

De la statistique aux probabilités en lycée

Octobre 2006



À propos de la couverture

L'image de la couverture a été réalisée avec le logiciel de modélisation 3D « POV-Ray » (téléchargeable gratuitement sur <http://www.povray.org/>) qui permet de décrire des scènes comportant des objets géométriques que l'on peut déplacer ou modifier en utilisant les transformations usuelles. L'ensemble sera éclairé puis « observé » à l'aide d'une caméra.

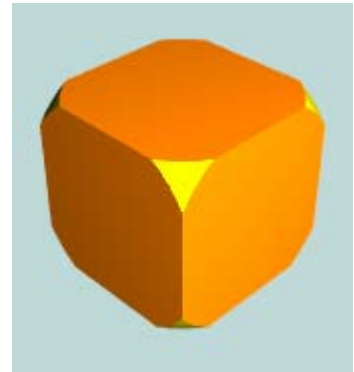
Ainsi, un dé est d'abord un cube...

```
camera {  
  location <10, 10, 10>  
  direction <1, 1, 1>  
  look_at <0, 0, 0>  
}  
  
box { <-2.0, -2.0, -2.0>, <2.0, 2, 2.0>  
  pigment { color Orange }  
}  
  
light_source { <10, 18, 20> color White }  
light_source { <10, 5, 10> color White }  
light_source { <0, 10, 10> color Yellow }
```



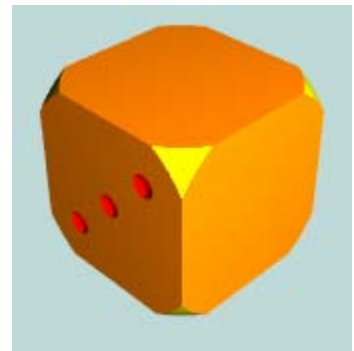
dont on adoucit les coins en prenant l'intersection avec une sphère...

```
camera {  
  location <10, 10, 10>  
  direction <1, 1, 1>  
  look_at <0, 0, 0>  
}  
intersection  
{  
  box { <-2.0, -2.0, -2.0>, <2.0, 2, 2.0>  
    pigment { color Orange }  
  }  
  sphere { <0, 0, 0>, 3.1  
    pigment { color Yellow }  
  }  
}  
light_source { <10, 18, 20> color White }  
light_source { <10, 5, 10> color White }  
light_source { <0, 10, 10> color Yellow }
```



et on creuse sur chaque face des trous hémisphériques pour indiquer le nombre de points.

```
camera {  
  location <10, 10, 10>  
  direction <1, 1, 1>  
  look_at <0, 0, 0>  
}  
difference  
{  
  intersection  
  {  
    box { <-2.0, -2.0, -2.0>, <2.0, 2, 2.0>  
      pigment { color Orange }  
    }  
    sphere { <0, 0, 0>, 3.1  
      pigment { color Yellow }  
    }  
  }  
  sphere { <2, 0.0, 0>, 0.25  
    pigment { color Red }  
  }  
  sphere { <2, 1.0, 1.0>, 0.25  
    pigment { color Red }  
  }  
  sphere { <2, -1.0, -1>, 0.25  
    pigment { color Red }  
  }  
}  
light_source { <10, 18, 20> color White }  
light_source { <10, 5, 10> color White }  
light_source { <0, 10, 10> color Yellow }
```



Voir la suite en 3^e page de couverture.

SOMMAIRE

INTRODUCTION	3
ANNEXE 1 : STATISTIQUE ET PROBABILITÉS, LE PARALLÈLE	5
ANNEXE 2	7
ANNEXE 3 : MODÉLISATION SIMULATION ADÉQUATION	8
I. STATISTIQUE DESCRIPTIVE	9
VRAIS ET FAUX HISTOGRAMMES	10
Un exemple de taille(s).....	10
STATISTIQUES ET INDUSTRIE	21
A. Temps de fabrication.....	21
B. Optimiser le stock.....	25
MÉDIANE, QUANTILES, LOGICIELS ET CALCULATRICES	28
Les définitions officielles.....	28
Les calculatrices et logiciels	29
II. NOMBRES AU HASARD.....	35
TABLES ET GÉNÉRATEURS DE NOMBRES « AU HASARD »	36
Exemples de générateurs de nombres au hasard	36
Quelques procédures de test pour des nombres présumés « au hasard »	40
Dans nos classes.....	44
Conclusion	46
Annexe 1 : test du poker, calcul des probabilités des différentes figures	47
Annexe 2 : test des séquences monotones (« runs up and down »)	49
Annexe 3 : nombre maximal de résultats consécutifs égaux dans une suite de 200 « pile ou face »	54
III. LA LOI DES GRANDS NOMBRES.....	59
INTRODUCTION	60
DES ÉNONCÉS VULGARISÉS	60
Type 1 : distribution de fréquences qui « tend » vers la loi de probabilité	60
Type 2 : la dispersion des fréquences observées diminue quand le nombre d'expériences augmente.....	63
Type 3 : la fréquence qui « tend » vers la probabilité.....	66
Type 4 : les fréquences se stabilisent (sans modèle a priori)	68
DES ÉNONCÉS NON VULGARISÉS.....	69
Loi forte des grands nombres.....	69
Loi faible des grands nombres	69
HISTORIQUE.....	70
ET QUE DIRE À NOS ÉLÈVES ?	71
Un exercice	71
Une activité	74
IV. PROBABILITÉS CONDITIONNELLES	77
INTRODUCTION.....	78
FRÉQUENCES : DES ARBRES AU TABLEAU	79
FRÉQUENCES : DU TABLEAU AUX ARBRES	81
TESTS MÉDICAUX	83
V. GÉNÉTIQUE ET PROBABILITÉS.....	87
UN PEU D'HISTOIRE	88
L'EXEMPLE DES MALADIES GÉNÉTIQUES.....	90
DANS NOS CLASSES	92
ANNEXE 1 : PETIT GLOSSAIRE DE GÉNÉTIQUE.....	100

VI. ADÉQUATION	101
ADÉQUATION D'UNE DISTRIBUTION DE FRÉQUENCES À UNE LOI DE PROBABILITÉS	102
Cas de l'équiprobabilité	102
TESTS D'ADÉQUATION AU BACCALAURÉAT	106
EXERCICE D'ÉVALUATION	111
EXERCICES DANS LES MANUELS	113
 VII. DE TOUT...AU HASARD	 123
DES DÉS ET DES POLYNÔMES	124
Les dés de Sicherman	124
Polynôme générateur d'une variable aléatoire	125
AUTOUR DU JEU DE L'OIE	128
UN JEU PAS SI ÉQUILIBRÉ QUE CA !	138
 VIII. SIMULATIONS	 145
LES PÉPITES DE CHOCOLAT	147
LES COLLECTIONS COMPLÈTES	148
LA CHASSE AUX CANARDS	149
LE LOTO	150
LES RECORDS	152
DEUX DÉS ET SOMME DES POINTS	153
ÉVOLUTION DE LA FRÉQUENCE DU « 6 » PENDANT QUE L'ON LANCE UN DÉ	154
PLANCHE DE GALTON	155
LANCER UN DÉ N FOIS	156
UNE ACTIVITÉ POUR LES ÉLÈVES	157
 BIBLIOGRAPHIE	 158

INTRODUCTION

Les programmes de seconde de l'année 2000 ont, pour la première fois, précisé quantitativement la place de la statistique dans l'horaire de mathématiques de la classe de seconde. Il convient d'y consacrer environ $\frac{1}{8}$ du temps disponible soit 4 semaines.

Il y avait jusque là des "résistances" à l'introduction réelle de la statistique dans l'enseignement des mathématiques, alors même que l'importance d'une approche scientifique de cette discipline dans la formation du citoyen n'échappe à personne. Reléguée en fin d'année, supprimée dès que le temps venait à manquer, la statistique, bien que présente dans les anciens programmes n'était que rarement enseignée. Plusieurs explications à ces attitudes peuvent être avancées :

- les professeurs ne sont que rarement formés (cela s'enseignait encore moins de leur temps) ;
- les collègues considèrent que ce ne sont pas vraiment des mathématiques ;
- si on s'en tient à des tâches techniques élémentaires, les statistiques favorisent les possesseurs de calculatrices performantes et accentuent les inégalités déjà mal vécues dans ce domaine ;
- les données récentes, motivantes, non contestables sont difficiles à trouver et le professeur risque d'avoir l'air de prendre parti ;
- les probabilités prenaient, dès la classe de première, définitivement le pas sur les aspects statistiques et étaient principalement enseignées à partir de leurs aspects combinatoires, sans aucun lien avec les questions statistiques éventuellement traitées en seconde et au collège.

Dans les nouveaux programmes des lycées, les contenus et l'esprit de l'enseignement de la statistique et des probabilités ont été profondément modifiés. L'introduction des modèles probabilistes doit être précédée d'une sensibilisation à l'aléatoire et les notions enseignées en statistique et en probabilités sont étroitement imbriquées comme le montre l'organigramme de l'annexe 2 ainsi que le « parallèle » de l'annexe 1 que l'on pourra proposer en lecture aux élèves de première. Le lien entre distribution de fréquences et loi de probabilités sera éclairé par un énoncé intuitif de la loi des grands nombres à laquelle nous consacrons le chapitre 3.

La grande nouveauté, présente aux trois niveaux, est l'utilisation de la simulation. Ce procédé peut laisser au professeur de mathématique une sensation de flou, voire un petit doute sur son caractère « scientifique ». La métaphore de la poule et de l'œuf est ici totalement justifiée : simuler une expérience (ce que l'on fera dès la seconde) c'est simuler un modèle de la dite expérience (les modèles probabilistes ne seront introduits qu'en première) et, en terminale, on adopte parfois comme modèles des résultats de simulation. Nous avons tenté de préciser les liens entre simulation, modélisation et problèmes d'adéquation dans le schéma de l'annexe 3.

L'un des problèmes est de rendre les simulations crédibles pour les élèves, c'est à dire de les persuader que si l'on avait fait l'expérience réellement les résultats auraient été « semblables » mais non identiques (la fluctuation étant « normale » dans les phénomènes

aléatoires). Pour simuler il faut disposer d'un générateur de «nombres au hasard ». Quelques uns des problèmes qu'ils soulèvent sont abordés dans le chapitre 2.

Un grand nombre d'expériences peut être simulé à partir des expériences de références (pièces, dés *etc.*) Des exemples (programmés sur un tableur) sont brièvement décrits dans le chapitre 8 et, implantés sous Excel, sont présents sur le CD.

Les activités statistiques proposées dans le cadre scolaire souffrent souvent d'un manque de réalisme. Nous avons veillé, notamment dans le chapitre 1 de statistique descriptive, à proposer des études réelles portant sur des populations nombreuses et nous avons observé, sur le terrain, l'interprétation qui en était donnée.

L'usage des arbres, qui ont acquis le statut de preuve acceptable, est encouragé. Nous présentons la notion de probabilités conditionnelles en investissant cet outil dans le chapitre 4. Les liens entre la génétique, les probabilités et la statistique nous semblaient mériter le chapitre 5. Enfin, quelques problèmes inclassables ailleurs ont été regroupés dans le chapitre 7 « De tout, au hasard ».

La brochure est complétée par un CDROM qui contient le texte intégral (aux formats Word et PDF), les tableaux de simulation, les programmes et commentaires officiels et quelques compléments.

De nombreuses publications de qualité (voir la bibliographie) sont parues depuis la mise en application des nouveaux programmes. Ce n'est donc là qu'une brochure de plus qui ne prétend ni à l'exhaustivité, ni à une originalité totale. Nous espérons néanmoins apporter quelques idées à nos collègues et contribuer ainsi à l'éveil de l'esprit critique des élèves face aux phénomènes aléatoires et aux statistiques publiées dans la presse et les autres médias.

Bernard KOCH

Membres du groupe :

Emmanuelle ACKER
Claire DUPUIS
Suzette ROUSSET-BERT
Mohamed ATLAGH
Jérôme AUDEUD
Bernard KOCH
Jean-Paul QUELEN
Dominique WEIL

ANNEXE 1 : STATISTIQUE ET PROBABILITÉS, LE PARALLÈLE

STATISTIQUE	PROBABILITÉS																								
La statistique étudie et met en forme des résultats expérimentaux, (statistique descriptive) et éventuellement, déduit de ces données des estimations ou prévisions (statistique inductive). Si on lance 100 fois un dé, on aura une liste de 100 résultats comme (1,5,3,3,6,...,1,4) c'est à dire une liste de 100 entiers compris entre 1 et 6. On peut s'intéresser alors au nombre de réalisations de certains événements : « obtenir 6 » « obtenir un résultat pair » On peut regrouper les données en comptant le nombre de fois où l'on a obtenu 1, 2, 3, 4, 5 et 6 c'est à dire déterminer les effectifs puis calculer les fréquences en divisant par l'effectif total (100 dans le cas de notre exemple). On obtient alors une « distribution de fréquences » comme par exemple : <table><tr><th>Valeurs</th><th>Effectifs</th><th>Fréquences</th></tr><tr><td>1</td><td>14</td><td>0.14</td></tr><tr><td>2</td><td>18</td><td>0.18</td></tr><tr><td>3</td><td>22</td><td>0.22</td></tr><tr><td>4</td><td>16</td><td>0.16</td></tr><tr><td>5</td><td>15</td><td>0.15</td></tr><tr><td>6</td><td>25</td><td>0.25</td></tr><tr><td>Total</td><td>100</td><td>1</td></tr></table> Les fréquences ont certaines propriétés. En notant $f(A)$ la fréquence d'un événement A on peut dire : $0 \leq f(A) \leq 1$ La somme des fréquences de toutes les valeurs possibles est 1.	Valeurs	Effectifs	Fréquences	1	14	0.14	2	18	0.18	3	22	0.22	4	16	0.16	5	15	0.15	6	25	0.25	Total	100	1	Le calcul des probabilités a pour objectif de fournir des modèles mathématiques pour l'étude des expériences aléatoires. Une expérience aléatoire est un processus dont on ne peut prévoir, avant son exécution, le résultat. (Exemples : lancer un dé, lancer deux dés, l'un rouge et l'autre noir). On ne peut prévoir le résultat d'un lancer d'un dé, mais on connaît l'ensemble des résultats possibles que l'on désignera par Ω : $\Omega = \{1, 2, 3, 4, 5, 6\}$ Certains sous-ensembles de Ω (souvent, au lycée, tous les sous-ensembles) seront appelés « événements ». « obtenir 6 » correspondra ainsi au sous-ensemble $\{6\}$, « obtenir un résultat pair » au sous ensemble $\{2, 4, 6\}$; « obtenir au moins 5 » au sous-ensemble $\{5, 6\}$. L'ensemble Ω lui même est un événement appelé événement certain, l'ensemble vide noté $\emptyset$ est appelé événement impossible. Les événements contenant 1 seul élément de Ω, comme $\{6\}, \{4\}$ sont appelés événements élémentaires. L'ensemble des événements sera noté E On associe à chaque événement un nombre réel compris entre 0 et 1 que l'on appellera sa probabilité en exigeant des probabilités qu'elles vérifient des propriétés analogues aux fréquences en statistique. En notant $p(A)$ la probabilité de l'événement A on impose : $0 \leq p(A) \leq 1$ $p(\Omega) = 1$
Valeurs	Effectifs	Fréquences																							
1	14	0.14																							
2	18	0.18																							
3	22	0.22																							
4	16	0.16																							
5	15	0.15																							
6	25	0.25																							
Total	100	1																							

La fréquence d'un événement est la somme des fréquences des valeurs qui le réalisent. Ainsi la fréquence de l'événement « obtenir un résultat pair » est $0.18+0.16+0.25=0.57$. Si on renouvelle l'expérience qui consiste à lancer 100 fois le même dé, les résultats seront presque sûrement différents de ceux obtenus la première fois. On dit qu'il y a une **fluctuation d'échantillonnage**. On peut obtenir par exemple la distribution de fréquences suivante :

Valeurs	Fréquences
1	0.12
2	0.17
3	0.21
4	0.19
5	0.10
6	0.21
Total	1.00

Il est presque certain qu'aucune des distributions de fréquences observées ne coïncidera avec la loi de probabilité (ici c'est évident puisque le nombre de lancers du dé n'est pas un multiple de 6).

Un théorème de probabilité connu sous le nom de « **loi des grands nombres** » précise dans quelle mesure les distributions de fréquences sont « proches » de la loi de probabilité qui modélise l'expérience observée.

Dans le cas du lancer d'un dé, cela signifie que les fréquences et les probabilités des deux tableaux ci-dessus seront proches, à condition que le dé ne soit pas truqué.

L'une des questions centrales en statistique est de tester si un modèle (une loi de probabilité) est adapté à la description d'une expérience. C'est le problème de **l'adéquation**.

Une technique utilisée dans le cas de l'équiprobabilité consiste à calculer la somme des carrés des écarts entre les fréquences et les probabilités.

Lorsque l'ensemble Ω est fini, la probabilité d'un événement est la somme des probabilités des événements élémentaires qui le composent.

Il suffit donc de connaître les probabilités des événements élémentaires pour pouvoir calculer toute probabilité. On définit ces probabilités dans un tableau appelé loi de probabilité définie sur Ω .

Si l'on pense que le dé est parfaitement symétrique, pour modéliser l'expérience qui consiste à lancer le dé on choisira sur $\Omega = \{1, 2, 3, 4, 5, 6\}$ la loi de probabilité :

Valeurs	Probabilités
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total	1

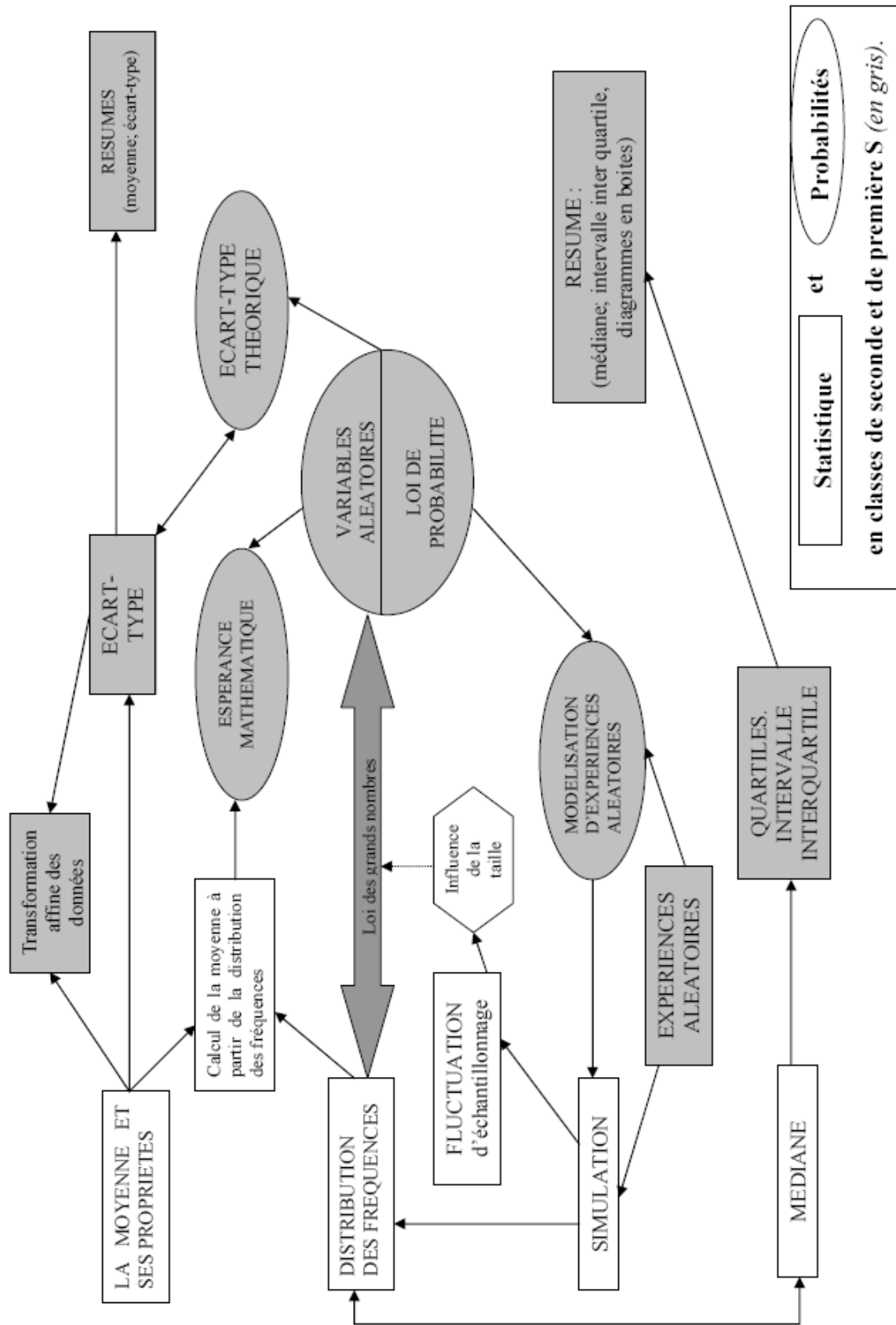
Un triplet $(\Omega, \mathcal{E}, P)$ vérifiant les propriétés (appelées axiomes) ci-dessus est appelé **espace probabilisé**. C'est un modèle mathématique abstrait dont l'étude est possible même dans le cas où il ne correspond pas, a priori, à une expérience aléatoire réelle.

Loi des grands nombres :

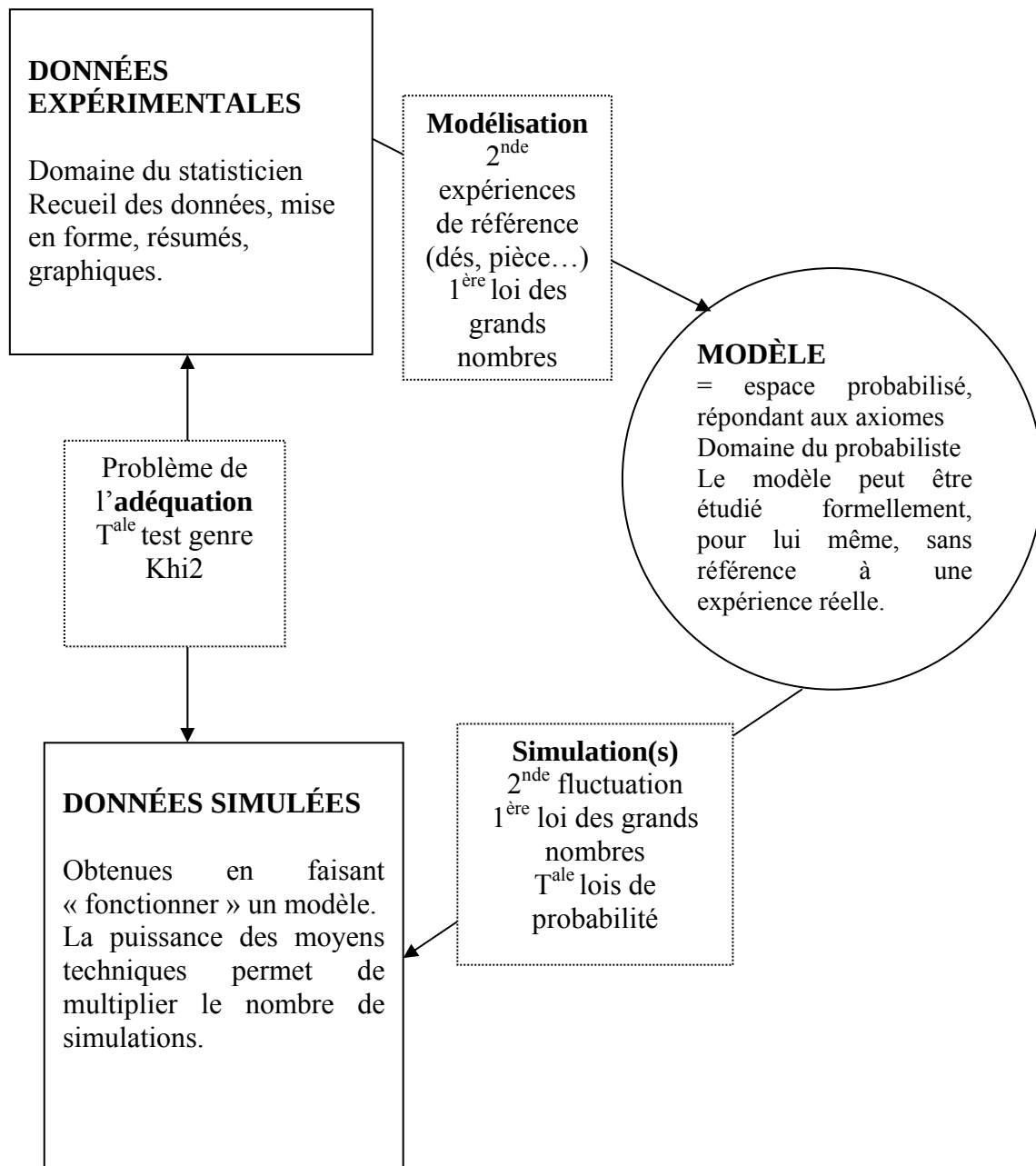
« Pour une expérience donnée, dans le modèle défini par une loi de probabilité P , les distributions des fréquences calculées sur des séries de taille n se rapprochent de P quand n devient grand. »

La somme des carrés des écarts entre les fréquences et les probabilités est aléatoire et suit, sous certaines conditions, une loi de probabilité appelée « loi du Chi deux ».

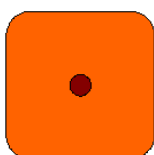
ANNEXE 2



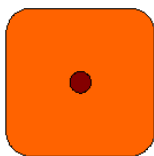
ANNEXE 3 : MODÉLISATION SIMULATION ADÉQUATION



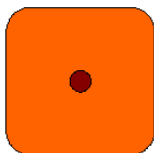
I. STATISTIQUE DESCRIPTIVE



Vrais et faux histogrammes



Statistique dans l'industrie



Médiane, quantiles, logiciels et calculatrices

VRAIS ET FAUX HISTOGRAMMES

Les histogrammes sont en principe réservés à la représentation des répartitions des caractères continus lorsqu'il y a regroupement en intervalles de classes. Dès qu'il est observé réellement, un caractère théoriquement continu devient discret puisqu'on observera un nombre fini de valeurs.

Les moyens modernes de calcul permettent de traiter rapidement un grand nombre de données et un regroupement en classe a posteriori ne se justifie plus puisqu'il représente une perte d'information et que les indicateurs (moyenne, médiane, quartiles, écart type...) ne peuvent plus être calculés précisément. Tout au plus saura-t-on en donner un encadrement. Si, pour des raisons de coût par exemple, les données ont été d'emblée recueillies dans des intervalles (voir l'exercice I des durées de vie ci-dessous) l'hypothèse d'une répartition uniforme le long des intervalles est généralement choisie et les calculs sont faits en utilisant les milieux des intervalles (moyenne, écart-type) ou l'interpolation linéaire (médiane, quartiles). Il faut être conscient que la répartition uniforme le long des intervalles de classes n'est pas toujours justifiée (voir l'exercice II des prix dans un super marché ci-dessous).

Pour ce qui concerne la représentation graphique par un histogramme, un regroupement en classes devient indispensable. Il se pose alors le problème du nombre convenable d'intervalles et de leurs amplitudes. L'utilisation d'histogrammes avec des intervalles d'amplitudes variables semble être une habitude (scolaire ?) franco-française. Les logiciels de statistiques ne proposent souvent que des histogrammes avec regroupement en intervalles de même longueur. Contentons nous d'explorer sur un exemple les conséquences du choix du nombre de classes (et/ou de leurs amplitudes)

Un exemple de taille(s)

Dans l'exemple illustrant ce paragraphe, on utilise les résultats d'une enquête réelle concernant les tailles (déterminée à 5mm près et exprimées en cm) de 2340 enfants de 5 à 6 ans (examen médical avant leur entrée à l'école primaire). On trouvera les données brutes dans le classeur TaillePoidsEnfants.xls.

Le calcul des indicateurs habituels à partir des données brutes ne pose aucun problème (à un ordinateur...)

Tailles de 2340 enfants.

Effectif total	Moyenne	Écart type		
2340	113.13	4.77		
Minimum	1^{er} quartile	Médiane	3^{ième} quartile	Maximum
96	110	113	116	133

Comme il n'y a qu'un nombre fini de valeurs on peut représenter la répartition par un diagramme en bâtons (réalisé ici avec EXCEL). Ce diagramme a le mérite de mettre en évidence un biais lors de la prise de mesures. Les opérateurs ont manifestement préféré arrondir les tailles à un entier plutôt qu'à un entier et demi ce qui donne un phénomène d'alternance entre bâton court et bâton long. Une perte d'information s'est donc produite dès les mesures et un regroupement en classes est souhaitable.

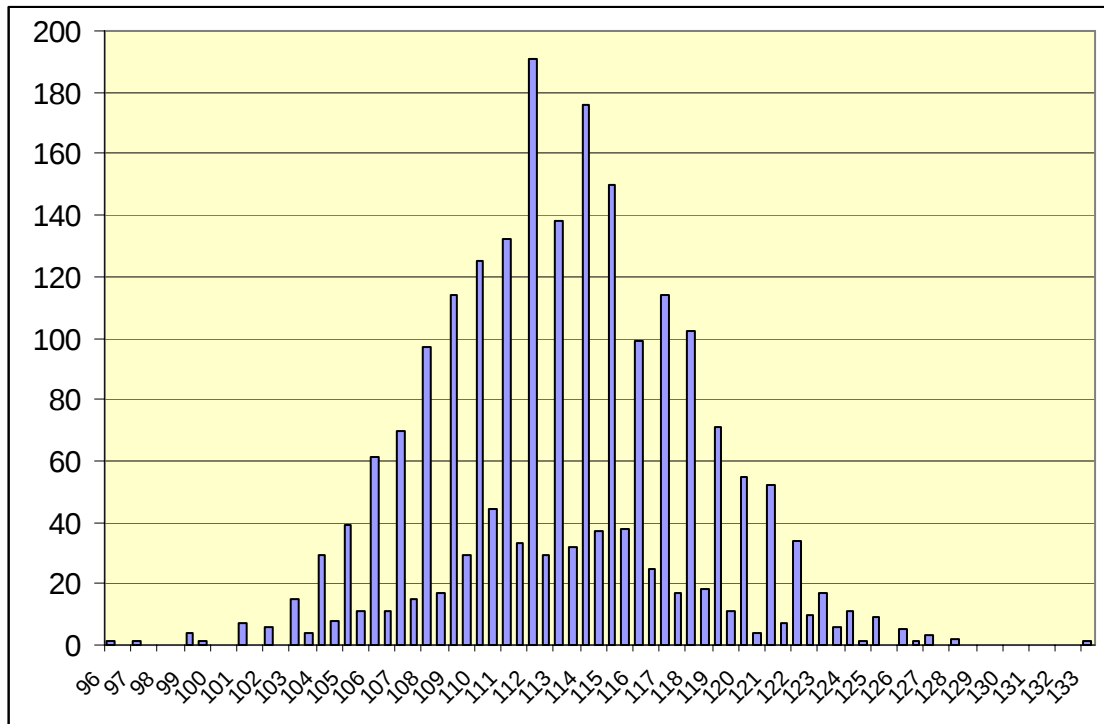


Diagramme en bâtons (sans regroupement en classes)

Une première idée est de regrouper les valeurs par deux : 96 et 96.5 puis 97 et 97.5 *etc.* On obtient un diagramme en bâtons qui gomme le biais précédemment détecté.

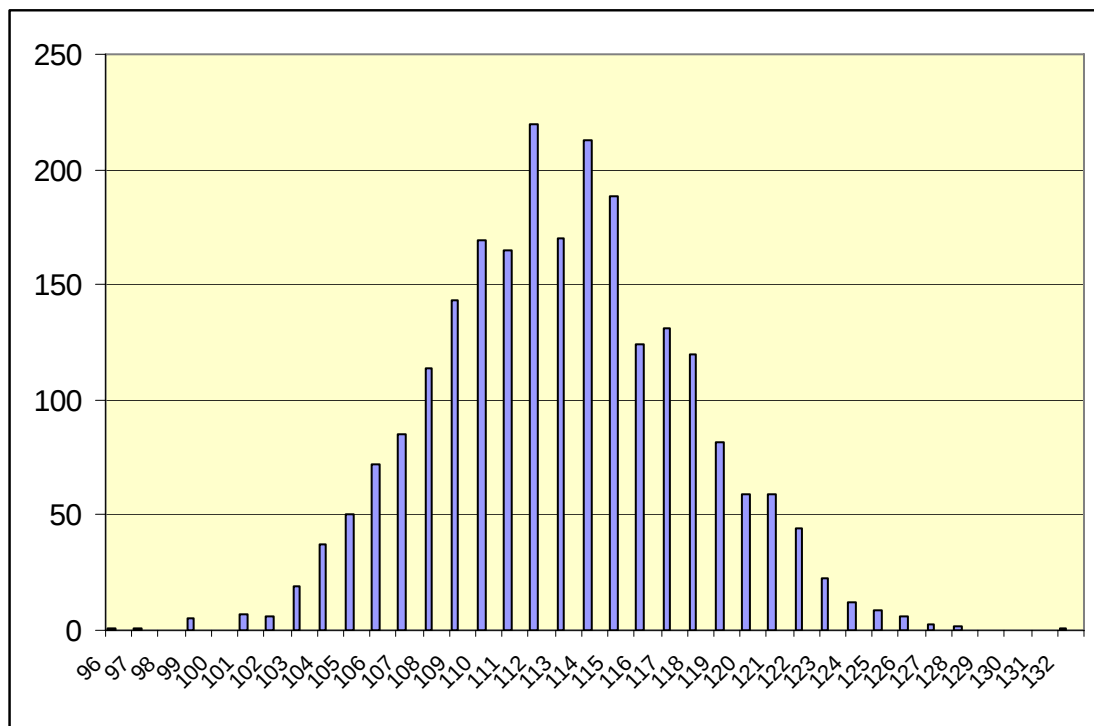


Diagramme en bâtons avec des valeurs groupées par deux

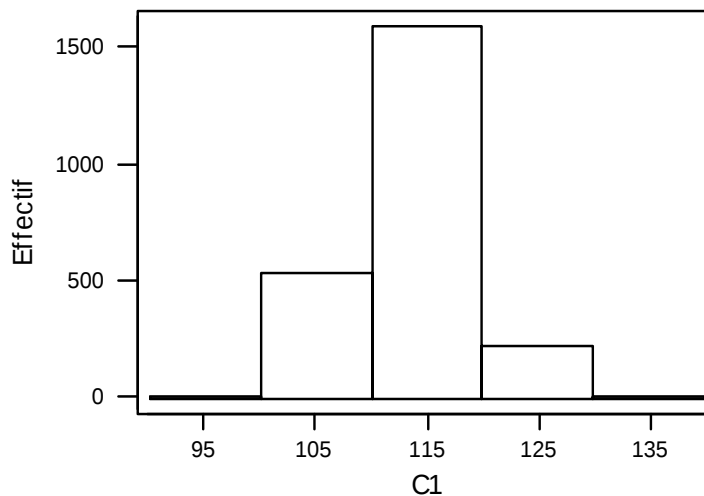
Le caractère taille étant théoriquement continu, les « lacunes » entre les bâtons n'ont aucune justification et le type de diagramme souhaitable est un histogramme. Le tableur EXCEL ne propose pas ce type de graphique. Pire, les diagrammes en bâtons sont appelés

improprement « histogramme » dans ce logiciel. Certes, en fixant le paramètre « Largeur de l'intervalle » (dans « Format de la série de données », « Options » à 0 on obtient des bâtons qui se touchent et le graphique ressemble à un histogramme.

Dans la suite nous utiliserons le logiciel Minitab qui prend en charge le regroupement en intervalles. La question de la meilleure procédure se pose. Comment en particulier choisir le nombre de classes et leurs amplitudes. Minitab utilise des intervalles de même amplitude ce qui rend la différence entre histogramme et diagramme à bâtons difficile à expliquer aux élèves.

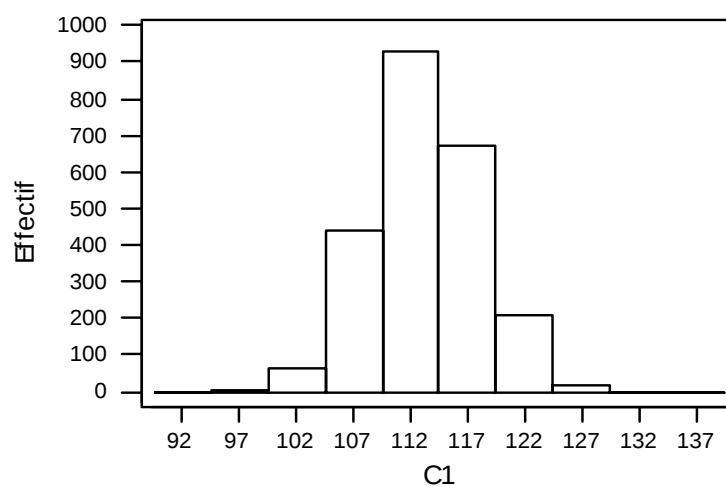
Examinons les histogrammes obtenus avec divers choix du nombre d'intervalles :

5 intervalles



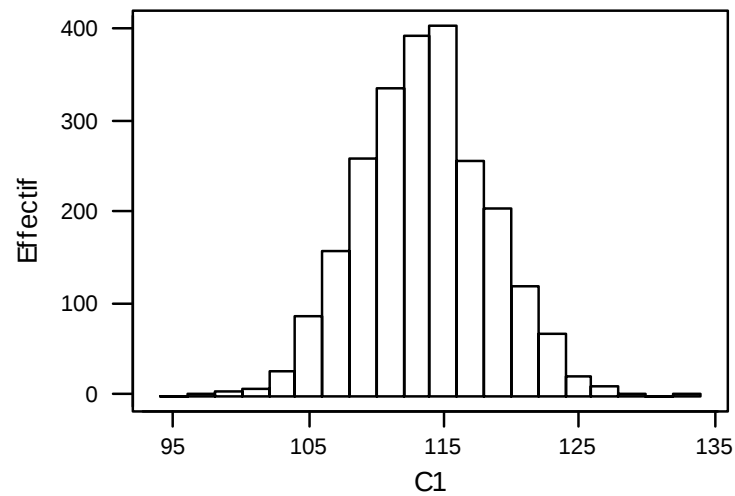
Manifestement la perte de précision est ici trop importante. Résumer 2340 données par un tel graphique représente un « gaspillage » excessif d'information.

10 intervalles



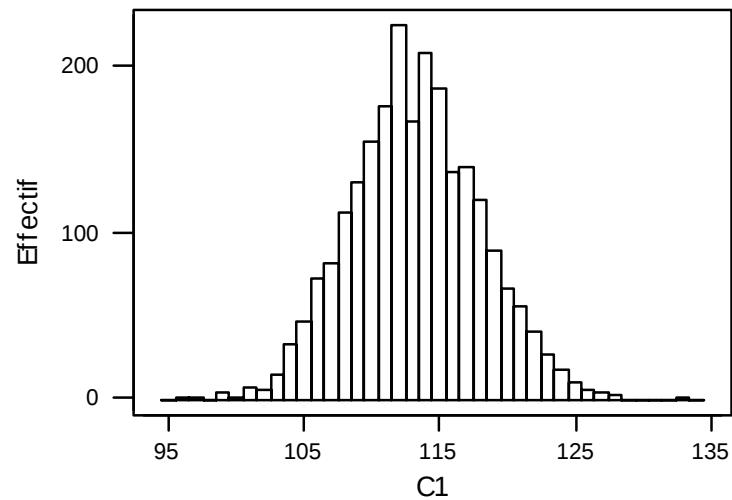
C'est un peu mieux.

20 intervalles



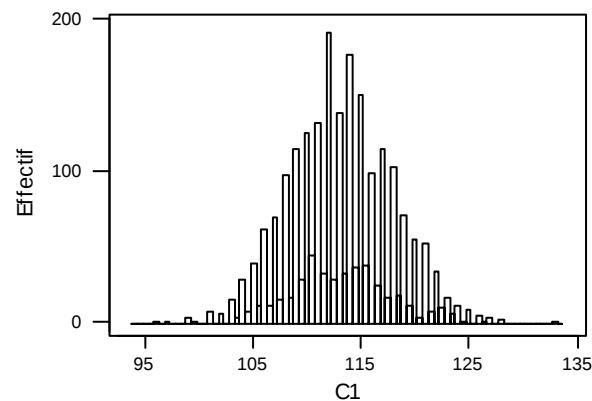
On obtient une allure de courbe « en cloche ».

40 intervalles



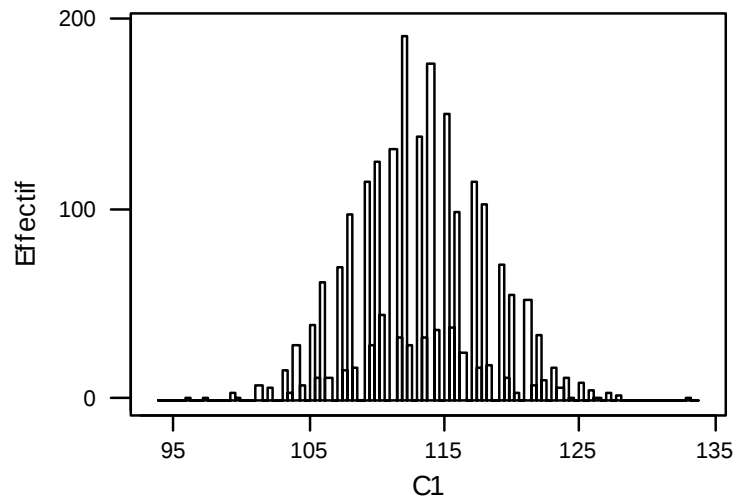
La nature hétérogène de la population étudiée par rapport au caractère taille apparaît. Il y a deux « maximum » (deux « intervalles modaux »). Il serait sans doute judicieux de séparer les filles et les garçons.

80 intervalles



Le phénomène d'alternance entre grands et petits rectangles du au désir des expérimentateurs d'arrondir plutôt à un entier réapparaît. Les valeurs observées (minimum 96 et maximum 133) se répartissent donc en fait en 75 valeurs distinctes ce qui montre qu'une répartition en 80 intervalles n'est pas raisonnable.

En poussant la déraison encore plus loin, en choisissant 100 intervalles (qui est le maximum admis par Minitab) on obtient :

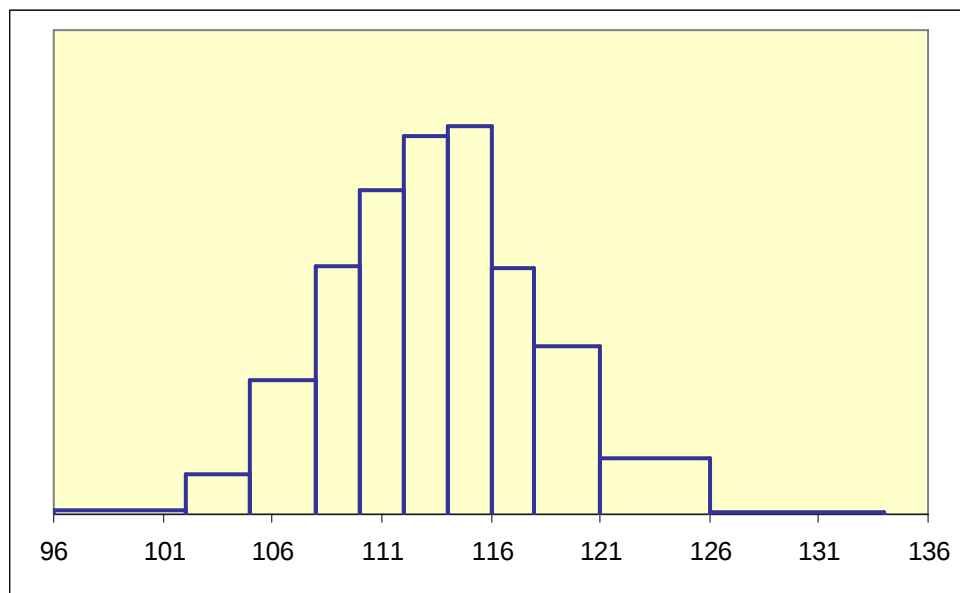


L'aspect devient chaotique. On voit bien que multiplier le nombre d'intervalles n'apporte plus rien au-delà d'une certaine limite.

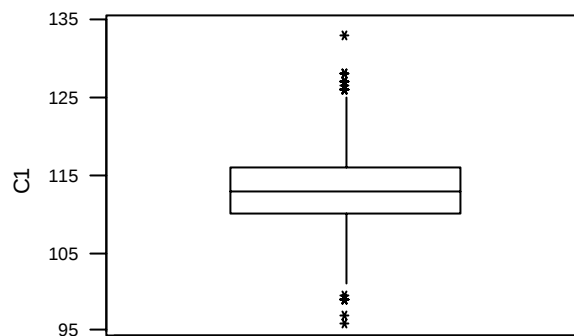
Lorsqu'on choisit l'option nombre d'intervalles « automatique », le logiciel Minitab choisit 40 intervalles de même amplitude.

Le classeur EXCEL nommé VraisHistogrammes.xls permet de fixer soi-même le nombre et les longueurs des intervalles. Voici un exemple d'utilisation de ce classeur :

Origines	Extrémités	Effectifs
96	102	14
102	105	62
105	108	207
108	110	257
110	112	334
112	114	390
114	116	401
116	118	255
118	121	261
121	126	147
126	134	12



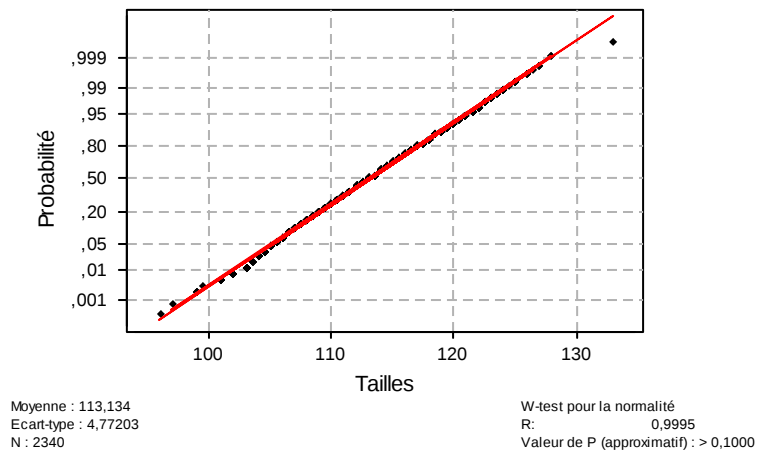
Le logiciel Minitab permet également de représenter cette même série par une boîte « à moustaches ». On remarquera que ce type de diagramme contient moins d'information qu'un histogramme (avec un nombre d'intervalles bien choisi)



Le même logiciel fournit différents tests de normalité pour une série de donnée. Voici par exemple le test de Ryan-Joiner appliqué successivement à la série des tailles des 2340 enfants puis en séparant garçons et filles.

Tailles de 2340 enfants :

Graphique de la courbe normale ou droite de Henry

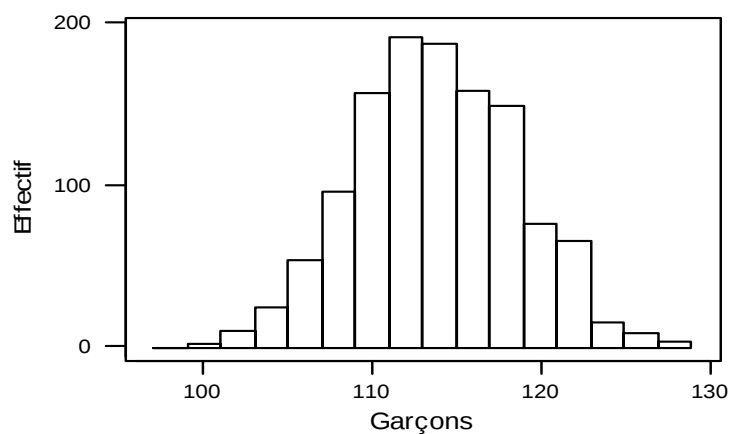


La P-valeur étant supérieure à 0.1 ce test ne permet pas de rejeter la normalité de la série des tailles des 2340 enfants au seuil habituel de 5%.

