000$, un nombre exceptionnellement bas de 8, ce qui, sans le discréditer totalement, le rend un peu suspect.



Un second contrôle peut être celui du ***poker***, où l'on regroupe consécutivement les chiffres par quatre et où l'on compare les fréquences des différentes configurations possibles à leur probabilité :

Configuration :	Proba :
Chiffres différents 6390	0,504
Une paire 9390	0,432
Deux paires 9393	0,027
3 chiffres idem 9399	0,036
4 chiffres idem 9999	0,001

Pour le générateur ALEA() d'Excel, on observe par exemple (sur deux expériences), pour 1000 groupes de quatre chiffres, les effectifs x_i suivants, à comparer aux valeurs théoriques t_i :

Configuration	expérience 1	expérience 2	t_i
Chiffres différents	$x_1 = 541$	497	504
Une paire	$x_2 = 409$	439	432
Deux paires	$x_3 = 18$	31	27
3 chiffres idem	$x_4 = 32$	32	36
4 chiffres idem	$x_5 = 0$	1	1

Une certaine fluctuation des observations est attendue, mais dans quelles limites ?

On peut mesurer l'adéquation des observations x_i aux valeurs théoriques correspondantes t_i en introduisant l'écart quadratique réduit :

$$\chi^2_{obs} = \sum_{i=1}^{5} \frac{(x_i - t_i)^2}{t_i} .$$

Pour étudier la variabilité de ce critère, on introduit les variables aléatoires X_i qui, à chaque échantillon de 1000 groupes de quatre chiffres consécutifs, associent le nombre de configurations de type i, ainsi que la variable aléatoire $T = \sum_{i=1}^{5} \frac{(X_i - t_i)^2}{t_i}$, avec $\sum_{i=1}^{5} X_i = 1000$.

La loi de T suit approximativement une loi tabulée et connue sous le nom de loi du χ^2 à quatre degrés de liberté (en effet la relation ci-dessus fait que la valeur de X_5 est déterminée dès que les valeurs de X_1 , X_2 , X_3 et X_4 sont connues).

La table permet alors d'obtenir :

$$P(T \leq 9{,}48) \approx 0{,}95.$$

On pourra alors considérer comme suspect d'observer une valeur de χ^2_{obs} supérieure à 9,48.

Pour les échantillons obtenus précédemment avec le générateur aléatoire d'Excel, on a :

$$\chi^2_{obs\,1} \approx 8{,}39 \text{ et } \chi^2_{obs\,2} \approx 1{,}25.$$

Un générateur de nombres aléatoires existe sur toutes les calculatrices sous la forme de la touche ***random*** [4] : Ran# (CASIO) ou rand (T. I.).

4 Le mot random signifie «hasard» en anglais, il vient du vieux français randon, que l'on retrouve dans randonnée.

c — Simuler d'autres distributions

A partir de la distribution uniforme sur [0 , 1], obtenue par la fonction random, on a recours à différentes techniques pour simuler toute autre distribution.

A l'aide de la partie entière (notée ici int), si l'instruction rand simule la loi U ([0 , 1]), l'instruction « int(10rand) » simule le tirage d'un chiffre entre 0 et 9, « 1 + int(6rand) » simule le lancer d'un dé et « int(rand + 0.5) » le lancer d'une pièce, codé par 0 ou 1.

Distribution de Bernoulli

Les résultats d'une variable aléatoire X suivant la loi de Bernoulli de paramètre p, notée B (1 , p) (c'est-à-dire telle que P($X = 1$) = p et P($X = 0$) = $1 - p$, avec $p \in$ [0 , 1]) sont simulés par l'instruction : « int(rand + p) ».

En effet, l'instruction rand + p correspond à une distribution uniforme sur [p ; 1 + p] et la partie entière d'un nombre choisi dans cet intervalle est 0 s'il appartient à [p ; 1[et 1 s'il appartient à [1 , 1 + p].

L'amplitude de l'intervalle conduisant au « succès » 1 est p tandis que celle de celui conduisant à « l'échec » 0 est $1 - p$.

Distribution binomiale B (n ; p)

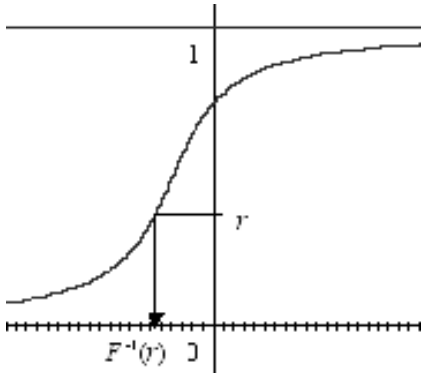
Cette distribution peut s'obtenir à l'aide de la somme de n variables de Bernoulli indépendantes de même paramètre p.

Il suffit donc de répéter et de sommer n fois l'instruction précédente.

Simulation théorique d'une distribution continue à partir de la loi U ([0;1])

Celle-ci repose sur le résultat suivant :

Si X est une variable aléatoire réelle, de fonction de répartition F continue, strictement croissante, alors la variable aléatoire Y = F(X) est uniformément distribuée sur [0 , 1].



Ainsi, si l'on tire n nombres au hasard r_i , parmi des nombres uniformément répartis sur [0, 1], un échantillon de la distribution de X sera donné par $F^{-1}(r_i)$. Cette méthode, où l'on retrouve l'idée d'une déformation de la loi uniforme, est dite «*de l'anamorphose*»

Distribution exponentielle E (λ)

Une variable aléatoire T correspondant au temps d'attente d'un succès dans un processus de Poisson suit une loi exponentielle de fonction de répartition définie sur [0 , + ∞ [par :

$$F(t) = P(T \leq t) = \int_0^t f(t)dt = \int_0^t \lambda e^{-\lambda t} dt$$

soit $F(t) = 1 - e^{-\lambda t}$.

Pour simuler les réalisations de T, il suffit, d'après la méthode de l'*anamorphose*, de calculer $-\frac{1}{\lambda}\ln(1 - r_i)$ où les valeurs r_i sont données par le *random*. Et comme les nombres $1 - r_i$ sont également uniformément distribués sur [0 ; 1], on peut dire que les réalisations d'une variable aléatoire suivant la loi exponentielle E (λ) sont simulées par l'instruction :

(- ln Ran#) ÷ λ .

Distribution de Poisson

Si T est une variable aléatoire de loi exponentielle E (1), la variable aléatoire X correspondant au nombre de réalisations de T durant l'intervalle [0 , λ] suit la loi de Poisson de paramètre λ.

On simule donc la distribution de la loi P (λ) en recherchant le plus grand entier n tel que $\sum_{i=1}^{n} -\ln r_i < \lambda \Leftrightarrow \prod_{i=1}^{n} r_i > e^{-\lambda}$ où les r_i sont donnés par le générateur de nombres aléatoires.

Distribution normale N (m ; σ)

• Simulation approchée :

Une réalisation d'une variable aléatoire suivant la loi normale N (m ; σ) peut être simulée par l'instruction suivante où l'on répète 12 fois l'instruction rand :

σ(rand + rand + rand + rand + rand
+ rand + rand + rand + rand + rand
+ rand + rand – 6) + m

Ce résultat repose sur le *théorème de la limite centrée*. La somme de n variables aléatoires X_i uniformes sur [0, 1] et indépendantes suit approximativement, pour n assez grand, une loi normale. Le choix de n = 12 tient au fait que la variance de la loi uniforme sur [0, 1] est 1/12, ce qui simplifie le calcul. On constate par ailleurs que n = 12 est «assez grand» dans cette situation.

• Simulation «exacte» :

Une simulation «exacte» de la loi N (m ; σ) est possible sous la forme de l'instruction suivante[5] (en mode Degrés) :

$m + \sigma \cos(360\ rand)\sqrt{}(-2\ln\ rand)$

L'idée repose sur un changement de variable des coordonnées polaires aux coordonnées cartésiennes. On montre que, si U et V sont deux variables aléatoires indépendantes de loi uniforme sur [0, 1], alors $\rho^2 = -2\ln U$ et $\theta = 2\pi V$ sont indépendantes, respectivement de loi exponentielle de paramètre 1/2 et de loi uniforme sur [0, 2p], puis on obtient que la variable aléatoire $X = \rho\cos\theta$ suit la loi normale N (0, 1).