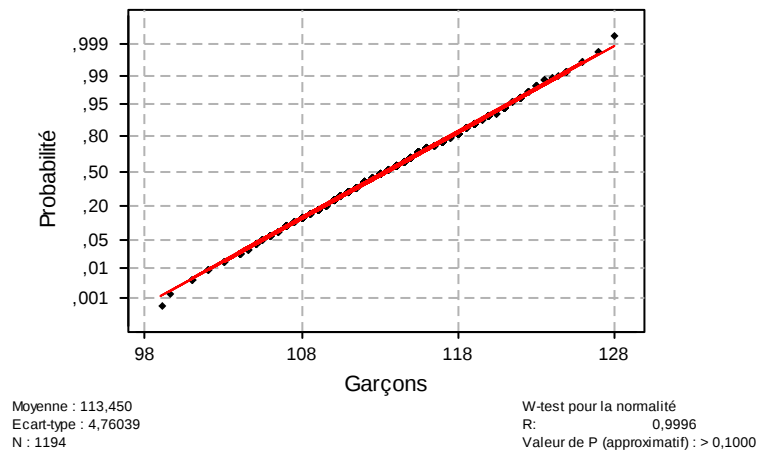
Tailles des garçons :

Effectif total	Moyenne	Écart type		
1194	113.453	4.76		
Minimum	1 ^{er} quartile	Médiane	3 ^{ème} quartile	Maximum
99	110	113	117	128

Tailles des garçons



Graphique de la courbe normale ou droite de Henry

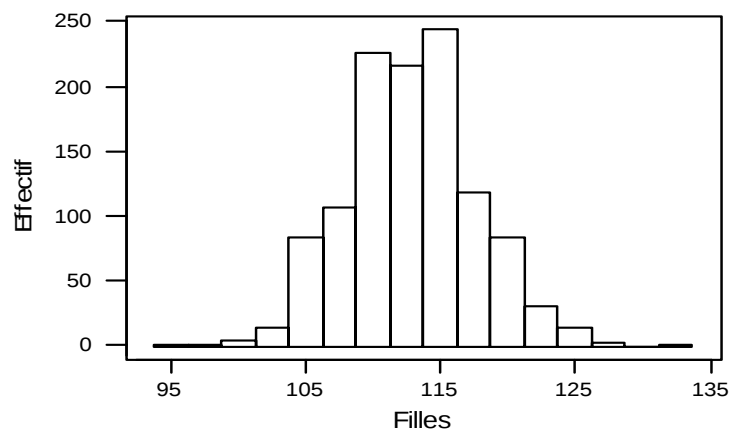


La P-valeur étant ici encore supérieure à 0.1 ce test ne permet pas de rejeter la normalité de la série des tailles des garçons au seuil habituel de 5%.

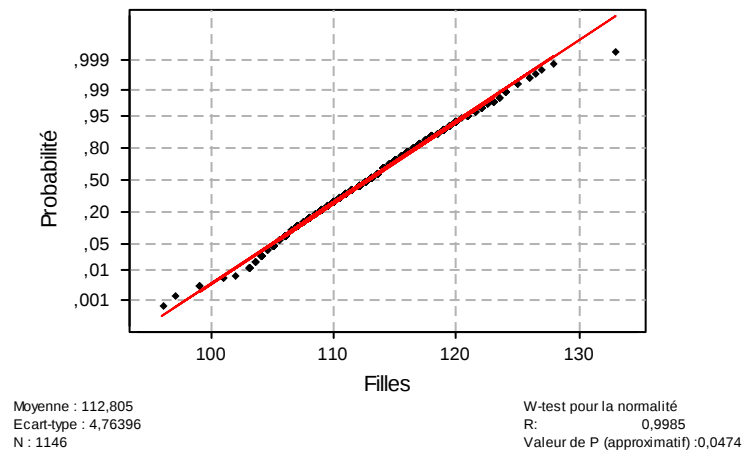
Tailles des filles :

Effectif total	Moyenne	Écart type		
1146	112.80	4.76		
Minimum	1 ^{er} quartile	Médiane	3 ^{ème} quartile	Maximum
96	109.5	113	116	133

Tailles des filles

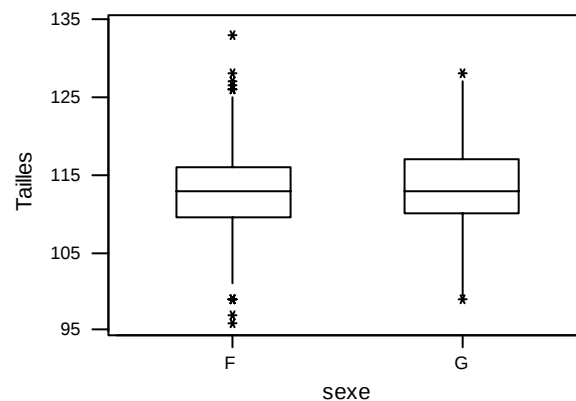


Graphique de la courbe normale ou droite de Henry



La P-valeur étant ici de 0.0474 ce test permet de rejeter la normalité de la série des tailles des filles au seuil habituel de 5%.

On peut représenter les deux séries par des boîtes à moustaches en parallèle :



Finalement, pour résumer des données nombreuses issues de l'observation d'un caractère continu, on devra trouver un compromis entre la perte inévitable d'information lors d'un regroupement en classes et la lisibilité des résultats.

Pour les calculs des caractéristiques centrales et de dispersion on utilisera les données réelles avec la précision avec laquelle elles ont été relevées.

Pour la représentation graphique sous forme d'histogramme un regroupement en classes est nécessaire. Le nombre de ces classes sera choisi raisonnablement par rapport au nombre de valeurs distinctes observées¹.

¹ Il existe une règle empirique attribuée à Sturges qui suggère d'utiliser, pour représenter n données, des classes de même amplitude dont le nombre est calculé par la formule $1 + \frac{10}{3} \log_{10}(n)$.

Exercices

I. Durée de vie

Pour étudier la durée de vie d'un lot de 120 moteurs électriques, on les branche simultanément et on relève, en fin de certaines journées, le nombre de moteurs arrêtés. On obtient le tableau suivant :

Jour	0	2	4	6	8	10	12	16	32
Arrêtés	0	2	6	14	30	70	90	110	120

- 1) Représenter graphiquement ces résultats.
- 2) Calculer un encadrement de la durée de vie moyenne des moteurs de ce lot.

Commentaires

Le caractère « durée de vie » est continu. L'observation a été, en revanche, très intermittente. On ne dispose donc d'emblée que d'une information imprécise correspondant à des intervalles. C'est la raison pour laquelle on ne peut trouver qu'un encadrement de la moyenne.

Le tableau numérique proposé contient en fait les effectifs cumulés croissants ce qui contraint les élèves à un petit traitement des données avant le calcul de la durée de vie moyenne. Le graphique attendu est plutôt un histogramme.

II. Des prix, des prix !

Le tableau ci-dessous contient le relevé des prix de tous les articles d'un présentoir « Maximum 5€ » dans un magasin (pour vous faciliter la tâche ils sont rangés dans l'ordre croissant).

0,49	1,50	2,80	3,99	4,99
0,49	1,75	2,80	4,50	4,99
0,50	1,75	2,90	4,80	4,99
0,50	1,75	2,95	4,80	4,99
0,50	1,80	2,99	4,80	4,99
0,80	1,90	2,99	4,90	4,99
0,80	1,90	2,99	4,90	4,99
0,80	1,95	2,99	4,90	4,99
0,80	1,95	2,99	4,90	4,99
0,90	1,95	2,99	4,90	4,99
0,90	1,95	2,99	4,95	4,99
0,90	1,95	2,99	4,95	4,99
0,99	1,95	2,99	4,95	4,99
0,99	1,99	3,90	4,95	4,99
0,99	1,99	3,95	4,95	4,99
0,99	1,99	3,95	4,95	4,99
0,99	2,50	3,95	4,95	
0,99	2,50	3,95	4,95	
1,50	2,50	3,95	4,95	
1,50	2,60	3,95	4,95	

1) Calculer la moyenne de ces prix.

2) Regrouper les prix observés en 5 classes

Intervalles de prix	[0,1[	[1,2[	[2,3[	[3,4[	[4,5[
Effectifs					

3) Calculer la moyenne que l'on obtiendrait en utilisant les milieux des intervalles.

Commentaire

Un exemple (assez réaliste) où la répartition à l'intérieur des intervalles est déséquilibrée vers la droite. Utiliser les milieux des intervalles conduit à une moyenne fortement sous évaluée.

STATISTIQUES ET INDUSTRIE

Parmi des reproches exprimés au sujet des anciens programmes de statistique et probabilité figurait l'affirmation que les problèmes posés étaient strictement scolaires et ne correspondaient en rien au travail réel des statisticiens sur le terrain. Nous avons donc voulu rechercher un certain nombre d'études non fictives afin de mesurer cette éventuelle distorsion entre la théorie et la pratique.

Les exemples auxquels nous avons eu accès utilisent surtout l'aspect descriptif des statistiques.

Dans les deux études détaillées ci-dessous, l'outil statistique est surtout un outil de communication dans l'entreprise utilisant les représentations graphiques et un outil d'aide à la décision. La décision finale tiendra compte non seulement des résultats statistiques mais de bien d'autres critères moins facilement quantifiables. Les textes en italique sont nos commentaires, le reste reprend les termes utilisés par les concepteurs des études.

A. Temps de fabrication

But : Étude du temps de fabrication de pièces et impact d'un ensemble d'améliorations sur le temps de fabrication

Problématique

- On mesure le temps nécessaire à la fabrication de pièces et on souhaite le réduire.
- Après les modifications apportées on effectue de nouvelles mesures et on veut confirmer que ces améliorations ont eu un impact positif sur le temps de fabrication.

1. Données à analyser (avant et après améliorations)

Semaine de livraison	N°commande	Semaine de commande	Année de commande	Date demandée	Temps de fabrication	Retard	Période
03/01	55737	15	2000	30/00	40	25	1
08/01	47260	21	2000	05/01	39	3	1
06/01	47346	39	2000	02/01	19	4	1
03/01	18113	40	2000	02/01	15	1	1
08/01	18114	40	2000	02/01	20	6	1
08/01	18115	40	2000	02/01	20	6	1
13/01	18117	43	2000	08/01	22	5	1
04/01	47358	43	2000	05/01	13	-1	1
05/01	47359	43	2000	02/01	14	3	1
52/01	18102	29	2000	49/00	75	55	1
03/01	18104	31	2000	49/00	24	6	1
04/01	18105	31	2000	03/01	25	1	1
03/01	18106	31	2000	03/01	24	0	1
04/01	18107	32	2000	50/00	24	6	1
03/01	18110	35	2000	02/01	20	1	1
03/01	47326	35	2000	48/00	20	7	1
03/01	47328	35	2000	50/00	20	5	1
08/01	47330	45	2000	09/01	15	-1	1
06/01	47331	36	2000	42/00	22	16	1
08/01	55808	44	2000	04/01	16	4	1
13/01	47362	45	2000	09/01	20	4	1
08/01	47363	45	2000	07/01	15	1	1
13/01	47364	45	2000	11/01	20	2	1
15/01	18119	46	2000	13/01	21	2	1
16/01	18120	46	2000	22/01	22	-6	1
08/01	18122	46	2000	09/01	14	-1	1
25/03	56024	19	2003	22/03	6	3	2
02/04	56025	29	2003	49/03	25	5	2
09/03	56026	6	2003	09/03	3	0	2

Dans le tableau ci-dessus les dates sont exprimées en semaines.

Exemple : la première commande a été passée en semaine 15 de l'année 2000 et la date de livraison demandée par le client était la semaine 30 de la même année.

Le temps de fabrication a été de 40 semaines donc le retard effectif est de 25 semaines et la commande a été livrée la semaine 3 de l'année 2001.

La période 1 correspond à la période avant améliorations, la période 2 à la période après améliorations.

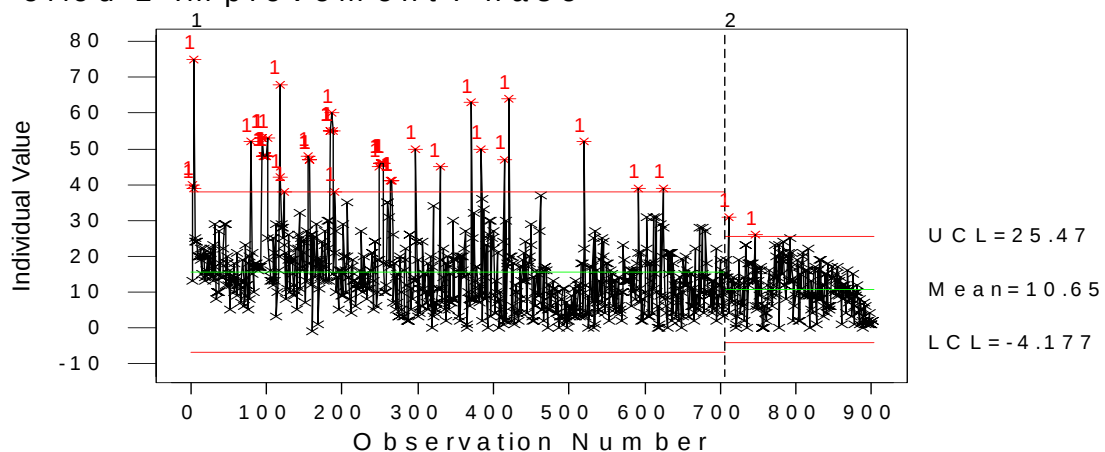
Remarque

Le tableau n'est pas présenté dans sa totalité (il y a en réalité 902 lignes). On trouve là la confirmation que l'outil statistique est surtout utilisé pour traiter un grand nombre de données. C'est l'une des idées qu'il conviendra de faire passer auprès des élèves car les exemples qu'ils traiteront eux-mêmes correspondront souvent à des effectifs plus petits pour que la charge de calcul ne devienne pas excessive.

2. Visualisation du temps de fabrication pour 902 observations

Le graphique ci-dessous, appelé « carte de contrôle des individus » représente en abscisses les commandes classées par ordre de livraison et en ordonnées les temps de fabrication (LT : lead time) exprimés en semaines.

I C h a r t f o r L T b y P e r i o d - P e r i o d 1 B a s e l i n e ;
P e r i o d 2 I m p r o v e m e n t P h a s e



Les 706 premières observations correspondent à la période avant améliorations (période 1)

Les 196 observations restantes correspondent à la période après améliorations (période 2)

UCL signifie : Upper Control Limit et correspond environ à $\bar{x} + 3\sigma$

LCL signifie : Lower Control Limit et correspond environ à $\bar{x} - 3\sigma$

Ces valeurs ont été calculées pour chacune des périodes et représentées par des traits horizontaux sur le graphique

Remarque

Le graphique utilisé n'est pas comme on pourrait s'y attendre un histogramme mais privilégie l'ordre chronologique. Il présente l'avantage de pouvoir être complété au fur et à mesure que l'on recueille les données.

3. Conclusion

On constate que :

- Le temps moyen de fabrication des pièces a baissé entre la période 1 et 2
- L'écart entre les limites supérieures et inférieures s'est réduit
- Il y a moins de points hors de l'intervalle [LCL ; UCL]

Donc les améliorations apportées (nouvelle organisation de l'atelier, modification du flux des pièces ...) ont amélioré le temps de fabrication des pièces.

Remarques

On peut noter que les conclusions restent prudemment qualitatives tout en portant à la fois sur la valeur centrale (moyenne) et la dispersion.

L'intervalle $[\bar{x} - 3\sigma, \bar{x} + 3\sigma]$ mis en évidence sur le graphique devrait, sous l'hypothèse de normalité, contenir 99,8% des observations.

Le choix de cette amplitude d'intervalle n'a rien d'universel dans le monde industriel et commercial. Une démarche récente connue sous le sigle « Six Sigma »² vise elle une performance de 99,99966% qui correspond à un intervalle de $[\bar{x} - 6\sigma, \bar{x} + 6\sigma]$.

Voici un tableau proposé dans le cadre de la démarche Six Sigma qui tente de mieux faire comprendre pourquoi, pour certaines applications, il est préférable de viser 99,99966% et pas seulement 99% de conformité à la norme.

99% (2.56 Sigma)	99.99966% (6 Sigma)
20,000 articles de poste perdus par heure	Sept articles perdus par heure
5,000 opérations chirurgicales incorrectes par semaine	1.7 opérations incorrectes par semaine
Deux atterrissages trop longs ou trop courts par jour	Un ou deux atterrissages trop courts/longs en cinq ans
200 000 mauvaises ordonnances de médicaments par an	68 mauvaises ordonnances par an
Pas d'électricité pendant sept heures par mois	Une heure sans électricité tous les 34 ans

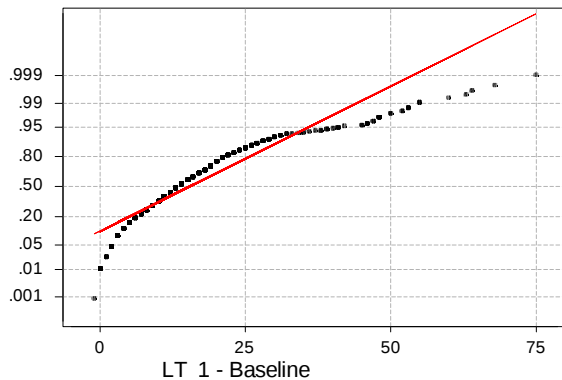
² Six Sigma Academy (SSA) a été fondée en 1994 par Dr. Mikel Harry et Mr. Richard Schroeder.

SSA a mis en œuvre avec beaucoup de succès la méthode six sigma dans 22 sociétés, dont Général Electric, AlliedSignal, Ford Motor Company, DuPont, et Dow Chemical, en appliquant cette démarche à tous les niveaux et secteurs de l'entreprise.

Étude de la normalité

Période 1 : avant améliorations

Normality Plot for Baseline Period (1)

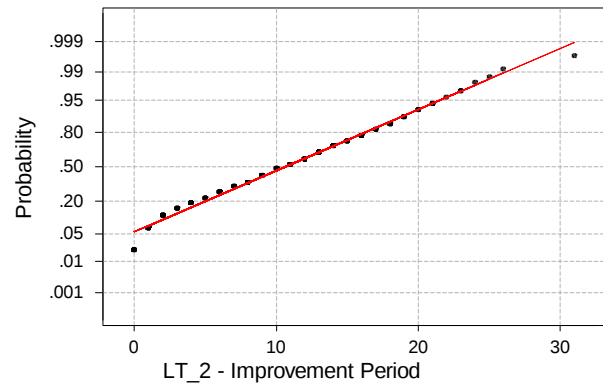


Average: 15.4802
StDev: 11.4551
N: 706

Anderson-Darling Normality Test
A-Squared: 20.080
P-Value: 0.000

Période 2 : après améliorations

Normality Plot for Improvement Period (2)



Average: 10.6480
StDev: 6.48498
N: 196

Anderson-Darling Normality Test
A-Squared: 0.913
P-Value: 0.020

Avec le postulat de départ que les données d'un processus maîtrisé obéissent à une loi normale, on confirme l'impact positif des modifications apportées.

Remarque

Dans les graphiques ci-dessus, la graduation de l'axe des ordonnées est « gaussienne ». Dans ce type de repère la fonction de répartition d'une loi normale est représentée par une droite (appelée droite de Henry). L'examen visuel des nuages de points (qui correspondent aux fréquences cumulées croissantes) montre que les points de la deuxième période sont bien plus proches d'une droite que ceux du premier graphique. La répartition des temps de fabrication de la deuxième période est donc plus proche d'une loi normale que celle de la première période.

Au-delà du constat graphique ci-dessus, le logiciel procède à un test de normalité d'Anderson Darling. La P-valeur calculée est la probabilité que la variable aléatoire A^2 dépasse la valeur observée sur l'échantillon. L'hypothèse de normalité est acceptée si on juge que la P-valeur est suffisamment grande (par exemple supérieure à 0.01).

Nous nous sommes longuement interrogés sur les raisons qui pourraient « justifier » le « postulat » invoqué. En quoi l'adéquation à la loi normale constitue-t-elle un critère de maîtrise d'un processus ? On peut penser que lorsque les écarts par rapport à la norme ne sont plus dus qu'à un grand nombre de petites causes indépendantes à effets additifs, le résultat suit une loi normale. C'est le phénomène observé par Gauss qui parlait de « lois des erreurs ». Lorsqu'une cause est dominante par rapport aux autres le résultat ne suivra plus une loi normale et remédier à cette cause dominante constituera une amélioration de l'efficacité du processus.

B. Optimiser le stock

Cadre : le stock de pièces dans l'usine est trop important, ce qui entraîne des coûts d'immobilisation à la société.

But : réduire le stock de pièces et améliorer le taux de rotation en modifiant les règles de mise en stock.

Problématique : il faut choisir les clients pour lesquels on mettra des pièces en stock c'est à dire les clients qui passent des commandes de montants réguliers et si possible élevés.

1. Identification des clients principaux

On recense environ 1200 clients, il s'agit dans un premier temps d'effectuer un premier tri basé sur une règle appelée :

Règle des 80-20 ou loi de Pareto et disant que 20% des causes représentent 80% des effets.

Dans notre situation cela signifierait que 20% des clients génèrent 80% du montant total des commandes.

C'est pour cela que l'on fait un diagramme de Pareto et que l'on effectue un premier tri en ne conservant que les clients permettant de réaliser 80% du montant total des commandes.

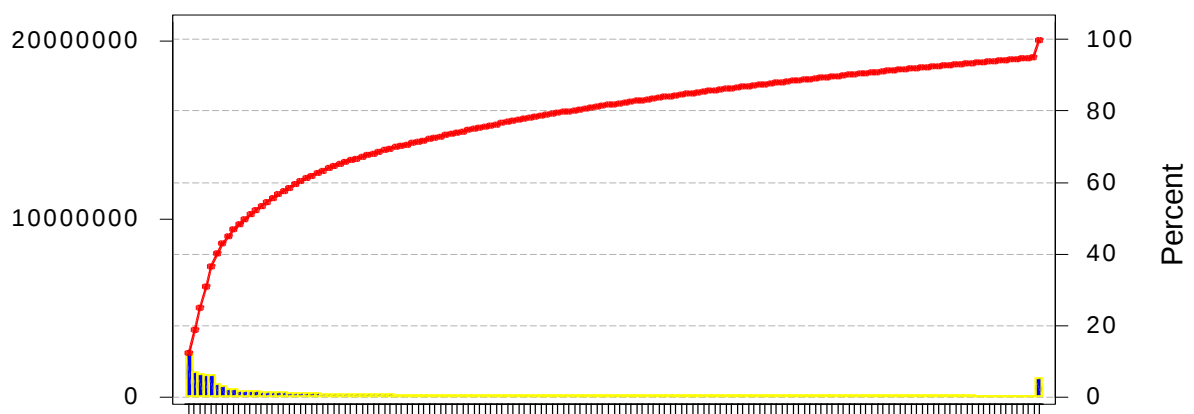
En réalité on ne prendra en compte que 43 clients correspondant en fait à 70 % du montant total des commandes.

En effet parmi les 153 clients qui représentent 80% du montant des commandes certains ont des montants de commande négligeables et l'on sait déjà que l'on ne mettra pas de pièces en stock pour de tels clients.

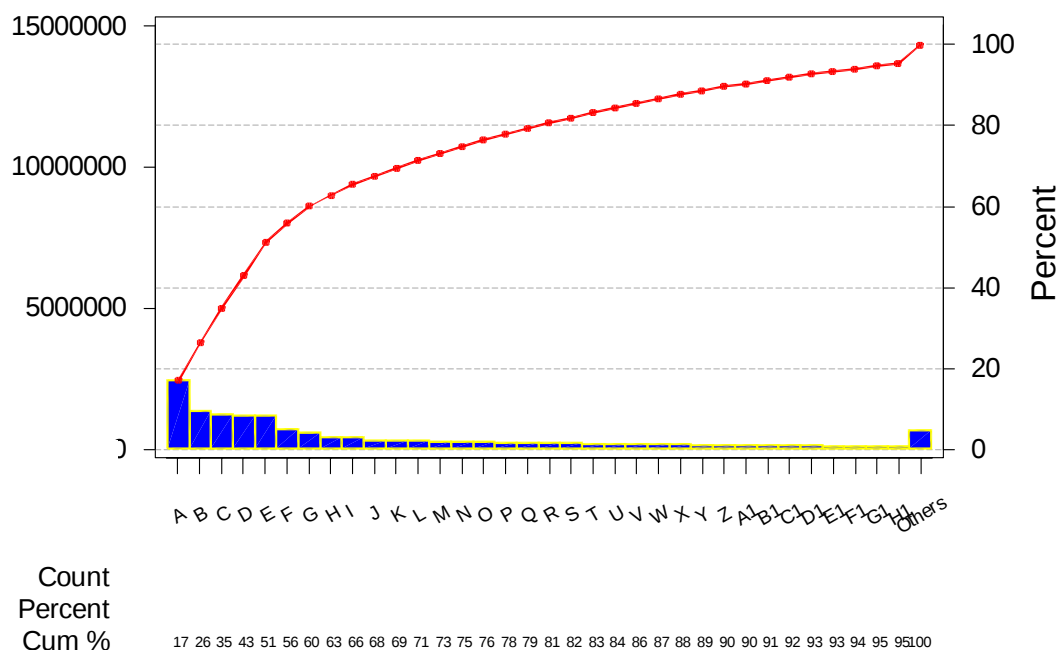
Construction d'un diagramme de Pareto :

Les clients sont classés par ordre décroissant de montant de commande (le montant de commande de chaque client est représenté par une barre) et on superpose au graphique obtenu le polygone des fréquences cumulées croissantes.

Pareto Chart for Client

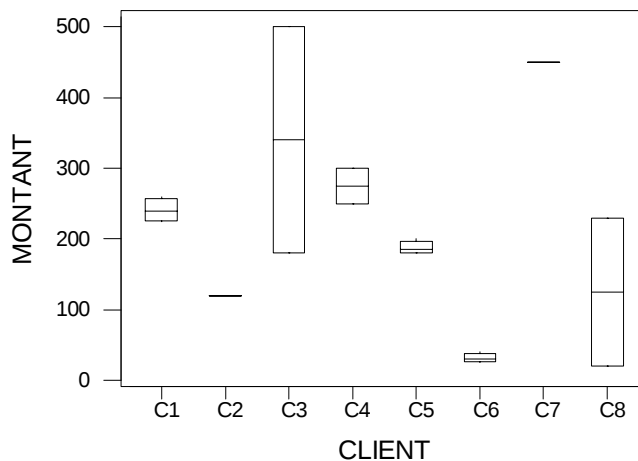


Pareto Chart for Client



CLIENTS	MONTANT_1999	MONTANT_2000	MONTANT_2001	MONTANT_2002
C1	250	225	230	260
C2	*	120	*	*
C3	180	*	500	*
C4	250	*	300	*
C5	180	189	200	182
C6	25	30	40	30
C7	*	*	450	*
C8	20	*	*	230

Diagramme en boîtes pour les huit clients sélectionnés :



Analyse .

- on gardera des pièces en stock pour les clients C1, C4 et C5 puisque ces clients passent des commandes de montants réguliers et élevés.
- On ne gardera pas de pièces en stock pour les clients C3 et C8 qui passent des commandes de montants trop irréguliers même si ces commandes sont de montants élevés.
- De même on ne gardera pas de pièces en stock pour les clients C2 et C7 qui n'ont passé qu'une unique commande.
- Pour le client C6 qui passe des commandes de montants réguliers mais peu élevés la décision appartiendra au commercial.

Remarques

Les diagrammes en boîtes représentés ici concernent des séries de montants de commandes dont l'effectif est au maximum 4. On peut s'interroger sur la pertinence de l'utilisation de cet outil graphique pour un si petit nombre de données. Ce graphique a cependant le mérite de mettre en évidence « au premier coup d'œil » la dispersion des montants des commandes qui est justement le critère de sélection principal adopté ici. De plus la réalisation du graphique est immédiate avec le logiciel utilisé (MiniTab).

Conclusion

Dans ces deux études pratiques ce sont les représentations graphiques qui sont surtout utilisées car elles permettent rapidement aux non initiés de comprendre la situation et de contribuer à la prise de décision. Les statistiques apparaissent principalement comme un outil de communication et d'aide à la décision.

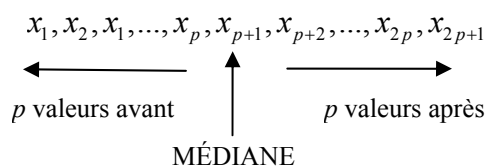
MÉDIANE, QUANTILES, LOGICIELS ET CALCULATRICES

Les définitions officielles

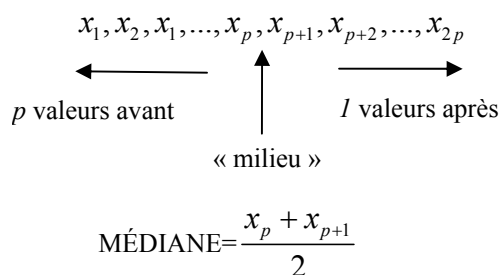
En classe de seconde (comme au collège déjà), la procédure préconisée par les documents officiels pour déterminer la valeur de la médiane d'une série statistique est la suivante :

Ranger les valeurs du caractère par ordre croissant. La médiane correspond alors :

- A la valeur située au milieu de la liste quand l'effectif total est impair



- A la moyenne des deux valeurs qui encadrent le milieu quand l'effectif total est pair.



En classe de première pour déterminer les valeurs des quartiles et déciles le programme préconise la procédure suivante :

Ranger les valeurs du caractère par ordre croissant puis calculer les fréquences cumulées croissantes :

- Le premier quartile (resp. $k^{\text{ième}}$ quartile) correspond à la première valeur du caractère dont la fréquence cumulée croissante est supérieure ou égale à 0,25 (respectivement $k*0.25$)
- Le premier décile (resp. $k^{\text{ième}}$ décile) correspond à la première valeur du caractère dont la fréquence cumulée croissante est supérieure ou égale à 0,1 (respectivement $k*0.1$)

Ainsi les quartiles et déciles sont, par définition, des valeurs de la série statistique ce qui ne sera pas toujours le cas de la médiane.

Par ailleurs, la médiane n'est égale au deuxième quartile que si l'effectif est impair ce qui nuit un peu à la cohérence.

On remarquera cependant que si l'on étudie un grand nombre de données (à l'inverse des « exemples d'école » analysés ci-dessous) ces différences s'estompent. Les valeurs x_p , $x_{p+1} \dots$ seront, sinon égales, du moins très proches.

Les calculatrices et logiciels

- Pour la série de valeurs : 1 ; 2 ; 3 ; 4

Caractère	Effectif	Fréquence	Fréquences cumulées croissantes
1	1	0,25	0,25
2	1	0,25	0,5
3	1	0,25	0,75
4	1	0,25	1

Médiane : 2,5

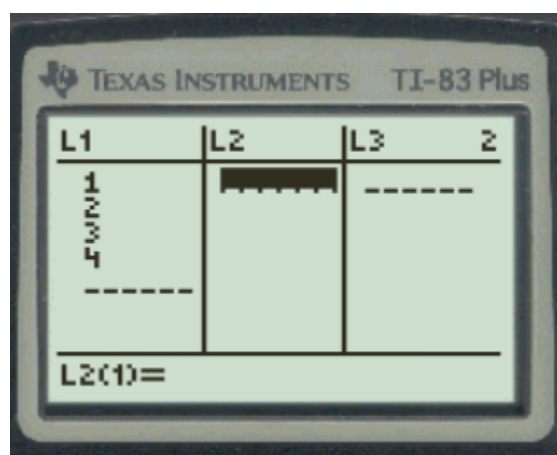
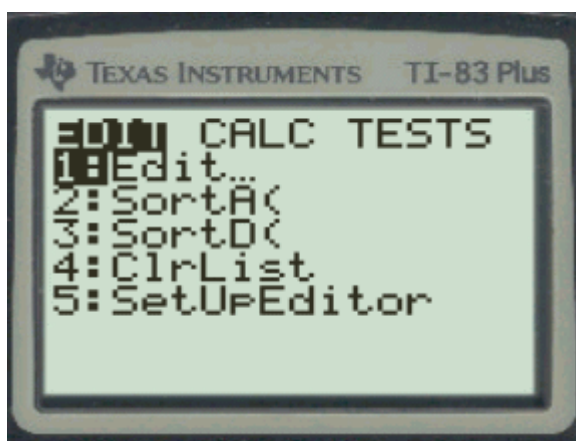
Premier quartile : 1

Troisième quartile : 3

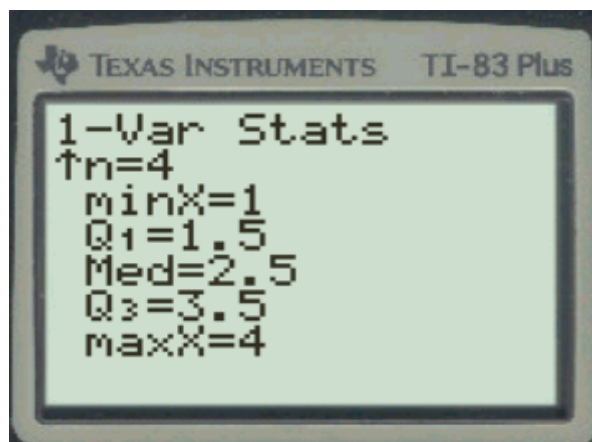
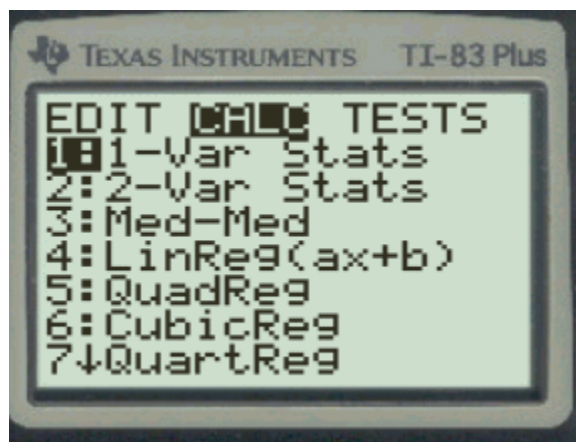
Les calculatrices

Les calculatrices récentes proposent des procédures de détermination de la médiane et des quartiles. Le principe de fonctionnement illustré ci-dessous sur l'exemple de la TI83 est sensiblement le même pour les autres modèles.

Il faut commencer par introduire (« éditer ») les données (valeurs et éventuellement effectifs ou fréquences) dans des listes.



On lance ensuite les calculs pour une série statistique à une variable :



On voit que la définition adoptée dans le programme ne correspond pas à celle de la calculatrice pour le calcul de quartiles.

La calculatrice semble procéder pour les quartiles comme pour la médiane en se limitant à la première « moitié » de la série (sans la médiane) pour Q_1 , et à la deuxième « moitié » de la série (sans la médiane) pour Q_3 .

$$\left. \begin{array}{l} 1 \\ 2 \end{array} \right\} Q_1 = \text{médiane de 1 et 2} = 1,5$$

$$\left. \begin{array}{l} 2,5 \\ 3 \\ 4 \end{array} \right\} Q_3 = \text{médiane de 3 et 4} = 3,5$$

Le logiciel MINITAB

Médiane = 2,5

$Q_1 = 1,25$

$Q_3 = 3,75$

On constate que la définition de la médiane est identique à celle du programme.

Concernant les quartiles :

Tout groupe de données a trois quartiles. Si vous triez les données par ordre croissant, le premier quart est inférieur ou égal au premier quartile. Le deuxième quartile correspond à la médiane. Les trois premiers quarts sont inférieurs ou égaux au troisième quartile et le dernier quart lui est supérieur ou égal.

Pour calculer les quartiles, utilisez Manip > Trier pour classer les données par ordre croissant. Le premier quartile (Q_1) est égal à l'observation en position $(n+1) / 4$, et le troisième (Q_3) à l'observation en position $3(n+1) / 4$, où n est le nombre d'observations. Si cette position n'est pas un entier, une interpolation est utilisée.

Supposez par exemple que $n=4$. Comme $(4 + 1)/4 = 1,25$ et Q_1 se situe au quart de la distance entre la première et la deuxième observation (appelons les x_1 et x_2). Donc $Q_1 = x_1 + 0,25(x_2 - x_1)$. Comme $3(4 + 1)/4 = 3,75$, $Q_3 = x_3 + 0,75(x_4 - x_3)$, où x_3 et x_4 sont la troisième et la quatrième observation.

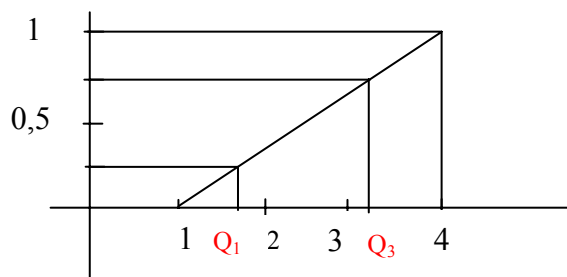
Le tableur EXCEL

EXCEL dispose d'une fonction QUARTILE() qui demande deux paramètres, la plage de données et le rang du quartile voulu. La formule =QUARTILE(A1:A4;1) donnera le premier quartile des valeurs situées dans la plage A1:A4. Les résultats fournis pour notre exemple sont là aussi en désaccord avec la définition du programme.

$$Q_1 = 1,75$$

$$Q_3 = 3,25$$

Pour déterminer les quartiles EXCEL semble procéder par interpolation comme illustré ci-contre.



Expérimentons deux autres exemples avec ces différents outils pour renforcer nos conjectures concernant les procédures utilisées.

- Pour la série de valeurs 1 ; 2 ; 3 ; 4 ; 5

Caractère	Effectif	Fréquence	Fréquence cumulée croissante
1	1	0,2	0,2
2	1	0,2	0,4
3	1	0,2	0,6
4	1	0,2	0,8
5	1	0,2	1

Médiane : 3

Premier quartile : 2

Troisième quartile : 4

Pour EXCEL

Médiane = 3

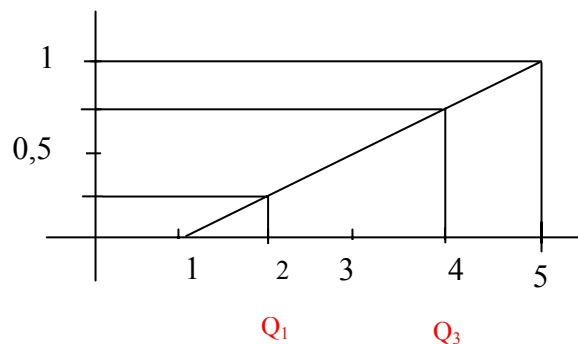
$$Q_1 = 2$$

$$Q_3 = 4$$

$$\left. \begin{array}{l} 1 \\ 2 \end{array} \right\} Q_1 = \text{médiane de 1 et 2} = 1,5$$

3

$$\left. \begin{array}{l} 4 \\ 5 \end{array} \right\} Q_3 = \text{médiane de 4 et 5} = 4,5$$



Pour MINITAB

Médiane = 3

$$Q_1 = 1,5$$

$$Q_3 = 4,5$$

- Pour la série de valeurs : 1 ; 2 ; 2 ; 3 ; 4

Caractère	Effectif	Fréquence	Fréquences cumulées croissantes
1	1	0,2	0,2
2	1	0,2	0,4
2	1	0,2	0,6
3	1	0,2	0,8
4	1	0,2	1

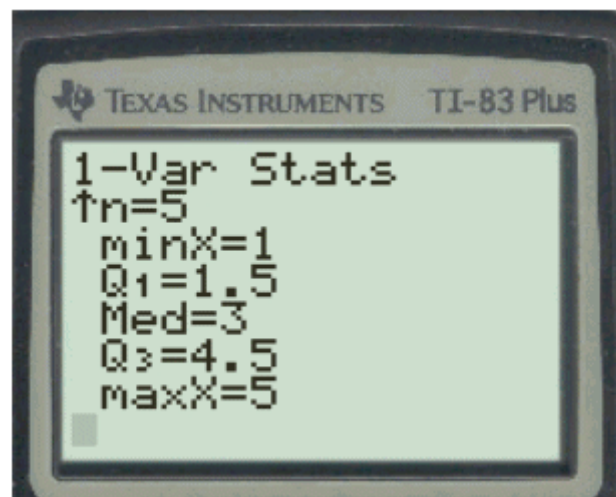
Médiane : 2

Premier quartile : 2

Troisième quartile : 3

La calculatrice (attention on ne peut utiliser deux fois une même valeur du caractère il faut lui affecter un effectif de 2)

1 }
 2 } $Q_1 = \text{médiane de 1 et 2} = 1,5$
 2
 3 }
 4 } $Q_3 = \text{médiane de 3 et 4} = 3,5$



MINITAB

Médiane = 2

$Q_1 = 1,5$

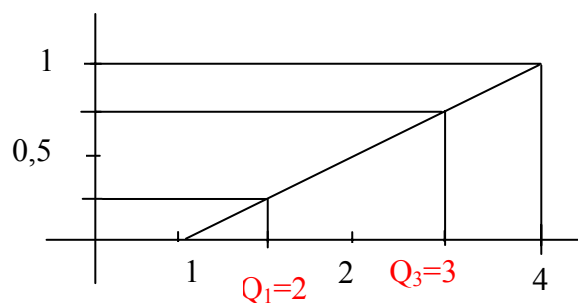
$Q_2 = 3,5$

EXCEL

Médiane = 2

$Q_1 = 2$

$Q_3 = 3$



Conclusion

Pour les outils testés, il semble y avoir une même définition de la médiane qui est celle suggérée par les textes officiels. En revanche, pour ce qui concerne les quartiles et déciles les procédures diffèrent et peuvent conduire à des résultats très différents pour des petites séries statistiques. Toutefois, dans le cas d'un effectif suffisamment grand, les écarts sur les quantiles obtenus sont minimes.

Le professeur amené à corriger un exercice d'évaluation dans lequel un calcul de quantiles est demandé, devra faire preuve de tolérance et accepter toute réponse conforme à la « notion de quantile ».

Dans la rédaction d'un énoncé, il évitera autant que possible de poser brutalement la question : « quelle est la médiane ? » ou « quel est le premier décile ? », et préférera demander le calcul d'une valeur répondant à une propriété explicitement formulée dans le contexte de l'exercice.

Exemple : Voici la répartition des habitations d'une petite ville selon le nombre de pièces :

Nombre de pièces	1	2	3	4	5	6	7	8	9	10
Nombre d'habitations	190	200	230	225	316	215	175	60	30	10

La municipalité décide d'attribuer une prime aux habitants des logements ayant le moins de pièces. Elle souhaite que la moitié environ des habitations soit concernée. A votre avis, quelles sont les habitations donnant droit à la prime ? Quel est le pourcentage d'habitations concernées ?

Quant à la définition des quartiles que l'on donnera dans le cours, on peut certes se référer à la recommandation du groupe de rénovation des programmes :

En statistique, pour toute série numérique de données à valeurs dans un intervalle I , on définit la fonction quantile Q , de $[0,1]$ dans I , par :

$$Q(u) = \inf\{x, F(x) \leq u\},$$

où $F(x)$ désigne la fréquence des éléments de la série inférieurs ou égaux à x .

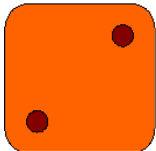
On peut cependant anticiper des difficultés de compréhension chez les élèves de la définition des quartiles ainsi obtenue :

« **le plus petit élément** q des **valeurs des termes de la série tel qu'au moins 25%** des données soient **inférieures ou égales** à q ».

Les difficultés résultent déjà de la structure de la phrase française dans laquelle l'ordre d'énonciation n'est pas congruent avec l'ordre des opérations successives qu'il faut effectuer pour trouver la valeur du quartile.

Pour les élèves il est peut-être préférable de donner des définitions algorithmiques des quantiles. La définition décrit ainsi le moyen de calculer le quantile correspondant et sa mise en pratique posera moins de problèmes.

II. NOMBRES AU HASARD



Tables et générateurs de nombres au hasard

TABLES ET GÉNÉRATEURS DE NOMBRES « AU HASARD »

La simulation nécessite que l'on dispose d'une « source de hasard ». Le problème qui se pose immédiatement est celui de la crédibilité de la simulation qui dépendra en premier lieu de la « qualité » du générateur de hasard utilisé. Imaginons que l'on décide de remplacer le traditionnel tirage du LOTO avec son urne sphérique contenant au départ les 49 boules que les téléspectateurs voient brassées généreusement avant chaque tirage par une procédure informatique où c'est un ordinateur qui proposerait les six numéros. On peut se demander si un joueur (perdant, et ce sont les plus nombreux) ferait confiance à l'ordinateur même si celui-ci est à même de garantir une « meilleure équiprobabilité » qu'un dispositif mécanique même vérifié par huissier.

Les réticences du grand public ne sont pas totalement injustifiées tant il est délicat de donner un sens précis aux termes cités entre guillemets dans le paragraphe ci-dessus. Les « sources de hasard » dont l'utilisation est suggérée par les programmes de lycée sont les tables de chiffres au hasard et les générateurs de nombres au hasard informatique.

Toutes les calculatrices actuelles ainsi que tous les langages informatiques proposent un « générateur de nombres au hasard ». C'est un programme qui calcule, par récurrence, une suite de nombres qui semblent « uniformément » répartis (c'est le minimum que l'on peut exiger d'eux) le long de l'intervalle $[0,1[$, mais chacun est, en fait, calculé à partir du précédent. Le processus est donc parfaitement déterministe et totalement prévisible dès que l'on connaît la valeur de départ (le germe). L'utilisateur final ignore la « formule » de calcul utilisée (secret souvent jalousement gardé par les concepteurs du programme) et pour lui les valeurs successives produites semblent tirées au sort.

Le caractère déterministe et prévisible d'une table de nombre au hasard est encore plus évident puisque dans ce cas tout est écrit et qu'utiliser la même table de la même manière donnera la même suite de nombres.

Exemples de générateurs de nombres au hasard

Une première génération de processus informatiques utilisés pour produire des nombres au hasard comportait les suites obtenues par congruence linéaire (1948, D.H. Lehmer).

$$u_{n+1} = au_n + b \pmod{c}$$

Exemple :

Programme RANDOM

$h \leftarrow$ donnée	(valeur de départ, germe)
Répéter n fois	(pour obtenir n nombres au hasard)
$h \leftarrow h * a + b$	
$h \leftarrow \text{FRAC}(h)$	(où $\text{FRAC}()$ est la fonction partie décimale)
Afficher h	
Fin répéter	

Pour nous faire une idée, explorons quelques exemples :

Le 147-générateur :

Choisir un nombre de départ (entre 0 et 1), par exemple zéro virgule le nombre de secondes écoulées depuis 0 heure ce matin ou encore zéro virgule le numéro de série de votre calculatrice ou encore zéro virgule votre date de naissance. Multiplier ce nombre initial par $a=147$ ($b=0$) et ne conserver que la partie décimale du résultat et recommencer avec elle.

Ce générateur donne des résultats satisfaisants si on examine, par exemple, la fréquence d'apparition des différents chiffres comme premier chiffre après la virgule des nombres engendrés. Voici le résultat d'une expérimentation sur une suite de 1000 nombres.

Chiffres	Effectifs
0	101
1	90
2	102
3	99
4	100
5	96
6	98
7	94
8	110
9	110

Pourquoi utiliser $a=147$ dans ce procédé ? Ne pourrait-on simplifier les calculs en utilisant par exemple le facteur multiplicatif 2, ou 6, ou 5 ? Là encore des calculs vont nous convaincre que certains facteurs multiplicatifs ne donnent pas une suite satisfaisante.

Pour $a=2$ on obtient pour 1000 tirages

Chiffres	Effectifs
0	955
1	6
2	5
3	6
4	4
5	7
6	5
7	5
8	4
9	3

En fait, la situation ne se dégrade vraiment qu'à partir d'une cinquantaine d'itérations qui mènent inmanquablement vers une partie décimale nulle. Le phénomène semble se confirmer pour toutes les valeurs paires de a .

Le facteur $a=5$ peut nous paraître suspect a priori en raison de l'utilisation de la base dix et pourtant le résultat d'une simulation sur 1000 itérations donne :

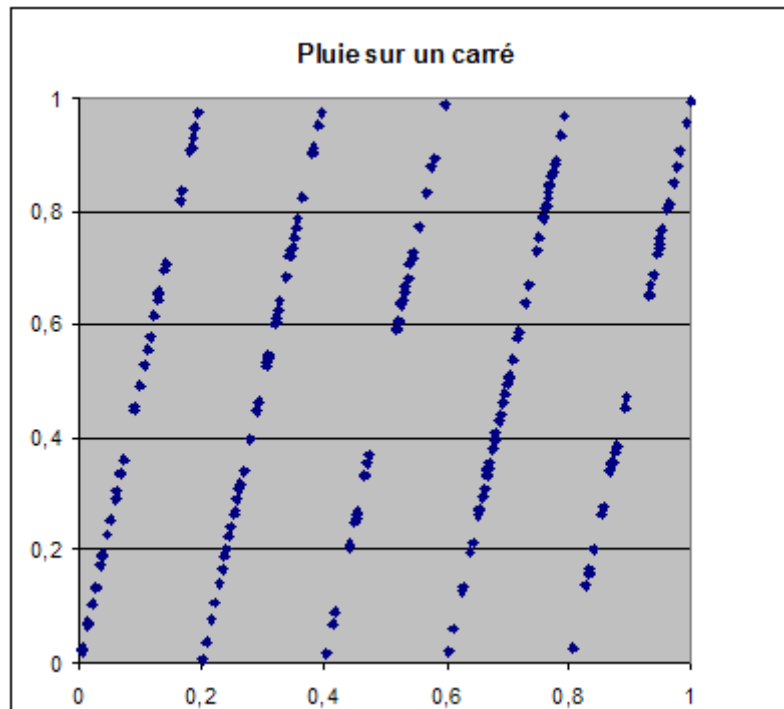
Chiffres	Effectifs
0	96
1	92
2	104
3	116
4	87
5	93
6	111
7	108
8	96
9	97

Tout facteur a premier semble donner un générateur satisfaisant du point de vue de la répartition du premier chiffre après la virgule.