3. Comment justifier la simulation

Est-ce que simuler permet toujours de bonnes conjectures, dans le cadre d'une situation aléatoire ? Combien de fois doit-on répéter les expériences ? Quelle incertitude a-t-on ? La réponse à ces questions, et la justification de la simulation, est fondée sur la loi des grands nombres, elle même précisée par le théorème limite central.

5 Certaines calculatrices possèdent une instruction randNorm.

La *loi des grands nombres* permet d'affirmer qu'en simulant une expérience aléatoire un grand nombre de fois, de façons indépendantes, les fréquences observées se rapprochent des probabilités à évaluer.

De façon plus précise, on a le théorème suivant.

Loi faible des grands nombres

Soit un événement A avec $P(A) = p$. Soit X_i , $1 \le i \le n$, des variables aléatoires de Bernoulli, indépendantes, de paramètre p (X_i vaut 1 si A est réalisé à l'expérience i et 0 sinon). On note $S_n = \sum_{i=1}^{i=n} X_i$ (qui suit la loi binomiale B (n , p)) et $F_n = \frac{1}{n} S_n$, la variable aléatoire correspondant à la fréquence d'observation de A sur les n expériences.

Alors, pour tout $t > 0$,

$$P\left(\left|F_n - p\right| > t\sqrt{\frac{p(1-p)}{n}}\right) \le \frac{1}{t^2} \quad .$$

Prenons l'exemple des 200 lancers de pile ou face, où l'on cherche à évaluer, par simulation, la probabilité de l'évènement A : « Sur 200 lancers, on a eu au moins une série de 6 lancers consécutifs égaux ». On a vu que $P(A) = p \approx 0{,}965$, mais bien sûr, lorsque l'on simule, on ignore cette valeur.

Déterminons, à l'aide de la loi des grands nombres, le nombre n de simulations de 200 lancers qu'il suffit d'effectuer, pour être pratiquement assuré que la fréquence f de l'évènement A sur les n simulations approchera p à 10^{-2} près. C'est de la statistique. On n'aura pas de certitude, mais un risque mesuré de se tromper, disons moins d'une chance sur 100.

On cherche donc, avec les notations du théorème, un nombre suffisant n de simulations tel que $P(\left|F_n - p\right| > \frac{1}{100}) \le \frac{1}{100}$. Comme la valeur de p, entre 0 et 1, est inconnue, on peut majorer $\sqrt{p(1-p)}$ par 1/2 .

On a alors : $P\left(\left|F_n - p\right| > \frac{t}{2\sqrt{n}}\right) \le \frac{1}{t^2}$. On prend $t = 10$, d'où $\frac{10}{2\sqrt{n}} = \frac{1}{100}$ et $n = 250000$.

En simulant 250 000 fois 200 lancers, on aura donc au moins 99% de chances d'obtenir une valeur de p à 10^{-2} près.

En fait, cette valeur de n est très exagérée, la majoration donnée par la loi des grands nombres, et résultant de l'inégalité de *Bienaymé-Tchebitchev*, étant très grossière. Le *théorème limite central* permet, lorsque cela est possible, d'utiliser la répartition de la loi normale.

Théorème limite central. *Soit X_i des variables aléatoires indépendantes, de même loi, de moyenne μ et d'écart type σ. Pour n suffisamment grand, la variable aléatoire* $\overline{X_n} = \frac{1}{n} S_n = \frac{1}{n} \sum_{i=1}^{i=n} X_i$ *suit approximativement la loi normale* $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Dans l'exemple qui nous concerne, les X_i sont des variables de *Bernoulli* de même para-

mètre p (d'espérance p et d'écart type $\sqrt{p(1-p)}$) alors, le théorème affirme que la variable aléatoire : $F_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{i=n} X_i$ (fréquence observée sur un échantillon de taille n) suit approximativement, pour n assez grand, la loi normale $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

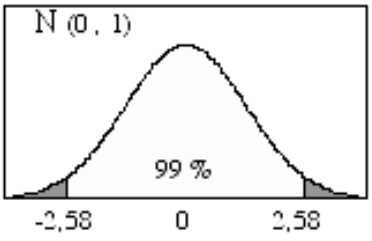
On cherche n tel que :

$$P(|F_n - p| \leq 0{,}01) \approx 0{,}99 .$$

On se ramène à la loi normale centrée réduite (tabulée) en posant $T = \frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}$.

On a alors :

$$P(|F_n - p| \leq 0{,}01) = P\left(|T| \leq \frac{0{,}01}{\sqrt{\frac{p(1-p)}{n}}}\right)$$



et l'on sait que, pour la loi normale centrée réduite, $P(|T| \leq 2{,}58) \approx 0{,}99$.

On en déduit $\sqrt{n} \approx 258\sqrt{p(1-p)} \leq \frac{258}{2}$

soit n de l'ordre de 16 600. Ceci est possible, avec la puissance de calcul de l'ordinateur.

4. Ce que peut apporter la simulation à l'enseignement de la statistique

a — La simulation permet de représenter la probabilité dans son aspect fréquentiste

Les probabilités définies, dans le cadre de l'équiprobabilité, par le rapport des cas favorables aux cas possibles, ne sont calculables que dans un cadre très limité, grosso modo, celui des jeux de hasard, dont les règles sont déterminées et la modélisation assez simple. Cette approche est inopérante en statistique. Ce n'est pas par un dénombrement exact que l'assureur évaluera la probabilité de naufrage d'un navire, ou le technicien la probabilité de panne d'une machine.

L'approche fréquentiste, fondée sur la loi des grands nombres de *Bernoulli*, consiste à lier la notion de probabilité à celle de fréquence observée après la répétition un grand nombre de fois d'une expérience[6]. La simulation permet, à peu de frais, d'y parvenir.

Emile Borel[7] affirme que *« les probabilités doivent être regardées comme analogues à la mesure des grandeurs physiques, c'est-à-dire qu'elles ne peuvent jamais être connues exactement, mais seulement avec une certaine approximation ».*

Cette démarche statistique, pour évaluer les probabilités «dans la vie», peut être assistée par la simulation. Quand cela est possible, les élèves effectuent quelques expériences réelles (avec une pièce de monnaie, des

6 La loi des grands nombres ne permet pas cependant de définir la probabilité, puisque faisant déjà appel à cette notion.

7 dans «Les probabilités et la vie».

dés...), pour se frotter à la réalité et accepter que le générateur de nombres aléatoires la prolonge.

La simulation permet également aux élèves de visualiser cette «convergence» vers la probabilité.

Le programme ci-contre montre, par exemple, comment, sur 500 lancers de pile ou face (en abscisses) évolue la proportion de «piles» (ici entre 0,4 et 0,6 en ordonnées). Sur l'écran apparaît la trajectoire des fréquences cumulées des piles sur quatre simulations de 500 lancers. On constate l'aspect chaotique des résultats du hasard sur les 100 premiers lancers, puis une convergence (en $\frac{1}{\sqrt{n}}$ selon le théorème limite central) vers la probabilité $p = 1/2$.

b — L'épreuve de l'expérience et l'attrait pour les nouvelles technologies

L'un des principaux intérêts pédagogiques de la simulation réside dans la ***nature expérimentale*** qu'elle donne à l'enseignement de la statistique et des probabilités, donnant davantage de sens aux concepts et motivant les élèves par l'aspect novateur de cette approche (utilisation des calculatrices programmables et de l'ordinateur). On voit comment la statistique fonctionne et cela rend les formules moins austères.