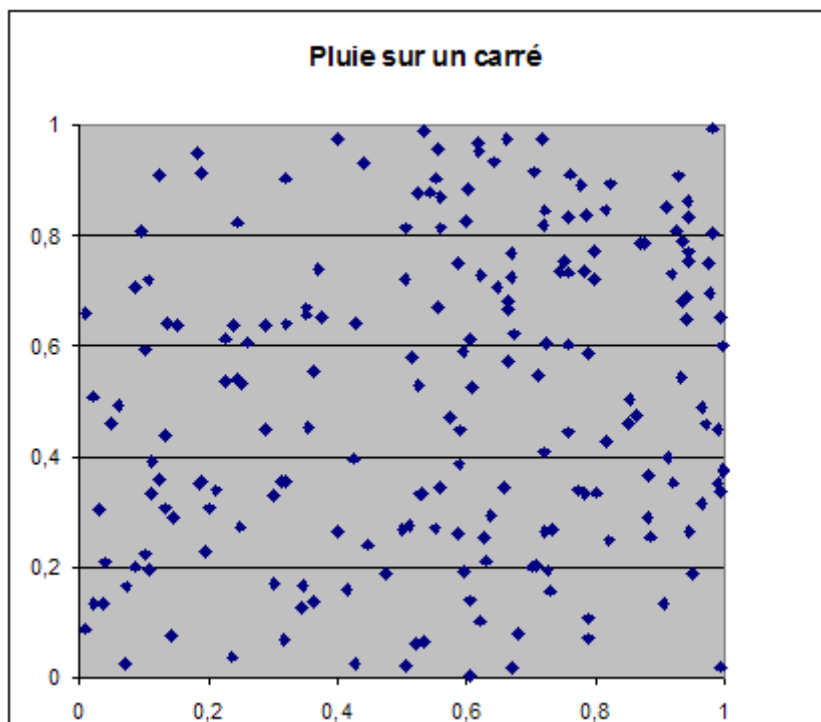
D. Knuth a démontré que, pour que le générateur $u_{n+1} = au_n + b \pmod{c}$ soit « de bonne qualité », les paramètres a , b et c doivent satisfaire au critères suivants :

- b et c doivent être premiers entre eux ;
- $a-1$ doit être un multiple de p , pour tout p nombre premier diviseur de c ;
- $a-1$ doit être un multiple de 4 si c est un multiple de 4 ;
- si c est une puissance de 2, le bit de poids faible des nombres produits vaut alternativement 0 et 1 (ce n'est d'ailleurs pas le seul cas où cela se produit).

Ces générateurs « affines » ne sont cependant pas utilisables pour toute simulation (même en prenant les précautions indiquées ci-dessus). Si l'on utilise le facteur $a=5$ pour simuler une pluie aléatoire de 200 points sur le carré de côté unité, en choisissant successivement l'abscisse et l'ordonnée du point on observe le phénomène ci-dessous qui s'explique aisément (pour a plus grand il y aura simplement plus de droites).



Bien entendu en utilisant pour le tirage de l'abscisse et de l'ordonnée deux valeurs de départ distinctes, c'est à dire deux programmes distincts (indépendants) la « pluie » devient plus satisfaisante, mais qu'est-ce donc qu'une pluie aléatoire satisfaisante ?



La définition de ce qu'est un « bon » générateur de nombres aléatoires n'est donc chose aisée, mais il est plus facile de se convaincre qu'un générateur est « mauvais ».

Quelques procédures de test pour des nombres présumés « au hasard »

Faute d'une définition parfaitement rigoureuse de ce qu'est une suite de nombres « aléatoire », des batteries de test existent. Nous en décrivons brièvement quelques-uns dans ce qui suit.

Les tests spectraux

Il s'agit de visualiser comme nous l'avons fait ci-dessus avec les « pluies sur un carré » des points du plan (ou de l'espace) dont les coordonnées sont les nombres à tester. L'apparition de motifs dans la densité des points selon les zones est un indicateur du caractère médiocrement aléatoire de la suite examinée.

Le test du Chi deux sur les fréquences des chiffres ...

Pour tester le caractère aléatoire d'une suite de nombre la comparaison de la fréquence de chaque chiffre à la probabilité de 1/10 constitue un premier test qui ne saurait suffire comme le montre la suite de 30 entiers 0 0 0 1 1 1 2 2 2 3 3 3 4...8 9 9 9 qui n'est guère aléatoire.

Voir dans le chapitre « Adéquation » l'application de ce test aux décimales de π .

Les tests de non corrélation

On calcule le coefficient de corrélation linéaire entre valeurs consécutives qui devrait être très proche de 0. Par exemple, pour 2500 couples de décimales de π on obtient un coefficient de corrélation linéaire de 0,00056863.

Le test du poker

Un test plus efficace est inspiré du jeu de poker : on regroupe la suite des chiffres par 5 et on détermine les fréquences de certaines « figures » que l'on compare aux probabilités que l'on sait calculer³.

Figure	Exemple	Fréquence	Probabilité
Chiffres distincts	79548		0.3024
paire	22167		0.5040
double paire	22887		0.1080
brelan	55567		0.0720
full	55999		0.0090
carré	33337		0.0045
poker	66666		0.0001

Voici, par exemple, le résultat du test du poker pour les 5000 premières décimales de π .

Les décimales de π ont été obtenues grâce au logiciel DERIVE. Le comptage des figures réalisé avec EXCEL est visible dans DécimalesPi.XLS).

	Poker	Carré	Full	Brelan	2 paires	1 paire	Rien	Total
Observés	0	4	10	61	127	467	331	1000
Théoriques	0,1	4,5	9	72	108	504	302,4	1000
Ecart	-0,1	-0,5	1	-11	19	-37	28,6	
Carrés	0,01	0,25	1	121	361	1369	817,96	
Carrés/Théoriques	0,1	0,05556	0,11111	1,68056	3,34259	2,71627	2,70489	10,711

³ Voir une démonstration en annexe 1

Dans la table du Chi deux, pour 6 degrés de liberté (nombre de modalités-1) et pour un seuil de 5% on lit la valeur 12,59. La valeur obtenue étant de 10.711, Le test du poker ne permet pas de rejeter l'hypothèse que les 5000 premières décimales de π sont réparties « au hasard ».

Les décimales de π semblent constituer une suite de chiffres au hasard acceptables (alors qu'elles sont parfaitement « prévisibles »).

Dans son ouvrage « Le fascinant nombre π »⁴, Jean-Paul Delahaye donne le même tableau pour 10 millions de décimales de π regroupées en 2 millions de mains de poker pour lesquelles le test semble encore plus convaincant.

	Poker	Carré	Full	Brelan	2 paires	1 paire	Rien	Total
Observés	200	8887	17891	144375	216520	1007151	604976	2000000
Théoriques	200	9000	18000	144000	216000	1008000	604800	2000000
Ecart	0	-113	-109	375	520	-849	176	
Carrés	0	12769	11881	140625	270400	720801	30976	
Carrés/Théoriques	0	1,41878	0,66006	0,97656	1,25185	0,71508	0,05122	5,07354

Jean-Paul Delahaye fait très justement remarquer que π n'a aucune raison d'être « aléatoire ». D'ailleurs son écriture dans d'autres bases de numération (deux et seize par exemple) fait apparaître des propriétés particulières.

Pour soumettre vos propres tables de chiffres au hasard au test du poker voir le fichier EXCEL Poker.XLS

Le test des séquences monotones (run test)

Ce test est basé sur le nombre de sous suites croissantes et décroissantes dans une suite de nombres (tous distincts) à tester. Par exemple pour les 12 nombres ci-dessous :

12 55 24 69 127 99 221 207 52 26 189 203

on compte 4 séquences croissantes :

12 ↗ 55 \ 24 ↗ 69 ↗ 127 \ 99 ↗ 221 \ 207 \ 52 \ 26 ↗ 189 ↗ 203

et 3 séquences décroissantes

12 / 55 ↘ 24 / 69 / 127 ↘ 99 / 221 ↘ 207 ↘ 52 ↘ 26 / 189 / 203

donc 7 séquences au total.

Si R est le nombre total de séquences monotones dans une suite de n nombres, on peut démontrer⁵ que l'espérance mathématique et la variance de la variable aléatoire R sont :

$$E(R) = \frac{2n-1}{3} \text{ et } V(R) = \frac{16n-29}{90}$$

et pour n assez grand, R suit approximativement une loi normale.

Soit

$$r = \frac{R - E(R)}{\sqrt{V(R)}}$$

Si $|r| > 1.96$ on peut rejeter au seuil de 5% l'hypothèse que l'échantillon est aléatoire.

⁴ *Le fascinant nombre π* , Jean-Paul Delahaye, Pour la Science Belin, 1997.

⁵ Voir une démonstration en annexe 2

Exemple

Prenons une suite de $n=1000$ nombres engendrés par la fonction ALEA() du tableur EXCEL.

Ils donnent par exemple 685 séquences monotones.

Les paramètres théoriques sont

$$E(R) = \frac{2 \times 1000 - 1}{3} \approx 666.33 \quad \text{et} \quad V(R) = \frac{16 \times 1000 - 29}{90} \approx 177.45$$

$$r \approx \frac{685 - 666.33}{\sqrt{177.45}} \approx 1.4$$

Comme $|r| < 1.96$ on ne peut rejeter l'hypothèse que les nombres sont aléatoires.

Pour soumettre vos propres séries de nombres à ce test utilisez le tableau EXCEL

SeqMonotonesTest.XLS

Le test de Kolmogorov-Smirnov

On compare la fonction de répartition correspondant à la loi uniforme aux fréquences cumulées croissantes observées dans l'échantillon de nombres dont on dispose.

Par exemple, pour des nombres « au hasard » de l'intervalle $[0,1]$ on a la fonction de répartition :

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x > 1 \end{cases}$$

et la suite, triée dans l'ordre croissant, des valeurs observées $(x_1, x_2, x_3, \dots, x_{n-1}, x_n)$ permet de définir la fonction en escalier :

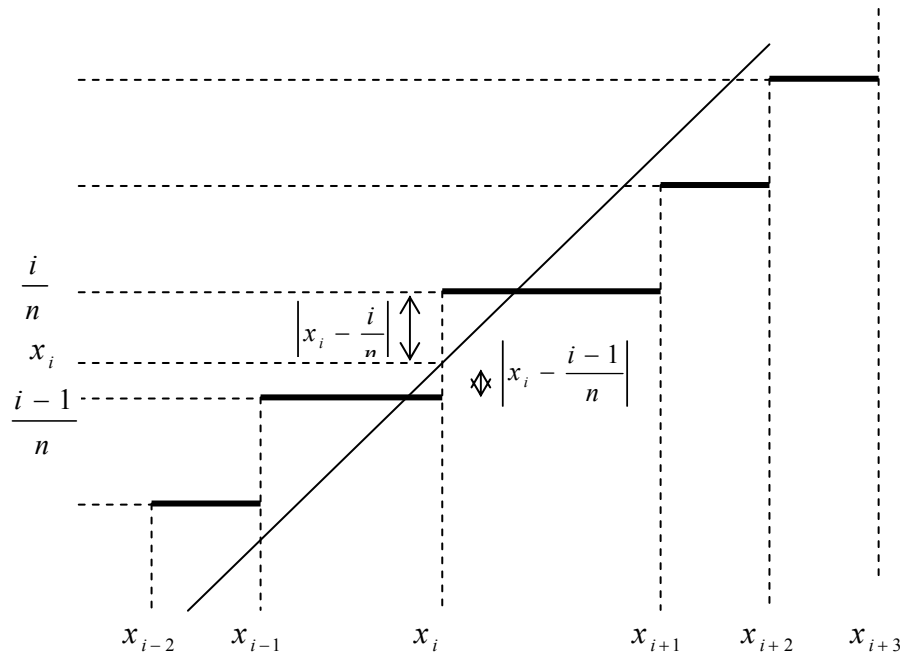
$$\hat{F}(x) = \begin{cases} 0 & \text{si } x < x_1 \\ \dots & \\ \frac{i}{n} & \text{si } x_i \leq x < x_{i+1} \\ \dots & \\ 1 & \text{si } x \geq x_n \end{cases}$$

On calcule ensuite la « distance de Kolmogorov Smirnov » entre ces deux fonctions.

$$d(F, \hat{F}) = \sup_{x \in \mathbb{R}} |F(x) - \hat{F}(x)|$$

Compte tenu du fait que l'une des fonctions est en escalier et que les deux fonctions coïncident à l'extérieur de l'intervalle $[0,1]$, on a :

$$d(F, \hat{F}) = \max_{i=1, \dots, n} \left\{ \left| x_i - \frac{i}{n} \right|, \left| x_i - \frac{i-1}{n} \right| \right\}$$



La valeur obtenue est alors comparée à celle lue dans la table ci-dessous correspondant au seuil choisi et au nombre de valeurs n .

Taille de l'échantillon	Niveau de signification pour la distance de Kolmogorov Smirnov				
Seuil	.20	.15	.10	.05	.01
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.392
17	.250	.266	.286	.318	.381
18	.244	.259	.278	.309	.371
19	.237	.252	.272	.301	.363
20	.231	.246	.264	.294	.356
25	.210	.220	.240	.270	.320
30	.190	.200	.220	.240	.290
35	.180	.190	.210	.230	.270
$n > 35$	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Pour soumettre vos propres séries de nombres au test de Kolmogorov Smirnov, utilisez le tableau EXCEL

KolmogorovSmirnov.XLS

Pour plus de détails sur les différents tests utilisés on peut consulter le superbe site suisse (en français)

<http://www.apprendre-en-ligne.net/random/>

ainsi que :

<http://www.hsc.fr/ressources/articles/nombre/index.html.fr#contenu>

Dans nos classes

Hasard « humain » et hasard informatique

Il n'est pas facile de produire « à la main », d'inventer une suite de chiffres aléatoires, d'imiter le vrai hasard.

Une des expériences les plus spectaculaires réalisable avec des personnes ayant une petite idée de la « loi des grands nombres » consiste à demander à une moitié de l'assistance de réaliser (ou de simuler avec une calculatrice, un ordinateur) une suite de 200 lancers de pile ou face en relevant les résultats. L'autre moitié de l'assistance « invente » une suite de 200 piles ou face en tentant de la rendre crédible.

On peut par exemple remettre aux élèves la fiche suivante :

NOM:

Voici une grille de 200 cases.

Si le jour de votre naissance est pair (comme moi qui suis né le 30 mars) remplissez « à la main » ces 200 cases de 0 ou de 1 réparties au hasard en égales proportions.

Si vous êtes né un jour impair, utilisez un dé (le lancer 200 fois, si on obtient 1, 2 ou 3 on note 0 sinon on note 1) ou une calculatrice ou une table de chiffres au hasard pour remplir les 200 cases.

Montrez votre fiche au professeur qui tentera de distinguer les grilles remplies à la main de celles remplies à l'aide de la calculatrice ou en utilisant un dé.

[illegible]

Le professeur recherche le nombre maximal de résultats consécutifs identiques dans les suites proposées. Les « tricheurs » qui ont tenté d'inventer une suite au hasard sont en général réticents à y inclure une séquence de résultats identiques d'une longueur de 5 ou 6. Pourtant, la probabilité d'obtenir 6 résultats consécutifs identiques dans une suite de 200 pile ou face est de 0.965 et l'absence d'une telle séquence permet de mettre en doute l'honnêteté du procédé employé pour la produire⁶.

⁶ Voir le calcul des probabilités dans l'annexe 3.

Le professeur qui propose cette expérience à sa classe attirera l'attention de ses élèves sur le fait qu'en matière de phénomène aléatoire tout ce qui n'est pas impossible peut théoriquement se produire mais qu'il y a néanmoins des lois. La réalisation d'un événement dont on peut démontrer qu'il est très rare devient ainsi un critère de suspicion.

Exercices autour de « vieux » générateurs de nombres au hasard

1) Mid-square

Au milieu du 20^{ième} siècle, Von Neumann proposa un des premiers générateurs de nombres aléatoires. Sa méthode était la suivante : partir d'un nombre initial de 4 chiffres, l'élever au carré et extraire les quatre chiffres du milieu.

En partant de 8965 par exemple et en appliquant l'algorithme, on obtient 3712, 7789, 6685, 6892 ...

- a) Calculer quelques termes supplémentaires. Cet algorithme vous semble-t-il satisfaisant ?
- b) Que se passe-t-il lorsque l'on choisit la valeur initiale 6783 ?

2) Fibonacci

On part de deux nombres de 4 chiffres u_0 et u_1 .

On définit la suite par $u_{n+2} = u_{n+1} + u_n$ (modulo 10000)

Comme dans la célèbre suite de Fibonacci, chaque terme est la somme des deux précédents mais ici on ne conserve que les 4 derniers chiffres (c'est le sens du modulo 10000)

En partant de 2145 et 7961 on obtient 0106, 8067, 8173, 6240,...

- a) Calculer 4 termes supplémentaires dans le cas ci-dessus. Parmi les 10 nombres de 4 chiffres ainsi engendrés, combien sont pairs ?
- b) Que se passe-t-il lorsque les deux nombres de départ sont pairs ?
- c) Cet algorithme vous semble-t-il satisfaisant pour engendrer des chiffres au hasard ?
- d) En fait, lorsqu'on ne conserve des nombres de 4 chiffres obtenus que les deux chiffres du milieu, la suite semble plus aléatoire. Expérimenter en engendrant 100 chiffres et en calculant la fréquence de chaque valeur.

Le classeur EXCEL MilieuCarré.xls présent sur le CD permet d'explorer les performances de ces deux générateurs.

Conclusion

Les suites de nombres pseudo aléatoires produites par des algorithmes (notamment ceux décrits dans ce chapitre) présentent certaines propriétés et satisfont aux tests statistiques usuels. Elles sont suffisantes pour des applications de simulation ou pour programmer des jeux. En raison de leur caractère parfaitement déterministe certains auteurs parlent à leur propos de « hasard faible ». La connaissance de l'algorithme et de la valeur de départ (le « germe ») permet de prédire l'ensemble des valeurs.

Les nombres aléatoires sont de plus en plus utilisés comme clé de codage dans des procédures de chiffrement. L'imprévisibilité est essentielle pour ces applications cryptographiques. Il doit être impossible de prédire quelle sera la valeur aléatoire suivante, connaissant complètement l'algorithme et le matériel qui engendre la suite ainsi que toutes les valeurs déjà produites. Cette qualité (au demeurant difficile à vérifier) ne peut être garantie qu'en observant un phénomène physique (désintégration radioactive, bruit thermique dans une jonction entre semi-conducteurs etc.) et en intégrant les états de ce phénomène dans le processus de génération des nombres aléatoires. Certaines sociétés fournissent des nombres aléatoires produits par de tels dispositifs⁷ et facturent leur prestation.

Une autre idée, issue de la théorie de l'information, est que le « vrai hasard » ne peut être compressé. Dans cette optique une suite de valeurs 0 ou 1 est aléatoire si pour en déterminer n termes il faut un programme dont la taille est d'au moins n bits. Une suite de nombres calculée par générateur affine ne répond manifestement pas à ce critère. L'algorithme utilisé est très simple et donc très court à décrire.

Le choix d'un générateur de nombres au hasard n'est donc pas aisé. On souhaitera, pour des applications de simulation, certaines qualités comme la rapidité (on simule pour gagner du temps), la robustesse (aucune valeur particulière ne doit provoquer de blocage ou de dégradation), l'absence de lien logique entre les valeurs engendrées (les tests décrits dans ce chapitre sont des vérifications partielles de cette dernière qualité).

Pour des besoins scolaires, les algorithmes implantés dans les calculatrices et les logiciels comme les tableurs s'avèreront suffisants. Rien ne permettra de distinguer des résultats de simulations à l'aide de ces générateurs de ceux produits par la réalisation effective des expériences.

⁷ Voir par exemple <http://www.fourmilab.ch/hotbits/> qui fournit des échantillons gratuits.

Annexe 1 : test du poker, calcul des probabilités des différentes figures

Appelons « main » tout quintuplet de cinq chiffres. Certes, dans le jeu de poker, l'ordre n'a aucune importance mais ici les dénombrements seront facilités si nous travaillons sur des quintuplets ordonnés.

Choisissons comme ensemble des possibles Ω l'ensemble de tous ces quintuplets. Le nombre d'éléments de Ω est donc $10^5 = 100000$.

Le nombre de quintuplets composés de 5 chiffres distincts du type XYZTU est : $10 \times 9 \times 8 \times 7 \times 6 = 30240$.

Le nombre de quintuplets avec 1 seul chiffre répété deux fois et 3 autres chiffres distincts (mains avec une paire du type XYZXT) est $\binom{5}{2} \times 10 \times 9 \times 8 \times 7 = 50400$, en effet $\binom{5}{2}$ est le nombre de positions possibles de la paire parmi les cinq chiffres, 10 le nombre de choix pour le chiffre doublé X et 9, 8, 7 les nombres de choix possibles pour les 3 autres chiffres distincts Y, Z et T.

Le nombre de quintuplets avec 2 chiffres répétés deux fois et 1 autre chiffre distinct des deux précédents (mains avec deux paires du type XXYYZ) est $\binom{5}{2} \times \binom{3}{2} \times 10 \times 9 \times 8 \times \frac{1}{2} = 10800$, en effet $\binom{5}{2}$ est le nombre de positions possibles de la première paire XX parmi les cinq chiffres, $\binom{3}{2}$ le nombre de positions possibles de la deuxième paire YY parmi les trois chiffres restants, 10 le nombre de valeurs possibles le chiffre X de la première paire, 9 le nombre de valeurs possibles pour le chiffre de la deuxième paire Y et 8 le nombre de choix possibles pour le chiffre isolé Z et le facteur $\frac{1}{2}$ corrige le fait que l'on a distingué artificiellement une « première » paire et une « deuxième » paire et que l'on a ainsi compté deux fois XXYYZ et YYXXZ qui sont identiques lorsqu'on permute les valeurs de X et Y.

Le nombre de quintuplets avec 1 chiffre répété trois fois et 2 autres chiffres distinct (mains avec un brelan XXYYXZ) est $\binom{5}{3} \times 10 \times 9 \times 8 = 720$, en effet $\binom{5}{3}$ est le nombre de positions possibles du chiffre triplé X parmi les cinq chiffres, 10, 9, 8 les nombres de choix possibles pour les valeurs des chiffres X, Y et Z.

Le nombre de quintuplets avec 1 chiffre répété trois fois et 1 autre chiffre répété deux fois (mains avec un full XXYYXY) est $\binom{5}{3} \times 10 \times 9 = 450$, en effet $\binom{5}{3}$ est le nombre de positions possibles du chiffre triplé X parmi les cinq chiffres, 10, 9 les nombres de choix possibles pour les valeurs des chiffres X et Y.

Le nombre de quintuplets avec 1 chiffre répété quatre fois et 1 autre chiffre (mains avec un carré XYXXXX) est $5 \times 10 \times 9 = 900$, en effet 5 est le nombre de positions possibles du

chiffre unique Y parmi les cinq chiffres, 10, 9 les nombres de choix possibles pour les valeurs des chiffres X et Y.

Le nombre de quintuplets avec 1 chiffre répété 5 fois (mains avec un poker XXXXX) est 10, le nombre de choix possibles pour la valeur du chiffre X.

On retrouve ainsi les probabilités annoncées dans le tableau du paragraphe consacré au test du poker.

Figures	Poker	Carré	Full	Brelan	2 paires	1 paires	Rien
Probabilités	0,0001	0,0045	0,0090	0,0720	0,1080	0,5040	0,3024

Annexe 2 : test des séquences monotones (« runs up and down »)

On considère une suite $(a_1, a_2, \dots, a_n)$ de n nombres distincts.

A toute permutation des termes de cette suite, on associe le nombre de « séquences monotones » dans la liste des éléments permutés.

Ainsi, si la liste des nombres est : 1 3 12 2 4 7 il y a deux séquences croissantes 1 3 12 et 2 4 7, et une séquence décroissante 12 2, soit 3 séquences monotones au total.

La variable aléatoire r , qui à chaque liste, donne le nombre de séquences monotones, peut servir à étudier la « qualité » d'une série de nombres présumés aléatoires.

Si les permutations sont choisies au hasard, la probabilité d'obtenir une liste donnée est $\frac{1}{n!}$.

On se propose de calculer ici l'espérance et la variance de r .

Remarque

On peut très bien, pour simplifier, remplacer la suite de départ par la suite des rangs (1, 2, 3, ..., n) puisqu'on ne s'intéresse qu'à l'ordre des nombres.

Soit X_i la variable aléatoire définie par :

$$X_i = \begin{cases} 1 & \text{si } a_i \text{ est le début d'une séquence monotone} \\ 0 & \text{sinon} \end{cases}$$

avec par convention $X_1 = 1$ et $X_n = 0$.

Dans l'exemple donné au début on a : $X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 0$ et $X_6 = 0$.

Le nombre de séquences monotones est dans ces conditions : $r = \sum_{i=1}^n X_i$.

Loi de probabilité de X_i et espérance mathématique de r

La valeur de X_i ne dépend que de a_{i-1} , a_i et a_{i+1} . Pour vérifier si a_i est le début d'une séquence monotone, il suffit de comparer a_i à a_{i+1} et à a_{i-1} .

De façon plus précise, si $a_i > a_{i-1}$ et $a_i > a_{i+1}$, alors a_i est le début d'une séquence décroissante et si $a_i < a_{i-1}$ et $a_i < a_{i+1}$, alors a_i est le début d'une séquence croissante. Toutes les permutations de a_{i-1} , a_i et a_{i+1} sont présentées dans le tableau suivant :

a_{i-1}	a_i	a_{i+1}	Interprétation	X_i
1	2	3	a_i n'est pas un début	0
1	3	2	a_i est le début d'une séquence décroissante	1
2	1	3	a_i est le début d'une séquence croissante	1
2	3	1	a_i est le début d'une séquence décroissante	1
3	2	1	a_i n'est pas un début	0
3	1	2	a_i est le début d'une séquence croissante	1

Il en résulte que pour $i = 2, \dots, n$, $p(X_i = 0) = \frac{1}{3}$ et $p(X_i = 1) = \frac{2}{3}$. D'où :

$$E(r) = E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i) = 1 + (n-2)\frac{2}{3} + 0 = \frac{2n-1}{3}.$$

Calcul de la variance de r

X_i étant la variable aléatoire définie précédemment, on a :

$$V(r) = E(r^2) - (E(r))^2 \text{ et } E(r^2) = E\left[\left(\sum_{i=1}^n X_i\right)^2\right].$$

Comme nous l'avons indiqué ci-dessus, pour déterminer X_i il suffit de comparer a_i à a_{i+1} et à a_{i-1} . La détermination de X_{i+3} nécessite l'examen de a_{i+2} , a_{i+3} et a_{i+4} .

$$\overbrace{a_{i-1} \ a_i \ a_{i+1}}^{\text{déterminent } X_i} \underbrace{a_{i+2} \ a_{i+3} \ a_{i+4}}_{\text{déterminent } X_{i+3}}$$

Les deux ensembles de nombres concernés étant disjoints, les variables aléatoires X_i et X_{i+3} sont indépendantes. A fortiori, X_i et X_j sont indépendantes dès que $j > i+2$.

D'où l'idée de décomposer $E(r^2)$ de la façon suivante :

$$E(r^2) = \sum_{i=1}^{n-1} E(X_i^2) + 2 \sum_{i=1}^{n-1} E(X_1 X_i) + 2 \sum_{i=2}^{n-2} E(X_i X_{i+1}) + 2 \sum_{i=2}^{n-3} E(X_i X_{i+2}) + 2 \sum_{i=2}^{n-4} \sum_{j=i+3}^{n-1} E(X_i X_j)$$

car $X_n = 0$ et donc $E(X_n) = 0$.

Or :

- $\sum_{i=1}^{n-1} E(X_i^2) = \sum_{i=1}^{n-1} E(X_i) = \frac{2n-1}{3}$ car $X_i^2 = X_i$ pour tout $i \leq n$
- $\sum_{i=1}^{n-1} E(X_1 X_i) = \sum_{i=1}^{n-1} E(X_i) = (n-2)\frac{2}{3}$ car $X_1 = 1$ pour tout $i \leq n$
- Comme dans le paragraphe précédent les 24 configurations de a_{i-1} , a_i , a_{i+1} et a_{i+2} sont données par le tableau suivant :

a_{i-1}	a_i	a_{i+1}	a_{i+2}	Interprétation	X_i	X_{i+1}	$X_i X_{i+1}$
1	2	3	4	$a_i \ a_{i+1}$ ne sont pas des débuts	0	0	0
1	2	4	3	a_{i+1} est un début mais pas a_i	0	1	0
1	3	2	4	$a_i \ a_{i+1}$ sont des débuts	1	1	1
1	3	4	2		0	1	0
1	4	3	2	a_i est un début mais pas a_{i+1}	1	0	0
1	4	2	3		1	1	1
2	1	3	4		1	0	0
2	1	4	3		1	1	1
2	3	1	4		1	1	1
2	3	4	1		0	1	0
2	4	3	1		1	0	0
2	4	1	3		1	1	1
3	2	1	4		0	1	0
3	2	4	1		1	1	1
3	1	2	4		1	0	0
3	1	4	2		1	1	1
3	4	1	2		1	1	1
3	4	2	1		1	0	0

4	2	3	1		1	1	1
4	2	1	3		0	1	0
4	3	2	1		0	0	0
4	3	1	2		0	1	0
4	1	3	2		1	1	1
4	1	2	3		1	0	0
					total :		
					10		

Ainsi $\sum_{i=2}^{n-2} E(X_i X_{i+1}) = (n-3) \frac{10}{24}$.

- De même les 120 configurations de a_{i-1} , a_i , a_{i+1} , a_{i+2} et a_{i+3} sont données par le tableau suivant⁸ :

a_{i-1}	a_i	a_{i+1}	a_{i+2}	a_{i+3}	Interprétation	X_i	X_{i+2}	$X_i X_{i+2}$
1	2	3	4	5	$a_i a_{i+2}$ ne sont pas des débuts	0	0	0
1	2	3	5	4	a_{i+2} est un début mais pas a_i	0	1	0
1	2	4	3	5		0	1	0
1	2	4	5	3		0	1	0
1	2	5	4	3		0	0	0
1	2	5	3	4		0	1	0
1	3	2	4	5		1	0	0
1	3	2	5	4	$a_i a_{i+2}$ sont des débuts	1	1	1
1	3	4	2	5		0	1	0
1	3	4	5	2		0	1	0
1	3	5	4	2		0	0	0
1	3	5	2	4		0	1	0
1	4	3	2	5		1	1	1
1	4	3	5	2		1	1	1
1	4	2	3	5		1	0	0
1	4	2	5	3		1	1	1
1	4	5	2	3		0	1	0
1	4	5	3	2		0	0	0
1	5	3	4	2		1	1	1
1	5	3	2	4		1	1	1
1	5	4	3	2		1	0	0
1	5	4	2	3		1	1	1
1	5	2	4	3		1	1	1
1	5	2	3	4		1	0	0
2	1	3	4	5		1	0	0
2	1	3	5	4		1	1	1
2	1	4	3	5		1	1	1
2	1	4	5	3		1	1	1
2	1	5	4	3		1	0	0
2	1	5	3	4		1	1	1
2	3	1	4	5		1	0	0
2	3	1	5	4		1	1	1
2	3	4	1	5		0	1	0
2	3	4	5	1		0	1	0
2	3	5	4	1		0	0	0
2	3	5	1	4		0	1	0
2	4	3	1	5		1	1	1

⁸ L'examen attentif de l'ensemble de ce tableau peut être omis en première lecture. (humour !)

2	4	3	5	1
2	4	1	3	5
2	4	1	5	3
2	4	5	1	3
2	4	5	3	1
2	5	3	4	1
2	5	3	1	4
2	5	4	3	1
2	5	4	1	3
2	5	1	4	3
2	5	1	3	4
3	2	1	4	5
3	2	1	5	4
3	2	4	1	5
3	2	4	5	1
3	2	5	4	1
3	2	5	1	4
3	1	2	4	5
3	1	2	5	4
3	1	4	2	5
3	1	4	5	2
3	1	5	4	2
3	1	5	2	4
3	4	1	2	5
3	4	1	5	2
3	4	2	1	5
3	4	2	5	1
3	4	5	2	1
3	4	5	1	2
3	5	1	4	2
3	5	1	2	4
3	5	4	1	2
3	5	4	2	1
3	5	2	4	1
3	5	2	1	4
4	2	3	1	5
4	2	3	5	1
4	2	1	3	5
4	2	1	5	3
4	2	5	1	3
4	2	5	3	1
4	3	2	1	5
4	3	2	5	1
4	3	1	2	5
4	3	1	5	2
4	3	5	1	2
4	3	5	2	1
4	1	3	2	5
4	1	3	5	2
4	1	2	3	5
4	1	2	5	3
4	1	5	2	3
4	1	5	3	2
4	5	3	1	2
4	5	3	2	1
4	5	1	3	2

1	1	1
1	0	0
1	1	1
0	1	0
0	0	0
1	1	1
1	1	1
1	0	0
1	1	1
1	1	1
1	0	0
0	0	0
0	1	0
1	1	1
1	1	1
1	0	0
1	1	1
1	0	0
1	1	1
1	1	1
1	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	0	0
1	1	1
1	0	0
1	1	1
1	1	1
1	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	0	0
0	1	0
0	0	0
0	1	0
1	1	1
1	0	0
1	1	1
1	1	1
1	0	0
1	1	1
1	1	1
1	0	0
1	1	1
1	1	1
1	0	0
1	1	1
1	0	0
1	1	1
1	1	1

4	5	1	2	3		1	0	0
4	5	2	1	3		1	1	1
4	5	2	3	1		1	1	1
5	2	3	4	1		1	1	1
5	2	3	1	4		1	1	1
5	2	4	3	1		1	0	0
5	2	4	1	3		1	1	1
5	2	1	4	3		0	1	0
5	2	1	3	4		0	0	0
5	3	2	4	1		0	1	0
5	3	2	1	4		0	1	0
5	3	4	2	1		1	0	0
5	3	4	1	2		1	1	1
5	3	1	4	2		0	1	0
5	3	1	2	4		0	0	0
5	4	3	2	1		0	0	0
5	4	3	1	2		0	1	0
5	4	2	3	1		0	1	0
5	4	2	1	3		0	1	0
5	4	1	2	3		0	0	0
5	4	1	3	2		0	1	0
5	1	3	4	2		1	1	1
5	1	3	2	4		1	1	1
5	1	4	3	2		1	0	0
5	1	4	2	3		1	1	1
5	1	2	4	3		1	1	1
5	1	2	3	4		1	0	0
						total :		54

Ainsi $\sum_{i=2}^{n-3} E(X_i X_{i+2}) = (n-4) \frac{54}{120}$ dès que $n > 4$ et vaut 0 pour $n = 4$.

- Si $j > i + 2$ alors X_i et X_j sont indépendants et par suite :

$$\sum_{i=2}^{n-4} \sum_{j=i+3}^{n-1} E(X_i X_j) = \sum_{i=2}^{n-4} \sum_{j=i+3}^{n-1} E(X_i) E(X_j) = [(n-5) + (n-4) + 0 + \dots + 0] \left(\frac{2}{3}\right)^2$$

dès que $n \geq 5$ et donc :

$$\sum_{i=2}^{n-4} \sum_{j=i+3}^{n-1} E(X_i X_j) = \frac{(n-5)(n-4)}{2} \left(\frac{2}{3}\right)^2 = (n-5)(n-4) \frac{2}{9}.$$

Dans ces conditions pour tout $n \geq 4$:

$$\begin{aligned} E(r^2) &= \frac{2n-1}{3} + 2(n-2) \frac{2}{3} + 2(n-3) \frac{10}{24} + 2(n-4) \frac{54}{120} + 2(n-5)(n-4) \frac{2}{9} \\ &= \frac{40n^2 - 24n - 19}{90} \end{aligned}$$

d'où :

$$V(r) = \frac{40n^2 - 24n - 19}{90} - \left(\frac{2n-1}{3}\right)^2 \text{ et finalement : } \boxed{V(r) = \frac{16n-29}{90}}.$$

Remarque

Cette formule n'est pas vraie pour $n = 2$ ou $n = 3$.

Annexe 3 : nombre maximal de résultats consécutifs égaux dans une suite de 200 « pile ou face »

Dans une série de 200 lancers de pile ou face, la présence d'une suite d'au moins 6 résultats consécutifs identiques (au moins 6 piles ou 6 faces consécutifs) est très probable. Ce résultat, qui heurte l'intuition de nombreuses personnes, permet de tester si des suites de 0 et de 1 ont bien été engendrées au hasard, ou « inventées » par un humain.

Nous proposons ci-dessous deux méthodes de calcul de la probabilité concernée. La première utilise les dénombrements et aboutit à une formule de récurrence sur le nombre de n -uplets ne comportant aucune suite de 6 résultats consécutifs identiques, la seconde permet d'établir directement une formule de récurrence sur les probabilités. Les deux procédés de calcul aboutissent bien entendu aux mêmes résultats.

Récurrence sur les dénombrements

Dans une série de n lancers de pile ou face, la présence d'une suite d'au moins 6 résultats consécutifs identiques (au moins 6 piles ou 6 faces consécutifs) est-elle exceptionnelle ?

Codons 0 et 1 les deux résultats pile ou face et considérons tous les n -uplets de 0 et de 1. Le nombre de tels n -uplets est 2^n . Notons s_n le nombre de n -uplets ne comportant aucune suite de 6 résultats consécutifs identiques.

Les premières valeurs de s_n sont aisées à calculer :

$$\begin{aligned}s_1 &= 2 \quad (\text{les deux singletons } 0 \text{ et } 1) \\s_2 &= 4 \quad (\text{les 4 couples } 00, 01, 10, 11) \\s_3 &= 8 \quad (\text{les 8 triplets } 000, 001, 010, 011, 1000, \dots)\end{aligned}$$

Comme ce n'est qu'à partir d'une série de 6 valeurs que l'on peut obtenir éventuellement une

suite de 6 résultats consécutifs identiques, on complète le calcul : $s_4 = 16$ et $s_5 = 32$.

Parmi les 64 séries de 6 valeurs il y en a deux qui comportent une suite de 6 résultats consécutifs identiques : 000000 et 111111. On a donc $s_6 = 62$.

La probabilité qui permettra de répondre à la question de départ est : $\frac{s_{200}}{2^{200}}$.

Pour $n \geq 6$ répartissons les s_n n -uplets $(b_1, b_2, \dots, b_{n-1}, b_n)$ ne comportant aucune suite de 6 résultats consécutifs identiques en 5 catégories :

1. ceux qui se terminent par deux valeurs distinctes donc tels que $b_n \neq b_{n-1}$ qui sont de deux types : $(b_1, b_2, \dots, b_{n-2}, 0, 1)$ ou $(b_1, b_2, \dots, b_{n-2}, 1, 0)$. Dans les deux cas, le $(n-1)$ -uplet $(b_1, b_2, \dots, b_{n-1})$ ne comporte aucune suite de 6 résultats consécutifs identiques. Il y a donc s_{n-1} tels $(n-1)$ -uplets et chacun d'eux se complète de manière unique en un n -uplet de la présente catégorie.
2. ceux qui se terminent par deux valeurs identiques exactement, donc tels que $b_n = b_{n-1} \neq b_{n-2}$ qui sont de deux types : $(b_1, b_2, \dots, b_{n-3}, 1, 0, 0)$ ou $(b_1, b_2, \dots, b_{n-3}, 0, 1, 1)$. Dans les deux cas, le $(n-2)$ -uplet $(b_1, b_2, \dots, b_{n-2})$ ne comporte aucune suite de 6 résultats consécutifs identiques. Il y a donc s_{n-2} tels $(n-2)$ -uplets

et chacun d'eux se complète de manière unique en un n -uplet de la présente catégorie.

3. ceux qui se terminent par trois valeurs identiques exactement donc tels que $b_n = b_{n-1} = b_{n-2} \neq b_{n-3}$ qui sont de deux types : $(b_1, b_2, \dots, b_{n-4}, 1, 0, 0, 0)$ ou $(b_1, b_2, \dots, b_{n-4}, 0, 1, 1, 1)$. Dans les deux cas, le $(n-3)$ -uplet $(b_1, b_2, \dots, b_{n-3})$ ne comporte aucune suite de 6 résultats consécutifs identiques. Il y a donc s_{n-3} tels $(n-3)$ -uplets et chacun d'eux se complète de manière unique en un n -uplet de la présente catégorie.
4.

On l'aura compris, un raisonnement analogue pour les n -uplets se terminant par 4 ou 5 valeurs identiques permet d'établir la relation de récurrence :

$s_n = s_{n-1} + s_{n-2} + s_{n-3} + s_{n-4} + s_{n-5}$ ce qui avec les valeurs initiales calculées directement ci-dessus permet le calcul de proche en proche des valeurs de s_n .

Formule de récurrence sur les probabilités

L'expérience qui consiste à répéter le lancer d'une pièce de monnaie peut être modélisée par une suite (X_n) de variables aléatoires indépendantes à valeurs dans $\{0, 1\}$ telle que P

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

La suite (Y_N, n) , stable dès qu'elle atteint la valeur N et qui donne, tant que N n'est pas atteint, la longueur de la dernière série de coups consécutifs égaux lors de n lancers est intéressante car elle permettra de répondre au problème posé.

Exemples :

Valeurs de $(Y_{3,n})$

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
X_n	0	1	1	0	1	0	0	1	1	0	0	1	1	1	1	0	0	1	0
Y_n	1	1	2	1	1	1	2	1	2	1	2	1	2	3	3	3	3	3	3

Valeurs de $(Y_{5,n})$ pour les mêmes valeurs de (X_n)

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
X_n	0	1	1	0	1	0	0	1	1	0	0	1	1	1	1	0	0	1	0
Y_n	1	1	2	1	1	1	2	1	2	1	2	1	2	3	4	1	2	1	1

$(Y_{N,n})$ est définie par :

$$Y_{N,1} = 1$$

$$Y_{N,n} = \begin{cases} N & \text{si } Y_{N,n} = N \\ Y_{N,n-1} + 1 & \text{si } Y_{N,n} < N \text{ et } X_n = X_{n-1} \\ 1 & \text{sinon} \end{cases}$$

$P(Y_{N,n} = N)$ est la probabilité pour que le nombre maximum de coups consécutifs égaux soit supérieur ou égal à N ; par contre $P(Y_{N,n} = k)$ n'est pas exploitable lorsque k est différent de N .

$(Y_{N,n})$ est à valeurs dans $\{1, 2, \dots, N\}$ et $Y_{N,n}$ ne dépend que de l'état antérieur $Y_{N,n-1}$ puisque (X_n) est une suite de variables aléatoires indépendantes. Une telle suite est encore appelée chaîne de Markov.

Soit $p_n = P(Y_{N,n} = N)$. On a : $p_{N,1} = p_{N,2} = \dots = p_{N,N-1} = 0$ et $p_{N,N} = \frac{2}{2^N}$.

Si on lance n fois la pièce avec $n > N$ alors $Y_{N,n} = N$ se réalise lorsque $Y_{N,n-1} = N$ ou lorsque les 3 conditions suivantes sont simultanément remplies :

- mêmes résultats pour les lancers de rang $n-1, n-2, \dots, n-N+1$; probabilité $1/2^{N-1}$
- résultat du dernier lancer identique au résultat du $n-1^{\text{e}}$ lancer ; probabilité $1/2$
- pas de série de N résultats identiques avant puisque ces séries sont déjà comptabilisées dans $\{Y_{N,n-1} = N\}$; probabilité : $1 - p_{n-N}$.

D'après la formule des probabilités totales on a donc pour tout $n > N$:

$$p_n = p_{n-1} + \frac{1}{2^N}(1 - p_{n-N}).$$

On peut ainsi calculer de proche en proche p_n pour $n > N$ sur tableur ou calculatrice, il suffit pour cela de garder en mémoire les N derniers p_n .

Quelques extraits du tableau des valeurs numériques établi avec EXCEL pour $N=6$:

n	s_n	$p_n = s_n/2^n$
1	2	1
2	4	1
3	8	1
4	16	1
5	32	1
6	62	0,96875
7	122	0,953125
8	240	0,9375
9	472	0,921875
10	928	0,90625
11	1824	0,890625
12	3586	0,87548828
13	7050	0,8605957
14	13860	0,84594727

15	27248	0,83154297
16	53568	0,81738281
17	105312	0,8034668
18	207038	0,78978729
19	407026	0,77634048

.....

97	3,2229E+28	0,20339238
98	6,336E+28	0,19992945
99	1,2456E+29	0,19652547
100	2,4488E+29	0,19317945
101	4,8143E+29	0,1898904
102	9,4647E+29	0,18665735
103	1,8607E+30	0,18347934

.....

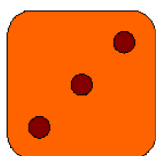
197	7,3359E+57	0,03652103
198	1,4422E+58	0,03589922
199	2,8353E+58	0,03528801
200	5,574E+58	0,0346872
201	1,0958E+59	0,03409662
202	2,1543E+59	0,03351609
203	4,2353E+59	0,03294545

.....

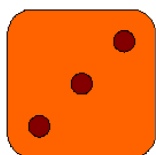
297	1,6698E+87	0,0065577
298	3,2827E+87	0,00644605
299	6,4536E+87	0,0063363
300	1,2688E+88	0,00622841

Dans une série de 200 lancers de pile ou face, la probabilité de trouver au moins une suite de 6 résultats consécutifs identiques est environ $1 - 0,0346872$ soit 0,965127. Compte tenu de l'ordre de grandeur des nombres concernés il est plus sage de se contenter de la valeur approchée 0,965. L'événement est donc très probable.

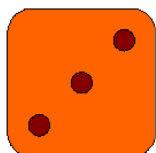
III. LA LOI DES GRANDS NOMBRES



Des énoncés vulgarisés



Des énoncés non vulgarisés



Et que dire à nos élèves ?

INTRODUCTION

Le programme officiel des classes de premières S et ES précise :

« Le lien entre loi de probabilité et distributions de fréquences sera éclairé par un énoncé vulgarisé de la loi des grands nombres. »

La consigne est brève et semble claire mais sa mise en œuvre pour les élèves est un défi énorme. Le théorème de la théorie des probabilités connu sous le nom de « loi des grands nombres » est l'un des plus difficiles à vulgariser qui soit. Il en existe principalement deux versions, mais ni la convergence presque sûre (version forte) ni la convergence en loi (version faible) ne sont aisées à rendre accessibles au public scolaire.

Nous contenter d'une version caricaturale (trop vulgarisée) risque d'entraîner chez nos élèves un renforcement de fausses conceptions et superstitions diverses. A l'inverse, vouloir être trop précis nous rendra incompréhensibles.

DES ÉNONCÉS VULGARISÉS

Nous avons classé les énoncés vulgarisés de la loi des grands nombres que nous citons en quatre types :

Type 1 : la distribution de fréquences « tend » vers la loi de probabilité.

Type 2 : la dispersion des fréquences observées diminue quand le nombre d'expériences augmente.

Type 3 : la fréquence « tend » vers la probabilité

Type 4 : les fréquences se stabilisent (sans modèle a priori)

Reprenons chacun de ces points de vue en l'illustrant d'exemples qui semblent s'y rattacher :

Type 1 : distribution de fréquences qui « tend » vers la loi de probabilité

C'est le point de vue du programme :

« Pour une expérience donnée, dans le modèle défini par une loi de probabilité P , les distributions des fréquences calculées sur des séries de taille n se rapprochent de P quand n devient grand. »

Ci-dessous un exemple de simulation qui illustre ce « rapprochement ».

On lance (un certain nombre de fois) deux dés, on calcule la somme des points obtenus puis on trace le diagramme en bâtons des fréquences en parallèle avec le diagramme des probabilités.

Répéter 25 fois l'expérience, ne permet pas de mettre en doute les modèles inadaptés parfois proposés par les élèves : Le modèle équiprobable (probabilité 1/12 puis 1/11 pour toutes les sommes possibles...) qu'un groupe d'élèves abandonnera assez rapidement et le modèle « commutatif » qui ne distingue pas les deux dés dont la remise en question est plus difficile à obtenir.