La simulation permet de mettre à l'épreuve de l'expérience certains résultats (parfois admis) du cours. Elle favorise le débat scientifique, obligeant les élèves à confronter leurs observations et à les analyser. Dans le cadre de travaux pratiques (un peu au sens de la phy-

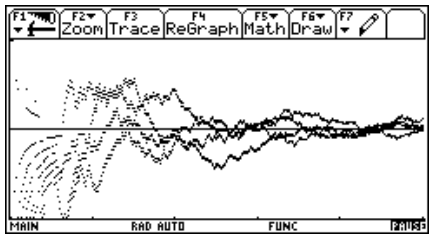
CASIO Graph 25 ou +

```
ViewWindow 0,500,100,0.4,
0.6,0.1↵
Graph Y= 0.5
For 1 → J To 4 ↵
0 → P ↵
For 1 → I To 500 ↵
Int(Ran#+0.5) →A↵
A ≠ 0 ⇒ Goto 1 ↵
P + 1 → P ↵
Lbl 1 ↵
Plot I , P ∏ I ↵
Next ↵
Next
```

T.I. 89 - 92

```
:FnOff
:ClrDraw
:PlotsOff
:0 → xmin
:500 → xmax
:100 → xscl
:0.4 → ymin
:0.6 → ymax
:0.1 → yscl
:DrawFunc 0.5
:For j , 1 , 4
:0 → p
:For i , 1 , 500
:int(rand( )+0.5) → a
:If a = 0
:p + 1 → p
:PtOn i , p/i
:EndFor
:EndFor
```

Exemple d'affichage sur TI 92 :

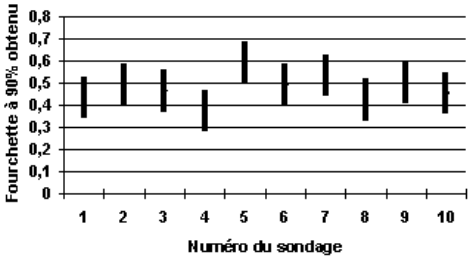


sique), on constate l'efficacité de la théorie, on donne une réalité aux formules. On peut également, en sens inverse, expérimenter d'abord, pour émettre des conjectures ou introduire une notion. Voyons deux exemples, au niveau seconde, puis B.T.S.

Expérimentation des « fourchettes » de sondage

La simulation permet d'effectuer de nombreux sondages et d'obtenir, à partir de chacun, une «fourchette» pour estimer un pourcentage p inconnu sur la population. L'observation de ces «fourchettes» permet de comprendre immédiatement leur dépendance par rapport à l'échantillon. Certaines questions posent problème, lorsqu'elles sont simplement présentées dans le cadre du cours.

Par exemple : «pour deux sondages différents, peut-on obtenir des fourchettes disjointes, ou très différentes» ou encore, «le pourcentage à estimer est-il nécessairement dans la fourchette». Ces questions, grâce à l'expérience, trouvent une réponse immédiate.



Sur la simulation ci-dessus, obtenue sur tableur, on observe par exemple que les fourchettes des sondages 4 et 5 sont disjointes (le seuil de confiance est ici 90%). On peut ensuite, profitant ici de l'interactivité du tableur, faire varier le seuil de confiance ou la taille de l'échantillon et en évaluer l'impact sur la qualité des «fourchettes» (taux d'erreur, précision de l'information...).

Expérimentation du théorème limite central

Dans les sections où la distribution de *Laplace-Gauss* est enseignée, l'élève peut se demander pourquoi, avec une expression analytique paradoxalement compliquée, elle est si répandue, au point d'être qualifiée de «normale» ?

La réponse à cette question est donnée par le théorème limite central : La somme de n variables aléatoires indépendantes de même loi suit approximativement, pour n assez grand, une loi normale. Mais ce résultat, alors qu'il est essentiel, est généralement admis. Il est donc instructif de l'expérimenter, par simulation [8].

En admettant que $n = 12$ est assez grand, l'instruction

$$\underbrace{\text{Ran\#+Ran\#+...+Ran\#}}_{12 \text{ fois}} - 6 + 4{,}5$$

doit approximativement simuler la loi normale N (4,5 ; 1).

Le programme de l'encadré ci-contre regroupe 100 résultats consécutifs de cette instruction en 9 classes : [0;1[; [1;2[... [8;9[, puis compare l'histogramme avec la courbe de densité de la loi normale N (4,5 ; 1).

On peut, dans ce programme, remplacer l'instruction Ran#, simulant la loi uniforme

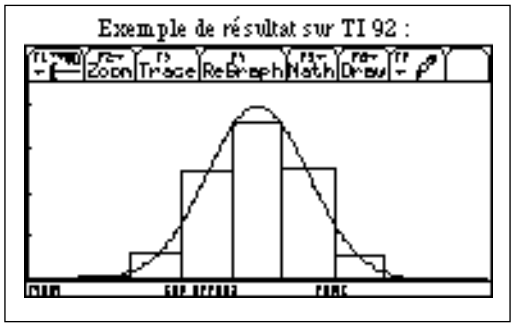
8 Une simulation physique de ce théorème est donnée par la planche de Galton.

CASIO Graph 40 ou +

Clrlist ↵
Seq(I,I,0,8,1) → List 1 ↵
Seq(0,J,1,9,1) → List 2 ↵
For 1 → K To 100 ↵
Ran#+Ran#+Ran#+Ran#+Ran#+Ran#+Ran#+Ran#+Ran#+Ran#+Ran#+Ran#+Ran#+Ran# – 1.5 → N ↵
1 + Int N → N ↵
N ≥ 1 And N ≤ 9 ⇒List 2[N]+1 → List 2[N]↵
Next ↵
List 2 p S-WindMan ↵
ViewWindow 0,9,1,0,45,10 ↵
0 → Hstart ↵
1 → Hpitch ↵
S-Gph1 DrawOn,Hist,List 1,List 2,Blue ↵
DrawStat ↵
Graph Y= (100÷√(2π))e(–.5(X–4.5)²)

TI 89 - 92

:DelVar L1 , L2
:seq (i,i,0,8,1) → L1
:seq (0,j,1,9,1) → L2
:For k , 1 , 100
:**rand( )+rand( )+rand( )+rand( )+rand( )+rand( )+rand( )+rand( )+rand( )+rand( )+rand( )+rand( ) – 1.5** → n
:int (n)+1 → n
:If n ≥ 1 and n ≤ 9
:L2[n] + 1 → L2[n]
:EndFor
:Disp L2
:Pause
:0 → xmin
:9 → xmax
:1 → xscl
:0 → ymin
:45 → ymax
:10 → yscl
:PlotsOn
:Newplot 1,4,L1,,L2,,,,1
:DrawFunc 100/(√(2π))×e^(–.5(x–4.5)^2)



sur [0, 1], par la simulation d'une autre loi. L'observation est analogue.

La conclusion à en tirer est que, lorsqu'un phénomène quelconque subit des variations dues à l'addition d'un grand nombre de perturbations aléatoires indépendantes (sans que l'une d'elle soit dominante), celles-ci suivront approximativement une loi normale.

c — Retrouver le hasard et l'ordre

« De façon apparemment paradoxale, l'accumulation d'événements au hasard aboutit à une répartition parfaitement prévisible des résultats possibles. Le hasard n'est capricieux qu'au coup par coup. »

«*Le Trésor*»
M. SERRES et *N. FAROUKI*,
article loi des grands nombres.

Un des principaux effets des activités de simulation est de réintroduire le «hasard» au cœur de notre enseignement de statistique et probabilités, lequel devient trop souvent du «dressage» aux techniques de résolution de problèmes (simple application de formules). Les

probabilités ne consistent-elles pas à mettre un peu d'ordre là où le néophyte ne voit que l'intervention du «hasard» ? Dès que l'on a décelé un certain ordre, on peut prévoir.

Donnons l'exemple de l'étude de pannes à taux d'avarie constant.

A partir d'un historique statistique de pannes (ici simulé), on recherche la nature de leur loi.