Modèle équiprobable :

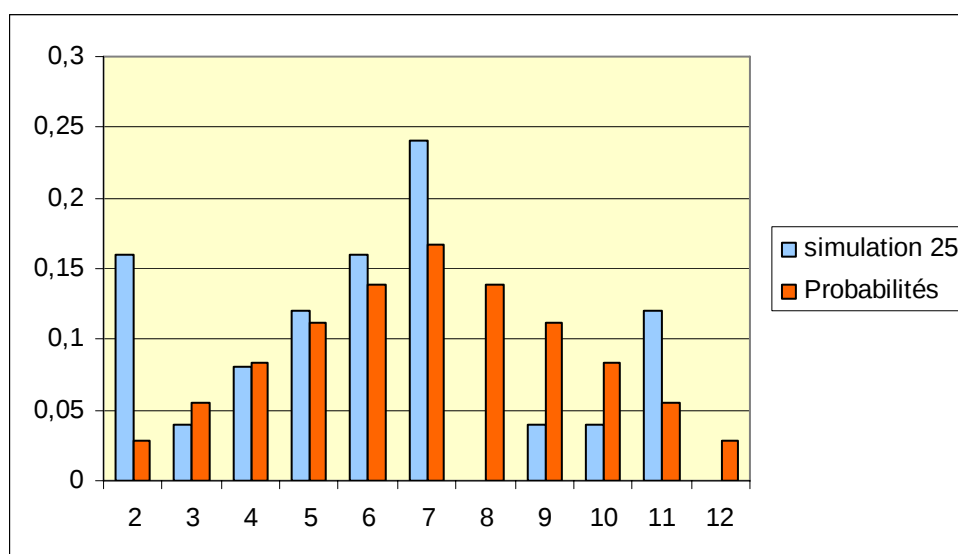
Somme	2	3	4	5	6	7	8	9	10	11	12
Probas	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11

Modèle « commutatif » :

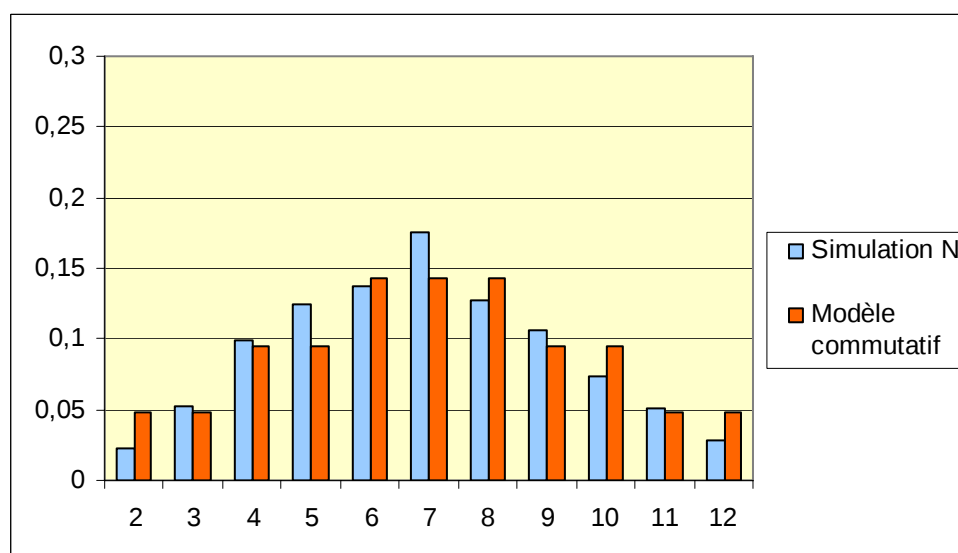
Somme	2	3	4	5	6	7	8	9	10	11	12
Probas	1/21	1/21	2/21	2/21	3/21	3/21	3/21	2/21	2/21	1/21	1/21

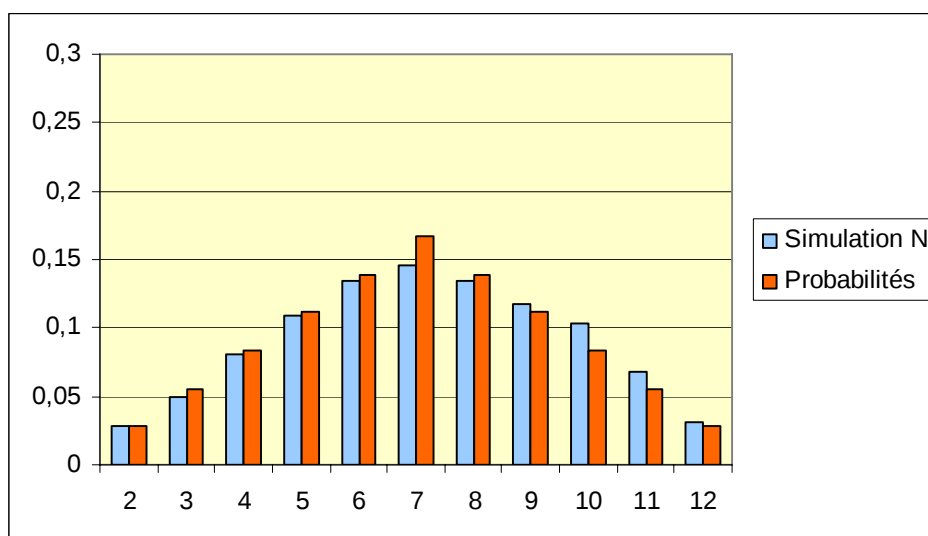
La simulation de 1000 expériences permet de convaincre les élèves que le « meilleur modèle » est celui où l'on distingue les deux dés.

25 lancers de dés



1000 lancers du dé





Un classeur EXCEL qui réalise ces simulations DeuxDesModele.XLS est disponible sur le CD d'accompagnement (il faut accepter d'activer les macros).

De nombreux manuels adoptent ce type d'énoncé, certains en reprenant mot à mot la formulation du programme officiel, d'autres en l'assortissant de précisions supplémentaires :

Hyperbole Nathan 1S

« Pour une expérience donnée, dans le modèle défini par une loi de probabilité, les distributions des fréquences calculées sur des séries de taille n se rapprochent de la loi de probabilité quand n devient grand »

Déclic Hachette 1S

« On peut démontrer (et on admet ici) que, pour une expérience donnée, dans le modèle défini par une loi de probabilité P , les distributions des fréquences calculées sur des séries de taille n se rapprochent de P quand n devient grand. »

Les auteurs mettent ici clairement en évidence le fait qu'il s'agit d'un théorème qui se démontre.

Belin 1S

« Propriété 7 (admise)

Soit un ensemble Ω sur lequel on a défini une loi de probabilité P . Les fréquences des éléments de Ω dans une suite de n expériences identiques tendent vers leur probabilité quand n augmente indéfiniment. »

Les « éléments de Ω » sont sans doute les événements élémentaires, le terme de « suite » induit une notion d'ordre qui n'est pas nécessaire ici où seul le nombre d'expériences est important.

Fractale 1S

« Propriété 2

La loi des grands nombres

Soit U un univers d'éventualités pour une expérience donnée,

$$U = \{e_1, e_2, e_3, \dots, e_n\},$$

et une loi de probabilité p , qui fixe les conditions de réalisation de chaque éventualité : $p(e_i) = p_i$. Si on réalise k fois l'expérience, on obtient une distribution des fréquences de réalisations de chaque éventualité $f(e_i) = f_i$. La fréquence de e_i se rapproche de p_i quand k devient grand. »

Ces deux dernières formulations sont très proches de celle du document d'accompagnement :

Lexique du document d'accompagnement

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

Type 2 : la dispersion des fréquences observées diminue quand le nombre d'expériences augmente

Ce point de vue est également cité dans la colonne commentaires du programme officiel :

On pourra aussi illustrer cette loi (des grands nombres) avec les diagrammes en boîtes obtenus en simulant par exemple 100 sondages de taille n pour $n=10$; 100 ; 1000.

C'est ce qu'exprime avec concision Michel Brissaud⁹

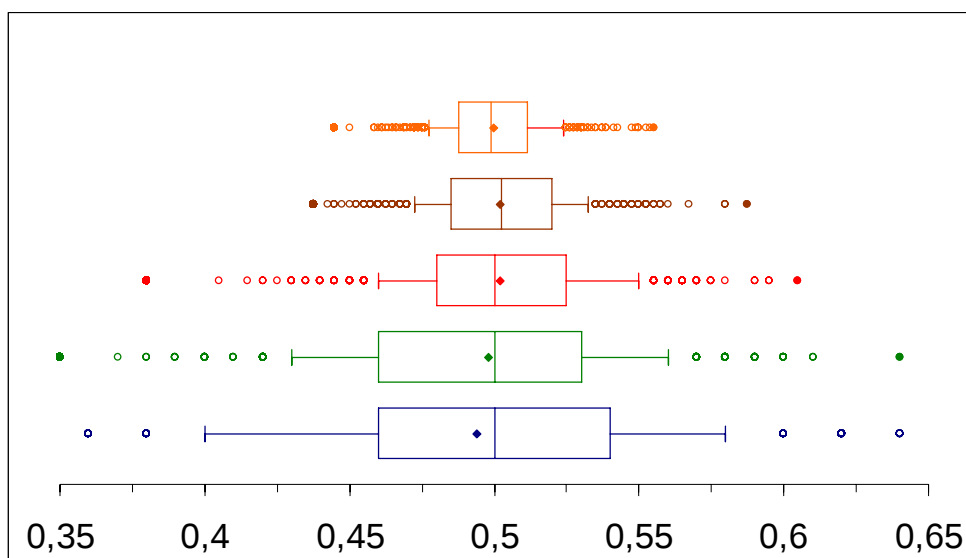
« On constate expérimentalement qu'en répétant N fois une même expérience la dispersion sur les fréquences diminue quand N augmente. »

Illustration par une simulation à l'aide d'un tableur : on répète 500 fois n lancers de pile ou face et on réalise les diagrammes en boîtes des répartitions des fréquences.

De bas en haut les diagrammes en boîtes obtenus pour $n=50, 100, 200, 400, 800$

Sur ces diagrammes : le losange représente la moyenne, les petits cercles donnent les valeurs extérieures à l'intervalle premier décile / neuvième décile, les petits disques les valeurs extrêmes, minimum et maximum qui ne sont plus visibles pour $n=50$.

⁹ http://www.ac-nancy-metz.fr/enseign/maths/apmep/PV_69_Brissaud.htm



Un classeur EXCEL qui réalise ces simulations et ces diagrammes en boîtes (dont nous devons le design à Michèle Chagnard, Bernard Langer et Gilbert Le Cam) nommé **DispersionDesFréquences.XLS** est disponible sur le CD d'accompagnement (il faut accepter d'activer les macros).

L'intérêt de cette formulation est de mettre en évidence qu'une seule fréquence, même calculée à partir d'un grand nombre d'expériences, n'a qu'une signification relative puisque des valeurs éloignées de la probabilité, bien que peu probables, n'en restent pas moins possibles.

Une difficulté pour les élèves réside dans l'indispensable double répétition puisqu'on reproduit N fois une expérience qui elle-même consiste à répéter n fois une épreuve.

Assez peu de manuels présentent les choses ainsi. Citons toutefois :

Bréal 1S et 1ES

« ...mais surtout on a vu en seconde que pour n assez grand, $f_E(A)$ a de fortes chances d'être une valeur approchée de la proportion de boules correspondant à A dans l'urne de modélisation, c'est à dire une valeur approchée de $P(A)$.

Plus précisément, on peut démontrer les résultats suivants :

la probabilité que $f_E(A)$ soit comprise entre $P(A) - \frac{1}{\sqrt{n}}$ et $P(A) + \frac{1}{\sqrt{n}}$:

- dépasse 90% pour tout n ;
- dépasse 93% pour tout n supérieur à 25 ;
- dépasse 95% pour tout n supérieur à 530.

C'est d'ailleurs cette propriété qui permet d'estimer statistiquement les probabilités des issues. C'est aussi elle qui fournit le critère ultime pour juger de la validité d'une modélisation.

Avec dans la marge :

C'est une forme de ce qu'on appelle le **théorème des grands nombres**. L'intervalle $\left[P(A) - \frac{1}{\sqrt{n}}; P(A) + \frac{1}{\sqrt{n}} \right]$ est l'**intervalle de dispersion**. »

Il est ici affirmé clairement qu'il y a une probabilité que la fréquence observée soit dans un intervalle de dispersion autour de la moyenne et l'amplitude de l'intervalle est quantifiée. L'affirmation « on peut démontrer » attribue clairement au résultat le statut de théorème. Elle fut également un défi pour nous.

Démonstration dans le cas du jeu de pile ou face : $f_E(A) = 0.5$

La probabilité que $f_E(A)$ soit comprise entre $P(A) - \frac{1}{\sqrt{n}}$ et $P(A) + \frac{1}{\sqrt{n}}$:

– dépasse 93% pour tout n supérieur à 25 ;

Le nombre de « pile » obtenu lors de 25 lancers d'une pièce est une variable aléatoire qui suit la loi binomiale $B(25 ; 0.5)$. Pour calculer les probabilités d'obtenir les différentes fréquences possibles recherchons la 25^{ème} ligne du triangle de Pascal (du moins la première moitié de cette ligne puisqu'elle est symétrique)

L'intervalle de dispersion est dans ce cas : $[0.3 ; 0.7]$ et la somme des probabilités donnant des fréquences de cet intervalle donne 0.9567

Cnp	1	25	300	2300	12650	53130	177100	480700	1081575	2042975	3268760	4457400	5200300
Fréquence	0	0,04	0,08	0,12	0,16	0,2	0,24	0,28	0,32	0,36	0,4	0,44	0,48
dans $[0.3, 0.7]$	0	0	0	0	0	0	0	0	0,03223	0,06089	0,09742	0,13284	0,15498

– dépasse 95% pour tout n supérieur à 530.

Le nombre d'expériences est suffisant pour utiliser l'approximation de la loi binomiale par une loi normale :

Le nombre de « pile » X suit la loi binomiale $B(530 ; 0.5)$ qui est « proche » de la loi normale de même moyenne et de même écart type soit $N(265 ; 11.51)$

L'intervalle de dispersion est dans ce cas $[0.45656 ; 0.54344]$ ce qui correspond à un nombre de « pile » dans $[242 ; 288]$

$$242 \leq X \leq 288 \Leftrightarrow \frac{242 - 265}{11.51} \leq \frac{X - 265}{11.51} \leq \frac{288 - 265}{11.51} \Leftrightarrow -2 \leq \frac{X - 265}{11.51} \leq 2$$

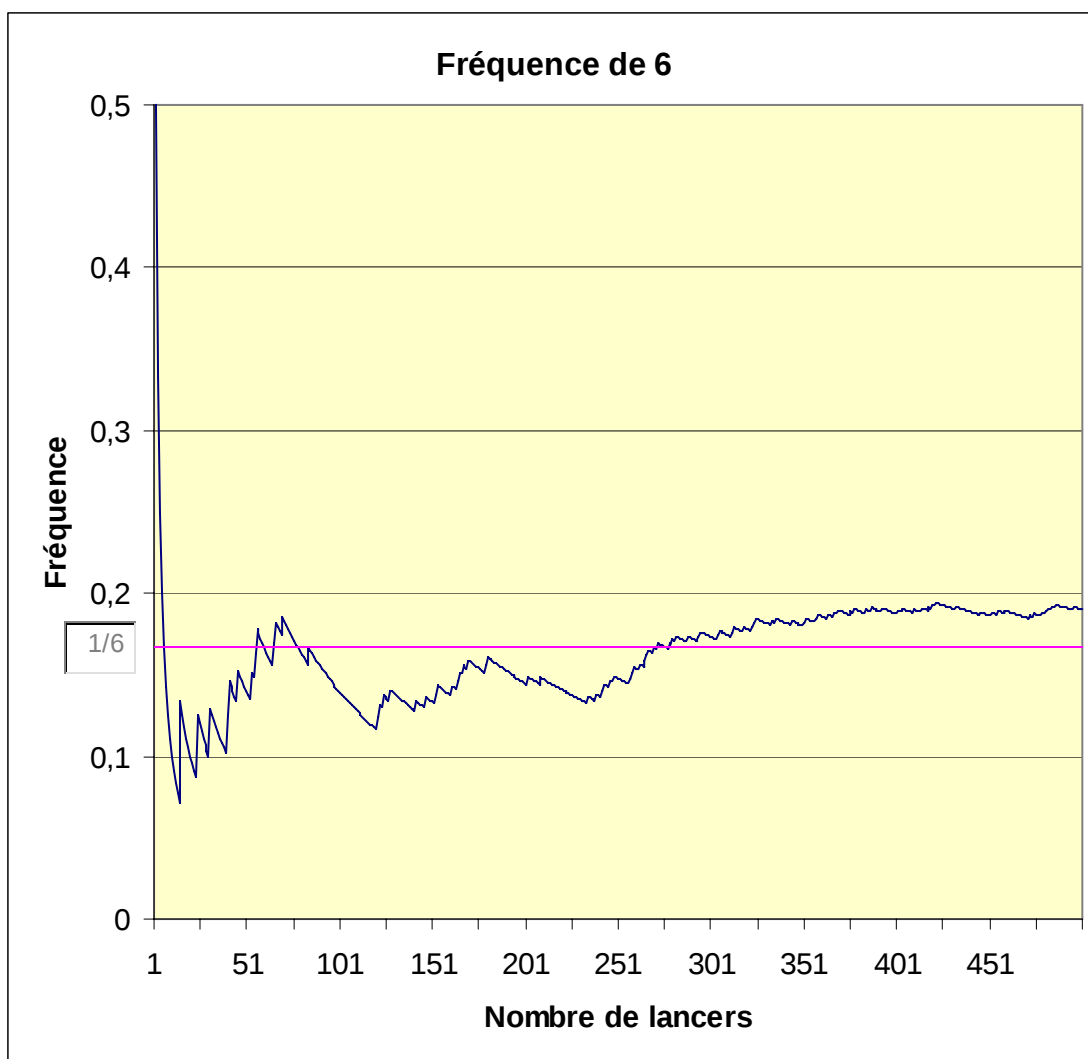
La probabilité qu'une variable aléatoire suivant la loi normale centrée réduite soit dans l'intervalle $[-2 ; +2]$ est égale à $2\Pi(2) - 1 \approx 2 \times 0.9772 - 1 = 0.954$ donc supérieure à 0.95.

(où Π est la fonction de répartition de la loi normale centrée réduite dont on trouve les valeurs dans une table)

Type 3 : la fréquence qui « tend » vers la probabilité

C'est une formulation qui semble très simple mais risque d'induire de nombreuses fausses conceptions. Ian Stewart, dans un article de la revue « Pour la Science » exprime la loi des grands nombres ainsi : « *la fréquence d'un événement tend vers sa probabilité quand le nombre d'expériences augmente indéfiniment* »

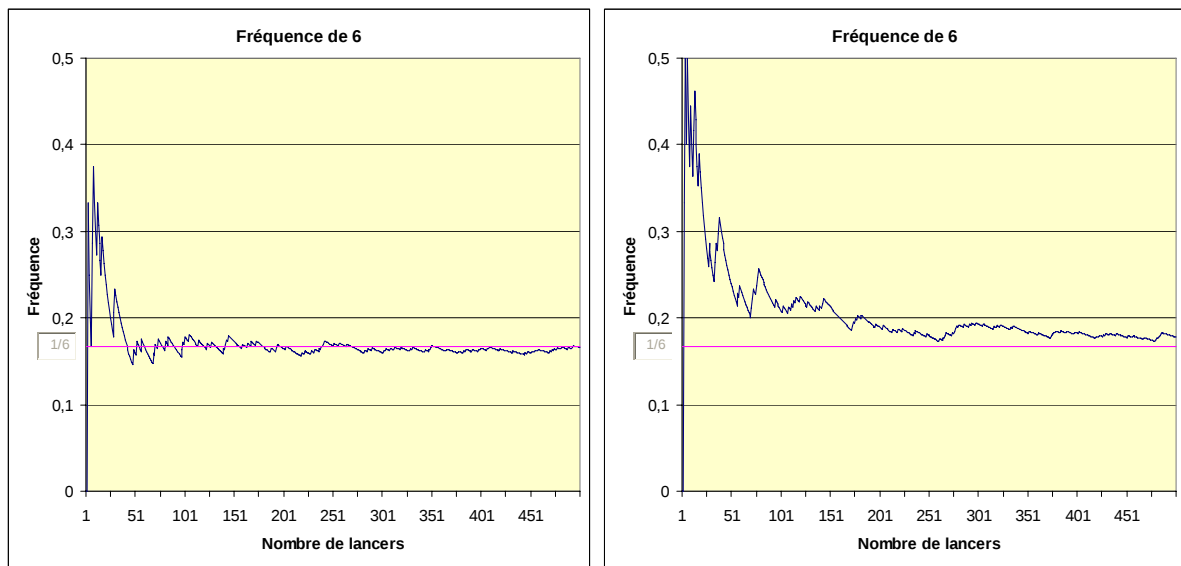
Une simulation simple permet une illustration de ce point de vue : pendant une répétition d'une expérience on calcule et on affiche l'évolution de la fréquence. Par exemple les graphiques ci-dessous donnent l'évolution de la fréquence de « 6 » lors de 500 lancers d'un dé.



La « convergence » est, d'après ce type de graphiques, assez « médiocre ». Dans le cas ci-dessus une probabilité de 0.2 semblerait aussi raisonnable que $1/6$.

Cette formulation, surtout si l'on utilise le verbe « tendre », risque de pousser les élèves à un amalgame avec la situation de l'asymptote horizontale, ou de la convergence d'une suite en analyse. Les idées fausses déjà présentes en analyse se retrouveront ici. La « convergence » de la fréquence sera imaginée comme nécessairement monotone, ou tout du moins de plus en plus proche de la probabilité.

La multiplication d'exemples (voir ci-dessous) dont l'aspect « fluctuera » beaucoup suffira peut-être à ébranler ces fausses conceptions.



Un classeur EXCEL qui réalise ces simulations **Fréquence du 6.XLS** est disponible sur le CD d'accompagnement (il faut accepter d'activer les macros).

Des énoncés de ce type sont présents dans certains manuels :

Transmath 1S

« La fréquence moyenne d'apparition d'un résultat dans une répétition d'épreuves tend vers la probabilité d'observer cette apparition dans une seule épreuve »

La phrase utilisée ne précise pas que c'est lorsqu'on répète un grand nombre de fois l'expérience et on peut se demander si l'utilisation du mot « apparition » est très heureuse. Par ailleurs que signifie exactement le terme « fréquence moyenne »

Transmath 1ES

(Les notions sont présentées sur un exemple : le lancer d'un dé)

« Loi des grands nombres

Notons F_n la fréquence de sortie du 1, par exemple, en n lancers.

Intuitivement, « plus le nombre n de lancers est grand, plus F_n se rapproche de la probabilité théorique de l'événement $\{1\}$, c'est-à-dire $\frac{1}{6}$. »

Cette propriété, présentée ici de manière intuitive, est appelée loi des grands nombres.

Elle peut permettre de contrôler si le modèle proposé a été bien choisi.

Parfois, on n'est pas en mesure de proposer un modèle a priori. C'est le cas, par exemple, d'un dé pipé. La loi des grands nombres permet alors, a posteriori, de proposer un modèle : on prend comme probabilité de chaque événement élémentaire un nombre vers lequel

semble « tendre » la fréquence de cet événement lorsque le nombre n d'expériences devient de plus en plus grand. »

Ce dernier paragraphe nous fournit une transition vers le quatrième type d'énoncés.

Type 4 : les fréquences se stabilisent (sans modèle a priori)

Dans cette optique, on insiste sur l'utilisation de la loi des grands nombres pour proposer un modèle dans les cas (fréquents) où la description de l'expérience ne permet pas de valider un modèle a priori. Le problème de l'adéquation du modèle se pose alors tout naturellement.

Ce point de vue est adopté dans plusieurs manuels :

Collection Terracher, Hachette 1èreS

« La loi des grands nombres :

Le résultat théorique suivant permet a posteriori de valider ou de remettre en question le choix d'un modèle $(p_1, \dots, p_r)$ associé à une expérience aléatoire : Si le modèle est bon, les fréquences f_i , calculées sur des séries de taille n , doivent se rapprocher des p_i lorsque n devient grand. »

Le problème de la validité du modèle est ici explicitement posé.

Dimathème 1S

« (Énoncé expérimental)

Quand on répète un grand nombre de fois une expérience aléatoire pouvant conduire à des résultats $a_1, \dots, a_n$, la fréquence de réalisation de chaque événement élémentaire $\{a_i\}$ se stabilise aux environs d'un nombre p_i compris entre 0 et 1. Ce nombre peut être considéré comme la probabilité de réalisation de l'événement $\{a_i\}$. »

Indice 1S

« La loi « des grands nombres »

Propriété

Si on répète n fois la **même** expérience de manière **indépendante** dans les **mêmes conditions**, et si n tend vers l'infini, la **fréquence** d'une issue tend vers une valeur théorique p , appelée **probabilité** de cette issue.

On dit que la distribution des fréquences $\{f_1, f_2, f_3, \dots, f_k\}$ tend vers la loi de probabilité $\{p_1, p_2, p_3, \dots, p_k\}$ »

Les précautions du début de la phrase « même expérience de manière indépendante dans les mêmes conditions » sont sans doute un peu redondantes et on a vraiment l'impression que l'on ne peut parler de probabilité qu'après avoir expérimenté.

On retrouve là le point de vue de l'économiste, mathématicien et philosophe des sciences Antoine Cournot (1801-1877) :

« En conséquence, la probabilité mathématique prise objectivement, ou conçue comme mesurant la possibilité des choses, ne peut en général être déterminée que par l'expérience. Si le nombre des épreuves d'un même hasard croissait à l'infini, elle serait déterminée exactement avec une certitude comparable à celle de l'événement dont le contraire est physiquement impossible. Pour un nombre très grand d'épreuves, la probabilité n'est encore donnée qu'approximativement ; mais on est autorisé à regarder comme extrêmement peu probable que la valeur réelle diffère totalement de la valeur conclue des observations. En d'autres termes, il arrivera très rarement que l'on commette une erreur notable en prenant pour la valeur réelle la valeur tirée des observations. »

DES ÉNONCÉS NON VULGARISÉS...

Convergence presque sûre

Soit un espace probabilisé $(\Omega, \mathcal{E}, P)$. On dit que la suite de variables aléatoires $(X_n)_{n \in \mathbb{N}}$ converge presque sûrement vers la variable aléatoire X , et on note $X_n \xrightarrow{p.s.} X$, si

$$\exists N \in \mathcal{E} \text{ avec } P(N) = 0, (\omega \in \Omega - N) \Rightarrow \left(\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega) \right)$$

En d'autres termes, l'ensemble des $\omega \in \Omega$ pour lesquels $X_n(\omega)$ ne tend pas vers $X(\omega)$ est de probabilité nulle.

Loi forte des grands nombres

Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires indépendantes, de même loi (avec une espérance finie μ) alors $\frac{1}{n} \sum_{k=1}^n X_k$ converge presque sûrement vers μ .

Loi faible des grands nombres

Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires indépendantes, de même loi (avec une espérance finie μ) alors

$$\forall \varepsilon > 0, \lim_{n \rightarrow \infty} \left(P \left(\left| \frac{1}{n} \sum_{k=1}^n X_k - \mu \right| > \varepsilon \right) \right) = 0$$

Pierre-Simon Laplace en 1819

« Concevons, par exemple, une urne qui renferme des boules blanches et des boules noires, et supposons qu'à chaque fois que l'on tire une boule, on la remette dans l'urne pour procéder à un nouveau tirage...

... La probabilité que le rapport du nombre de boules blanches extraites au nombre total des boules sorties ne s'écarte pas, au delà d'un intervalle donné, du rapport du nombre des boules blanches au nombre total des boules contenues dans l'urne, approche indéfiniment de la certitude par la multiplication indéfinie des événements, quelque petit que l'on suppose cet intervalle. »

HISTORIQUE

La loi faible des grands nombres est due à Jacques Bernoulli qui l'aurait énoncée vers 1689. Elle n'a été publiée qu'après sa mort en 1713 dans l'*Ars Conjectandi* par son neveu Nicolas. Évidemment le terme "faible" n'apparaît qu'au XIX^{ème} siècle après la découverte par Poisson de la loi dite "forte".



Jacques Bernoulli
1654-1705
Suisse



Siméon Denis Poisson
1781-1860
France

ET QUE DIRE À NOS ÉLÈVES ?

Nous n'aurons pas la prétention d'apporter une réponse définitive à cette question et nous ne chercherons pas à départager les différents types de formulations cités. La phrase (ou les phrases) finalement retenues ne devraient être données aux élèves qu'après de nombreuses expérimentations et simulations susceptibles d'en éclairer le sens et d'en faire sentir la portée et les limites.

Une attention particulière sera nécessaire pour éviter que ne se pérennisent de fausses conceptions :

Une première erreur consiste à penser que les effectifs convergent vers des effectifs « théoriques », que si l'on joue un grand nombre de fois à pile ou face le nombre de piles deviendra proche du nombre de faces. C'est faux ! Ce ne sont pas les effectifs qui « convergent » mais les fréquences. Le nombre de pile peut, par exemple, s'écarter de 1000000 du nombre de faces (en jouant suffisamment longtemps cet événement est même certain !). En revanche, dans le cas du jeu de pile ou face, où il n'y a que deux éventualités, le retour (à un moment donné), à l'égalité des fréquences est certain si on poursuit à l'infini. Ce n'est plus le cas pour le lancer d'un dé, si on poursuit à l'infini, la probabilité de retrouver un même nombre d'apparitions de chacune des faces n'est plus que de l'ordre de 0.35.

Une autre idée erronée qui relève presque de la superstition est qu'après 10 résultats « pile » successifs le résultat « face » devient plus probable. C'est faux ! La pièce, comme tout processus aléatoire indépendant n'a pas de « mémoire » et il n'y a pas de « compensation ». Le nombre des « longues » séries de résultats identiques sont très probables (voir à ce sujet « Valeurs consécutives égales » dans le chapitre « Nombres au hasard ») et n'influencent pas la suite de l'expérience.

Il convient donc, par une pratique fréquente de simulations de donner à nos élèves une bonne idée de ce qu'est un phénomène aléatoire, de les convaincre qu'il y a des fluctuations, de leur faire sentir que la répétition d'un grand nombre d'expériences réelles ou simulées permet de dégager une certaine régularité mais aucune certitude.

Est-il possible d'imaginer des activités et des exercices invoquant la loi des grands dont nous n'aurons donné qu'une version intuitive ? Voici deux exemples.

Un exercice

Énoncé : (tel qu'il a été trouvé dans un livre)

Un appareil fabriqué en très grande série peut être défectueux en raison de deux défauts seulement qui seront désignés dans la suite par A et B.

Dans un lot de 10000 appareils testés, on a constaté que :

1000 appareils présentaient le défaut A (et peut-être le défaut B),

800 présentaient le défaut B (et peut-être le A)

300 présentaient simultanément les défauts A et B.

Un client achète l'un des appareils produits. Quelle est la probabilité :

- 1) qu'il présente le défaut A seulement ?
- 2) qu'il présente le défaut B seulement ?
- 3) que l'appareil ne présente aucun défaut ?

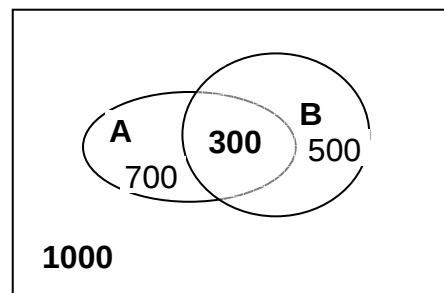
L'exercice est tout d'abord très utile pour introduire les trois types de schémas souvent efficaces dans la résolution d'exercices de statistique ou de probabilités : les diagrammes, arbres et tableaux. Les trois représentations sont utilisables ici avec un léger avantage au tableau à double entrée dans lequel il est possible d'inscrire tous les nombres de l'énoncé avant tout calcul.

Il s'agit de visualiser la répartition des 10000 appareils testés selon les défauts qu'ils présentent : on est dans le domaine de la statistique descriptive.

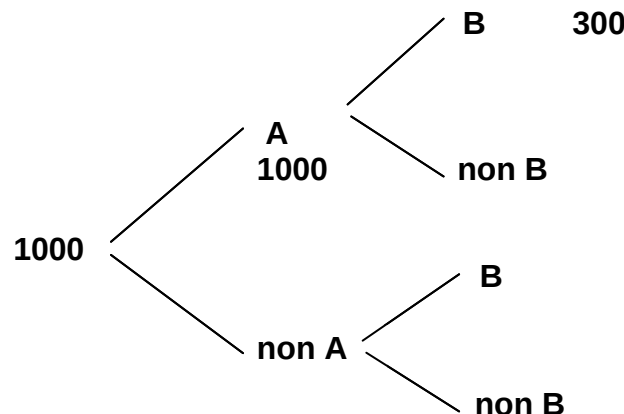
Tableau (les nombres donnés dans l'énoncé sont en gras)

	A	non A	Total
B	300	500	800
non B	700	8500	9200
Total	1000	9000	10000

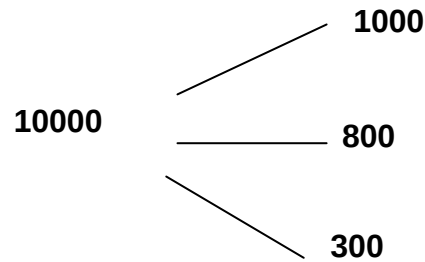
Diagramme



Arbre



Les élèves se contentent généralement d'une lecture linéaire en notant les nombres donnés ce qui rend l'interprétation assez malaisée. On voit parfois des ébauches de schémas inadaptés comme le pseudo arbre :



Après avoir souffert pour interpréter clairement l'énoncé les élèves sont facilement convaincus de l'utilité et de l'efficacité de ces schémas et le professeur peut commencer à les rendre sensibles aux règles régissant leur utilisation.

Pour ce qui concerne les questions de probabilité qui suivent, plusieurs problèmes peuvent être soulevés et débattus :

1. Quelle est l'expérience aléatoire ?
2. Connaît-on vraiment l'ensemble Ω des résultats possibles ?
3. Quelles hypothèses supplémentaires permettent de proposer des réponses ?

L'expérience aléatoire c'est l'achat d'un des appareils produits que l'on peut assimiler à un tirage au hasard d'un appareil parmi les appareils produits.

L'univers est donc l'ensemble Ω des appareils produits dont on sait malheureusement peu de choses. La précision « fabriqué en très grande série » nous permet tout au plus de supposer que Ω a « beaucoup » d'éléments. Il est raisonnable de penser que tous les appareils produits n'ont pas été testés (sinon pourquoi mettre en vente des appareils dont on sait qu'ils sont défectueux ?) et même que les appareils testés ne sont plus vendables.

Le modèle est donc très incomplet et des hypothèses supplémentaires nécessaires pour pouvoir proposer des réponses aux questions posées. On est face à un problème de statistique inductive : estimer des probabilités dans l'ensemble des appareils produits à partir de fréquences observées sur un échantillon (les appareils testés).

L'échantillon étant de « grande » taille (10000), la loi des « grands » nombres permet de « justifier » que l'on utilise les fréquences observées dans l'échantillon comme des probabilités dont elles sont, presque sûrement, très proches.

C'est ce que font souvent spontanément les élèves les plus rapides qui donnent les réponses attendues avant que le professeur ne soulève le problème. L'exercice, avec son énoncé original, est de ce fait excellent pour lancer le débat sur toutes ces questions en classe.

Pour proposer l'exercice aux élèves livrés à eux mêmes, en travail à la maison ou en évaluation, une phrase supplémentaire est nécessaire dans l'énoncé.

On ajoutera par exemple : « **Nous admettons qu'en vertu de la loi des grands nombres, les fréquences observées sur l'échantillon des appareils testés peuvent être assimilées aux probabilités dans la population des appareils produits** ».

Une activité

Une personne choisit au hasard l'une des 5 cases blanches de la grille ci-contre, selon l'un des protocoles suivants :

- Protocole 1 : choix au hasard d'une case blanche parmi les cinq ;
- Protocole 2 : choix d'une ligne au hasard puis d'une case blanche au hasard dans cette ligne ;
- Protocole 3 : choix d'une colonne au hasard puis d'une case blanche au hasard dans cette colonne.

①		
②		③
	④	⑤

Objectif : Simuler l'expérience un grand nombre de fois avec chacun des trois protocoles et proposer les lois de probabilité respectives.

1. Travail préliminaire (sur feuille avec un bilan en commun)

- Quelles sont les expériences aléatoires effectuées et quel est l'ensemble des issues possibles, pour chacun des trois protocoles ?
- Quelle loi de probabilité peut-on proposer pour l'expérience selon le 1^{er} protocole ?
- Les issues sont-elles équiprobables pour les protocoles 2 et 3 ?
- Décomposer ces deux dernières expériences en deux étapes, de telle façon que les résultats à chaque étape soient équiprobables.

2. Simulation sous Excel

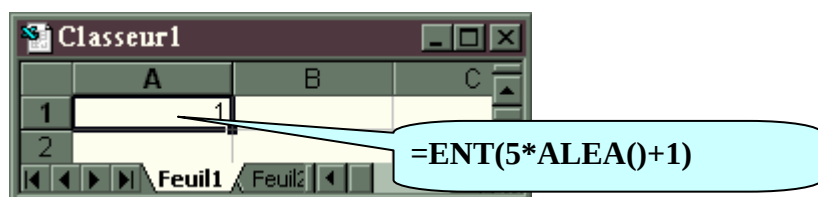
Simuler chacune des trois expériences pour :

- dans le protocole 1, vérifier que les fréquences observées se rapprochent de la loi de probabilité modèle proposée ;
- dans les protocoles 2 et 3, proposer une loi de probabilité cohérente avec les fréquences observées.

Rappel : Quelques fonctions de calcul du tableur (Excel)

- La fonction ALEA() donne un nombre décimal aléatoire entre 0 et 1 (1 exclu).
- La fonction ENT(a) donne la partie entière de a .
- La formule ENT($n \cdot \text{ALEA}() + p$) donne un nombre entier entre p et $p+n-1$.
- La fonction NB.SI(A1:B1000; a) compte le nombre de fois où le caractère a apparaît dans les cases de A1 à B1000 (dans cet exemple).
- La fonction SI($x ; y ; z$) teste la condition x , renvoie y si x est vraie et renvoie z sinon.

Procédure



- **Protocole 1** (dans la feuille 1)

Cellule A1 : Afficher un nombre entier au hasard entre 1 et 5 et recopier la formule jusqu'à A1000 pour avoir 1000 expériences.

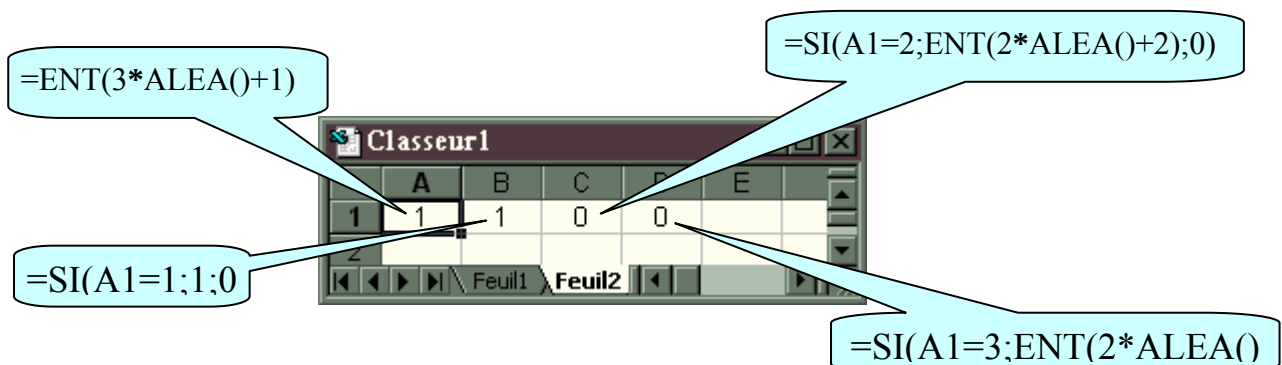
- **Protocole 2** : (dans la feuille 2)

Cellule A1 : simuler le choix d'une ligne entre 1 et 3.

Cellule B1 : Si la ligne choisie est la 1^{ère}, on prend la case 1, sinon on affiche 0.

Cellule C1 : Si la ligne choisie est la 2^{ème}, on choisit au hasard 2 ou 3, sinon on affiche 0.

Cellule D1 : Si la ligne choisie est la 3^{ème}, on choisit au hasard 4 ou 5, sinon on affiche 0.



Formulas shown in callouts:

- Cellule A1: `=ENT(3*ALEA()+1)`
- Cellule B1: `=SI(A1=1;1;0)`
- Cellule C1: `=SI(A1=2;ENT(2*ALEA()+2);0)`
- Cellule D1: `=SI(A1=3;ENT(2*ALEA())`

Recopier ces quatre formules jusqu'à la ligne 1000 pour avoir 1000 expériences.

- **Protocole 3** : (dans la feuille 3)

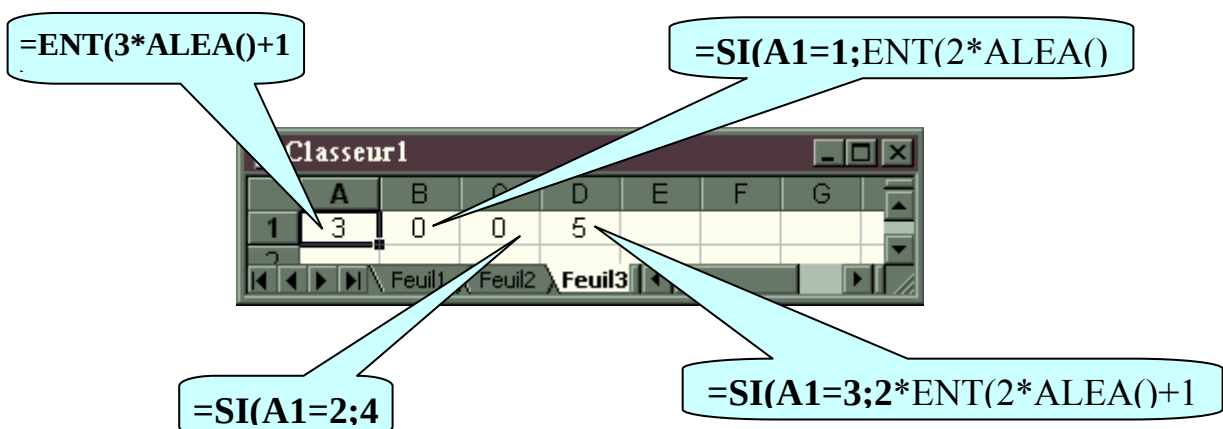
Cellule A1 : simuler le choix d'une colonne entre 1 et 3.

Cellule B1 : Si la colonne choisie est la 1^{ère}, on choisit au hasard la case 1 ou 2, sinon on affiche 0.

Cellule C1 : Si la colonne choisie est la 2^{ème}, on prend la case 4, sinon on affiche 0.

Cellule D1 : Si la colonne choisie est la 3^{ème}, on choisit au hasard les nombres 1 ou 2 qu'on transforme en 3 et 5 en les multipliant par deux puis en leur ajoutant 1, sinon on affiche 0.

Recopier ces quatre formules jusqu'à la ligne 1000 pour avoir 1000 expériences.

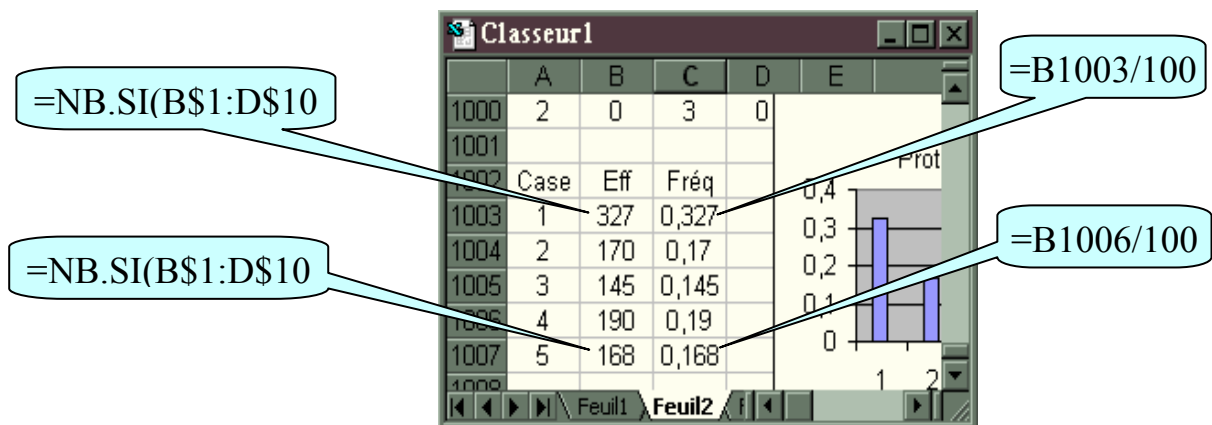


Formulas shown in callouts:

- Cellule A1: `=ENT(3*ALEA()+1)`
- Cellule B1: `=SI(A1=1;ENT(2*ALEA()` (truncated)
- Cellule C1: `=SI(A1=2;4`
- Cellule D1: `=SI(A1=3;2*ENT(2*ALEA()+1`

- Pour chaque protocole, calculer la fréquence du choix de chaque case et afficher le diagramme en bâton pour la distribution de fréquences correspondante.

Exemple pour le protocole 2



Remarque : chaque appui de la touche **F9** effectue une nouvelle simulation.

- Imprimer sur une page (ou relever) les trois distributions avec leur diagramme.

3. Interprétation des résultats

- La simulation suivant le protocole 1 est-elle conforme au modèle ?
- Pour le protocole 2, quel semble être le rapport entre les fréquences de la distribution ? Proposer une loi de probabilité pour le protocole 2 qui soit conforme à vos constatations.
- Même questions pour le protocole 3.

4. Pour aller plus loin

Cumuler plusieurs séries de 1000 expériences pour obtenir un échantillon de très grande taille. Vérifier que les lois de probabilité proposées pour les protocoles 2 et 3 semblent correctes, c'est-à-dire que les fréquences observées sont, en vertu de la loi des grands nombres, proches des probabilités

Procédure :

- Permettre le cumul des données en allant dans le menu **Outils** puis **Options...** et dans l'onglet **Calcul**, cocher les cases : **Sur ordre**, **itération** et entrer **1** pour **nb maximal d'itérations**.
 - Dans la cellule A1001, mettre un compteur avec la fonction : **=A1001+1000**
 - Modifier les cellules B1003 à B1007 en calculant les effectifs cumulés par :

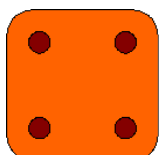
$$=NB.SI(B\$1:D\$1000;1)+B1003 \quad (...)$$

$$=NB.SI(B\$1:D\$1000;5)+B1007$$
 - Modifier les cellules C1003 à C1007 en calculant les fréquences par :

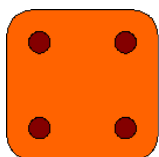
$$=B1003/(A\$1001) \quad (...)$$

$$=B1007/(A\$1001)$$
 - Entrer également dans les cellules D1003 à D1007 les valeurs des probabilités et dans les cellules E1003 à E1007 l'écart entre la fréquence et la probabilité.
- Appuyer sur F9 pour lancer le cumul et observer si les fréquences se rapprochent des probabilités.
 - Quel est l'écart maximal que vous observez entre les fréquences et les probabilités lorsque le nombre d'expériences est 10 000, 100 000, 1 000 000 ?
 - Conclure.

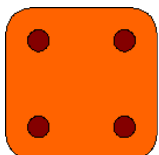
IV. PROBABILITÉS CONDITIONNELLES



Fréquences : des arbres au tableau



Fréquences : du tableau aux arbres

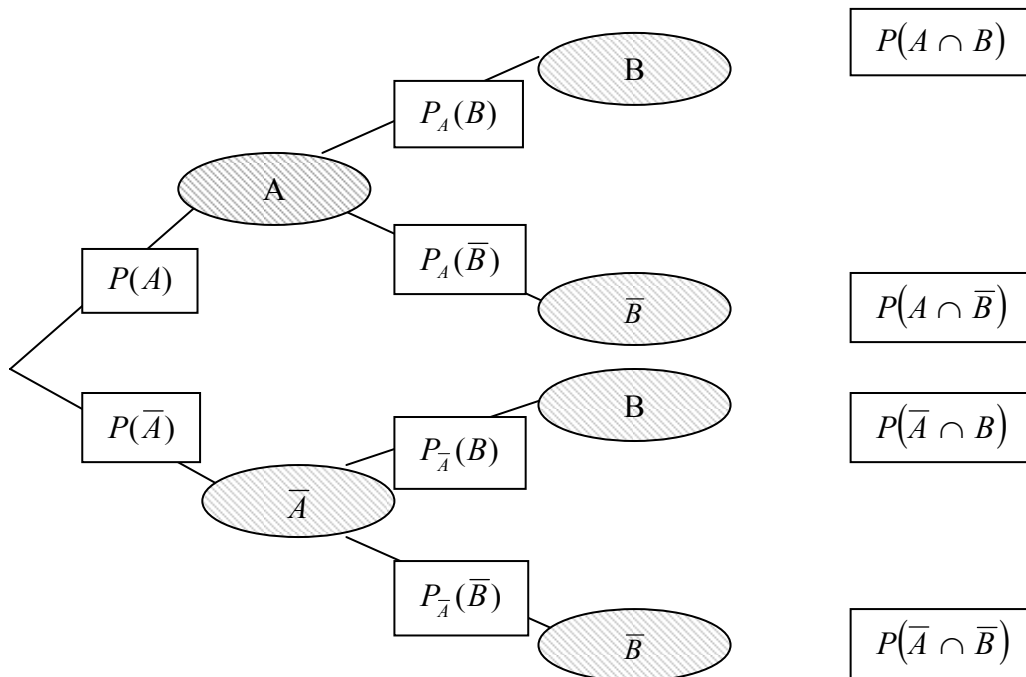


Tests médicaux

INTRODUCTION

L'arbre est un outil graphique qui permet de trier et d'organiser les données. Il constitue une aide à la mémorisation des formules (définition des probabilités conditionnelles et formule des probabilités totales) et le programme précise que, bien construit, il constitue une preuve acceptable.

Dans le cas générique d'une population triée suivant deux critères A et B on obtient l'arbre suivant :



Règle 1 : La somme des probabilités situées sur les branches issues d'un même nœud est 1 se traduit par : $P(A) + P(\bar{A}) = 1$, $P_A(B) + P_A(\bar{B}) = 1$ etc.

Règle 2 : L'événement terminal (A et B) a une probabilité qui s'obtient en faisant le produit des probabilités sur les branches qui y mènent : $P(A \cap B) = P(A) \times P_A(B)$.

Règle 3 : La probabilité d'un événement est la somme des probabilités des branches aboutissant à cet événement : $P(B) = P(A \cap B) + P(\bar{A} \cap B)$.

Dans la même situation, un tableau à double entrée nous donne :

	A	$\bar{A}$	
B	$P(A \cap B)$	$P(\bar{A} \cap B)$	$P(B)$
$\bar{B}$	$P(A \cap \bar{B})$	$P(\bar{A} \cap \bar{B})$	$P(\bar{B})$
	$P(A)$	$P(\bar{A})$	

Les deux modes de représentation conduisent bien sûr aux mêmes résultats et leur utilisation dépend essentiellement des données du problème à traiter. Dans la pratique, il convient d'habituer les élèves à choisir le mode le plus adapté et à l'exploiter complètement.

FRÉQUENCES : DES ARBRES AU TABLEAU

L'activité qui suit peut être proposée aux élèves en début de terminale (avant d'aborder le chapitre).

On réalise une enquête sur une population en s'intéressant à deux critères : Le sexe des personnes interrogées et leur tranche d'âge (moins de vingt ans et plus de vingt ans).

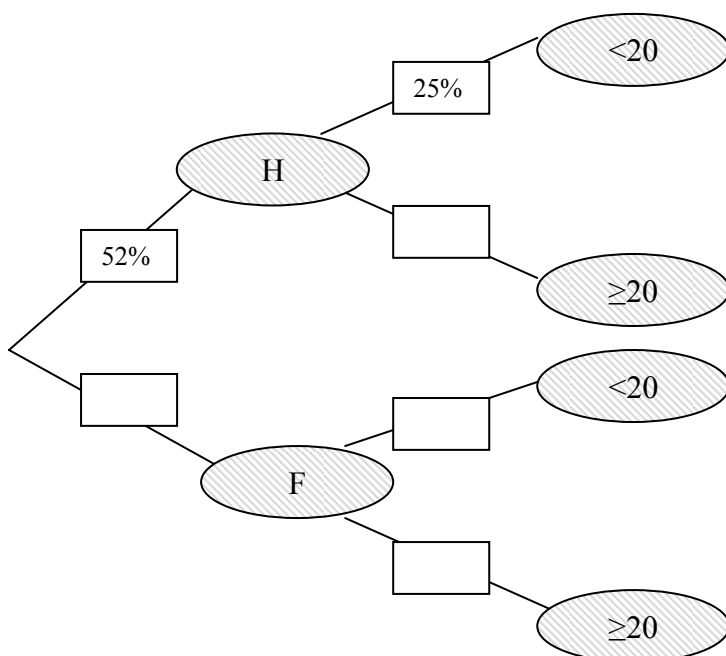
On constate que

52% des personnes interrogées sont des hommes

25% des hommes interrogés ont moins de vingt ans.

La fréquence du résultat « la personne interrogée est une femme de moins de vingt ans » est 16%.

1. Compléter le premier des deux arbres suivants.
2. Quelle est la fréquence du résultat : « la personne interrogée est un homme de moins de vingt ans. » On justifiera dans le détail l'opération effectuée.
3. Quel pourcentage des femmes a moins de vingt ans ?
4. Quel est le pourcentage des moins de vingt ans ?
5. Compléter le deuxième arbre.
6. Résumer cette enquête par un tableau à double entrée.

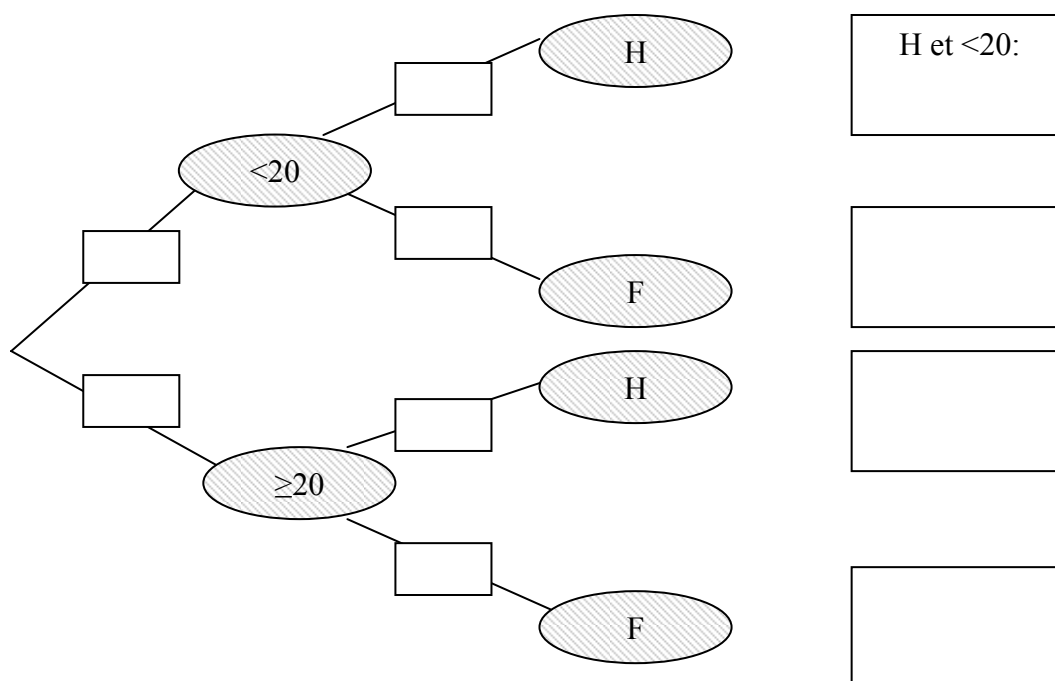


H et <20:

H et ≥20:

F et <20:
16%

F et ≥20:



Commentaires

- L'objectif de cette activité "à la maison" est d'introduire la notion de probabilité conditionnelle et la formule des probabilités totales en se basant sur des calculs de fréquences.
- Bien que ce ne soit pas un exercice de probabilité, lors de sa correction le professeur peut passer du domaine statistique au domaine des probabilités : Il suffit pour cela de considérer que l'on prélève au hasard un individu de la population.
- La correction de la question 1 est l'occasion de préciser les règles en vigueur pour construire correctement un arbre de probabilité.
- Pour la question 2, on peut attendre une justification du type : Si N est le nombre de personnes interrogées, alors comme "52% des personnes interrogées sont des hommes", le nombre d'hommes sondé est $0,52N$. Comme "25% des hommes interrogés ont moins de vingt ans" le nombre d'hommes de moins de vingt ans est $0,25 \times 0,52 N = 0,13 N$ et ainsi 13% des sondés sont des hommes de moins de vingt ans.
- La question 3 peut être corrigée en utilisant directement le premier arbre et permet d'introduire la formule $P_B(A) = \frac{P(A \cap B)}{P(B)}$, la formule des probabilités totales apparaît alors dans la question 4.
- Le tableau à double entrée constitue un résumé de la situation, cependant, si les probabilités des événements (marges) et celles des intersections y apparaissent, ce n'est pas le cas des probabilités conditionnelles qui doivent être calculées.
- L'utilité du deuxième arbre peut être mise en évidence en demandant de calculer la proportion de femmes chez les moins de vingt ans. Cette inversion dans l'ordre des

critères de tri est ici naturelle aux yeux des élèves : Ce ne sera pas toujours le cas dans d'autres situations.

- Enfin, on peut demander aux élèves laquelle des deux représentations est la plus adaptée pour les faire réfléchir sur ce point.

FRÉQUENCES : DU TABLEAU AUX ARBRES

Voici une deuxième activité, toujours pour élèves de terminale qui permet de travailler le passage du tableau à double entrée à l'arbre de probabilité. Elle constitue en plus une introduction possible à la notion d'indépendance.

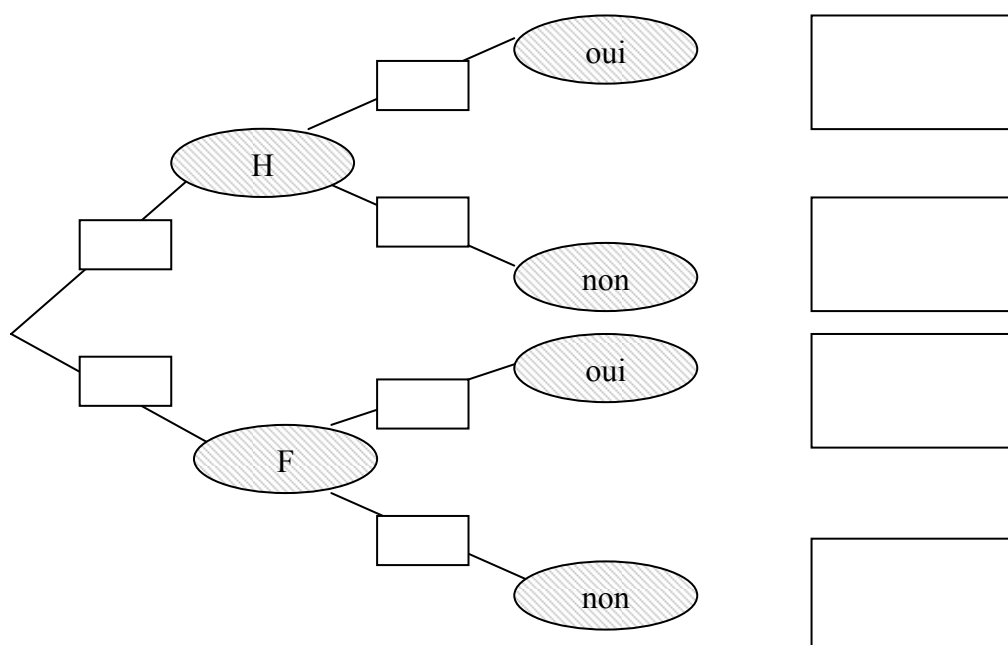
On effectue un sondage sur la vente des produits issus du commerce équitable en demandant aux clients d'un hypermarché, s'ils ont acheté dans le mois au moins un produit issu du commerce équitable.

- 9% des personnes interrogées sont des hommes qui ont acheté dans le mois au moins un produit issu du commerce équitable.
- 16% des personnes interrogées sont des hommes qui n'ont pas acheté dans le mois au moins un produit issu du commerce équitable.
- 27% des personnes interrogées sont des femmes qui ont acheté dans le mois au moins un produit issu du commerce équitable.

On souhaite savoir si les hommes et les femmes ont le même comportement d'achat en ce qui concerne les produits issus du commerce équitable.

Pour cela :

1. Compléter un tableau à double entrée.
2. Quels sont les pourcentages d'hommes et de femmes dans l'échantillon interrogé ?
3. Quel est le pourcentage de personnes interrogées ayant acheté dans le mois au moins un produit issu du commerce équitable.
4. Montrer que le pourcentage d'hommes ayant acheté dans le mois au moins un produit issu du commerce équitable est le même que le pourcentage de femmes ayant acheté dans le mois au moins un produit issu du commerce équitable. Pouvait-on arriver à cette conclusion sans calculer les pourcentages, en observant le tableau ?
5. Représenter la situation en complétant l'arbre ci-dessous. La conclusion est-elle clairement visible sur l'arbre ?



Commentaires

Dans cette situation, les données correspondent à des fréquences d'intersections d'événements. La représentation par un tableau étant naturelle, on peut choisir de ne pas poser la question 1 aux élèves, en particulier si le début de l'activité est traité en classe.

Le pourcentage d'hommes et celui de personnes ayant acheté au moins une fois un produit issu du commerce équitable se trouvent directement par addition. Les autres peuvent alors se trouver par différence. Le tableau obtenu est le suivant.

	H	F	Total
E	9%	27%	36%
non E	16%	48%	64%
Total	25%	75%	100%

Les réponses aux questions 2 et 3 découlent alors directement du tableau.

Dans la question 4, il s'agit de mettre en évidence la proportionnalité des colonnes du tableau.

Lors de la correction, on interprète $9/25$ comme une probabilité conditionnelle (on modifie l'énoncé en prenant au hasard une personne parmi les sondés). On obtient alors, entre autre, la formule : $P_H(E) = P(E)$ qui peut être choisie pour définir l'indépendance de deux événements de probabilités non nulles.

L'intérêt de la représentation par un arbre apparaît à la question 5. Les élèves pourront retenir que l'indépendance se visualise sur l'arbre par l'égalité des poids sur les branches.

TESTS MÉDICAUX

Cette troisième activité peut-être proposée aux élèves en fin de chapitre sur le conditionnement. Elle répond à l'une des exigences du programme : "Application à la problématique des tests de dépistage en médecine."

Partie A

Un laboratoire produit un test pour dépister une maladie. La population à tester comporte 1% de malades. Des études ont montré que lorsqu'une personne est atteinte par la maladie, il y a une probabilité de 95% que le test soit positif alors que, lorsqu'elle est saine la probabilité que le test soit positif est $p = 4\%$.

1. On teste au hasard une personne de la population. Déterminer la probabilité que le test soit positif.
2. Le test est positif. Quelle est la probabilité que la personne soit saine ? Ce résultat vous semble-t-il satisfaisant ?

Partie B

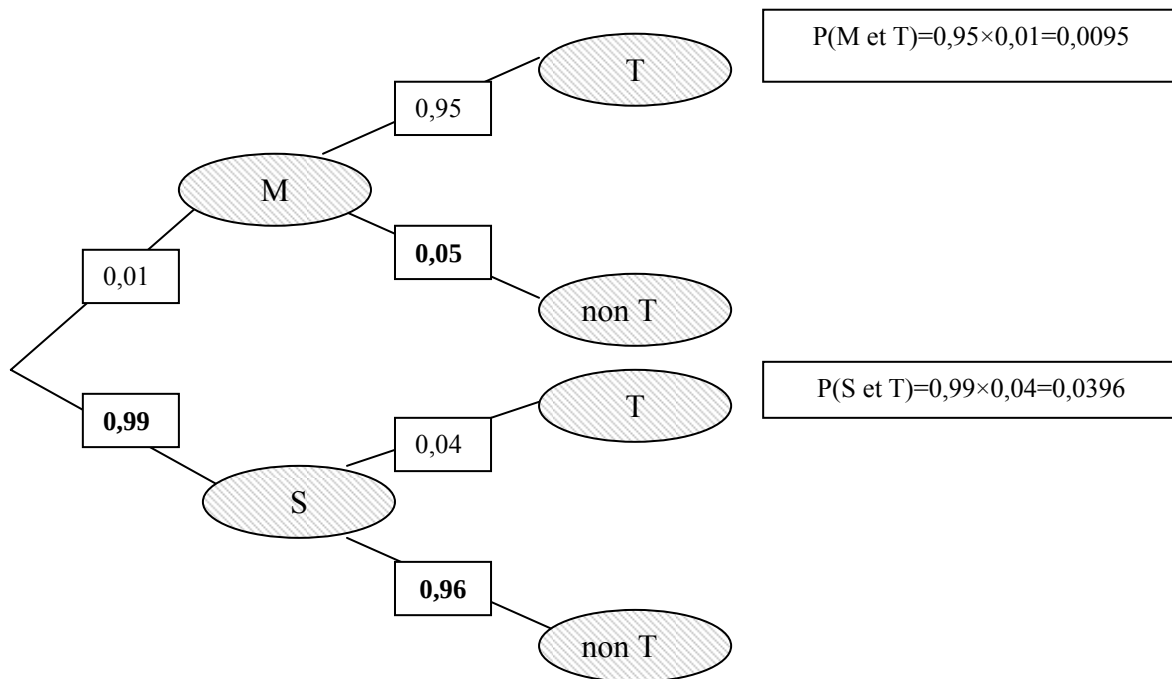
Lorsqu'une personne est positive au test, on lui prescrit un traitement. Ce dernier est fort heureusement sans conséquence sur la santé du patient si celui-ci est sain, mais par contre il est très onéreux. Le laboratoire a le choix entre deux stratégies :

- Améliorer le test précédent.
 - Concevoir un deuxième test qui sera utilisé uniquement sur les personnes positives au premier test.
1. **Stratégie 1**
On décide d'améliorer le premier test. On suppose que l'amélioration ne porte que sur la probabilité p qu'une personne saine soit déclarée positive.
 - a) Exprimer en fonction de p la probabilité qu'une personne soit saine sachant que le test est positif. On notera $f(p)$ cette probabilité.
 - b) Donner des valeurs approchées de $f(0,03)$, $f(0,02)$ et $f(0,01)$.
 2. **Stratégie 2**
On décide de concevoir un deuxième test (uniquement pour les positifs au premier). On supposera que ce deuxième test a les mêmes caractéristiques que le premier et que, dans chacune des deux catégories de personnes testées (saines et malades) ses résultats sont indépendants de ceux du premier test. Pour les calculs, on prendra $P(S) = 0,8$
 - a) Calculer la probabilité α qu'une personne déclarée positive à la suite du deuxième test soit saine.
 3.
 - b) Pour quelles valeurs de p a-t-on $f(p) \leq \alpha$? Commenter.

Commentaires

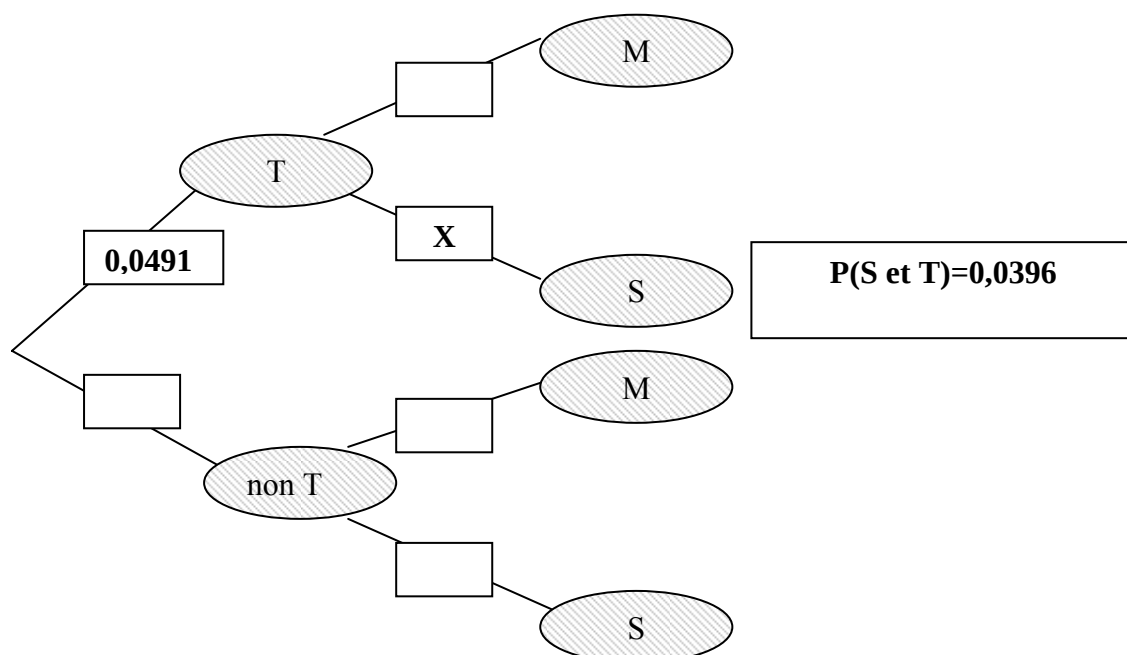
Cette activité répond à l'une des demandes du programme officiel : "On appliquera entre autre la formule des probabilités totales à la problématique des tests de dépistage."

Notons respectivement M, S et T les événements "la personne testée est malade", "saine", "testée positive". Les données du problème sont alors $P(M) = 0,01$, $P_M(T) = 0,95$ et $P_S(T) = 0,04$. On obtient l'arbre suivant :



Ainsi, $P(T) = 0,01 \times 0,95 + 0,99 \times 0,04 = 0,0491$.

Il ne reste plus qu'à inverser l'arbre pour répondre à la question 2 :



On en déduit que $P_T(S) = \frac{0,0396}{0,0491} \approx 0,8$. Ainsi près de 80% des personnes positives au test seront saines, ce qui ne peut pas être considéré comme satisfaisant.

L'étude de la première stratégie se traite comme la première partie mais avec $P_S(T) = p$.

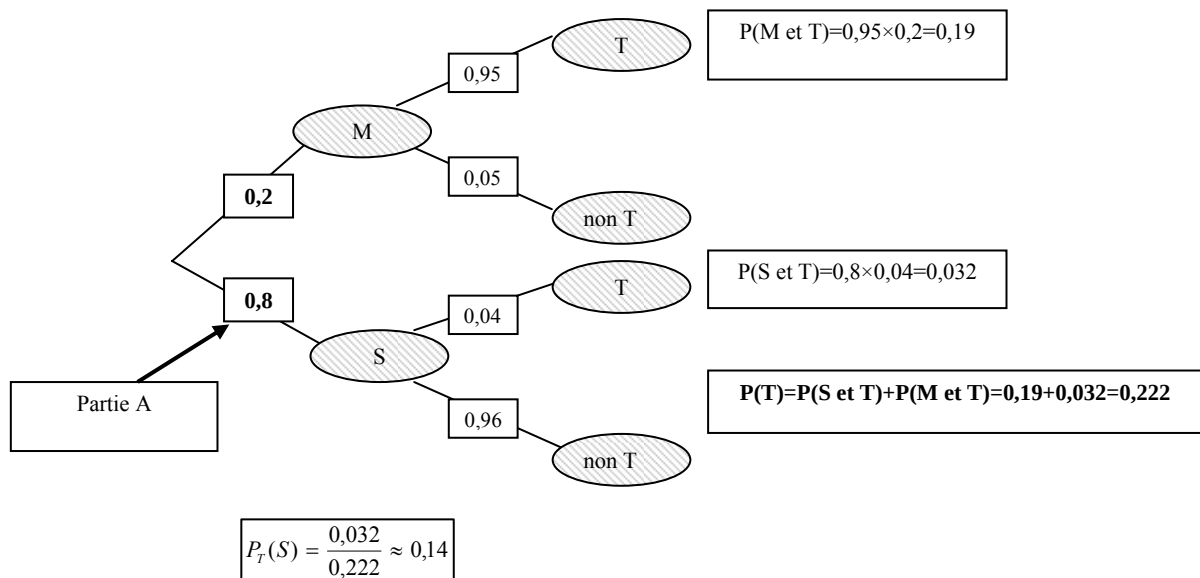
On trouve alors $P(T) = 0,99p + 0,0095$ et $P_T(S) = f(p) = \frac{99p}{99p + 0,95}$.

Il est bien sûr possible de faire étudier la fonction f mais dans un premier temps un tableau de valeurs est suffisant pour comprendre que la première stratégie n'est pas très efficace :

p	f(p)
0,04	80,65%
0,038	79,84%
0,036	78,95%
0,034	77,99%
0,032	76,93%
0,03	75,77%
0,028	74,48%
0,026	73,04%
0,024	71,44%
0,022	69,63%
0,02	67,58%
0,018	65,23%
0,016	62,51%
0,014	59,33%
0,012	55,57%
0,01	51,03%
0,008	45,46%
0,006	38,47%
0,004	29,42%
0,002	17,25%

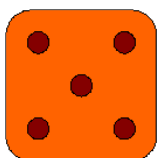
Lorsque $p=0,01$, plus de la moitié des personnes testées positives sont saines

Stratégie 2 : on teste une deuxième fois les positifs

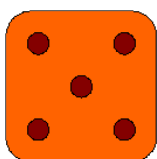


Ainsi $a \approx 0,14$ et $f(p) \leq a$ pour $p \leq \frac{16}{99} \times 10^{-2} \approx 0,0016$.

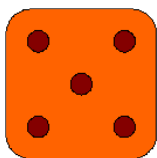
V. GÉNÉTIQUE ET PROBABILITÉS



Un peu d'histoire



Les maladies génétiques



Loi de Hardy – Weinberg

UN PEU D'HISTOIRE



Depuis ses débuts, la génétique entretient avec la statistique des rapports étroits.

La découverte de Gregor Mendel et les enseignements qu'il sut en tirer reste un modèle de nos jours. Il reste le premier à avoir utilisé le raisonnement statistique pour étayer sa théorie corpusculaire de l'hérédité et réfuter la conception erronée du « mélange des sangs » en vogue précédemment.

Rappelons brièvement l'historique de la découverte de Mendel au travers de ses études portant sur les pois réalisées au milieu du 19^{ième} siècle.

Il s'intéressa en particulier à deux caractères l'aspect lisse ou ridé et la couleur jaune ou verte de la graine.

En croisant des lignées pures de pois ayant des graines rondes avec des lignées pures de pois ayant des graines ridées il remarqua que les pois hybrides ainsi obtenus

avaient tous des graines rondes. Il en déduisit que le caractère « rond » était dominant sur le caractère « ridé » qualifié lui, de récessif. En croisant les hybrides entre eux, Mendel observa à nouveau le caractère récessif « ridé », chez une partie des descendants.

En croisant des lignées pures de pois ayant des graines jaunes et rondes avec des lignées pures de pois ayant des graines vertes et ridées, il nota que les pois hybrides ainsi obtenus étaient à graines jaunes et rondes. Mendel en conclut que la couleur jaune devait être dominante sur la couleur verte (comme la forme ronde l'était sur la forme ridée)

En croisant ces hybrides à graines jaunes et rondes entre eux, il observa la réapparition des caractères récessifs, c'est-à-dire qu'une partie des descendants étaient à graines vertes et ridées.

Le tableau ci-dessous donne les résultats observés par Mendel :

Aspect des graines	Jaunes Rondes	Jaunes Ridées	Vertes Rondes	Vertes Ridées
Effectifs	315	101	108	32

Les conclusions de Mendel ont été depuis pleinement confirmées par la découverte des gènes portés par les chromosomes.

Dans le cas des pois le caractère couleur est codé par un gène¹⁰ présentant deux formes allèles J et j correspondant aux couleurs jaune et vert. Le jaune est dominant, le vert récessif. Le caractère rond ou ridé, est porté par un autre gène à deux allèles R (rond dominant) et r (ridé récessif). Si on croise deux individus dont le génotype est JjRr, on obtient 16 génotypes équiprobables comme le montre le tableau ci-dessous :

¹⁰ Voir en annexe 1 un petit glossaire des termes utilisés en génétique

		Parent 1			
		R		r	
		J	j	J	j
Parent 2	R	J			
		j			
	r	J			
		j			

Les descendants seront jaunes et ronds dans 9 cas sur 16, jaunes et ridés dans 3 cas sur 16, verts et ronds dans 3 cas sur 16 finalement verts et ridés dans 1 cas sur 16. Dans le cas du croisement de 556 pois réalisé par Mendel les effectifs théoriques devraient être ceux du tableau ci-dessous :

Aspect des graines	Jaunes Rondes	Jaunes Ridées	Vertes Rondes	Vertes Ridées
Effectifs théoriques	312.75	104.25	104.25	34.75

Comparons des effectifs observés et des effectifs théoriques à l'aide d'un test du Khi deux¹¹ pour lequel les calculs sont détaillés dans le tableau ci-dessous :

Aspect des graines	Jaunes Rondes	Jaunes Ridées	Vertes Rondes	Vertes Ridées	Totaux
Effectifs observés	315	101	108	32	556
Effectifs théoriques	312,75	104,25	104,25	34,75	556
Écarts	2,25	-3,25	3,75	-2,75	0
Carrés	5,0625	10,5625	14,0625	7,5625	37,25
Carrés/théoriques	0,016187	0,101319	0,134892	0,217626	0,470024
d²=					0,470024

Pour pouvoir rejeter l'hypothèse d'adéquation d^2 devrait dépasser 7,82 (valeur lue dans la table du Chi deux pour 3 degrés de liberté). La valeur observée est très inférieure à ce seuil.

En fait, elle est même inférieure au premier décile ce qui signifie que, dans l'hypothèse où le modèle proposé est le bon, un tel résultat expérimental avait, a priori, une probabilité inférieure à 1/10 de se produire. Ce résultat, presque « trop beau pour être vrai », a valu à Mendel d'être soupçonné d'avoir manipulé ses chiffres pour qu'ils confirment plus nettement sa théorie.

¹¹ Le principe du test d'adéquation du Chi deux est présenté dans le chapitre 6.

L'EXEMPLE DES MALADIES GÉNÉTIQUES

Maladies génétiques causées par un gène récessif

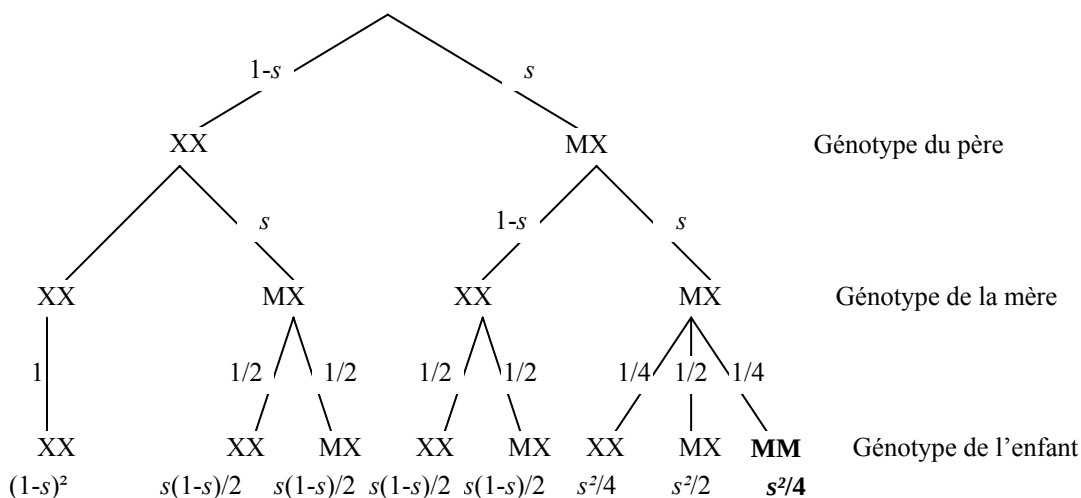
Considérons une maladie génétique humaine due à la présence d'un gène spécifique que nous noterons M réparti indépendamment du sexe. Supposons que seuls les porteurs de la combinaison homozygote MM développent la maladie. Les porteurs de la combinaison MX (où X désigne un allèle autre que M) sont des porteurs sains.

Supposons en outre que la gravité de la maladie est telle que les personnes atteintes ne peuvent avoir d'enfants. La maladie ne pourra apparaître que lorsque les deux parents sont hétérozygotes de type MX .

		Père	
		M	X
Mère	M	MM (malade)	MX (porteur sain)
	X	MX (porteur sain)	XX (non porteur)

Un premier problème est de calculer la proportion s de porteurs sains dans la population, connaissant la proportion f de malades.

Si tous les couples de parents sont équiprobables on peut construire l'arbre ci-dessous :



On lit en particulier que le génotype MM a pour probabilité $f = \frac{s^2}{4}$ et donc que $s = 2\sqrt{f}$.

La mucoviscidose est une maladie correspondant approximativement au modèle ci-dessus.

En France environ 1 enfant sur 2000 en est atteint, c'est-à-dire $f = \frac{1}{2000}$. Il en résulte que

$s = 2\sqrt{\frac{1}{2000}} \approx 0.0224$ c'est-à-dire que les porteurs sains du gène responsable représentent 2.24% de la population totale.

Dans le cas où les personnes malades peuvent avoir des enfants et en supposant encore que tous les couples de parents sont équiprobables, il faudrait ajouter une troisième branche MM aux deux premiers niveaux de l'arbre ci-dessus. Résumons plutôt les situations possibles et leurs probabilités dans un tableau :

		Génotype du père		
		MM	MX	XX
Génotype de la mère	MM	f^2 1	fs 1/2	$f(1-f-s)$ 0
	MX	fs 1/2	s^2 1/4	$s(1-f-s)$ 0
	XX	$f(1-f-s)$ 0	$s(1-f-s)$ 0	$(1-f-s)^2$ 0

Dans chaque cellule de ce tableau on lit la probabilité de la combinaison de génotype des parents correspondante (première ligne), ainsi que la probabilité que l'enfant soit malade sachant que l'on est dans cette combinaison de génotype des parents.

La probabilité qu'un enfant soit malade dans ce cas est : $f^2 + fs + \frac{s^2}{4}$.

On doit donc avoir :

$$f^2 + fs + \frac{s^2}{4} = f$$

que l'on peut écrire sous la forme habituelle d'une équation du second degré d'inconnue s :

$$\frac{s^2}{4} + fs + f^2 - f = 0$$

dont la résolution donne :

$$\Delta = f^2 - 4 \times \frac{1}{4} (f^2 - f) = f \quad \text{et} \quad s = \frac{-f + \sqrt{f}}{\frac{1}{2}} = 2(\sqrt{f} - f) \quad (\text{seule solution positive}).$$

L'hémochromatose génétique est une maladie correspondant à ce mode de transmission. Elle provoque une accumulation progressive du fer dans l'organisme. Il existe un traitement simple, efficace et peu coûteux (saignées) d'où l'importance d'un dépistage efficace. La fréquence de cette maladie est de l'ordre de 5/1000.

La proportion de porteurs sains est donc : $s = 2(\sqrt{f} - f) = 2(\sqrt{0.005} - 0.005) \approx 0.131$.

La maladie étant assez rare, l'utilisation de la formule du paragraphe précédent donnerait $s = 2\sqrt{f} = 2\sqrt{0.005} \approx 0.141$ soit un résultat encore assez proche de la bonne valeur.

DANS NOS CLASSES

Activité sur la loi de Hardy-Weinberg en terminale S

Ce texte a été donné à une classe de Terminale S dans le cadre d'un TD. La partie I a été traitée correctement en une heure environ et la partie II, donnée à faire à la maison de manière facultative, n'a été abordée que par quelques élèves de bon niveau.

Dans les cas simples un gène est formé de 2 allèles pouvant prendre les états A et a et un individu peut donc avoir l'un des trois génotypes suivants : AA, Aa, aa.

Un enfant lors de la conception hérite d'un gène de chacun de ses parents, chacun d'eux étant choisi au hasard. Ainsi si le père est du type AA et la mère de type Aa, les enfants peuvent être du type AA ou Aa.

I. On considère une population dont les proportions des génotypes sont données par le tableau suivant :

AA	Aa	aa
$p_0=0,4$	$q_0=0,3$	$r_0=0,3$

1. Si on considère que les couples se forment au hasard quant aux 3 génotypes considérés, calculer la probabilité pour qu'un enfant ait

- a.** Deux parents de type AA
- b.** Un parent de type AA et un parent de type Aa
- c.** Deux parents de type Aa.

2. Quelle est la probabilité pour que l'enfant soit de type AA sachant que :

- a.** les deux parents sont de type AA ?
- b.** un des deux parents est de type AA et l'autre du type Aa ?
- c.** les deux parents sont de type Aa ?

3. a. Calculer la probabilité p_1 pour que l'enfant soit de type AA.

b. Calculer de même la probabilité r_1 pour que l'enfant soit de type aa.

c. En déduire la probabilité q_1 pour que l'enfant soit de type Aa.

4. Calculer de même les probabilités p_2 , q_2 et r_2 pour qu'un enfant de seconde génération soit de type AA, Aa et aa. Que remarque-t-on ?

II. Généralisation

p_0 , q_0 et r_0 sont maintenant des nombres positifs de somme égale à 1.

1. a. Montrer que la probabilité pour qu'un enfant soit de type AA est $p_1 = \left(p_0 + \frac{1}{2} q_0 \right)^2$.

b. Sans calcul, donner une expression de r_1 .

2. Soit $\alpha = p_0 - r_0$.

a. Montrer que $p_1 = \frac{(1+\alpha)^2}{4}$.

b. En déduire les expressions de r_1 puis de q_1 en fonction de α . Que peut-on dire de $p_1 - r_1$?

3. Donner les expressions de p_2 , q_2 et r_2 en fonction de α .

4. On a étudié les génotypes de 100 enfants choisis au hasard parmi les enfants de la première génération puis les génotypes de 100 enfants choisis au hasard parmi les enfants de la seconde génération. Les fréquences d'enfants de type AA, Aa, aa étant

respectivement $(\hat{p}_1, \hat{q}_1, \hat{r}_1)$ et $(\hat{p}_2, \hat{q}_2, \hat{r}_2)$ on remarque que : $\frac{\hat{q}_1^2}{\hat{p}_1 \hat{r}_1} \approx 3,91$ et que

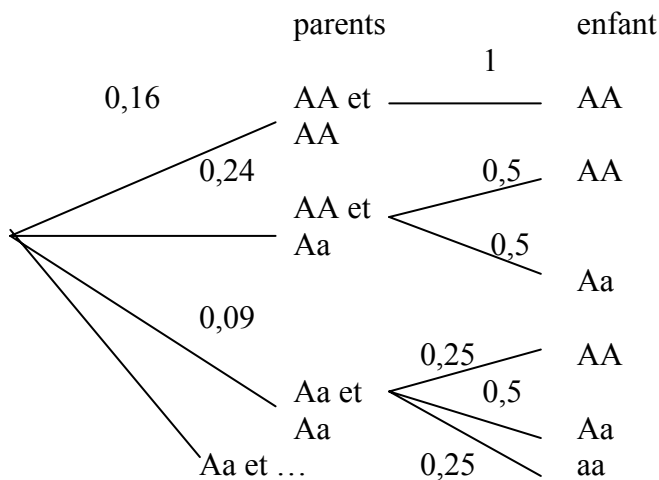
$\frac{\hat{q}_2^2}{\hat{p}_2 \hat{r}_2} \approx 3,95$. Qu'en pensez-vous ?

Corrigé

I.1. La probabilité pour qu'un enfant ait ses deux parents de type :

- AA est $0,4 \times 0,4 = 0,16$ car les deux événements sont indépendants
- AA et Aa est $2 \times 0,4 \times 0,3 = 0,24$
- Aa est $0,3 \times 0,3 = 0,09$

2. L'enfant ne peut être de type AA si un des parents est de type aa. Ainsi il est possible de représenter la situation à l'aide de l'arbre suivant :



3. a. On utilise ici la formule des probabilités totales :

$$p_1 = 0,16 \times 1 + 0,24 \times 0,5 + 0,09 \times 0,25 = 0,3025.$$

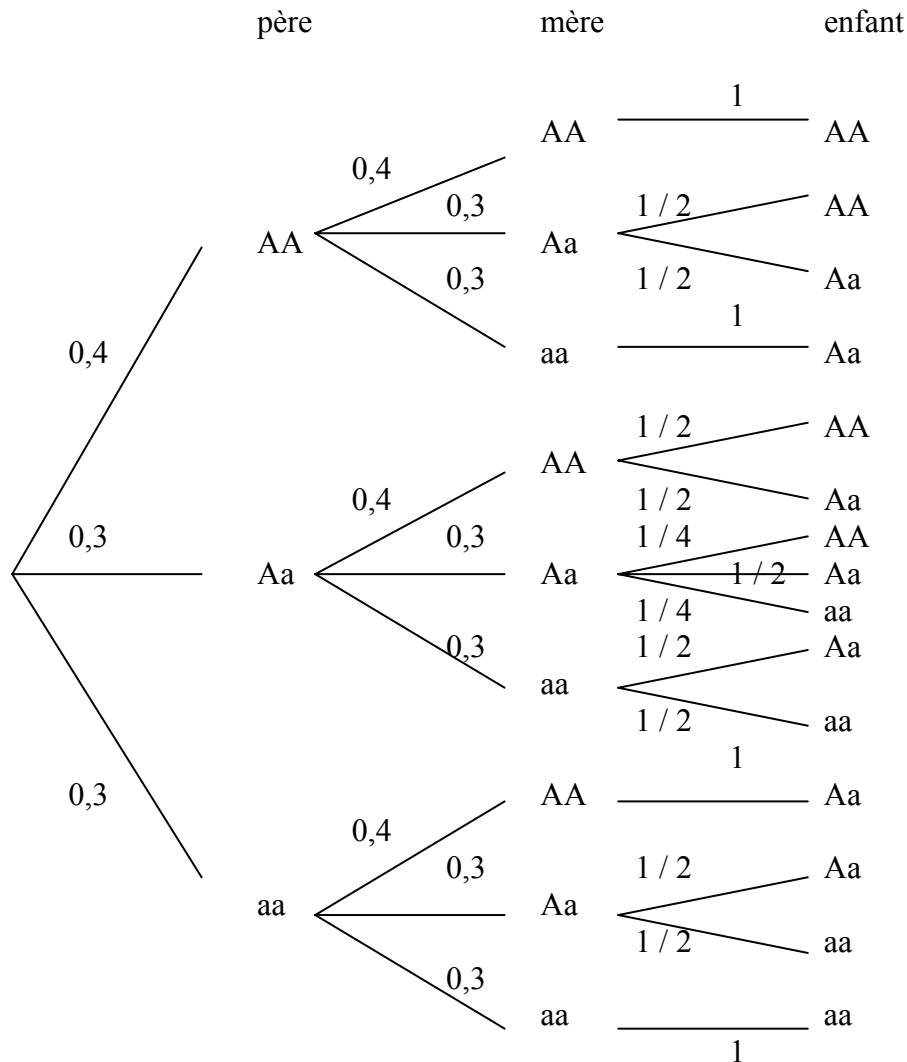
b. A l'aide d'un arbre du même type :

$$r_1 = 0,09 \times 1 + 0,18 \times 0,5 + 0,09 \times 0,25 = 0,2025.$$

c. D'où q_1 :

$$q_1 = 1 - p_1 - r_1 = 0,495.$$

Remarquons qu'un arbre complet permet de répondre directement aux questions 1., 2. et 3. et de retrouver ainsi p_1 , q_1 et r_1 .



4. Grâce à des arbres du même type on a comme dans la question 3. :

$$p_2 = p_1^2 \times 1 + 2 p_1 q_1 \times 0,5 + q_1^2 \times 0,25 = 0,3025,$$

$r_2 = r_1^2 \times 1 + 2 r_1 q_1 \times 0,5 + q_1^2 \times 0,25 = 0,2025$ et par conséquent $q_2 = 0,495$. On remarque que, dès la génération 1, il y a stabilisation des probabilités.

II. 1. a. On obtient de la même façon l'identité $p_1 = \left(p_0 + \frac{1}{2} q_0 \right)^2$ à partir de la formule des probabilités totales : $p_1 = p_0^2 \times 1 + 2 p_0 q_0 \times 0,5 + q_0^2 \times 0,25$.

b. Il est intéressant de remarquer ou de faire remarquer que p_0 et r_0 jouent un rôle symétrique. Ainsi : $r_1 = \left(r_0 + \frac{1}{2} q_0 \right)^2$.

2. a. On sait que $q_0 = 1 - p_0 - r_0$. Si on remplace dans $p_1 = \left(p_0 + \frac{1}{2} q_0 \right)^2$ on obtient :

$$p_1 = \left(p_0 + \frac{1}{2} (1 - p_0 - r_0) \right)^2 = \left(\frac{1}{2} + \frac{1}{2} p_0 - \frac{1}{2} r_0 \right)^2 = \frac{(1 + \alpha)^2}{4}.$$

b. On peut encore remarquer que p_0 et r_0 jouent un rôle symétrique et que $r_0 - p_0 = -\alpha$.

Ainsi : $r_1 = \frac{(1-\alpha)^2}{4}$ et dans ces conditions :

$$q_1 = 1 - p_1 - r_1 = 1 - \frac{(1+\alpha)^2}{4} - \frac{(1-\alpha)^2}{4} = \frac{1-\alpha^2}{2}.$$

3. Comme $p_1 - r_1 = \frac{(1+\alpha)^2}{4} - \frac{(1-\alpha)^2}{4} = \alpha$, les calculs précédents sont encore valables pour l'indice suivant et on a : $p_2 = p_1$, $q_2 = q_1$, $r_2 = r_1$.

4. Formulée comme cela, cette question n'a trouvée aucune réponse, aussi bien auprès des élèves de TS qu'auprès des stagiaires de notre formation. Seul a été mis en évidence le fait que :

$\frac{\hat{q}_1^2}{\hat{p}_1 \hat{r}_1} \approx \frac{\hat{q}_2^2}{\hat{p}_2 \hat{r}_2}$, ce qui est normal lorsqu'il y a stabilisation des fréquences dès la première génération. En fait il est normal de trouver des valeurs proches de 4 puisque :

$$\frac{q_1^2}{p_1 r_1} = \frac{\left(\frac{1-\alpha^2}{2}\right)^2}{\frac{(1+\alpha)^2}{4} \frac{(1-\alpha)^2}{4}} = 4. \text{ On aurait dû demander d'abord de calculer } \frac{q_1^2}{p_1 r_1} \text{ puis d'en}$$

tirer ensuite une conclusion.

Notons enfin cette question peut être abordée par des élèves n'ayant fait que la partie I.

En conclusion

Ce modèle a été mis en évidence de façon indépendante au début du XX^e siècle par Hardy, mathématicien et Weinberg, médecin. Cet équilibre des probabilités dès la première génération suppose réalisées les conditions suivantes : la population est panmictique (les couples se forment au hasard au sens des gènes et leurs gamètes se rencontrent au hasard), il ne doit avoir ni sélection, ni mutation ni migration et enfin il ne doit pas avoir non plus de croisement entre générations successives.

On peut remarquer que si on mélange tous les allèles des individus de la population initiale on obtient alors un ensemble contenant une proportion $p_0 + \frac{1}{2}q_0$ d'allèles A et une proportion $r_0 + \frac{1}{2}q_0$ d'allèles a. Si l'ensemble est de taille suffisante, les tirages au hasard de deux allèles pour former un gène peuvent être considérés comme indépendants et on retrouve ainsi dans ces conditions les probabilités p_1 et r_1 de la question II 1.

Par ailleurs si A est l'allèle dominant et si aa est le gène d'une maladie, les individus de type aa sont malades et les individus de type Aa sont des porteurs sains.

Si la population est grande et homogène on doit observer des valeurs p_k , q_k et r_k constantes de générations en générations.

Pour les professeurs uniquement

Et s'il n'y a pas de stabilisation ? Et bien il y a plusieurs causes possibles : consanguinité, migration d'une partie de la population, mortalité accrue des porteurs de l'allèle aa avant l'âge de la procréation. Reste à savoir comment tester l'hypothèse de la stabilité des différentes proportions. La première idée est d'utiliser le χ^2 .

On suppose qu'on a analysé 939 génotypes d'un échantillon d'une classe d'âge donnée. Le tableau suivant résume les résultats obtenus :

AA	Aa	aa	Total
322	410	207	939

Peut-on dire au vu de ces données si les proportions de génotypes seront identiques à la génération suivante ? Ceci suppose que les proportions ne dépendent pas du sexe et que l'homme et la femme donnent un allèle au hasard et de façon indépendante.

On note $\hat{p}_0$, $\hat{q}_0$ et $\hat{r}_0$ les fréquences d'obtention d'individus de type respectif AA, Aa et aa dans un échantillon de taille N.