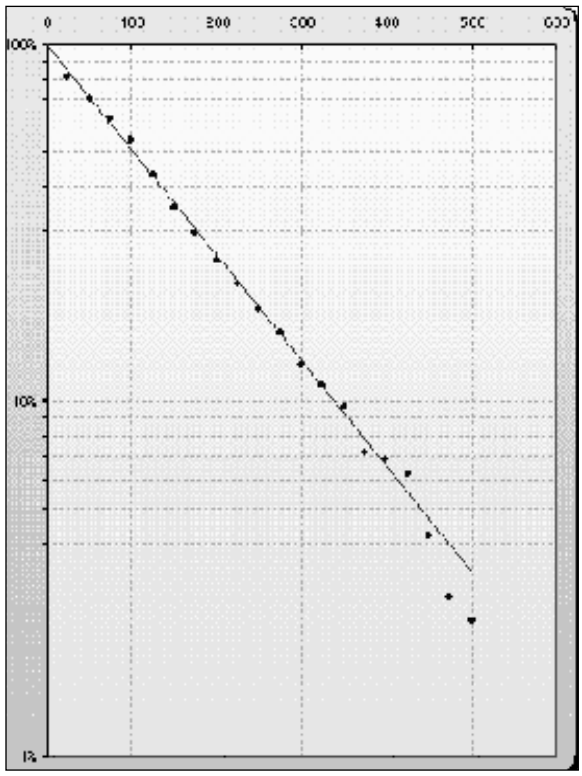
Une pièce est supposée avoir un taux d'avarie par heure de 0,007 , c'est à dire que, pour toute durée d'une heure choisie au hasard, la probabilité de panne est 0,007.

Un simple programme sur calculatrice, permet de simuler le temps de bon fonctionnement de la pièce (le temps n'est pas continu mais «saute» d'heure en heure, et, chaque heure l'instruction int(rand + 0.007) simule la probabilité de panne).

Lorsque les élèves expérimentent, avec ce programme, les temps de bon fonctionnement successifs, ils constatent la très forte présence du hasard. On passe ainsi, par exemple, de 192 heures sans panne, à 37 heures, puis à 407...

En regroupant un nombre important d'observations (200 par exemple), en classes, et en reportant les résultats sur papier semi-logarithmique, on verra surgir un ordre... celui du modèle de la loi exponentielle.

On pourra dès lors déterminer l'espérance des temps de bon fonctionnement de cette pièce.



d — Explorer, quand les outils mathématiques font défaut

De la même manière que le statisticien, l'économiste ou l'ingénieur exploite la puissance des ordinateurs pour étudier des situations aléatoires pour lesquelles le calcul est impossible ou trop compliqué, on peut, avec les élèves, explorer par simulation des situations riches, pour lesquelles ils ne possèdent pas les outils mathématiques d'un traitement complet.

En seconde, l'étude mathématique des séries de lancers consécutifs à pile ou face n'est pas possible. En revanche leur simulation, outre l'étude des fluctuations d'échantillonnage, mettra en évidence une propriété non triviale du hasard.

En B.T.S. par exemple, la théorie des files d'attente ou de la gestion de stocks, n'est pas au programme. Leur simulation permettra l'étude, dans un contexte intéressant et pratique, de lois figurant au programme (loi de Poisson, exponentielle, normale...).

Bibliographie

BOULEAU Nicolas – *«Probabilités de l'ingénieur : variables aléatoires et simulation»* – Hermann 1986.

BOREL Emile – *«Les probabilités et la vie»* – *«Que sais-je»* – P.U.F. 1947.

CHAITIN Gregory – *«Les suites aléatoires»* – *Dossier «Pour la science» : «Le hasard»* – Hors série avril 96.

Commission Inter-IREM Lycées technologiques
– *«Simulations d'expériences aléatoires – Une expérimentation du hasard de la première au BTS»* – IREM Paris-Nord - 1998.
– *«Simulation et statistique en seconde»* – IREM Paris-Nord - 2000.

Commission Inter-IREM Statistique et probabilités
– *«Enseigner les probabilités au lycée»* – 1997.

DELAHAYE Jean-Paul – *«Aléas du hasard informatique»* – Revue *«Pour la science»* Mars 1998.

DEWDNEY Alexander – *«Les hasards simulés»* – *Dossier «Pour la science» : «Le hasard»* – Hors série avril 96.

SAPORTA Gilbert – *«Probabilités, analyse des données et statistique»* – TECHNIP 1990.

Sur Internet : fonctions de simulation développées par l'INRIA de Rocquencourt sur http://www-rocq.inria.fr/scilab/doc/demos_html/node280.html