$\hat{p}_0$, $\hat{q}_0$ et $\hat{r}_0$ étant des estimateurs de p_0 , q_0 et r_0 on détermine p_1 , q_1 et r_1 en utilisant les formules trouvées dans la partie II. :

$$p_1 = \left(\hat{p}_0 + \frac{1}{2} \hat{q}_0 \right)^2, \quad r_1 = \left(\hat{r}_0 + \frac{1}{2} \hat{q}_0 \right)^2 \quad \text{et} \quad q_1 = 1 - p_1 - r_1.$$

Sous hypothèse de stabilité p_0 , q_0 et r_0 doivent être égaux respectivement à p_1 , q_1 et r_1 et la quantité $d^2 = \frac{(\hat{p}_0 - p_1)^2}{p_1} + \frac{(\hat{q}_0 - q_1)^2}{q_1} + \frac{(\hat{r}_0 - r_1)^2}{r_1}$ doit être faible.

On démontre que $d^2 = \frac{(\hat{p}_0 \hat{r}_0 - \frac{1}{4} \hat{q}_0^2)^2}{(\hat{p}_0 + \frac{1}{2} \hat{q}_0)^2 (\hat{r}_0 + \frac{1}{2} \hat{q}_0)^2}$ (1) et que N d^2 suit asymptotiquement

une loi du χ^2 à 1 degré de liberté. On peut alors par une table ou par simulation déterminer un seuil, par exemple 10% et décider de rejeter l'hypothèse de stabilité si N d^2 dépasse ce seuil. Ici :

$$\hat{p}_0 = \frac{322}{939} \quad \hat{q}_0 = \frac{410}{939} \quad \hat{r}_0 = \frac{207}{939}$$

et on obtient :

$$N d^2 = 939 \frac{(322 \times 207 - \frac{1}{4} 410^2)^2}{(322 + \frac{1}{2} 410)^2 \times (207 + \frac{1}{2} 410)^2} \approx 12,08.$$

Par simulation¹² on obtient un neuvième décile de 2,7 environ et on peut ainsi rejeter l'hypothèse d'une stabilité des proportions avec un risque d'erreur inférieur à 10%.

¹² On trouvera un programme de simulation en ligne sur :
<http://perso.wanadoo.fr/jpq/proba/hardy-weinberg/hwkh2.htm>

Une autre façon d'envisager le problème

On sait que sous hypothèse de stabilité $\frac{q_k^2}{p_k r_k} = 4$ pour tout $k \neq 0$, (II. 4.) et donc la valeur

de $\Delta = \frac{\hat{q}_0^2}{\hat{p}_0 \hat{r}_0}$ doit être proche de 4 si N est grand.

La simulation¹³ permet d'approcher sous cette hypothèse la distribution de cette variable aléatoire.

Si le risque choisi est r on définit deux valeurs a et b de telle façon que $r / 2$ valeurs de cette distribution soient inférieures à a et que $r / 2$ valeurs soient supérieures à b . Ainsi le rejet de l'hypothèse de situation stable dès que Δ est à l'extérieur de l'intervalle $[a ; b]$ se fera avec un risque d'erreur inférieur à r . Si par exemple 5% des valeurs simulées sont inférieures à a et si 5% des valeurs simulées sont supérieures à b alors le risque de rejet à tort sera de 10%. Le test utilisé ici est encore appelé test bilatéral.

Loi de Hardy-Weinberg et test d'indépendance

Dire que les proportions des génotypes sont constantes de générations en générations revient à supposer que ces proportions sont identiques chez l'homme et la femme, ceux-ci donnant lors de la conception un allèle au hasard et de façon indépendante. Les données du paragraphe précédent doivent être compatibles avec le modèle suivant :

Le père donne l'allèle	A	a
La mère donne l'allèle		
A	p	$\frac{1}{2}q$
a	$\frac{1}{2}q$	r

où p, q, r sont les proportions des génotypes AA, Aa, aa.

Sous cette hypothèse le tableau des données pourra être représenté sous la forme :

Le père donne l'allèle	A	a	totaux
La mère donne l'allèle			
A	322	205	527
a	205	207	412
totaux	527	412	939

De façon plus générale si les effectifs sont représentés par le tableau suivant :

¹³ Simulation disponible sur
<http://perso.wanadoo.fr/jpq/proba/hardy-weinberg/hw.htm> et
<http://perso.wanadoo.fr/jpq/proba/hardy-weinberg/hw1.htm>

Le père donne l'allèle	A	a	totaux
La mère donne l'allèle			
A	a	b	n
A	c	d	N-n
totaux	m	N-m	N

Avec $b = c$ et donc $m = n$ alors on démontre que $z = N \frac{(ad - bc)^2}{nm(N - n)(N - m)}$ suit

approximativement pour N assez grand une loi du khi2 à 1 degré de liberté (voir document d'accompagnement – Annexe, probabilités et statistique Séries ES et S page 148 et 149).

Or on peut remarquer que z est égal à la quantité $N d^2$ du paragraphe précédent et que par conséquent le premier test rencontré dans ce document n'est rien d'autre qu'un test d'indépendance.

Démonstration de l'égalité (1) $d^2 = \frac{(\hat{p}_0 \hat{r}_0 - \frac{1}{4} \hat{q}_0^2)^2}{(\hat{p}_0 + \frac{1}{2} \hat{q}_0)^2 (\hat{r}_0 + \frac{1}{2} \hat{q}_0)^2} :$

d^2 est définie par l'égalité $d^2 = \frac{(\hat{p}_0 - p_1)^2}{p_1} + \frac{(\hat{q}_0 - q_1)^2}{q_1} + \frac{(\hat{r}_0 - r_1)^2}{r_1}$.

Si on note α la différence $\hat{p}_0 - \hat{r}_0$ alors d'après le II. du TD :

$\alpha = \hat{p}_0 - \hat{r}_0 = p_1 - r_1$. On a donc $\hat{p}_0 - p_1 = \hat{r}_0 - r_1$ et par suite $\hat{q}_0 - q_1 = -2(\hat{p}_0 - p_1)$. Ainsi d^2 s'écrit encore :

$$\begin{aligned} d^2 &= (\hat{p}_0 - p_1)^2 \left(\frac{1}{p_1} + \frac{4}{q_1} + \frac{1}{r_1} \right) \\ &= (\hat{p}_0 - p_1)^2 \left(\frac{q_1 r_1 + 4 p_1 r_1 + p_1 q_1}{p_1 q_1 r_1} \right) \\ &= (\hat{p}_0 - p_1)^2 \left(\frac{q_1 (1 - q_1) + 4 p_1 r_1}{p_1 q_1 r_1} \right) \\ &= (\hat{p}_0 - p_1)^2 \left(\frac{1}{p_1 r_1} \right) \end{aligned}$$

car $q_1^2 = 4 p_1 r_1$.

Or $p_1 = (\hat{p}_0 + \frac{1}{2} \hat{q}_0)^2$ et $r_1 = (\hat{r}_0 + \frac{1}{2} \hat{q}_0)^2$. Par suite $\hat{p}_0 - p_1 = \hat{p}_0 - \hat{p}_0^2 - \hat{p}_0 \hat{q}_0 - \frac{1}{4} \hat{q}_0^2$

et on a donc bien

$$\frac{(\hat{p}_0 - p_1)^2}{p_1 r_1} = \frac{(\hat{p}_0 \hat{r}_0 - \frac{1}{4} \hat{q}_0^2)^2}{(\hat{p}_0 + \frac{1}{2} \hat{q}_0)^2 (\hat{r}_0 + \frac{1}{2} \hat{q}_0)^2}.$$

ANNEXE 1 : PETIT GLOSSAIRE DE GÉNÉTIQUE¹⁴

ADN :

Acide désoxyribonucléique. Macromolécule biologique constitutive des chromosomes qui est le support moléculaire de l'information génétique et de l'hérédité.

Allèle :

Variante, état d'un gène. Tout gène peut avoir plusieurs allèles (au moins deux) qui déterminent souvent l'apparition de caractères héréditaires différents.

Caractère :

Particularité physique ou biologique d'un organisme commandée par un ou plusieurs gènes. L'ensemble des caractères constitue le phénotype.

Chromosome :

Au moment des divisions cellulaires, l'ADN se compacte sur lui-même. Les morceaux d'ADN du génome deviennent visibles au microscope : ce sont les chromosomes, qui correspondent donc à un état provisoire de l'ADN. Chaque chromosome possède un bras court appelé p et un bras long appelé q, reliés au niveau du centromère.

Gène :

Élément du patrimoine génétique correspondant à une portion d'ADN sur un chromosome et participant à l'apparition d'un caractère.

Hérédité :

Transmission des caractères génétiques d'une génération aux suivantes.

Hétérozygote :

Présence de deux allèles différents (un allèle dominant et un allèle récessif ou deux allèles codominants) pour un gène donné sur les chromosomes homologues.

Homozygote :

Présence de deux allèles identiques pour un gène donné sur les chromosomes homologues.

Locus :

Emplacement d'un gène sur un chromosome. On le désigne par le n° du chromosome puis sa position sur le bras chromosomique.

Phénotype :

Ensemble des caractères observables résultant de l'expression du génotype.

Récessif (allèle) :

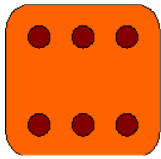
Version d'un gène qui ne s'exprime pas en présence de l'allèle dominant. Cet allèle ne s'exprime qu'à l'état homozygote.

¹⁴ Pour plus de précision on peut consulter les sites :

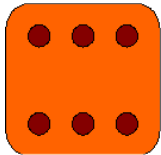
<http://www.science-generation.com/biotlexi.php>

<http://gerard.chevrier.m2osw.com/gcartable/genetique/genlexique.htm>

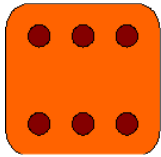
VI. ADÉQUATION



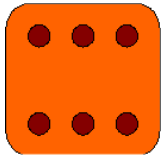
Adéquation d'une distribution de fréquences à une loi de probabilités



Tests d'adéquation au baccalauréat



Exercice d'évaluation



Exercices dans les manuels

on calculera plutôt

$$d_{obs}^2 = \left(f_1 - \frac{1}{6}\right)^2 + \left(f_2 - \frac{1}{6}\right)^2 + \left(f_3 - \frac{1}{6}\right)^2 + \left(f_4 - \frac{1}{6}\right)^2 + \left(f_5 - \frac{1}{6}\right)^2 + \left(f_6 - \frac{1}{6}\right)^2$$

Par de nombreuses simulations on étudie la distribution de la variable aléatoire d^2 obtenue en remplaçant dans la formule précédente les fréquences observées f_i par les fréquences obtenues par simulation de n lancers avec un dé « normal ».

Si la valeur de d_{obs}^2 n'est dépassée que très rarement avec un dé « normal » (par exemple dans moins de 5% des cas) on rejette l'hypothèse que le dé testé est « normal ».

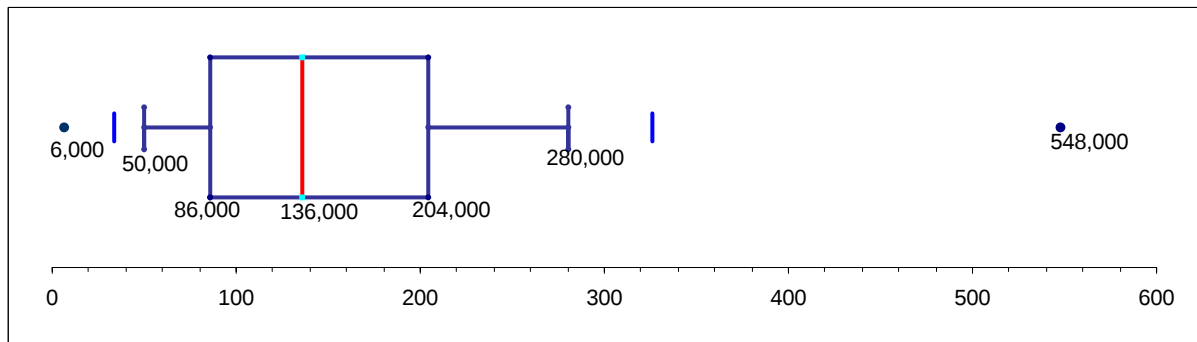
Bien entendu comme $f_i = \frac{e_i}{n}$ on a $d_{obs}^2 = \frac{1}{n^2} D_{obs}^2$ et les deux procédures sont équivalentes.

Généralisations

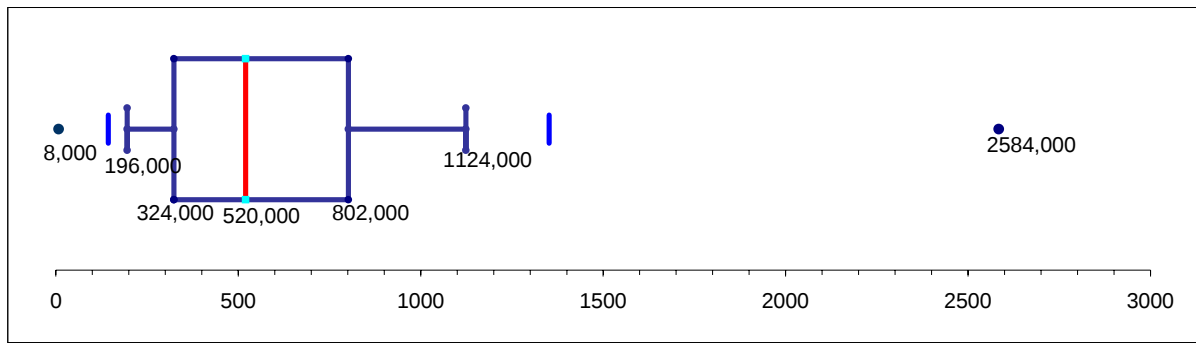
Si on désire généraliser cette procédure un premier inconvénient est que D_{obs}^2 (ainsi que d_{obs}^2) dépend du nombre n de fois où l'on répète l'expérience réelle. Pour chaque valeur de n il faudra renouveler les simulations pour déterminer un maximum acceptable pour D_{obs}^2 ou d_{obs}^2 .

Les trois boîtes à moustaches ci-dessous représentent la répartition des valeurs observées de D_{obs}^2 lors de 1000 simulations de n lancers d'un dé normal pour trois valeurs de n .

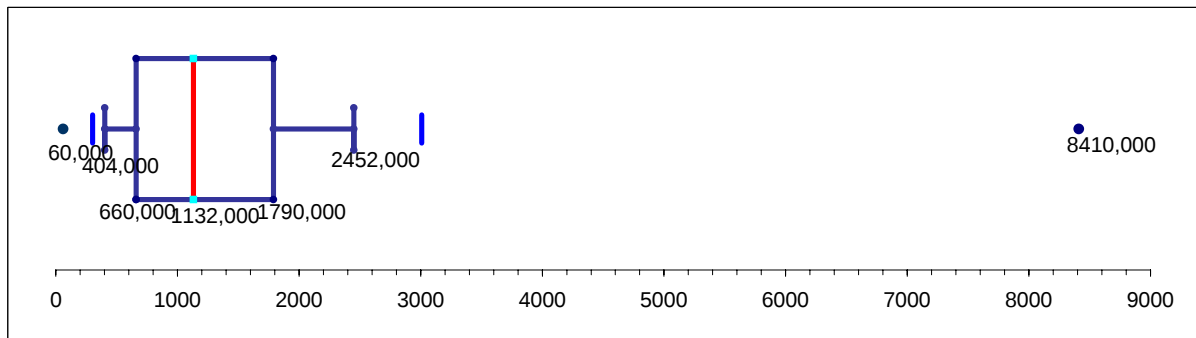
1000 fois 180 lancers ($n=180$)



1000 fois 720 lancers ($n=720=4*180$)



1000 fois 1620 lancers ($n=1620=9*180$)



n	Q 0.05	Décile 1	Quartile 1	Médiane	Quartile 3	Décile 9	Q 0.95
180	30	50	86	136	204	280	325
720	128	196	324	520	802	1124	1440
1620	290	404	660	1132	1790	2452	3020

L'examen de ces diagrammes et du tableau ci-dessus incite à conjecturer que les quantiles de D_{obs}^2 sont « proportionnels » à n . D'où l'idée de calculer plutôt $\frac{D_{obs}^2}{n} = n \times d_{obs}^2$.

On démontre en effet que dans le cas du test de l'équiprobabilité sur 6 valeurs $6 \times \frac{D^2}{n}$ et donc aussi $6 \times n \times d^2$ suivent asymptotiquement une loi du Khi-deux à 5 degrés de liberté.

La consultation d'une table du Khi-deux à 5 degrés de liberté donne les quantiles suivants :

Deg. lib.	Q 0.05	Décile 1	Quartile 1	Médiane	Quartile 3	Décile 9	Q 0.95
5	1.15	1.61	2.67	4.35	6.63	9.24	11.1

En multipliant ces valeurs par n et en divisant par 6 on obtient effectivement des quantiles très proches de ceux obtenus par simulation :

Deg. lib.	Q 0.05	Décile 1	Quartile 1	Médiane	Quartile 3	Décile 9	Q 0.95
5	1.15	1.61	2.67	4.35	6.63	9.24	11.1
*180/6	34	48	80	130	199	277	333
*720/6	138	193	320	522	796	1109	1332
*1620/6	310	435	721	1174	1790	2495	2997

Bien entendu, lorsque l'on veut tester l'équiprobabilité sur k valeurs c'est la quantité $k \times \frac{D_{obs}^2}{n} = k \times n \times d_{obs}^2$ qu'il convient de comparer aux quantiles du Khi-deux à $k-1$ degrés de liberté. La loi de probabilité de $k \times \frac{D^2}{n} = k \times n \times d^2$ ne dépend donc plus de n (assez grand tout de même !) mais uniquement du nombre k de valeurs différentes prises par le caractère à tester.

Généralisation au cas où la loi de probabilité n'est pas l'équiprobabilité (hors programme au lycée)

On suppose que l'on a la distribution des fréquences observées sur n expériences :

Valeurs	x_1	x_2	...	x_i	...	x_k	Total
Fréquences	f_1	f_2	...	f_i	...	f_k	1

On veut tester si cette distribution de fréquence peut correspondre à la loi de probabilité :

Valeurs	x_1	x_2	...	x_i	...	x_k	Total
Probabilités	p_1	p_2	...	p_i	...	p_k	1

C'est alors la quantité :

$$\chi_{obs}^2 = n \sum_{i=1}^{i=k} \frac{(f_i - p_i)^2}{p_i} = \sum_{i=1}^{i=k} \frac{(e_i - np_i)^2}{np_i}$$

qu'il convient de comparer aux quantiles du Khi-deux à $k-1$ degrés de liberté.

On notera que dans le cas de l'équiprobabilité : $\forall i = 1..k, p_i = \frac{1}{k}$ et la formule devient :

$$\chi_{obs}^2 = n \sum_{i=1}^{i=k} \frac{(f_i - p_i)^2}{p_i} = n \sum_{i=1}^{i=k} \frac{\left(f_i - \frac{1}{k}\right)^2}{\frac{1}{k}} = nk \sum_{i=1}^{i=k} \left(f_i - \frac{1}{k}\right)^2$$

qui est bien celle utilisée dans les cas particuliers étudiés précédemment.

On trouvera en annexe 3 une table du Khi deux.

Un test d'adéquation du Khi deux, utilisable dans le cas le plus général vous est proposé dans le classeur EXCEL Chi2_freq_probas.XLS disponible sur le CD.

TESTS D'ADÉQUATION AU BACCALAURÉAT

Plusieurs sujets du Bac 2003 comportaient des exercices de probabilités comme celui qui figure en tête du sujet de ES (commun à tous les candidats)¹⁵. Il s'agit de tester l'adéquation d'une distribution de fréquences observées à un modèle théorique équiprobable.

On désire mesurer l'écart, une « **distance** » entre les deux répartitions. L'énoncé invite à calculer :

$$d_{obs}^2 = \left(f_1 - \frac{1}{5}\right)^2 + \left(f_2 - \frac{1}{5}\right)^2 + \left(f_3 - \frac{1}{5}\right)^2 + \left(f_4 - \frac{1}{5}\right)^2 + \left(f_5 - \frac{1}{5}\right)^2$$

et donne une approximation de la distribution de la variable aléatoire d^2 obtenue par de nombreuses simulations

Si, lors de ces simulations la valeur de d_{obs}^2 est dépassée très rarement (en l'occurrence dans moins 10% des cas) on rejette l'hypothèse d'équiprobabilité.

L'énoncé

Baccalauréat ES 2003

Exercice I (4 points)

Commun à tous les candidats

Les guichets d'une agence bancaire d'une petite ville sont ouverts au public cinq jours par semaine : les mardi, mercredi, jeudi, vendredi et samedi.

Le tableau ci-dessous donne la répartition journalière des 250 retraits d'argent liquide effectués aux guichets une certaine semaine.

Jour de la semaine	mardi	mercredi	jeudi	vendredi	samedi
Rang i du jour	1	2	3	4	5
Nombre de retraits	37	55	45	53	60

On veut tester l'hypothèse « le nombre de retraits est indépendant du jour de la semaine ».

On suppose donc que le nombre des retraits journaliers est égal à $\frac{1}{5}$ du nombre des retraits de la semaine.

On pose $d_{obs}^2 = \sum_{i=1}^5 \left(f_i - \frac{1}{5}\right)^2$ où f_i est la fréquence des retraits du $i^{\text{ème}}$ jour.

1. Calculer les fréquences des retraits pour chacun des cinq jours de la semaine.

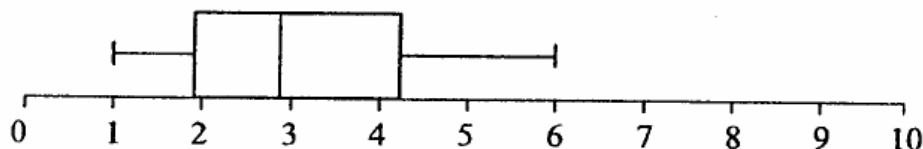
2. Calculer alors la valeur de 1000 d_{obs}^2 (la multiplication par 1000 permet d'obtenir un résultat plus lisible).

¹⁵ Voir en annexe 1 le sujet de ES posé au centre de Pondichery qui est très analogue à celui analysé ici.

3. En supposant qu'il y a équiprobabilité des retraits journaliers, on a simulé 2000 séries de 250 retraits hebdomadaires.

Pour chaque série, on a calculé la valeur du 1000 d_{obs}^2 , correspondant. On a obtenu ainsi 2000 valeurs de 1000 d_{obs}^2 .

Ces valeurs ont permis de construire le diagramme en boîte ci-dessous où les extrémités des « pattes » correspondent respectivement au premier décile et au neuvième décile.



Lire sur le diagramme une valeur approchée du neuvième décile.

4. En argumentant soigneusement la réponse, dire si pour la série observée au début, on peut affirmer, avec un risque d'erreur inférieur à 10%, que « le nombre de retraits est indépendant du jour de la semaine » ?

Un corrigé commenté

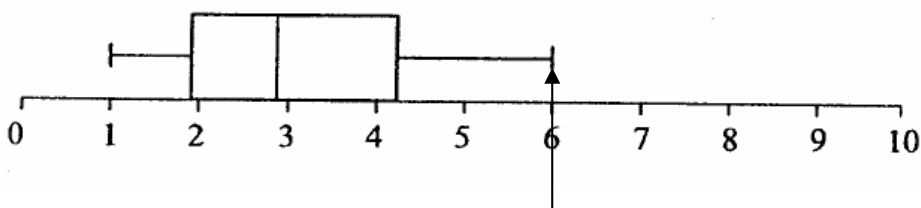
Questions 1 et 2 :

Les calculs ne sont pas difficiles mais bien plus commodes si on dispose d'un tableau.

Jour	mardi	mercredi	jeudi	vendredi	samedi	Total
rang	1	2	3	4	5	
retraits	37	55	45	53	60	250
fréquence	0,148	0,22	0,18	0,212	0,24	1
écart	-0,052	0,02	-0,02	0,012	0,04	
carré	0,002704	0,0004	0,0004	0,000144	0,0016	0,005248
					1000*d ² obs	5,248

La multiplication par 1000 peut sembler arbitraire. Ce changement d'échelle n'est évidemment pas nécessaire si on procède par simulation dans la suite et le diagramme en boîte donné dans la question 3 pouvait tout aussi bien comporter en abscisse les valeurs de d^2 mais leur ordre de grandeur les rendrait moins lisibles.

Question 3 :



L'extrémité de la « moustache » de droite correspond au neuvième décile qui est donc d'environ 6.

Question 4 :

La valeur $1000 d^2_{obs}$ observée est de 5,248. Elle est inférieure au neuvième décile qui est 6. On ne peut donc rejeter l'hypothèse d'équipartition.

L'énoncé souhaite une argumentation soignée et utilise le terme de « risque d'erreur ». Si on interprète cette expression comme « probabilité de prendre une mauvaise décision » on se heurte à une difficulté majeure qui rend, dans le cas de l'énoncé, toute réponse impossible.

Il y a en effet deux risques d'erreur :

Lorsque l'on teste une hypothèse (traditionnellement appelée « hypothèse nulle » et notée H_0), on procède à des calculs de probabilités en supposant qu'elle est vraie. Les seules probabilités que l'on peut alors calculer sont des probabilités calculées dans le modèle où l'hypothèse H_0 est vraie. On peut ainsi calculer la probabilité de **rejeter à tort** l'hypothèse H_0 (alors qu'elle est vraie)¹⁶ en adoptant une certaine règle de décision. La probabilité d'**accepter à tort** l'hypothèse H_0 (alors qu'elle est fausse)¹⁷ ne pourrait se calculer que dans le vrai modèle inconnu dans la pratique.

Dans le cas du test d'équiprobabilité concernant les cinq jours proposés dans l'énoncé, la procédure consiste à simuler 250 tirages, à calculer les fréquences et la quantité $1000 d^2_{sim}$.

La règle de décision, suggérée par l'énoncé, consiste à rejeter l'hypothèse d'équiprobabilité si $d^2_{obs} > 6$. Cette dernière valeur étant le neuvième décile des 2000 valeurs de $1000 d^2$ obtenues par une simulation dans le modèle équiprobable. Le calcul de $1000 d^2_{obs}$ avec les fréquences observées donne un résultat inférieur à ce neuvième décile. On ne peut donc rejeter l'hypothèse d'équiprobabilité.

Si le nombre de retraits n'est pas indépendant du jour de la semaine, seule la connaissance de la répartition réelle du nombre des retraits selon les jours permet de calculer (ou tout du moins d'estimer par simulation) la probabilité d'accepter à tort l'hypothèse d'équiprobabilité en appliquant la règle de décision choisie.

Exemples

Supposons que les proportions réelles (inconnues dans la pratique) soient :

1	2	3	4	5
0.22	0.22	0.21	0.21	0.14

Ce n'est pas l'équipartition, sans en être trop grossièrement éloigné !

Avec ce modèle, 2000 simulations de 250 tirages avec calcul de $1000 * d^2_{sim}$ donnent la répartition suivante :

0	1	2	3	4	5	6	7	8	9	10	plus	Total
	13	60	109	162	205	213	200	183	168	158	529	2000
											<6	762

On remarque que 762 tirages ont donné $1000 d^2_{sim}$ inférieur à 6. La probabilité que l'on a, avant de faire le test, d'accepter (à tort) l'hypothèse d'équiprobabilité est de $762/2000$ soit 0,382 alors que la probabilité de rejeter (à raison) l'équiprobabilité est $1-0,382$ soit 0.618.

¹⁶ appelé risque de première espèce et souvent noté α dans la théorie des tests d'hypothèses.

¹⁷ appelé risque de deuxième espèce et souvent noté β dans la théorie des tests d'hypothèses.

Plus la répartition réelle est proche de l'équirépartition, plus la probabilité d'accepter l'hypothèse d'équiprobabilité est élevée.

Dans le cas de proportions réelles pratiquement équilibrées comme, par exemple :

1	2	3	4	5
0.205	0.205	0.200	0.195	0.195

Avec ce modèle, 2000 simulations de 250 tirages avec calcul de $1000 \cdot d_{sim}^2$ donnent la répartition suivante :

0	1	2	3	4	5	6	7	8	9	10	plus	Total
	252	419	394	315	226	138	97	54	29	32	44	2000
											<6	1744

Cette fois ce sont 1744 tirages qui ont donné $1000 \cdot d_{sim}^2$ inférieur à 6. La probabilité que l'on a, avant de faire le test, d'accepter (à tort) l'hypothèse d'équiprobabilité est de 0,872. Il est évidemment normal que cette valeur soit très proche de 0,9 qui est la probabilité d'accepter l'équiprobabilité alors qu'elle est vraie.

Dans le cas d'une répartition réelle plus nettement déséquilibrée comme, par exemple :

1	2	3	4	5
0.25	0.25	0.20	0.15	0.15

On observe une répartition des $1000 \cdot d_{sim}^2$ qui permet d'estimer la probabilité d'accepter à tort l'équiprobabilité à 0,09.

0	1	2	3	4	5	6	7	8	9	10	plus	Total
	1	7	14	30	43	85	97	88	134	149	1352	2000
											<6	180

Le risque d'erreur (la probabilité de prendre une mauvaise décision à savoir l'acceptation à tort) lorsque la répartition n'est pas uniforme dépend donc de la répartition réelle qui est, par essence, inconnue dans le problème pratique. Plus la répartition réelle est déséquilibrée, plus le test a des chances de le détecter comme l'illustre le tableau récapitulatif des exemples précédents :

Répartition	22/22/21/21/14	205/205/200/195/195	25/25/20/15/15
Probabilité d'acceptation (à tort)	0.382	0.872	0.090

Pour procéder à d'autres simulations de ce type vous pouvez utiliser le classeur EXCEL banqueBacES.XLS disponible sur le CD.

La seule majoration générale que l'on peut donner dans le cas de cet exemple c'est que la probabilité d'acceptation à tort est toujours inférieure à 0,9.

La prudence incite donc à donner une réponse comme :

« On ne peut rejeter l'hypothèse d'équiprobabilité »
« Rien ne permet de rejeter l'hypothèse d'équiprobabilité »
en laissant dans le flou le « risque d'erreur » dans ce cas. La formulation de la question n'incitait malheureusement pas les candidats à cette prudence.

EXERCICE D'ÉVALUATION

(d'après n°4 Hyperbole Nathan)

Nos supposons que dans certaines courses en couloirs, les concurrents sont affectés par tirage au sort dans les couloirs numérotés de 1 à 8.

Pour étudier si certains couloirs présentent un avantage par rapport aux autres on a relevé le numéro du couloir du gagnant lors de 144 courses. On a obtenu les résultats suivants :

Couloir du vainqueur	1	2	3	4	5	6	7	8
Nombre de courses	29	19	18	25	17	10	15	11

a) Calculer la quantité $d_{obs}^2 = \left(f_1 - \frac{1}{8}\right)^2 + \left(f_2 - \frac{1}{8}\right)^2 + \dots + \left(f_7 - \frac{1}{8}\right)^2 + \left(f_8 - \frac{1}{8}\right)^2$

où $f_1, f_2, \dots, f_7, f_8$ sont les fréquences associées à chaque couloir.

b) On a simulé 200 fois une série de 144 expériences modélisable par la loi équirépartie sur l'ensemble $\{1, 2, 3, 4, 5, 6, 7, 8\}$. On a obtenu ainsi 200 valeurs de d_{obs}^2 qui sont données, triées en ordre croissant, dans le tableau ci-dessous :

0,000388	0,000888	0,001150	0,001338	0,001588	0,001863	0,002225	0,002550	0,002888	0,003700
0,000450	0,000913	0,001163	0,001375	0,001613	0,001875	0,002250	0,002563	0,002900	0,003763
0,000475	0,000913	0,001175	0,001375	0,001638	0,001875	0,002250	0,002600	0,002938	0,003800
0,000525	0,000925	0,001188	0,001375	0,001675	0,001900	0,002263	0,002638	0,002950	0,003838
0,000525	0,000938	0,001188	0,001388	0,001688	0,001988	0,002275	0,002638	0,003050	0,003838
0,000550	0,000938	0,001188	0,001400	0,001688	0,002038	0,002275	0,002638	0,003088	0,003850
0,000588	0,000950	0,001200	0,001400	0,001688	0,002038	0,002288	0,002650	0,003200	0,003850
0,000650	0,000950	0,001225	0,001413	0,001700	0,002050	0,002300	0,002675	0,003200	0,004000
0,000675	0,000988	0,001225	0,001425	0,001725	0,002050	0,002300	0,002688	0,003338	0,004013
0,000700	0,001000	0,001238	0,001438	0,001725	0,002075	0,002375	0,002738	0,003363	0,004175
0,000700	0,001000	0,001250	0,001450	0,001750	0,002075	0,002400	0,002763	0,003375	0,004213
0,000738	0,001013	0,001263	0,001450	0,001750	0,002075	0,002413	0,002775	0,003475	0,004350
0,000750	0,001063	0,001263	0,001488	0,001775	0,002088	0,002450	0,002800	0,003488	0,004438
0,000763	0,001075	0,001275	0,001513	0,001800	0,002113	0,002475	0,002800	0,003500	0,004938
0,000763	0,001100	0,001288	0,001513	0,001800	0,002125	0,002500	0,002800	0,003513	0,005000
0,000775	0,001100	0,001288	0,001525	0,001825	0,002138	0,002500	0,002838	0,003550	0,005113
0,000800	0,001100	0,001313	0,001525	0,001825	0,002138	0,002500	0,002850	0,003625	0,005300
0,000813	0,001113	0,001313	0,001525	0,001825	0,002150	0,002525	0,002850	0,003663	0,005400
0,000825	0,001113	0,001313	0,001563	0,001850	0,002163	0,002525	0,002875	0,003675	0,005425
0,000888	0,001150	0,001325	0,001563	0,001863	0,002188	0,002538	0,002888	0,003688	0,007525

Représenter ces résultats par un diagramme en boîte.

c) On décide de considérer que certains couloirs présentent un avantage par rapport à d'autres si la valeur d_{obs}^2 calculée à partir des observations réelles dépasse le neuvième décile des valeurs de d_{obs}^2 obtenues par simulation. Quelle décision va-t-on prendre ?

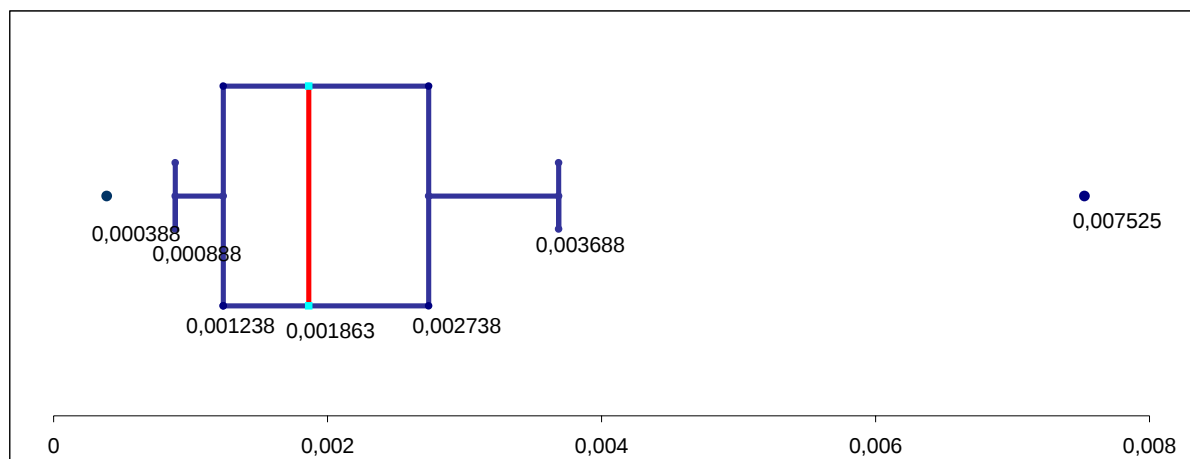
Corrigé

a) Le calcul de d^2 est détaillé dans le tableau ci dessous.

Valeurs	Effectifs	Fréquences	Probabilités	Écarts	Carrés
1	29	0,2013889	0,125	0,0763889	0,0058352623
2	19	0,1319444	0,125	0,0069444	0,0000482253
3	18	0,1250000	0,125	0,0000000	0,0000000000
4	25	0,1736111	0,125	0,0486111	0,0023630401
5	17	0,1180556	0,125	-0,0069444	0,0000482253
6	10	0,0694444	0,125	-0,0555556	0,0030864198
7	15	0,1041667	0,125	-0,0208333	0,0004340278
8	11	0,0763889	0,125	-0,0486111	0,0023630401
	144	1	1	$d^2=$	0,0141782407

b) Diagramme en boîte :

Effectif	200,000
Minimum	0,000388
Maximum	0,007525
Médiane	0,001863
Quartile 1	0,001238
Quartile 3	0,002738
Décile 1	0,000888
Décile 9	0,003688
Moyenne	0,002112



c) Le neuvième décile est 0,003688. La valeur obtenue à partir des observations réelles dépasse largement ce décile. On considérera que certains couloirs présentent un avantage par rapport à d'autres.

Remarque :

L'habillage de l'exercice est assez artificiel. Les sportifs sont unanimes à considérer que les couloirs centraux sont les « meilleurs ». L'attribution des couloirs ne se fait généralement pas par tirage au sort mais en fonction des temps de référence obtenus lors d'épreuves antérieures (séries par exemple) et ce dans l'ordre 4-5-3-6-2-7-1-8.

EXERCICES DANS LES MANUELS

Certains manuels (assez peu nombreux à vrai dire) proposent des exercices intéressants sur le thème de l'adéquation. Nous en citons quelques uns assortis de commentaires et d'une proposition de corrigé.

Hachette, collection Terracher

Ce manuel indique en préambule de cette série d'exercices qu'il convient de calculer la

quantité $d^2 = \sum_{i=1}^k \frac{(e_i - np_i)^2}{np_i}$ et fournit pour certains nombres de degré de liberté les

valeurs à ne pas dépasser (en fait un extrait d'une table du Khi deux sans la nommer)

N°70

La pièce acceptable

On lance 100 fois une pièce de monnaie. Entre quelles valeurs doit être compris le nombre de faces obtenu pour que la pièce soit considérée comme « équilibrée ».

Cet exercice est proposé dans le paragraphe consacré au test du Khi deux. En fait l'exemple correspond à 1 seul degré de liberté. Le test d'adéquation du Khi deux est dans ce cas équivalent au test sur une proportion.

Corrigé

Si on note P le nombre de piles obtenu, le nombre de faces est alors $100 - P$, l'effectif théorique attendu est 50 donc :

$$d^2 = \frac{(P-50)^2}{50} + \frac{(100-P-50)^2}{50} = \frac{2(P-50)^2}{50} = \frac{(P-50)^2}{25}$$

Pour 1 degré de liberté, la valeur maximale (fournie dans l'introduction de cette série d'exercices) est $\nu_{19} = 3,84$. On considèrera la pièce comme « équilibrée » si

$$\frac{(P-50)^2}{25} < 3,84 \text{ c'est à dire } (P-50)^2 < 96 \text{ soit } |P-50| < \sqrt{96}.$$

On obtient les encadrements $-9,8 < P-50 < 9,8$ donc $40,2 < P < 59,8$ et comme P est entiers : $41 \leq P \leq 59$.

N°71

L'égalité des sexes

Dans une commune, en 2001, on a enregistré 418 naissances de garçons et 388 naissances de filles. Le maire annonce : « Dans ma commune, il naît autant de filles que de garçons. »

Dans tout le pays, en 2001, 41800 garçons et 38800 filles sont nés. Le président annonce : « Dans mon pays, il naît autant de filles que de garçons. »

L'un des élus a une formation en Statistique. Peut-on deviner lequel ?

Le problème correspond encore à un seul degré de liberté et a le mérite de mettre en évidence les effets d'une homothétie sur une répartition.

Corrigé

Dans la commune il y a eu 806 naissances donc : $d^2 = \frac{(418-403)^2}{403} + \frac{(388-403)^2}{403} \square 1.12$

Pour 1 degré de liberté, la valeur maximale ν_{19} est 3,84. L'affirmation du maire est donc « statistiquement » acceptable. Dans le pays, il faut multiplier tous les nombres utilisés par 100 ce qui donne $d^2 = \frac{(41800-40300)^2}{40300} + \frac{(38800-40300)^2}{40300} \square 111.66$ qui était prévisible sans faire le calcul puisque les numérateurs sont multipliés par 100^2 et les dénominateurs par 100. La valeur obtenue est très supérieure à $\nu_{19} = 3,84$. Le président à tort.

N°72

Le test des paires

Après avoir lancé 600 fois un dé, on a obtenu les résultats suivants :

numéro	1	2	3	4	5	6
Nombre d'occurrences	85	99	115	98	86	117

1) Calculer d^2 . Peut-on rejeter la régularité du dé

2) On regroupe les résultats par paires, suivant les trois modalités $\{1,5\}, \{2,4\}, \{3,6\}$.

Répondre aux mêmes questions.

3) Alors ?

Corrigé

1) On a un effectif théorique de 100 donc : $d^2 = \frac{15^2 + 1^2 + 15^2 + 2^2 + 14^2 + 17^2}{100} = 9,4$

Pour 5 degrés de liberté, la valeur maximale ν_{19} est 11,07. On ne peut donc rejeter la régularité du dé.

2) Le regroupement par paires donne

Paire	$\{1,5\}$	$\{2,4\}$	$\{3,6\}$
Effectif	171	197	232

Dans ce cas : $d^2 = \frac{29^2 + 3^2 + 32^2}{200} = 9,37$ mais pour 2 degrés de liberté, la valeur maximale

ν_{19} n'est plus que 5,99.

3) Alors ? On serait amené à rejeter l'équiprobabilité sur les trois paires. Le procédé qui consiste à regrouper, a posteriori, en paires en choisissant le regroupement qui donnera les plus gros écarts est très contestable.

Il y a ici 15 répartitions en paires possibles et parmi ces 15 regroupements 5 seulement qui conduisent à rejeter l'équiprobabilité (dont bien sûr, celle opportunément choisie dans l'énoncé).

En statistique on évite de choisir a posteriori la procédure de test en sélectionnant celle qui confirme ce que l'on cherchait à démontrer !

Calculs pour tous les regroupements par paires possibles :

Les quinze regroupements par paires possibles

1	2	3	4	5	6		
85	99	115	98	86	117		
1	2	3	4	5	6		
	184		213		203		
	-16		13		3	d²	
	256		169		9	434	2,17
1	2	3	5	4	6		
	184		201		215		
	-16		1		15	d²	
	256		1		225	482	2,41
1	2	3	6	5	4		
	184		232		184		
	-16		32		-16	d²	
	256		1024		256	1536	7,68
1	3	2	4	5	6		
	200		197		203		
	0		-3		3	d²	
	0		9		9	18	0,09
1	3	2	5	4	6		
	200		185		215		
	0		-15		15	d²	
	0		225		225	450	2,25
1	3	2	6	4	5		
	200		216		184		
	0		16		-16	d²	
	0		256		256	512	2,56
1	4	2	3	5	6		
	183		214		203		
	-17		14		3	d²	
	289		196		9	494	2,47
1	4	2	5	3	6		
	183		185		232		
	-17		-15		32	d²	
	289		225		1024	1538	7,69
1	4	2	6	5	3		
	183		216		201		
	-17		16		1	d²	
	289		256		1	546	2,73

1	5	2	3	4	6		
	171		214		215		
	-29		14		15		d²
	841		196		225	1262	6,31
1	5	2	4	3	6		
	171		197		232		
	-29		-3		32		d²
	841		9		1024	1874	9,37
1	5	2	6	3	4		
	171		216		213		
	-29		16		13		d²
	841		256		169	1266	6,33
1	6	2	3	4	5		
	202		214		184		
	2		14		-16		d²
	4		196		256	456	2,28
1	6	2	4	3	5		
	202		197		201		
	2		-3		1		d²
	4		9		1	14	0,07
1	6	2	5	3	4		
	202		185		213		
	2		-15		13		d²
	4		225		169	398	1,99

Nathan Collection Hyperbole

Contrairement à Terracher la définition adoptée dans cet ouvrage pour d_{obs}^2 que l'on calcule pour tester l'adéquation d'une distribution de fréquences $(f_1, f_2, \dots, f_{k-1}, f_k)$ à une loi équiprobable chacune des k éventualités ayant donc une probabilité $p = \frac{1}{k}$ est celle qui est citée dans le document d'accompagnement des programmes :

$$d_{obs}^2 = (f_1 - p)^2 + (f_2 - p)^2 + \dots + (f_{k-1} - p)^2 + (f_k - p)^2$$

Cette formule n'est pas rappelée dans tous les exercices. Les valeurs obtenues dépendent du nombre de fois où l'on répète l'expérience qui doit donc être le même dans la simulation destinée à déterminer la valeur maximale acceptable.

N°1

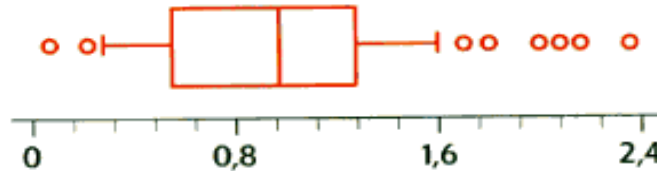
Pour tester une table de nombre au hasard, on a recensé parmi 250 chiffres consécutifs le nombre de 0, de 1,.... Les résultats sont rassemblés dans le tableau suivant.

Chiffre	0	1	2	3	4	5	6	7	8	9
Effectif	22	24	28	23	18	33	29	17	31	25

On se propose de savoir si ces résultats sont en adéquation avec la loi équirépartie sur l'ensemble $E = \{0,1,2,3,4,5,6,7,8,9\}$.

a) Calculer la valeur de d_{obs}^2 .

b) On a répété 1000 fois une simulation de 250 expériences modélisable par la loi équirépartie sur E et on a obtenu une série de 1000 valeurs de $250d_{obs}^2$ représentée par son diagramme en boîte.



Donner une valeur approchée du neuvième décile D_9 , de cette série.

c) Peut-on considérer que les fréquences des différents chiffres de cette table sont en accord avec une loi équirépartie ?

Corrigé

a) Dans le tableau ci-dessous détaillé le calcul de d_{obs}^2 . La valeur obtenue est 0,004032.

b) Pour la simulation on a relevé les valeurs de $250 * d_{obs}^2$. Avec un peu de flair, malgré une graduation déroutante on lit sur le diagramme le neuvième décile : 1,6.

Comme $250 * d_{obs}^2 = 1,008$ est inférieur à ce décile on ne peut rejeter l'accord avec une loi équirépartie.

Chiffres	Effectif	Fréquences	Probabilité	Écarts	Carrés
0	22	0,088	0,1	-0,012	0,000144
1	24	0,096	0,1	-0,004	0,000016
2	28	0,112	0,1	0,012	0,000144
3	23	0,092	0,1	-0,008	0,000064
4	18	0,072	0,1	-0,028	0,000784
5	33	0,132	0,1	0,032	0,001024
6	29	0,116	0,1	0,016	0,000256
7	17	0,068	0,1	-0,032	0,001024
8	31	0,124	0,1	0,024	0,000576
9	25	0,100	0,1	0,000	0,000000
Total	250	1,000	1	$d^2=$	0,004032
				$250d^2=$	1,008000

Remarque

Ici c'est $10 * 250 * d_{obs}^2 = 10,08$ qui est une réalisation d'un Khi deux à 9 degrés de liberté.

La table du Khi deux fournit un neuvième décile de 14,7 (assez proche du 16 obtenu par simulations ?)

N°13

Résultats trop parfaits

Un professeur a demandé à ses élèves de jeter 6000 fois un dé et de noter leurs résultats.

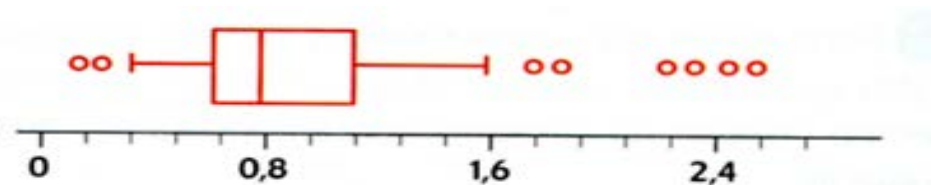
Matthieu propose le tableau suivant.

face	1	2	3	4	5	6
effectif	991	1005	1015	994	1007	988

Les résultats semblent très probants et Matthieu s'attend à des félicitations.

a) Calculer la valeur de d_{obs}^2 dans l'hypothèse d'un dé bien équilibré.

b) La simulation de cette expérience 1000 fois a donné une série de 1000 valeurs de $6000d^2$.



Pourquoi le professeur peut-il légitimement penser que Matthieu a truqué ses résultats et n'a vraisemblablement pas lancé le dé 6000 fois comme demandé ?

Cet exercice a le mérite de présenter un exemple d'utilisation d'un test d'adéquation pour mettre en évidence que des valeurs observées trop proches des valeurs théoriques attendues sont aussi rares (et donc aussi suspectes) que des valeurs trop éloignées.

Corrigé

Face	Effectifs	Fréquences	Probabilités	Écarts	Carrés
1	991	0,16516667	0,16666667	-0,00150000	0,00000225
2	1005	0,16750000	0,16666667	0,00083333	0,00000069
3	1015	0,16916667	0,16666667	0,00250000	0,00000625
4	994	0,16566667	0,16666667	-0,00100000	0,00000100
5	1007	0,16783333	0,16666667	0,00116666	0,00000136
6	988	0,16466667	0,16666667	-0,00200000	0,00000400
Sommes	6000	1	1		0,00001556
				6000*d²=	0,09333333

a) Dans le tableau ci-dessus est détaillé le calcul de $6000*d_{obs}^2$. La valeur obtenue est 0,093.

b) La lecture du diagramme montre que la valeur de $6000*d_{obs}^2$ fournie par les données de Matthieu est très inférieure à toutes les valeurs observées lors de 1000 simulations avec un dé normal. Trop beau pour être vrai !

Remarque

Ici c'est $6*6000*d_{obs}^2 = 0,56$ qui est une réalisation d'un Khi deux à 5 degrés de liberté. La table du Khi deux fournit un premier décile de 1,61.

Annexe 1

L'énoncé de Pondichery

Un pisciculteur possède un bassin qui contient 3 variétés de truites : communes, saumonées et arc-en-ciel. Il voudrait savoir s'il peut considérer que son bassin contient autant de truites de chaque variété. Pour cela il effectue, au hasard, 400 prélèvements d'une truite avec remise et obtient les résultats suivants :

Variétés	Commune	Saumonée	Arc-en-ciel
Effectifs	146	118	136

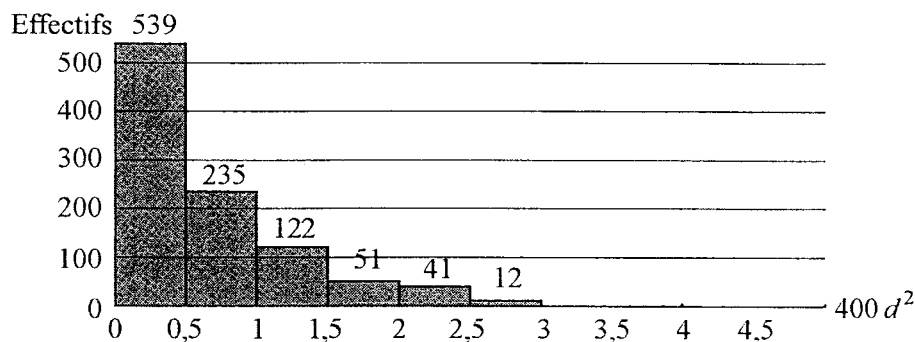
1. a. Calculer les fréquences de prélèvement f_c d'une truite commune, f_s d'une truite saumonée et f_a d'une truite arc-en-ciel. On donnera les valeurs décimales exactes.

b. On pose $d^2 = \left(f_c - \frac{1}{3}\right)^2 + \left(f_s - \frac{1}{3}\right)^2 + \left(f_a - \frac{1}{3}\right)^2$

Calculer $400 d^2$ arrondi à 10^{-2} ; on note $400 d^2_{obs}$ cette valeur.

2. À l'aide d'un ordinateur, le pisciculteur simule le prélèvement au hasard de 400 truites suivant la loi équirépartie. Il répète 1000 fois cette opération et calcule à chaque fois la valeur de $400 d^2$.

Le diagramme à bandes ci-dessous représente la série des 1000 valeurs de $400 d^2$, obtenues par simulation.



Déterminer une valeur approchée à 0,5 près par défaut, du neuvième décile D9 de cette série.

3. En argumentant soigneusement la réponse dire si on peut affirmer avec un risque d'erreur inférieur à 10 % que « le bassin contient autant de truites de chaque variété ».

4. On considère désormais que le bassin contient autant de truites de chaque variété. Quand un client se présente, il prélève au hasard une truite du bassin.

Trois clients prélèvent chacun une truite. Le grand nombre de truites du bassin permet d'assimiler ces prélèvements à des tirages successifs avec remise.

Calculer la probabilité qu'un seul des trois clients prélève une truite commune.

Éléments de corrigé

Question 1 :

Les calculs sont détaillés dans le tableau ci-dessous.

Variétés	Commune	Saumonée	Arc en ciel	Total
Effectifs	146	118	136	400
Fréquences	0,365	0,295	0,340	1
Écarts	0,03166667	0,03166667	0,03166667	d^2
Carrés	0,00100278	0,00100278	0,00100278	0,00300833
			$400*d^2$	1,20

Question 2 :

Le cumul, en partant de la fin, des effectifs indiqués donne 12, 53 puis 104 alors que le dixième de l'effectif total est 100. Le neuvième décile est donc dans l'intervalle [1.5 ; 2] et 1.5 en est une valeur approchée à 0,5 près par défaut.

Question 3 :

La valeur $400 d_{obs}^2$ observée est de 1,2. Elle est inférieure (assez nettement) au neuvième décile arrondi qui est 1,5. On ne peut donc rejeter l'hypothèse d'équirépartition.

Annexe 2

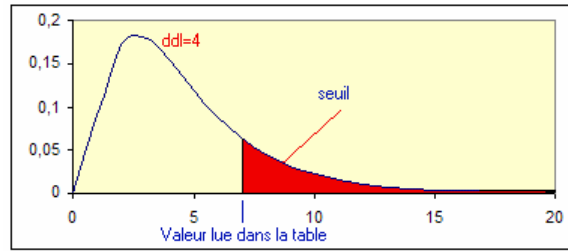
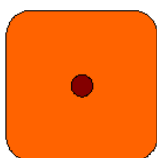
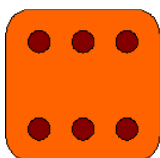


TABLE DU KHI DEUX

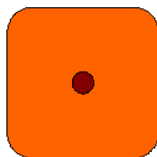
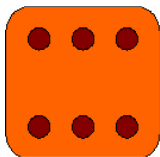
D9 Q3 Med Q1 D1

Degrés de liberté	Seuil →								
	0,01	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,99
1	6,6349	3,8415	2,7055	1,3233	0,4549	0,1015	0,0158	0,0039	0,0002
2	9,2104	5,9915	4,6052	2,7726	1,3863	0,5754	0,2107	0,1026	0,0201
3	11,3449	7,8147	6,2514	4,1083	2,3660	1,2125	0,5844	0,3518	0,1148
4	13,2767	9,4877	7,7794	5,3853	3,3567	1,9226	1,0636	0,7107	0,2971
5	15,0863	11,0705	9,2363	6,6257	4,3515	2,6746	1,6103	1,1455	0,5543
6	16,8119	12,5916	10,6446	7,8408	5,3481	3,4546	2,2041	1,6354	0,8721
7	18,4753	14,0671	12,0170	9,0371	6,3458	4,2549	2,8331	2,1673	1,2390
8	20,0902	15,5073	13,3616	10,2189	7,3441	5,0706	3,4895	2,7326	1,6465
9	21,6660	16,9190	14,6837	11,3887	8,3428	5,8988	4,1682	3,3251	2,0879
10	23,2093	18,3070	15,9872	12,5489	9,3418	6,7372	4,8652	3,9403	2,5582
11	24,7250	19,6752	17,2750	13,7007	10,3410	7,5841	5,5778	4,5748	3,0535
12	26,2170	21,0261	18,5493	14,8454	11,3403	8,4384	6,3038	5,2260	3,5706
13	27,6882	22,3620	19,8119	15,9839	12,3398	9,2991	7,0415	5,8919	4,1069
14	29,1412	23,6848	21,0641	17,1169	13,3393	10,1653	7,7895	6,5706	4,6604
15	30,5780	24,9958	22,3071	18,2451	14,3389	11,0365	8,5468	7,2609	5,2294
16	31,9999	26,2962	23,5418	19,3689	15,3385	11,9122	9,3122	7,9616	5,8122
17	33,4087	27,5871	24,7690	20,4887	16,3382	12,7919	10,0852	8,6718	6,4077
18	34,8052	28,8693	25,9894	21,6049	17,3379	13,6753	10,8649	9,3904	7,0149
19	36,1908	30,1435	27,2036	22,7178	18,3376	14,5620	11,6509	10,1170	7,6327
20	37,5663	31,4104	28,4120	23,8277	19,3374	15,4518	12,4426	10,8508	8,2604
21	38,9322	32,6706	29,6151	24,9348	20,3372	16,3444	13,2396	11,5913	8,8972
22	40,2894	33,9245	30,8133	26,0393	21,3370	17,2396	14,0415	12,3380	9,5425
23	41,6383	35,1725	32,0069	27,1413	22,3369	18,1373	14,8480	13,0905	10,1957
24	42,9798	36,4150	33,1962	28,2412	23,3367	19,0373	15,6587	13,8484	10,8563
25	44,3140	37,6525	34,3816	29,3388	24,3366	19,9393	16,4734	14,6114	11,5240
26	45,6416	38,8851	35,5632	30,4346	25,3365	20,8434	17,2919	15,3792	12,1982
27	46,9628	40,1133	36,7412	31,5284	26,3363	21,7494	18,1139	16,1514	12,8785
28	48,2782	41,3372	37,9159	32,6205	27,3362	22,6572	18,9392	16,9279	13,5647
29	49,5878	42,5569	39,0875	33,7109	28,3361	23,5666	19,7677	17,7084	14,2564
30	50,8922	43,7730	40,2560	34,7997	29,3360	24,4776	20,5992	18,4927	14,9535

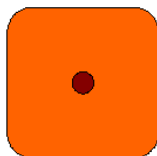
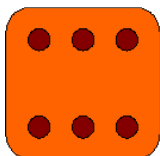
VII. DE TOUT... AU HASARD



Des dés et des polynômes



Autour du jeu de l'oie



Un jeu pas si équilibré que ça

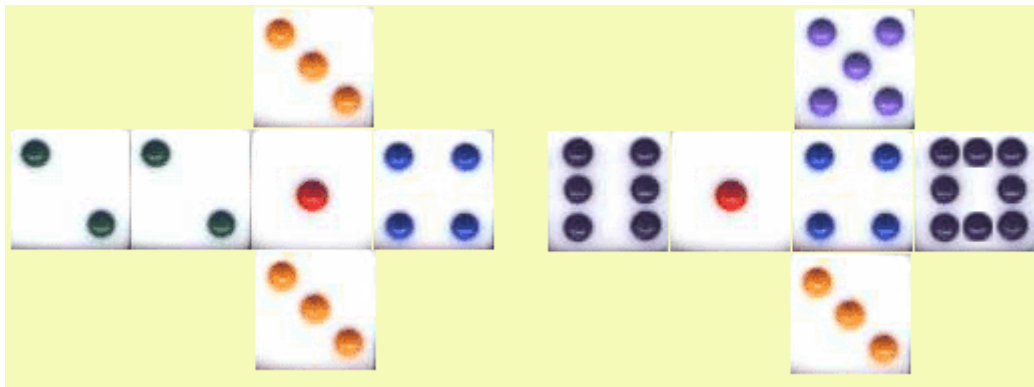
DES DÉS ET DES POLYNÔMES

Les dés de Sicherman

Lorsque l'on lance deux dés cubiques normaux et que l'on calcule la somme des points on obtient une variable aléatoire dont la loi est :

Somme	2	3	4	5	6	7	8	9	10	11	12
Probabilité	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Sicherman a proposé d'utiliser deux dés dont les faces portent respectivement les nombres de points suivant 1, 2, 2, 3, 3 et 4 pour le premier, 1, 3, 4, 5, 6 et 8 pour le second.



Lorsque l'on lance ces deux dés pour faire la somme des points, les 36 résultats équiprobables sont résumés dans le tableau ci-dessous :

	1	2	2	3	3	4
1	1	3	3	4	4	5
3	4	5	5	6	6	7
4	5	6	6	7	7	8
5	6	7	7	8	8	9
6	7	8	8	9	9	10
8	9	10	10	11	11	12

Et la loi de probabilité de la variable aléatoire somme des points est la même que pour deux dés ordinaires.

Cela nous amène, en particulier, à poser deux problèmes :

- 1) Existe-t-il d'autres manières de numérotter les faces de deux dés cubiques pour qu'en les lançant ensemble on obtienne pour la somme des points la même loi de probabilité qu'en lançant deux dés ordinaires ?
- 2) Existe-t-il une manière de numérotter les faces de deux dés cubiques pour qu'en les lançant ensemble on obtienne pour la somme des points la loi équiprobable sur $2, \dots, 12$?

Polynôme générateur d'une variable aléatoire

Pour répondre aux deux questions ci-dessus nous allons associer à toute variable aléatoire prenant un nombre fini de valeurs entières un polynôme.

Par exemple :

Le nombre de points obtenus lorsqu'on lance un dé cubique normal est une variable aléatoire X dont la loi est bien connue :

Points	1	2	3	4	5	6
Probabilités	1/6	1/6	1/6	1/6	1/6	1/6

Associons à cette variable aléatoire le polynôme :

$$P(x) = \frac{1}{6}x^1 + \frac{1}{6}x^2 + \frac{1}{6}x^3 + \frac{1}{6}x^4 + \frac{1}{6}x^5 + \frac{1}{6}x^6.$$

dans lequel les exposants sont les valeurs de la variable aléatoire et les coefficients les probabilités correspondantes.

Plus généralement, si X est une variable aléatoire de loi :

Valeurs	a_1	a_2	$\dots$	$\dots$	a_{n-1}	a_n
Probabilités	p_1	p_2	$\dots$	$\dots$	p_{n-1}	p_n

On lui associe le polynôme $P(x) = \sum_{i=1}^n p_i x^{a_i}$. Nous dirons que $P(x)$ est le polynôme générateur de la variable aléatoire X .

Lorsqu'on multiplie deux polynômes, les exposants (valeurs des variables aléatoires) s'ajoutent et les coefficients (probabilités) se multiplient. On constate donc que le polynôme générateur de la somme de deux variables aléatoires indépendantes est le produit des deux polynômes générateurs.

Polynôme générateur d'un dé cubique

A quelles conditions un polynôme peut-il être le polynôme générateur de la variable aléatoire associée à un dé cubique dont les 6 faces portent chacune un nombre entier de points ?

Les coefficients sont tous positifs ou nuls et de somme 1 (condition nécessaire pour tout polynôme générateur)

Chaque coefficient est égal à un nombre entier de sixièmes (entre 0 et 6/6) car chaque valeur est présente sur un certain nombre de faces du dé.

Factorisation des polynômes générateurs correspondants aux deux situations décrites ci-dessus.

Le polynôme générateur de la variable aléatoire somme des points pour deux dés normaux :

$$P(x) = \frac{1}{36}x^2 + \frac{2}{36}x^3 + \frac{3}{36}x^4 + \frac{4}{36}x^5 + \frac{5}{36}x^6 + \frac{6}{36}x^7 + \frac{5}{36}x^8 + \frac{4}{36}x^9 + \frac{3}{36}x^{10} + \frac{2}{36}x^{11} + \frac{1}{36}x^{12}$$

a pour factorisation (rationnelle) :

$$P(x) = \frac{1}{36}x^2(x+1)^2(x^2+x+1)^2(x^2-x+1)^2$$

dont les facteurs s'associent d'une seule manière en produit de deux polynômes générateurs de dés :

$$P(x) = \frac{1}{6}(x^8 + x^6 + x^5 + x^4 + x^3 + x) \frac{1}{6}(x^4 + 2x^3 + 2x^2 + x)$$

Cette factorisation correspond bien aux deux dés proposés par Sicherman.

Les dés de Sicherman sont donc les seuls qui se comportent comme deux dés normaux lorsqu'on fait la somme des points obtenus.

Pour examiner la possibilité d'obtenir l'équiprobabilité sur 2,...,12 en utilisant deux dés cubiques, il faut tenter de factoriser le polynôme :

$$E(x) = \frac{1}{11}(x^2 + x^3 + x^4 + x^5 + x^6 + x^7 + x^8 + x^9 + x^{10} + x^{11} + x^{12})$$

dont la seule factorisation rationnelle est

$$E(x) = \frac{1}{11}x^2(1 + x + x^2 + x^3 + x^4 + x^5 + x^6 + x^7 + x^8 + x^9 + x^{10})$$

Avant même d'examiner cette factorisation, on pouvait prévoir que la constante $\frac{1}{11}$ ne saurait se factoriser en sixièmes.

Il n'y a donc aucune manière de numérotiser les faces de deux cubiques pour que la somme des points obtenus ait pour loi l'équiprobabilité sur 2,...,12.

Les remarques précédentes peuvent donner l'idée de rechercher des polynômes dont les coefficients sont des multiples entiers de $\frac{1}{36}$.

Quelques essais nous ont permis de découvrir :

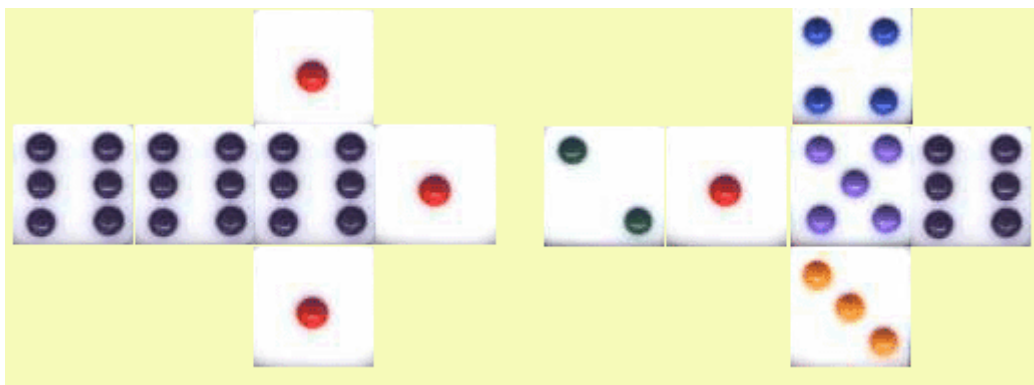
Des sommes de 2 à 12 presque toutes équiprobables

$$Q(x) = \frac{3}{36}x^2 + \frac{3}{36}x^3 + \frac{3}{36}x^4 + \frac{3}{36}x^5 + \frac{3}{36}x^6 + \frac{6}{36}x^7 + \frac{3}{36}x^8 + \frac{3}{36}x^9 + \frac{3}{36}x^{10} + \frac{3}{36}x^{11} + \frac{3}{36}x^{12}$$

qui se factorise :

$$Q(x) = \frac{1}{6}(3x^6 + 3x) \frac{1}{6}(x^6 + x^5 + x^4 + x^3 + x^2 + x)$$

Ce qui montre qu'en utilisant un dé ordinaire et un dé dont trois faces sont numérotées 6 et les trois autres 1, on obtient des sommes de 2 à 12 et seule la somme 7 aura une probabilité double de celles des autres.



L'impossible tirage au hasard d'un chiffre

La même technique permet de montrer qu'avec deux dés cubiques il sera impossible d'obtenir l'équiprobabilité sur 0, 1, ..., 9 c'est-à-dire le tirage au hasard d'un chiffre.

Un dé à neuf faces équiprobables

Pour l'équiprobabilité sur 1, 2, 3...9 il existe une factorisation exploitable du polynôme générateur :

$$\begin{aligned} H(x) &= \frac{4}{36} (x + x^2 + x^3 + x^4 + x^5 + x^6 + x^7 + x^8 + x^9) \\ &= \frac{2}{6} (x^3 + x^2 + x) \frac{2}{6} (x^6 + x^3 + 1) \end{aligned}$$

qui permet de montrer que lancer ensemble deux dés dont les faces portent respectivement 3,3,2,2,1,1 et 6,6,3,3,0,0 équivaut au tirage au hasard d'un entier entre 1 et 9, un dé à neuf faces équiprobables en quelque sorte.

Un dé à 18 faces équiprobables

Là encore une factorisation du polynôme générateur permet de trouver une solution intéressante :

$$\begin{aligned} K(x) &= \frac{2}{36} (x + x^2 + x^3 + x^4 + x^5 + x^6 + x^7 + x^8 + x^9 + x^{10} + x^{11} + x^{12} + x^{13} + x^{14} + x^{15} + x^{16} + x^{17} + x^{18}) \\ &= \frac{2}{6} (x^{12} + x^6 + 1) \frac{1}{6} (x + x^2 + x^3 + x^4 + x^5 + x^6) \end{aligned}$$

Un dé « ordinaire » et un dé dont les faces portent 12, 12, 6, 6, 0, 0 donneront pour somme un entier de 1 à 18 avec équiprobabilité.



AUTOUR DU JEU DE L'OIE

La case 58 fatale !



Au jeu de l'oie les joueurs déplacent des pions sur un plateau comportant 63 cases en lançant un dé cubique standard. Malheur à celui qui tombera sur la case n°58 car il devra se replacer sur la case départ. Mais au fait, quelle est la probabilité de tomber sur cette case fatale ?

Étude asymptotique

Imaginons que le plateau de jeu comporte un nombre infini de case. Soit p_n la probabilité de passer par la case n . Étant donné que le nombre moyen de cases parcourues à chaque lancer de dé tend vers $\frac{7}{2}$ (espérance du nombre de points obtenu avec un dé), si p_n admet une limite lorsque n tend vers $+\infty$ alors cette limite est $\frac{2}{7}$.

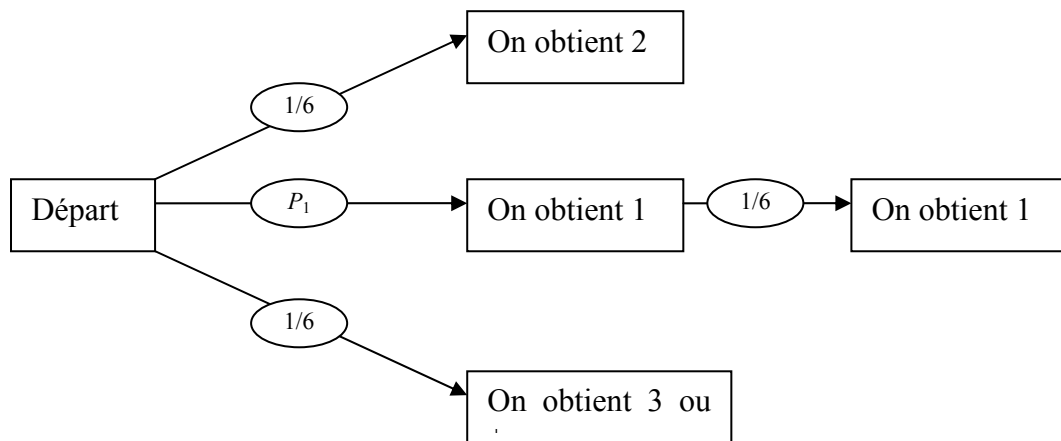
Formule de récurrence

Pour n entier naturel non nul, notons A_n l'événement "tomber sur la case n " et $p_n = P(A_n)$.

Calcul de p_i , pour $i \leq 6$:

On a $p_1 = \frac{1}{6} \approx 0,167$.

Pour le calcul de p_2 on peut visualiser la situation par un arbre :



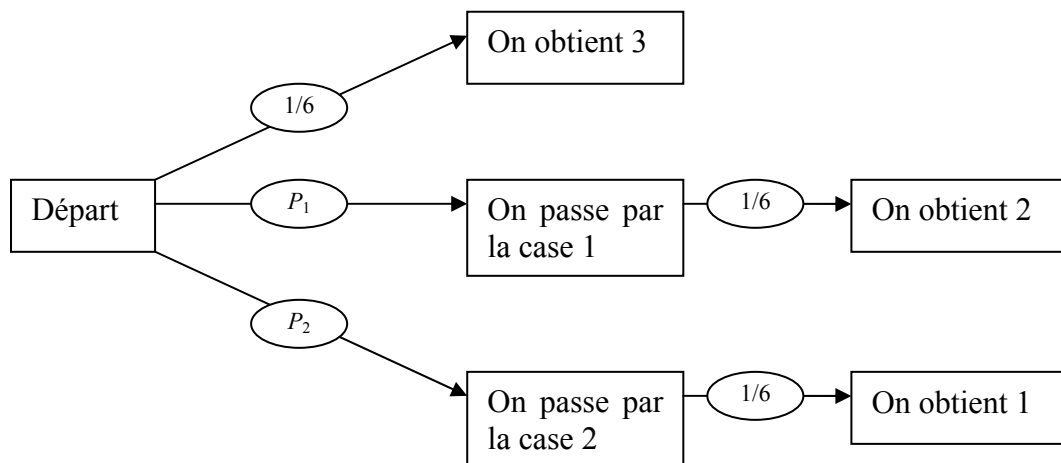
Ainsi, d'après la formule des probabilités totales $p_2 = \frac{1}{6} + \frac{1}{6} p_1 = \frac{7}{36} \approx 0,194$.

De même, il y a trois manières d'arriver sur la case 3 :

- Directement en lançant 3.
- Passer par la case 1 et lancer 2.
- Passer par la case 2 et lancer 1.

On trouve cette fois $p_3 = \frac{1}{6} + \frac{1}{6} p_1 + \frac{1}{6} p_2 = \frac{49}{216} \approx 0,227$.

Le calcul est illustré par l'arbre suivant.



Des raisonnements analogues nous donnerons successivement :

$$p_4 = \frac{1}{6} + \frac{1}{6} p_1 + \frac{1}{6} p_2 + \frac{1}{6} p_3 = \frac{343}{1296} \approx 0,265.$$

$$p_5 = \frac{1}{6} + \frac{1}{6}p_1 + \frac{1}{6}p_2 + \frac{1}{6}p_3 + \frac{1}{6}p_4 = \frac{2401}{7776} \approx 0,309.$$

$$p_6 = \frac{1}{6} + \frac{1}{6}p_1 + \frac{1}{6}p_2 + \frac{1}{6}p_3 + \frac{1}{6}p_4 + \frac{1}{6}p_5 = \frac{16807}{46656} \approx 0,360.$$

On peut remarquer que les six premiers termes de (p_n) sont ceux d'une suite géométrique de raison $\frac{7}{6}$.

En effet : $p_{n+1} = p_n + \frac{1}{6}p_n = \frac{7}{6}p_n$ pour $n \leq 5$ et donc $p_n = \frac{1}{6} \times \left(\frac{7}{6}\right)^{n-1}$ pour $1 \leq n \leq 6$.

On pourra consulter en annexe 1 une autre manière d'aborder le problème.

Calcul de p_n , pour $n \geq 7$

A_n est la réunion disjointe des événements $A_n \cap A_{n-1}, B_n, C_n, D_n, E_n, F_n$ où :

B_n est l'événement : "on était sur la case $n-2$ et on a obtenu un 2".

C_n est l'événement : "on était sur la case $n-3$ et on a obtenu un 3".

D_n est l'événement : "on était sur la case $n-4$ et on a obtenu un 4".

E_n est l'événement : "on était sur la case $n-5$ et on a obtenu un 5".

F_n est l'événement : "on était sur la case $n-6$ et on a obtenu un 6".

La formule des probabilités totales nous donne alors :

$$p_n = \frac{1}{6}p_{n-1} + \frac{1}{6}p_{n-2} + \frac{1}{6}p_{n-3} + \frac{1}{6}p_{n-4} + \frac{1}{6}p_{n-5} + \frac{1}{6}p_{n-6}.$$

En particulier, on trouve :

$$p_7 = \frac{1}{6} \times \left(\left(\frac{7}{6} \right)^6 - 1 \right).$$

Pour le calcul effectif de p_n avec $n \geq 8$, on peut éventuellement remarquer que

$$p_n - p_{n-1} = \frac{1}{6}(p_{n-1} - p_{n-7})$$

et donc que

$$p_n = \frac{7}{6}p_{n-1} - \frac{1}{6}p_{n-7}.$$

Voici un tableau de valeurs approchées des premiers termes de (p_n)

1	0.16666	11	0.29339	21	0.28596
2	0.19444	12	0.29083	22	0.28594
3	0.22685	13	0.27926	23	0.28575
4	0.26466	14	0.28353	24	0.28559
5	0.30877	15	0.28611	25	0.28559
6	0.36023	16	0.28707	26	0.28574
7	0.25360	17	0.28670	27	0.28576
8	0.26809	18	0.28558	28	0.28573
9	0.28036	19	0.28471	29	0.28570
10	0.28928	20	0.28562	30	0.28569

En poursuivant ce calcul, on constate que p_{58} est une valeur approchée de $\frac{2}{7}$ avec une précision de l'ordre de 10^{-9} .

Convergence de (p_n)

Les résultats concernant les suites récurrentes linéaires nous permettent de démontrer que la suite (p_n) converge.

Le polynôme caractéristique de la suite (p_n) est :

$$P(X) = X^6 - \frac{1}{6}(X^5 + X^4 + X^3 + X^2 + X + 1).$$

Une de ses racines est 1 et les autres, dont a priori on ne peut que déterminer des valeurs approchées, sont en module strictement inférieures à 1.

En effet si λ est une racine avec $|\lambda| > 1$, on a $|\lambda|^6 = \frac{1}{6}|\lambda^5 + \lambda^4 + \dots + 1| < \frac{1}{6}(|\lambda|^5 + \dots + 1) \leq |\lambda|^5$

C'est-à-dire $|\lambda|^6 \leq |\lambda|^5$ et il y a contradiction.

D'autre part, on vérifie sans trop de problème que le polynôme caractéristique est premier avec son polynôme dérivé (par exemple, à l'aide de l'algorithme d'Euclide) ce qui prouve que les racines $\{1, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$ de $P(X)$ sont toutes simples et nous montre que p_n peut s'écrire sous la forme :

$$p_n = a + b\lambda_1^n + c\lambda_2^n + \dots + f\lambda_5^n.$$

On en déduit que (p_n) converge vers a . Comme l'étude asymptotique nous a montré que la seule limite possible de (p_n) est $\frac{2}{7}$, on en déduit que $a = \frac{2}{7}$.

Pour estimer la rapidité de convergence, on peut vérifier à l'aide d'un logiciel de calcul que les modules des valeurs propres autres que 1 sont majorés par 0,74. Comme $0,74^{58}$ est de l'ordre de 10^{-8} cela reste cohérent avec la précision observée au paragraphe précédent.

Autre démonstration de la convergence

Utilisons le fait que p_{n+6} est la moyenne (l'isobarycentre) de $\{p_n; \dots; p_{n+5}\}$.

Soit m_n et M_n les minimum et maximum de $\{p_n; p_{n+1}; \dots; p_{n+5}\}$.

Comme $p_{n+6} \in [m_n; M_n]$, on en déduit que (m_n) est croissante et que (M_n) est décroissante (au sens large). Ces deux suites étant bornées, elles convergent vers m et M avec $m \leq M$.

Supposons que $m < M$. Pour $\varepsilon > 0$ il existe un entier n_0 à partir duquel

$$m - \varepsilon < m_n \leq p_n \leq M_n < M + \varepsilon.$$

Comme p_{n+6} est la moyenne de six nombres dont le plus grand est M_n et dont les cinq autres sont minorés par m_n on a :

$$\frac{5}{6}m_n + \frac{1}{6}M_n \leq p_{n+6}.$$

Mais alors pour $n \geq n_0$, on a

$$\frac{5}{6}(m - \varepsilon) + \frac{1}{6}M \leq \frac{5}{6}m_n + \frac{1}{6}M_n \leq p_{n+6}$$

Cette dernière inégalité est vérifiée en particulier pour $p_{n_0+6}, \dots, p_{n_0+11}$ donc pour le plus petit d'entre eux, m_{n_0+6} :

$$\frac{5}{6}(m - \varepsilon) + \frac{1}{6}M < m_{n_0+6} \leq m,$$

dont on tire

$$5m - 5\varepsilon + M < 6m$$

et finalement

$$M - m < 5\varepsilon.$$

La dernière inégalité étant vérifiée pour tout $\varepsilon > 0$, on en déduit que $m = M$ et que la suite (p_n) converge vers cette limite commune qui d'après l'étude asymptotique est $\frac{2}{7}$.

Généralisation à d'autres types de dés

Sous certaines réserves exprimées plus loin le problème reste le même quelque soit le type de dés utilisés pour se déplacer. Examinons tout d'abord trois cas particuliers avant de préciser les hypothèses et de démontrer les résultats avancés.

Exemple 1 : Lancer de deux dés équilibrés.

Un lancer de deux dés peut être considéré comme le lancer d'un seul dé non équilibré à onze faces (de 2 à 12). Dans ce cas le nombre moyen de cases parcourues à chaque lancer est 7 et (p_n) va tendre vers $\frac{1}{7}$.

Exemple 2 : Lancer d'un dé équilibré comportant deux faces marquées d'un 2, deux faces marquées d'un 4 et deux faces marquées d'un 6.

Cette fois on voit que les cases portant un numéro impair ne peuvent pas être atteintes. Par contre la suite (p_{2n}) tend vers $\frac{1}{3}$. Ce résultat se démontrerait de la même manière que les autres après avoir "gommé" les cases portant un numéro impair.

Cette situation se généralise à tous les cas où le PGCD des nombres portés par les faces du dé est différent de 1.

Exemple 3 : Lancer d'un autre "dé".

Considérons un dé cubique équilibré dont les faces porteraient les nombres 6, 6, 6, 10, 10 et 15. Nous sommes donc dans une situation où le PGCD est égal à 1.

Certaines des premières cases du plateau ne peuvent pas être atteintes en lançant un tel dé. Plus précisément la case n pourra être atteinte s'il existe un triplet d'entiers naturels (a, b, c) tel que $n = 6a + 10b + 15c$.

En fait, la case numéro 29 est la dernière des cases qui ne peut pas être atteinte. En effet 29 ne peut pas être obtenu sous la forme $6a + 10b + 15c$ avec a, b, c entiers naturels mais pour les six nombres suivants on vérifie que :

$$30 = 5 \times 6 \text{ (la case 30 est atteinte si on obtient, entre autres, 5 fois un 6);}$$

$$31 = 6 + 10 + 15,$$

$$32 = 2 \times 6 + 2 \times 10,$$

$$33 = 3 \times 6 + 15,$$

$$34 = 6 \times 4 + 10 \text{ et}$$

$$35 = 2 \times 10 + 15.$$

Enfin tout nombre M supérieur ou égal à 36 pourra s'écrire sous la forme $N + 6K$ avec N et K entiers naturels et $30 \leq N \leq 35$. Ainsi la case M pourra en particulier être atteinte en passant par la case N puis en réalisant K fois un 6.

Par ailleurs, l'espérance mathématique de la variable aléatoire associée à ce dé est :

$$E(X) = \frac{1}{6}(6 + 6 + 6 + 10 + 10 + 15) = \frac{53}{6}.$$

Ce résultat correspond au nombre moyen de cases parcourues à chaque lancer et (p_n) converge vers $\frac{6}{53} \approx 0,1132$.

Cas général

Considérons plus généralement un dé à k faces portant des entiers naturels premiers dans leur ensemble notés $d_1, \dots, d_k$ rangés en ordre croissant, et tels que $q_i = P(D = d_i) \neq 0$. Montrons tout d'abord que toutes les cases du plateau peuvent être atteintes, excepté certaines des premières. Ceci revient à montrer qu'il existe une suite de d_1 entiers consécutifs qui peuvent tous s'écrire comme somme d'éléments de $\{d_1, \dots, d_k\}$.

Démonstration

Comme $\{d_1, \dots, d_k\}$ sont premiers entre eux dans leur ensemble, il existe $(x_1, \dots, x_k)$ tel que $\sum_{j=1}^k x_j d_j = 1$. Posons $n = \sum_{j=1}^k x_j d_j$ avec $x = -\inf(x_j)$. Alors pour tout i tel que $0 \leq i \leq d_1$, on a $n + i = \sum_{j=1}^k x d_1 d_j + i \sum_{j=1}^k x_j d_j = \sum_{j=1}^k (x d_1 + i x_j) d_j$.

Comme $d_1 \geq i$ et que $x \geq -x_j$, dans cette dernière somme les coefficients des d_j sont tous positifs et ainsi les $(d_1 + 1)$ nombres $(n + i)$ peuvent tous s'écrire comme somme d'éléments de $\{d_1, \dots, d_k\}$. Notons tout de même que ce n'est pas, en général, la plus petite suite d'entiers réalisant cette condition.

Étude de la convergence

Dans les conditions énoncées plus haut, on sait que la probabilité p_n d'atteindre la case n est strictement positive dès que n est suffisamment grand. Notons D la variable aléatoire associée à notre dé. Le nombre moyen de cases parcourues à chaque lancer de dé est alors

$$E(D) = \sum_{i=1}^k q_i d_i = m \text{ et si } (p_n) \text{ converge sa limite sera } \frac{1}{m}.$$

Pour $n > d_k$, A_n est la réunion disjointe des événements $A_{n-d_i} \cap (D = d_i)$ et ainsi :

$$p_n = \sum_{i=1}^k q_i p_{n-d_i}.$$

Le polynôme caractéristique est cette fois $P(x) = x^{d_k} - \sum_{j=1}^k q_j x^{d_k-d_j}$.

A nouveau, comme dans le cas étudié plus haut, 1 est une racine.

Soit λ une autre racine. Si $|\lambda| > 1$,

$$|\lambda|^{d_k} = \left| \sum_{j=1}^k q_j \lambda^{d_k-d_j} \right| \leq \sum_{j=1}^k q_j |\lambda|^{d_k-d_j} \leq |\lambda|^{d_k-d_1} \sum_{j=1}^k q_j \leq |\lambda|^{d_k-d_1},$$

qui est contradictoire avec $|\lambda| > 1$. On en déduit que $|\lambda| \leq 1$.

Supposons que $|\lambda| = 1$. Alors, $1 = |\lambda|^{d_k} = \left| \sum_{j=1}^k q_j \lambda^{d_k-d_j} \right| \leq \sum_{j=1}^k q_j |\lambda|^{d_k-d_j} = 1$. L'inégalité précédente

est donc une égalité, ce qui implique que pour tout j , l'argument de $\lambda^{d_k-d_j}$ est constant (c'est donc celui de $\lambda^{d_k-d_k} = 1$). Ainsi pour tout j , $\lambda^{d_k-d_j} = 1$ et donc $\lambda^{d_k} = \lambda^{d_j}$. De plus

$0 = P(\lambda) = \lambda^{d_k} - \sum_{j=1}^k q_j$ et donc $\lambda^{d_k} = 1$. λ est racine (d_j -ème) de l'unité pour tout j . Mais comme $\{d_1, \dots, d_k\}$ sont premiers entre eux dans leur ensemble $\lambda = 1$. Ainsi la seule racine de module 1 est 1. On vérifie que le polynôme dérivé ne s'annule pas en 1 car $P'(1) = d_k - \sum_{j=1}^k (d_k - d_j)q_j = \sum_{j=1}^k d_j q_j$. Ceci a pour conséquence que 1 est racine simple et donc que (p_n) est bien une suite convergente.

Pour terminer, détaillons cette étude dans le cas de l'exemple 3.

Ici $k=3$. $d_1 = 6$, $d_2 = 10$, $d_3 = 15$, $q_1 = \frac{1}{2} = p_6$, $q_2 = \frac{1}{3} = p_{10}$, $q_3 = \frac{1}{6} = p_{15}$.

$p_{12} = \frac{1}{4}$ (obtenir deux fois de suite 6). Les autres valeurs de p_n pour $n < 15$ sont nulles et

pour $n > 15$ on a $p_n = \frac{1}{2} p_{n-6} + \frac{1}{3} p_{n-10} + \frac{1}{6} p_{n-15}$.

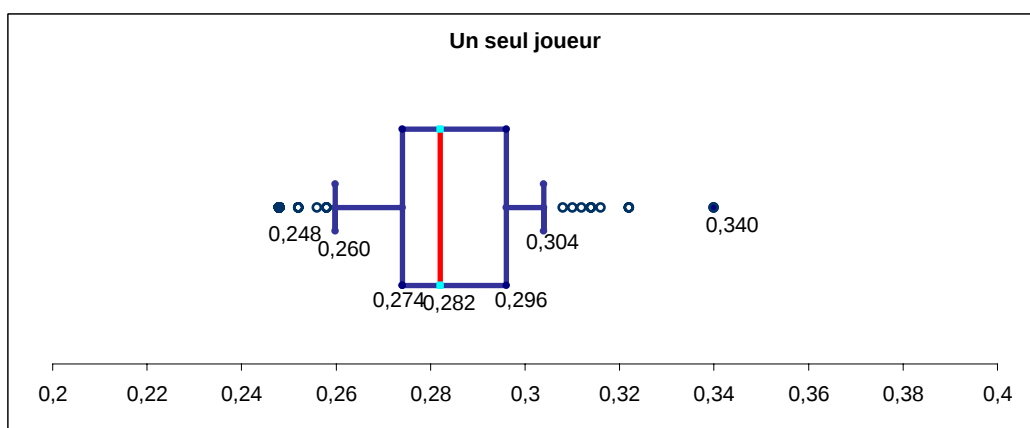
Le calcul nous permet de constater que pour $n \geq 132$: $\left| p_n - \frac{6}{53} \right| < 10^{-3}$

Le fichier EXCEL "dé à 3 faces.XLS" fournit un tableau de valeurs et une simulation.

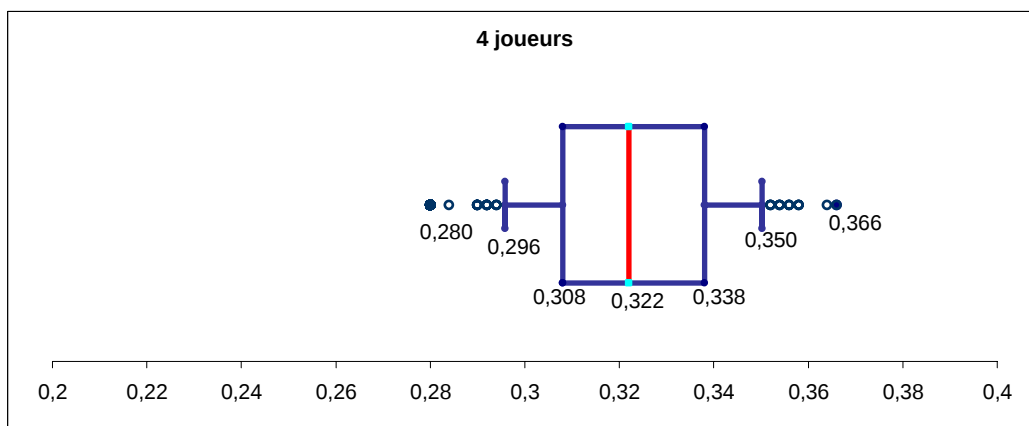
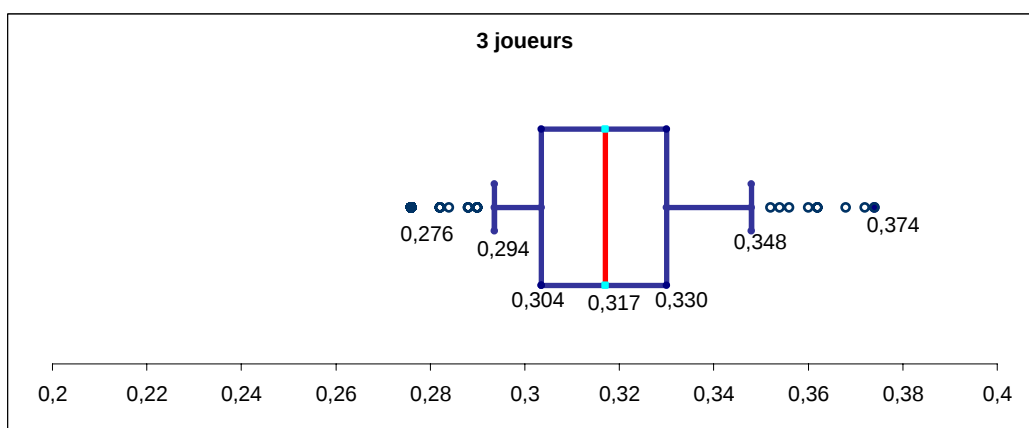
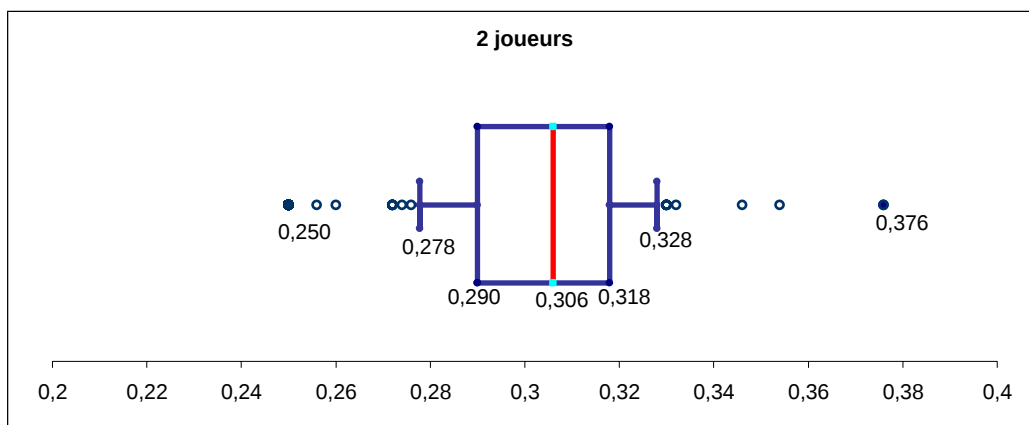
Jouons à plusieurs

Supposons que l'on joue à plusieurs et que le jeu s'arrête dès qu'un joueur atteint ou dépasse la case 58. On note p_{58} la probabilité qu'un des joueurs tombe sur la case 58. A priori, le calcul de p_{58} n'est pas simple. Explorons cette situation par des simulations.

Considérons d'abord le cas d'un seul joueur : Nous pouvons dans un premier temps simuler 500 parties et compter le nombre de ces parties où le joueur passe par la case 58. Un calcul de fréquence doit nous donner un nombre proche de $\frac{2}{7}$. En réalisant 100 fois cette simulation nous obtenons la boîte à moustache suivante :



En procédant de même avec 2, 3 et 4 joueurs nous obtenons cette fois :



On pourra se référer au fichier "case58 à 3 joueurs avec commentaires.xls" pour plus de précision sur la simulation.

Pour terminer voici le résultat d'une simulation portant sur 1000 parties avec un nombre n de joueurs tel que $1 \leq n \leq 20$. Dans le tableau figure le nombre N de parties où un joueur tombe sur la case 58

n	1	2	3	4	5	6	7	8	9	10
N	286	328	323	316	323	341	315	320	321	333
n	11	12	13	14	15	16	17	18	19	20
N	345	342	352	360	317	349	377	366	361	379

Malgré les quelques irrégularités, on peut conjecturer que N croît avec n mais ceci reste à démontrer.

Annexe 1 : Un autre procédé de calcul de p_i , pour $i \leq 6$

Pour ce calcul, on peut s'intéresser au problème plus général de la décomposition d'un entier en somme d'entiers :

De combien de manière peut-on décomposer un entier non nul n en somme de k entiers non nuls en tenant compte de l'ordre ?

Notons $f(k, n)$ ce nombre.

On trouve par exemple que $f(3;5)=6$ car 5 est somme des 6 triplets suivants: (3;1;1); (1;3;1); (1;1;3); (2;2;1); (2;1;2) et (1;2;2).

Plus généralement :

- Si $k > n$ alors $f(k; n) = 0$.
- $f(n; n) = 1$ et $f(1; n) = 1$, pour tout n est en particulier pour $n = 2$.
- Pour $k \leq n$, on a $f(k+1; n+1) = f(k; n) + f(k+1; n)$. En effet les $(k+1)$ -uplets de somme $n+1$, $(a_1; \dots; a_{k+1})$ sont de deux types différents:
 - *ceux pour lesquels $a_{k+1} = 1$ à qui on associe bijectivement les k -uplets $(a_1; \dots; a_k)$ de somme n . Il y en a donc $f(k; n)$.
 - ** ceux pour lesquels $a_{k+1} > 1$ à qui on associe bijectivement les $(k+1)$ -uplets de somme n $(a_1; \dots; a_k; a_{k+1} - 1)$. Il y en a donc $f(k+1; n)$.

Cette dernière relation associée avec $f(1;2) = f(2;2) = 1$ montre que pour tout k et n entiers,

$$f(k; n) = \binom{n-1}{k-1}.$$

Ainsi on passera par la case i ($i \leq 6$) avec la probabilité :

$$p_i = \sum_{k=1}^i f(k; i) \left(\frac{1}{6}\right)^k = \frac{1}{6} \sum_{k=1}^i \binom{i-1}{k-1} \left(\frac{1}{6}\right)^{k-1} = \frac{1}{6} \left(1 + \frac{1}{6}\right)^{i-1} = \frac{1}{6} \times \left(\frac{7}{6}\right)^{i-1}$$

Annexe 2 : Le plateau d'un vrai jeu de l'oie traditionnel



UN JEU PAS SI ÉQUILIBRÉ QUE ÇA !

Ce petit futé d'Alain a réussi à convaincre Bernard de jouer avec lui.

"Ce n'est pas compliqué, chacun d'entre nous choisit un nombre entier ; si les deux sont pairs, je te dois 3€, si les deux sont impairs, je t'en dois un, enfin s'ils n'ont pas la même parité tu me donne 2 €."

Au début Bernard, sûr de gagner deux fois sur trois se laisse convaincre mais révisé rapidement son jugement suite aux premières parties dont le bilan est globalement négatif.

Une analyse un peu plus fine pourrait lui faire croire que le jeu est équilibré. En effet, on peut représenter la situation par le tableau :

	Pair	Impair
Pair	+3	-2
Impair	-2	+1

Deux fois sur quatre on perd 2€, une fois sur quatre on en gagne 3 et une fois sur quatre on en gagne un. Comme $\frac{2}{4} \times 2 - \frac{1}{4} \times 3 - \frac{1}{4} \times 1 = 0$, le jeu est bien équilibré.

Ce n'est malheureusement pas avec ce type d'argument que Bernard regagnera de quoi emmener sa petite amie au cinéma comme nous allons le voir par la suite.

Une surface à notre secours

En réalité, le calcul présenté est celui de l'espérance de gain de Bernard dans le cas où chacun des deux joueurs choisit Pair ou Impair de manière équiprobable. Il n'en sera plus de même si l'un ou les deux joueurs adoptent une stratégie différente.

Plaçons nous tout d'abord dans le cas où Alain joue Pair une fois sur deux (en lançant une pièce de monnaie pour décider) et Bernard joue Pair avec une probabilité p :

	Pair (p)	Impair ($1-p$)
Pair (0,5)	+3	-2
Impair (0,5)	-2	+1

Le calcul de l'espérance de gain de Bernard nous donne :

$$3 \times \frac{p}{2} - 2 \times \frac{1-p}{2} - 2 \times \frac{p}{2} + \frac{1-p}{2} = p - \frac{1}{2}.$$

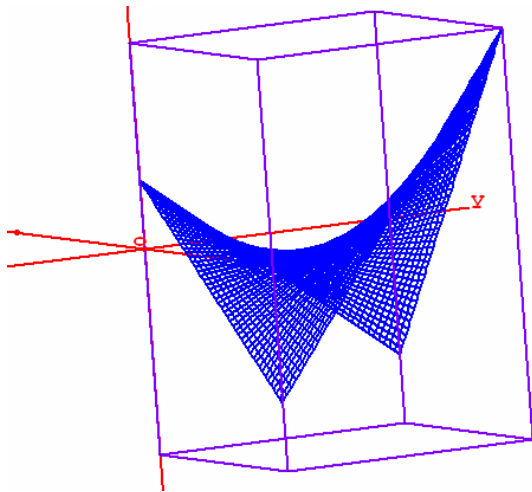
Ainsi Bernard a tout intérêt à jouer plus souvent Pair qu'Impair. On peut remarquer que le problème est symétrique et que si Bernard joue de manière équilibrée, Alain a intérêt à jouer plus souvent Impair. Évidemment ce type de stratégie suppose que l'adversaire ne se rend compte de rien et qu'il n'a pas fait la même analyse que nous.

Que se passe-t-il dans le cas où les deux joueurs modifient simultanément la probabilité de jouer Pair ?

	Pair (p)	Impair ($1-p$)
Pair (q)	+3	-2
Impair ($1-q$)	-2	+1

G étant la variable aléatoire qui à une partie associe le gain de Bernard (c'est-à-dire la perte d'Alain), on obtient :

$$E(G) = 3pq - 2(p(1-q) + q(1-p)) + (1-p)(1-q) = 8pq - 3(p+q) + 1.$$



La représentation ci-contre de la surface (S) d'équation $z = 8xy - 3(x+y) + 1$ avec x et y dans $[0;1]$ peut alors nous éclairer.

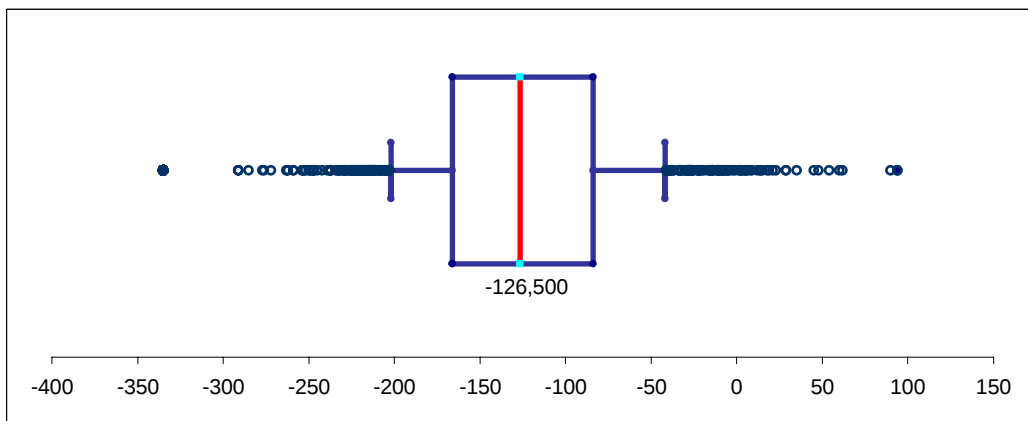
Les points de cette surface situés "sous" le plan (Oxy) correspondent à des stratégies mixtes perdantes pour Bernard. Le choix de ce dernier (la valeur de p) peut être optimisé s'il connaît la valeur de q :

$E(G) = (8q - 3)p + (1 - 3q) = f_q(p)$
est fonction affine de p et :

- Si $q > \frac{3}{8}$, f_q est croissante et le maximum $(5q-2)$ est atteint pour $p = 1$.
- Si $q < \frac{3}{8}$, f_q est décroissante et le maximum $(1-3q)$ est atteint pour $p = 0$.
- Si $q = \frac{3}{8}$, f_q est constante avec $f_q(p) = -\frac{1}{8}$.

Dans ce dernier cas, quoi que fasse Bernard, il sera perdant ; ce sera le cas également dès que $\frac{1}{3} < q < \frac{2}{5}$ puisqu'alors le maximum de f_q est négatif. Ces situations s'interprètent graphiquement par le fait que les segments intersection de (S) avec les plans $X = q$ sont entièrement situés sous le plan (Oxy). Au passage, on peut constater que (S) est un morceau de parabolôïde hyperbolique (PH ou « selle de cheval »). On retrouve que le PH est une surface réglée c'est-à-dire engendrée par le déplacement d'une droite dans l'espace.

D'autre part, la mise sous forme canonique : $E(G) = 8\left(p - \frac{3}{8}\right)\left(q - \frac{3}{8}\right) - \frac{1}{8}$ permet de voir que jouer Pair avec la probabilité $3/8$ est la meilleure (ou la moins pire) des stratégies pour les deux joueurs. Une simulation de 1000 fois 1000 parties (avec $p=0,5$ et $q=3/8$) a donné, pour la répartition des gains de Bernard, le diagramme en boîte suivant :



Gagner à tout prix

On peut imaginer la situation où Bernard désirerait jouer (et gagner) malgré tout. Cette situation n'est pas du tout extravagante puisque de nombreuses personnes jouent régulièrement au loto où l'espérance de gain est nettement plus défavorable.

Plaçons nous dans le cas $p = 1/2$ et $q = 3/8$ et supposons que Bernard puisse arrêter le jeu lorsqu'il le désire. Par exemple Bernard s'arrête dès que son gain est strictement positif ou dès qu'il perd une somme supérieure ou égale à 3€ (dans la suite, nous appellerons jeu une telle suite de parties). Les parties jouées sont modélisées par des variables aléatoires indépendantes X_i dont la loi commune est donnée par :

$$P(X_i=3) = 3/16, \quad P(X_i=1) = 5/16 \quad P(X_i=-2) = 1/2.$$

Considérons également les variables aléatoire S_n et T . S_n est le gain total après la n -ième partie (Si $S_n > 0$ ou si $S_n \leq -3$ alors $S_k = S_n$ pour tout $k \geq n$). T est la variable aléatoire à valeurs dans $\mathbb{N} \cup \{\infty\}$ qui à tout jeu associe sa longueur c'est-à-dire le nombre de parties jouées.

On remarque qu'alors S_1 a la même loi que X_1 et que $P(T=1) = 1/2$.

- S_2 prend ses valeurs dans l'ensemble $\{-4; -1; 1; 3\}$:

$$P(S_2 = -4) = P_{(X_1=-2)}(X_2 = -2) \times P(X_1 = -2) = \frac{1}{4} \text{ (par indépendance de } X_1 \text{ et } X_2)$$

$$\text{On trouve de même } P(S_2 = -1) = \frac{5}{32} \quad P(S_2 = 3) = \frac{3}{16}$$

$$\text{et } P(S_2 = 1) = \frac{5}{16} + \frac{3}{32} = \frac{13}{32}.$$

En particulier (ce qui commence à intéresser Bernard) $P(S_2 > 0) = \frac{19}{32}$ et

$$P(T = 2) = \frac{11}{32}.$$

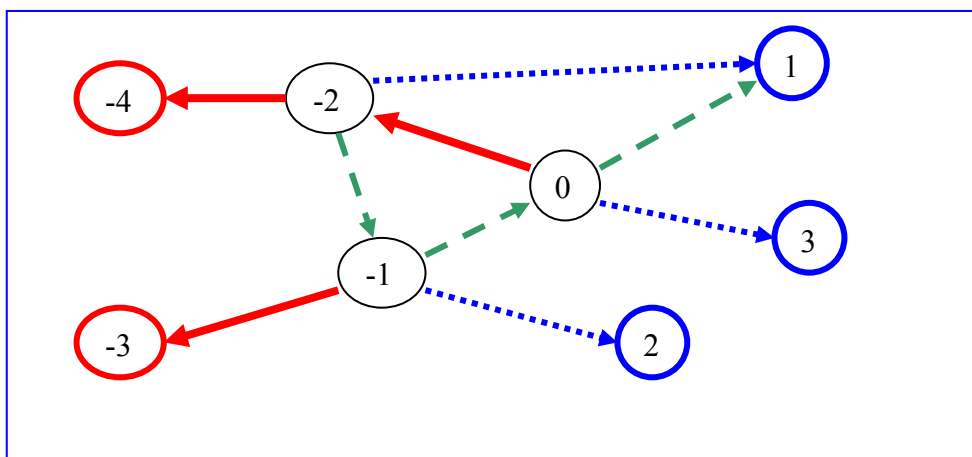
Par contre, on vérifie que $E(S_2) = -\frac{3}{16}$.

- Le jeu ne continue que dans le cas où $S_2 = -1$ et La loi de S_3 est donnée par :

s	-4	-3	0	1	2	3
$P(S_3 = s)$	1/4	5/64	25/512	13/32	15/512	3/16

$$P(S_3 > 0) = \frac{319}{512} \approx 0,623, \quad P(T=3) = \frac{55}{512}$$

Plus généralement la situation peut être décrite par le diagramme de transition suivant :



Les flèches pleines sont empruntées avec la probabilité $1/2$, celles en pointillés avec probabilité $3/16$ et celles tiretées avec probabilité $5/16$.

On peut alors constater qu'après 3 parties jouées, soit le jeu est terminé, soit les deux joueurs sont à égalité. Ainsi, on en déduit que :

$$P(S_{3n} = 0) = (P(S_3 = 0))^n = \left(\frac{25}{512}\right)^n \text{ et, pour } s \in \{-4; -3; 1; 2; 3\}$$

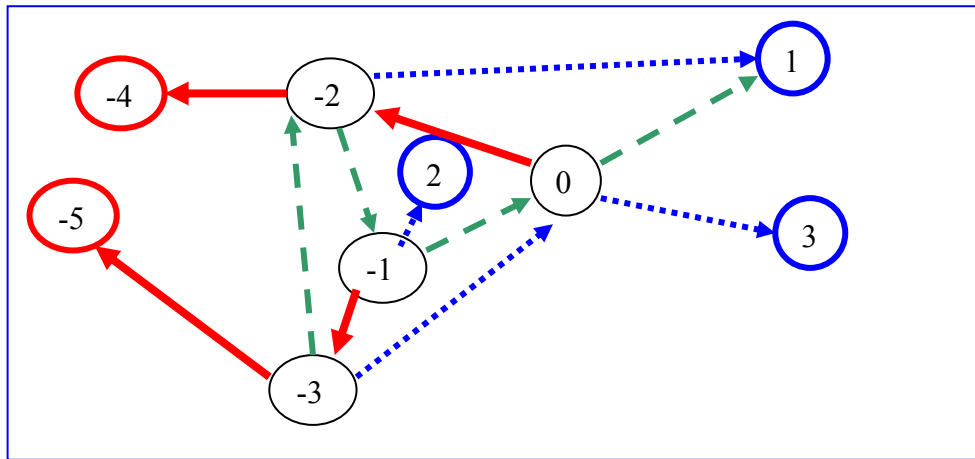
$$P(S_{3n+3} = s) = P(S_{3n} = s) + P((S_{3n} = 0) \cap (S_{3n+3} = s)) = P(S_{3n} = s) + P(S_3 = s) \times P(S_{3n} = 0)$$

On trouve alors $P(S_{3n} = s)$ comme somme des termes d'une suite géométrique de raison $(25/512)$ et lorsque n tend vers l'infini $P(S_n = s)$ a une limite donnée dans le tableau suivant :

s	-4	-3	0	1	2	3
$P(S_3 = s)$	128/487	40/487	0	208/487	15/487	96/487

On en déduit que la probabilité de gagner est égale à $319/487 \approx 0,655$ mais que l'espérance de gain reste négative : $-109/487 \approx -0,22$.

Si Bernard décide de s'arrêter de jouer lorsqu'il perd 4€ (à la place de 3) la situation est plus complexe :



On peut dans ce cas considérer les suites :

$a_n = P(S_n = -5)$, $b_n = P(S_n = -4)$, ..., $i_n = P(S_n = 3)$, qui vérifient les relations de récurrence:

$$\begin{cases} a_{n+1} = a_n + \frac{1}{2}c_n & b_{n+1} = b_n + \frac{1}{2}d_n \\ c_{n+1} = \frac{1}{2}e_n & d_{n+1} = \frac{5}{16}c_n + \frac{1}{2}f_n \\ e_{n+1} = \frac{5}{16}d_n & f_{n+1} = \frac{3}{16}c_n + \frac{5}{16}e_n \\ g_{n+1} = \frac{3}{16}d_n + \frac{5}{16}f_n + g_n \\ h_{n+1} = \frac{3}{16}e_n + h_n & i_{n+1} = \frac{3}{16}f_n + i_n \end{cases}$$

Un logiciel de calcul peut nous permettre d'évaluer les probabilités des différents états pour n assez grand. Pour $n = 100$, on trouve que $P(T > 100) < 10^{-30}$ (ce qui semble garantir que le

jeu à peu de chances de se prolonger indéfiniment) et le tableau de valeurs approchées suivantes :

k	-5	-4	2	3
$P(S_{100}=k)$	0,0440	0,2816	0,0330	0,2009

Ainsi Bernard gagnera un peu plus de deux fois sur trois puisque $P(S_{100}>0) \approx 0,6744$ cependant son espérance de gain restera évidemment négative: $E(S_{100}) \approx -0,247$.

Pour étudier d'autres situations de la même veine, il est préférable d'avoir recours à une simulation (Fichier Excel : Gagner à tout prix):

C est le capital de Bernard et il s'arrête dès qu'il gagne (1, 2 ou 3 €) ou dès qu'il a tout perdu (C ou C+1 €); à chaque fois on a simulé 1000 jeux:

C	4	5	6	7	8	9	10	15	20	25	50	100
Nombre de jeux gagnés	675	700	756	770	779	805	828	854	877	879	894	916

On peut également se placer dans la situation où Bernard dispose d'un capital fixe (100 € par exemple). Son problème est alors de savoir quelle stratégie il doit adopter pour avoir une chance raisonnable de gagner. Il décidera de s'arrêter dès qu'il gagne une somme G ou dès qu'il a perdu ses 100€. Voici les résultats des simulations effectuées (toujours sur 1000 jeux).

G	20	10	9	8	7	6
Nombre de jeux gagnés	265	488	548	583	609	685

Dans ce cas, il semble que pour avoir 2 chances sur 3 de gagner (comme dans la situation précédente), il doive s'arrêter de jouer dès qu'il gagne au moins 6€. D'autres simulations avec $G=6$ donnent une moyenne de 669 parties gagnées sur 1000. Malheureusement pour lui, on peut dans ce cas estimer son espérance de gain à environ -29€ ce qui n'est évidemment guère brillant.

Pièges à gogos

Revenons à notre problème initial et cherchons à quelles conditions sur les sommes mises en jeu la situation est favorable à Alain.

Le jeu sera alors décrit par le tableau suivant où a, b, c, d sont des réels positifs

	Pair (p)	Impair ($1-p$)
Pair (q)	a	$-b$
Impair ($1-q$)	$-c$	d

En notant $S=a+b+c+d$ et en mettant l'espérance de gain de Bernard $E(p, q)$ sous forme canonique, on obtient :

$$E(p; q) = S \left(p - \frac{b+d}{S} \right) \left(q - \frac{c+d}{S} \right) + \frac{ad-bc}{S}$$

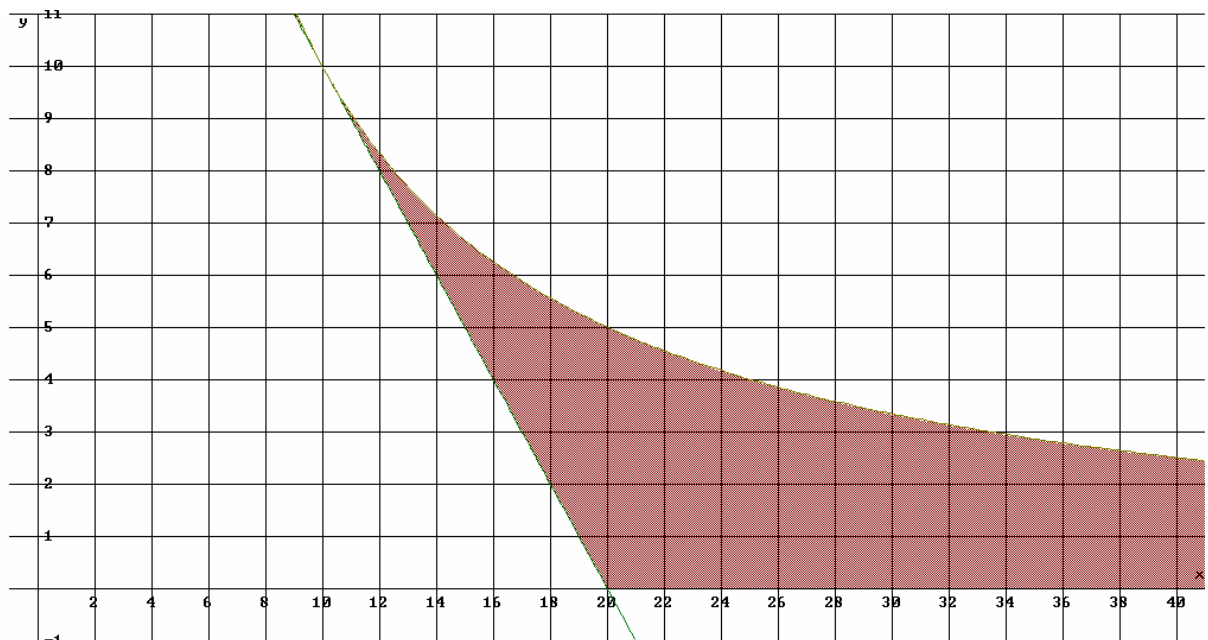
Le jeu sera intéressant pour Alain dès que $ad-bc < 0$. En ne considérant que le cas $c=b$ (on ne distingue pas pair - impair de impair - pair ce qui est bien pratique), on trouve :

$$E(p; q) = S \left(p - \frac{b+d}{S} \right) \left(q - \frac{b+d}{S} \right) + \frac{ad - b^2}{S}$$

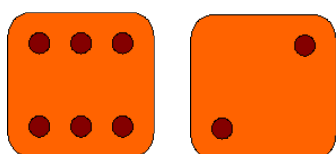
Ainsi en particulier un jeu en apparence équilibré (tel que $a+d=2b$) sera toujours favorable à Alain (le maximum de xy sous la contrainte $x+y=k$ est obtenu pour $x=y$).

En prenant $b=10$, terminons notre étude par la recherche des "pièges à gogos" c'est-à-dire des jeux en apparence gagnants pour Bernard ($a+d>20$) et qui lui sont en réalité défavorables ($ad<100$).

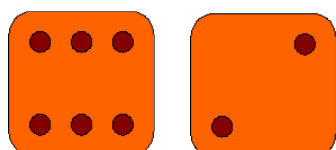
La représentation graphique ci-dessous nous montre les couples qui répondent à la question : Par exemple (16;6) qui procure une espérance de gain de $2/21$ à Alain à condition qu'il joue Pile avec probabilité $p=8/21$.



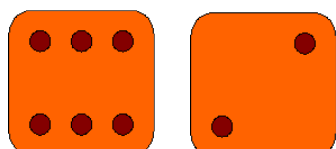
VIII. SIMULATIONS



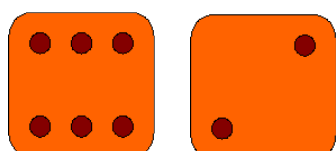
Les pépites de chocolat



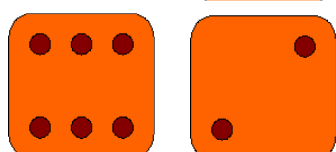
Les collections complètes



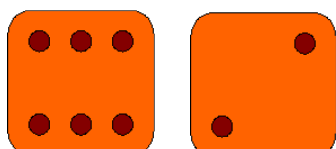
La chasse aux canards



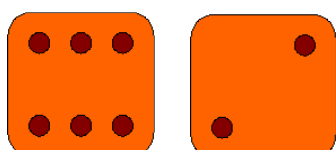
Le loto



Les records



Deux dés et somme des points



Évolutions de la fréquence du « 6 » pendant que l'on lance un dé

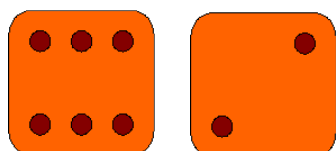
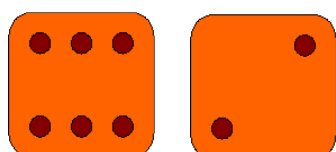
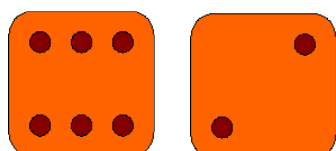


Planche de Galton



Lancer un dé N fois



Une activité pour les élèves

Nous donnons dans ce chapitre une brève présentation des simulations que nous avons implantées sous EXCEL. La présence de ce tableur dans la quasi-totalité des lycées nous a incité à choisir cette option. Les classeurs décrits sont sur le CD d'accompagnement de la brochure.

Pour utiliser ces simulations avec EXCEL il est indispensable de configurer le logiciel en adoptant les options suivantes :

1. Le **calcul « sur ordre »** uniquement. (Outils – Options – Onglet Calcul – cocher « Sur ordre »). Le re-calcul automatique, qui est généralement l'option par défaut, provoque en effet un nouveau tirage au hasard à chaque modification. Cette option étant sélectionnée, il faut savoir que l'ordre de calcul est la touche F9.
2. **L'autorisation des itérations avec un nombre maximal d'itérations fixé à 1.** (Outils – Options – Onglet Calcul – cocher « itération » et fixer « Nb maximal d'itérations à 1 »). Cette précaution permet de cumuler les résultats de simulations successives car elle autorise une cellule à calculer sa nouvelle valeur en fonction de son ancienne sans provoquer d'erreur du type « référence circulaire ».
3. **L'autorisation des macros** lors de l'ouverture des fichiers proposés est indispensable pour assurer leur bon fonctionnement.

Les classeurs proposés sont structurés de manière analogue :

- Une première feuille intitulée « Aide » décrit l'expérience simulée et donne quelques conseils d'utilisation du classeur.
- La simulation proprement dite est sur une, voire plusieurs, autres feuilles.
- Des compléments éventuels, concernant notamment le modèle probabiliste correspondant à l'expérience, sont regroupés sur une dernière feuille.

LES PÉPITES DE CHOCOLAT

Chocolat.XLS

L'expérience aléatoire

Un pâtissier verse 100 pépites de chocolat dans la pâte destinée à la fabrication de 100 croissants¹⁸. Ces pépites se répartissent aléatoirement.

Combien y aura-t-il, en moyenne de croissants sans pépite ?

Le principe de la simulation

La simulation consiste à choisir 100 nombres au hasard de 00 à 99 qui représentent les numéros des croissants qui reçoivent chacune des pépites puis à compter combien de pépites ont abouties dans chaque petit pain.

Le modèle théorique

Le nombre X de pépites aboutissant dans un petit pain quelconque fixé est une variable aléatoire de loi binomiale $B(0.01;100)$. Ce qui permet de calculer les probabilités :

$$P(X = k) = \binom{100}{k} 0.01^k 0.99^{100-k}$$

Nombre de pépites	Probabilités
0	0,3660323
1	0,3697296
2	0,1848648
3	0,0609992
4	0,0149417
5	0,0028978
6	0,0004635
7	0,0000629
8	0,0000074
9	0,0000008
10	0,0000001
>10	0,0000000

¹⁸ Une recette équivalente d'un point de vue probabiliste est proposée par Arthur Engel dans « Les certitudes du hasard » (grains de raisin dans des petits pains).

LES COLLECTIONS COMPLÈTES

CollectionComplete.XLS

Série complète.XLS

L'expérience aléatoire

Ce classeur simule la quête d'une collection complète : on tire (avec remise) une boule dans une urne contenant des boules numérotées jusqu'à ce que l'on ait obtenu, au moins une fois, chacun des numéros possibles. Combien faut-il en moyenne de lancers pour parvenir à ce résultat ?

Le principe de la simulation

On simule le lancer d'un dé jusqu'à ce que l'on ait obtenu tous les résultats possibles c'est à dire la collection complète 1, 2, 3, 4, 5 et 6. Pour détecter le moment où la collection est complète, on compte après chaque lancer, le nombre de résultats différents obtenus jusque là.

La simulation proposée (touche F9 !) renouvelle 100 fois l'expérience de la recherche d'une collection complète et donne la distribution des nombres de lancers nécessaires (à plus de 60 lancers on abandonne).

Le modèle

Soit X_n la variable aléatoire égale au nombre de faces distinctes obtenu après n lancers.

$P(X_1 = 1) = 1$ et $P(X_1 = i) = 0$ pour $i = 2, 3, 4, 5, 6$

$P(X_2 = 1) = \frac{1}{6}$ et $P(X_2 = 2) = \frac{5}{6}$ et $P(X_2 = i) = 0$ pour $i = 3, 4, 5, 6 \dots$

...

Posons $a_n = P(X_n = 1)$, $b_n = P(X_n = 2)$, ..., $f_n = P(X_n = 6)$.

On obtient alors

$a_{n+1} = 1/6 a_n$	$b_{n+1} = 1/3 b_n + 5/6 a_n$	$c_{n+1} = 1/2 c_n + 2/3 b_n$
$d_{n+1} = 2/3 d_n + 1/2 c_n$	$e_{n+1} = 5/6 e_n + 1/3 d_n$	$f_{n+1} = f_n + 1/6 e_n$

Ainsi a_n est géométrique de raison $1/6$, b_n est somme de suites géométriques de raisons $1/6$ et $1/3$, c_n est somme de suites géométriques de raisons $1/6, 1/3$ et $1/2$, ... et e_n somme de suites géométriques de raisons $1/6, 1/3, 1/2, 2/3$ et $5/6$.

Le calcul nous donne en particulier pour $n \geq 2$:

$$e_n = 5 \left(\left(\frac{5}{6} \right)^{n-1} - 4 \left(\frac{2}{3} \right)^{n-1} + 6 \left(\frac{1}{2} \right)^{n-1} - 4 \left(\frac{1}{3} \right)^{n-1} + \left(\frac{1}{6} \right)^{n-1} \right)$$

Soit T la variable aléatoire égale au nombre de lancer nécessaire pour avoir une collection complète. La série complète est obtenue après n lancers si $(T=n)$ soit $(X_n=6)$ et $(X_{n-1}=5)$.

En conditionnant, on obtient que la probabilité de cet événement est $1/6 e_{n-1}$.

Il ne reste plus qu'à calculer $E(T)$ dont la valeur exacte est **14,7**.

On utilise pour cela $\sum_{n=1}^{\infty} n x^{n-1} = \frac{1}{(1-x)^2}$.

LA CHASSE AUX CANARDS

Chasse aux canards.XLS

L'expérience aléatoire

Ce classeur simule une chasse impitoyable¹⁹. Des tireurs infailibles choisissent leur cible au hasard et on s'intéresse au nombre moyen de survivants.

La feuille "canards" permet de simuler l'expérience qui consiste pour dix chasseurs à tirer sur un vol de dix canards en choisissant au hasard leur victime. Elle compte le nombre de survivants et cumule les résultats des simulations successives.

Le principe de la simulation

On détermine au hasard, pour chaque chasseur, le numéro du canard qu'il vise. Il ne reste plus qu'à compter le nombre de canards qu'aucun chasseur n'a choisi pour cible.

Le modèle

A l'habillage près, il s'agit du même problème que celui des pépites de chocolat. Le nombre X de chasseurs visant un canard quelconque fixé est une variable aléatoire de loi binomiale $B(0.1;10)$. Ce qui permet de calculer les probabilités :

$$P(X = k) = \binom{10}{k} 0.1^k 0.9^{10-k}$$

En particulier la probabilité pour un canard donné de survivre est

$$P(X = 0) = 0.9^{10} \approx 0.3487$$

L'espérance mathématique du nombre de survivants parmi les dix canards est donc : 10×0.3487 soit environ 3.5.

¹⁹ L'idée de ce carnage est encore empruntée à Arthur Engel qui, dans « Les certitudes du hasard », en analyse le triste bilan.

LE LOTO

LOTOSimul.XLS

L'expérience aléatoire

Un jeu de la Française des Jeux bien connu de tous : on tire successivement, sans remise, 6 boules dans une urne qui contient au départ 49 boules numérotées de 1 à 49.

Le principe de la simulation

Le procédé utilisé peut être adapté pour tout tirage avec remise. Il consiste à associer à chacune des 49 boules de l'urne de départ un nombre aléatoire. On décide que les six boules associées aux 6 plus petits nombres aléatoires sont les boules tirées. Il suffit donc de trier ce que le tableur fait très bien.

La feuille « Statistique officielle » donne les résultats des tirages du LOTO depuis le 19 mai 1976 jusqu'au 12 février 2005²⁰.

Le modèle

Un tirage est une combinaison de 6 nombres pris parmi 49. Le nombre de tirages possibles est donc $\binom{49}{6} = 13983816$. Le nombre de tirages contenant un numéro x donné est :

$\binom{48}{5} = 1712304$. La probabilité de sortie de chaque numéro est donc $\frac{1712304}{13983816} = \frac{6}{49}$.

Numéros consécutifs

Un événement suffisamment fréquent pour intriguer les profanes concerne la présence, parmi les numéros tirés, d'une paire, c'est-à-dire de deux entiers consécutifs. La feuille « LOTO PAIRES » permet de simuler cet événement.

Le principe de la simulation

On trie le tirage obtenu et l'on calcule les écarts entre deux valeurs voisines dans la combinaison triée. La présence d'au moins un écart de 1 signifie que le tirage contient (au moins) deux entiers consécutifs.

Le modèle

Codons un tirage de la manière suivante : dans la suite croissante des 49 numéros possibles. Remplaçons les numéros tirés par des barres. Par exemple pour la combinaison

5 14 17 29 35 41

nous obtenons :

1 2 3 4 / 6 7 8 9 10 11 12 13 / 15 16 / 18 19 20 21 22 23 24 25 26 27 28 /
30 31 32 33 34 / 36 37 38 39 40 / 42 43 44 45 46 47 48 49

²⁰ Une version actualisée de cette statistique est proposée sur le site http://www.fdjeux.com/jeux/lotto/lotto_s_palmares.php

Les 43 boules non tirées apparaissent réparties en 7 groupes dont certains pourraient être vides. Remarquons que les numéros des boules ne sont pas indispensables pour retrouver la combinaison codée, les nombres de boules contenues dans chacun des groupes suffisent. Le tirage précédent peut ainsi se coder :

$$4 / 8 / 2 / 11 / 5 / 5 / 8$$

et, réciproquement la répartition :

$$10 / 0 / 12 / 1 / 8 / 0 / 12$$

correspond à la combinaison

$$11 \ 12 \ 25 \ 27 \ 36 \ 37.$$

Tirer 6 boules parmi 49 équivaut à répartir 43 objets dans 7 groupes. Le nombre de combinaisons possibles est donc le nombre de façon de répartir 43 objets (identiques) en 7 groupes (dont certains éventuellement vides).

Plus généralement, répartir n objets (identiques) en k groupes (certains pouvant être vides) équivaut à choisir $k-1$ emplacements parmi $n+k-1$, les emplacements des « séparateurs » symbolisés ici par des « / ».

Ceci peut se faire de $\binom{n+k-1}{k-1}$ manières.

A quoi correspond sur le codage la présence de deux numéros consécutifs ? Au fait qu'il y a un groupe vide (un 0 dans le codage) qui, de plus, n'est ni le premier ni le dernier.

Passons à l'événement contraire : « la combinaison tirée ne contient pas deux numéros consécutifs ». Une telle combinaison correspondra à une répartition dans laquelle aucun des groupes intérieurs n'est vide. Calculons le nombre de répartitions de 43 objets en 7 groupes aucun des 5 groupes intérieurs n'étant vide. Pour obtenir une telle répartition on peut commencer par placer 1 objet dans chacune des cinq boîtes intérieures puis répartir les 38 objets restant sans autre contrainte dans les 7 groupes ce qui d'après le résultat général

ci-dessus peut se faire de $\binom{44}{6}$ manières.

La probabilité d'obtenir une combinaison sans numéros consécutifs est donc :

$$\frac{\binom{44}{6}}{\binom{49}{6}} = \frac{7\,059\,052}{13\,983\,816} \approx 0.50480155$$

La probabilité d'avoir au moins deux numéros consécutifs parmi les 6 tirés vaut donc 0.49519845.

Une probabilité si proche de 1/2, permet de poser le problème des sondages dont le résultat est très proche de 50/50.

LES RECORDS

LesRecords.XLS

L'expérience aléatoire

Dans une suite de nombres on appelle record tout nombre supérieur à tous les précédents.

Le principe de la simulation

On tire les valeurs au hasard et on détermine à chaque fois le record actuel c'est-à-dire le plus grand nombre obtenu jusque là.

Une première feuille intitulée « Records 10 » permet de simuler 100 tirages de 10 chiffres au hasard et détermine dans chaque suite de 10 chiffres le nombre de « records » au sens défini ci-dessus.

La feuille « Records 20 » permet de simuler 100 tirages de 20 nombres au hasard dans l'intervalle $[0,1[$ et compte dans chaque suite de 20 nombres le nombre de « records » au sens défini ci-dessus.

Le modèle

Pour n un entier naturel tel que $0 \leq n \leq 9$ et i un entier naturel tels que $2 \leq i \leq 9$. On considère les événements suivants :

$A(1,n)$: "le premier des chiffres est n ".

$B(i,n)$: "le plus grand des i premiers chiffres est inférieur ou égal à n ".

$A(i,n)$: "le plus grand des i premiers chiffres est n ".

Soit la variable aléatoire T qui à tout tirage associe le nombre de records et pour $1 \leq k \leq 10$, T_k les variables aléatoires de loi de Bernoulli telles que $T_k=1$ si le k -ème chiffre est un record.

Ainsi $T = \sum_{k=1}^{10} T_k$ et donc $E(T) = \sum_{k=1}^{10} p_k$ en posant $p_k = P(T_k)$.

Remarque : les variables T_k ne sont pas indépendantes.

On a $p_1=1$ et pour $1 \leq k \leq 9$

$$p_{k+1} = \frac{9}{10} P(A(k,0)) + \frac{8}{10} P(A(k,1)) + \dots + \frac{1}{10} P(A(k,8)).$$

D'autre part pour $n \geq 1$, $A(k,n) = B(k,n) - B(k,n-1)$ (différence d'ensembles)

et donc $P(A(k,n)) = P(B(k,n)) - P(B(k,n-1)) = \left(\frac{n}{10}\right)^k - \left(\frac{n-1}{10}\right)^k = \frac{1}{10^k} (n^k - (n-1)^k)$.

Ainsi $p_{k+1} = \frac{1}{10^{k+1}} (9(1^k) + 8(2^k - 1^k) + 7(3^k - 2^k) + \dots + (9^k - 8^k)) = \frac{1}{10^{k+1}} \sum_{i=1}^9 i^k$.

On obtient ainsi une valeur approchée de $E(T)$: 2,51546

DEUX DÉS ET SOMME DES POINTS

deuxdemodele.XLS

deuxdesNfois.XLS

L'expérience aléatoire

On lance deux dés cubiques normaux et on calcule la somme des points.

Le principe de la simulation

Le premier dé est dans la colonne A, le second dans la colonne B, la somme dans la colonne C.

Les simulations se font par rafales de 25 lancers. La feuille « Simulation » affiche, en parallèle, les probabilités et les fréquences correspondant aux 25 derniers lancers et pour le cumul des lancers.

Le modèle

Il est bien connu des professeurs $1/36$, $2/36$, $3/36$ *etc.*

La simulation proposée ici permet de convaincre les élèves qui seraient tentés d'en proposer un autre.

ÉVOLUTION DE LA FRÉQUENCE DU « 6 » PENDANT QUE L'ON LANCE UN DÉ

Fréquence du 6.XLS

L'expérience aléatoire

On lance 500 fois un dé cubique normal et on calcule, après chaque lancer, la fréquence du résultat « 6 ».

Le principe de la simulation

On simule 500 lancers d'un dé dans une colonne et on totalise dans une colonne voisine le nombre de « 6 » obtenus jusque là. Il ne reste plus qu'à diviser par le nombre de lancers pour calculer la fréquence et la représenter graphiquement.

Le modèle

On espère constater que la fréquence « tend » plus ou moins vers la probabilité, mais cette « convergence » est assez médiocre et pas très convaincante. C'est néanmoins l'une des formes de la loi des grands nombres proposées par les manuels.

PLANCHE DE GALTON²¹

Galton.XLS

L'expérience aléatoire

Des billes s'écoulent le long d'un plan incliné et rencontrent, à plusieurs niveaux, des obstacles (des clous) qu'elles contournent en choisissant au hasard la gauche ou la droite.

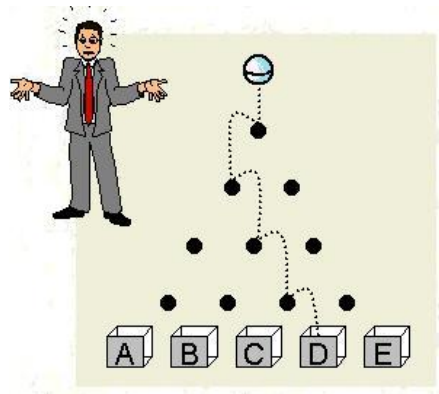
Le principe de la simulation

On simule la chute d'un nombre paramétrable de billes dans une planche de Galton à 5 niveaux.

La feuille EXCEL ne sert en fait ici que d'interface de visualisation, la simulation est programmée en Visual Basic dans un module

Le modèle

Le numéro de la case (de 0 à 5 pour une planche à 5 niveaux) dans laquelle finit une boule suit la loi binomiale $B(5, 1/2)$.



²¹ On trouvera d'autres programmes de simulation de la machine de Galton dans de nombreux sites comme <http://perso.wanadoo.fr/jpg/proba/galton/> ou http://perso.wanadoo.fr/therese.eveilleau/pages/truc_mat/textes/galton.htm

LANCER UN DÉ N FOIS

UnDéCumul.XLS

L'expérience aléatoire

On lance un dé cubique normal.

Le principe de la simulation

On simule un dé sous EXCEL par la formule : =ENT(6*ALEA())+1

Le modèle

Les six résultats sont équiprobables de probabilité 1/6.

UNE ACTIVITÉ POUR LES ÉLÈVES

Certains organismes (SOFRES, INSEE...) réalisent des sondages pour essayer de prévoir les résultats d'un vote (nous prendrons comme exemple un référendum). Pour cela un certain nombre d'électeurs (échantillon) sont interrogés et la proportion de votants pour le "OUI", éventuellement corrigée²², sera le résultat du sondage. En raison de la fluctuation d'échantillonnage la proportion obtenue ne sera qu'une approximation des intentions réelles de vote des français (que l'on ne connaît pas). Dès lors, il est important de savoir la précision de cette approximation.

Pour cela, il est possible d'effectuer une démarche inverse: On part d'une population connue dont on extrait des échantillons afin d'étudier les variations des résultats obtenus. La réalisation de tels sondages étant matériellement impossible à réaliser on effectuera une simulation à l'aide du logiciel EXCEL.

1. Réalisation de la feuille de calcul.

a) Comme il s'agit d'une simulation, il faut avant tout choisir un nombre P représentant la proportion de votants pour le OUI dans la population. On impose pour cela $0,3 < P < 0,7$. Cette valeur sera placée dans la cellule **E1**. (Pour toute référence ultérieure à cette cellule : **\$E\$1**)

b) Dans la cellule A4, on génère un nombre aléatoire à l'aide de la fonction **ALEA()**

c) Dans la cellule B4, on mettra le résultat simulé du premier votant : Si le nombre aléatoire obtenu en A est inférieur à P , il y a vote pour OUI sinon il y a vote pour NON. Pour cela, on utilisera la fonction :

SI("condition"; " valeur si condition réalisée"; " valeur sinon")

d) On recopie 1000 fois ces deux cellules

e) Plus loin, dans la colonne D par exemple, on calculera la proportion de votants pour OUI dans les 20, puis les 400, puis les 1000 premiers électeurs simulés à l'aide de la fonction **NB.SI("plage"; " valeur")**.

On appellera f_{20} , f_{400} et f_{1000} les résultats obtenus pour 20, 400 et 1000 "votes".

2. Utilisation des résultats (la touche **F9** permet de tout recalculer. Une utilisation de cette touche génère ainsi une nouvelle simulation)

a) P fixé. (par exemple $P=0,55$)

Que constate-t-on sur les valeurs f_{20} , f_{400} et f_{1000} lorsqu'on réalise plusieurs simulations ?

b) Fait-on les mêmes constatations avec d'autres valeurs de P .

c) On choisit cette fois $P=0,49$. La population considérée vote donc en majorité pour le NON. Appuyer 20 fois de suite sur la touche "F9" et noter parmi les 20 simulations effectuées combien donnent effectivement le NON gagnant. Commenter.

d) La théorie des probabilités dit que dans 95% des cas $|P - f_n| \leq \frac{1}{\sqrt{n}}$. On considère

généralement que cette marge d'erreur est satisfaisante.

On sonde nous donne $f_{400}=0,55$. A quel intervalle appartient (normalement) P ?

On suppose ici $P=0,49$. A quel intervalle peut-on s'attendre à ce que f_{1000} appartienne ?

Quels sont les résultats possibles de f_{1000} qui garantissent (à 95%) une victoire de OUI ?

²² Par exemple les instituts de sondage ont constaté que les électeurs de droite sont moins enclins à répondre aux sondages que ceux de gauche et corrigent leurs estimations en fonction de cette tendance.

BIBLIOGRAPHIE

Nous avons choisi de ne citer que des ouvrages que nous avons consultés lors de la préparation de cette brochure et pour nos enseignements.

1) Publimath

Le lecteur intéressé trouvera de nombreuses références dans la base de données bibliographiques sur l'enseignement des mathématiques

<http://publimath.irem.univ-mrs.fr/>

2) Commission Inter-IREM Statistique et Probabilités

La Commission Inter-IREM Statistique et Probabilités a publié plusieurs ouvrages comprenant une très abondante bibliographie.

Commission Inter-IREM Statistique et Probabilités. (CHAPUT Brigitte ; COURTEBRAS Bernard ; DANTAL Bernard ; GIRARD Jean-Claude ; HENRY Michel ; PICHARD Jean-François ; THIÉNARD Jean-Claude), Autour de la modélisation en probabilités, *Presses universitaires de Franche-Comté, Besançon*, 2001.

Commission Inter-IREM Statistique et Probabilités. (CHAPUT Brigitte ; DANTAL Bernard ; GIRARD Jean-Claude ; GRANGÉ Jean-Pierre ; HENRY Michel ; PARZYSZ Bernard ; PICHARD Jean-François ; RAYMONDAUD Hubert), *Probabilités au lycée*, Brochure APMEP **143**, Association des Professeurs de Mathématiques de l'Enseignement Public, Paris, 2003.

3) IREM

Plusieurs IREM ont publié des brochures de Probabilités et Statistique, dont :

Groupe Probabilités et Statistique de l'IREM de Franche-Comté, Lois continues, test d'adéquation, Une approche pour non spécialistes, *Presses universitaires de Franche-Comté, Besançon*, 2005.

4) Articles sur les statistiques et leur enseignement classés par revue

Bulletin de l'Association des Professeurs de Mathématiques de l'Enseignement Public

Louis-Marie BONNEVAL, *Intervalle : de confiance ?*, Bulletin de l'APMEP **427**, 161-170, 2000.

Louis-Marie BONNEVAL, *Test d'équipartition, quel risque d'erreur ?*, Bulletin de l'APMEP **456**, 31-45, 2005.

Michel HENRY, *Lois continues en Terminale S, quelle approche ?*, Bulletin de l'APMEP **454**, 688-694, 2004.

Repères IREM

Michel HENRY, *Des lois de probabilités continues en terminales, pourquoi et pourquoi faire*, Repères IREM **51**, 5-25, TOPIQUES éditions, Metz, 2003.

Québec Science

Sur le site www.quebecscience.qc.ca, on trouve de nombreux articles en ligne dont Isabelle MONTPETIT, *Vos vraies chances de gagner au loto*, 2000.

Pour la Science

Jean Paul DELAHAYE, *Aléas du hasard informatique*, Pour la Science, **245**, mars 1998.

Ian STEWART, *Abrogeons la loi des moyennes, Subtilités mathématiques du jeu de pile ou face*, Pour la Science, **248**, juin 1998.

Tangente

Le numéro 77 de la revue Tangente est entièrement consacré aux statistiques, Archimède Argenteuil, 2000.

5) Ouvrages « autour » des statistiques

Claudine ROBERT est l'auteur de plusieurs ouvrages et articles, dont

Claudine ROBERT, *Contes et décomptes de la statistique. Une initiation par l'exemple, Vuibert, Paris, 2003.*

Il s'agit d'une nouvelle édition, légèrement remaniée, de *L'Empereur et la Girafe* paru en 1995 aux Éditions Diderot.

Arthur ENGEL, *Les certitudes du hasard, Aleas, Lyon, 1990.*

Jean-Jacques DROESBEKE et Philippe TASSI, *Histoire de la statistique, PUF, Que sais-je ?, 2527, 1990.*

6) Manuels et sites de référence

Jean-Jacques DROESBEKE, *Éléments de statistique, Ellipses, 2001.*

Jacques HARTONG a publié plusieurs ouvrages dont un cours téléchargeable de Calcul des probabilités, Association Libre Cours (www.librecours.org), Date d'envoi 18 février 2004 (notre collègue Jacques HARTONG est décédé depuis).

Richard ISAAC, *Une initiation aux Probabilités, Vuibert Springer, Paris, 2005.*

Michel JANVIER, *Statistique descriptive avec ou sans tableur, cours et exercices corrigés, Dunod, Paris, 1999.*

Michel LEJEUNE, *Statistique, La théorie et ses applications, Springer-Verlag, Paris, 2005.*

Gilbert SAPORTA, *Probabilités, analyse des données et statistiques, Technip, 1992.*

Thomas H. WONNACOTT et Ronald J. WONNACOTT, *Statistique : économie, gestion, sciences, médecine, Economica, Paris, 1991.*

Le projet SEL (pour Statistique En Ligne) propose une initiation interactive à la statistique et se trouve à l'adresse <http://www.inrialpes.fr/sel/>.

Il existe aussi une version SMEL, Statistique Médicale En Ligne, à l'adresse <http://www.math-info.univ-paris5.fr/smel/>

7) Publications en Statistiques et Probabilités, des groupes et des animateurs de l'IREM de Strasbourg depuis 1998

Tous ces renseignements (ou presque) se trouvent sur le site <http://irem.u-strasbg.fr> dans les rubriques : « Publications », « L'Ouvert en ligne », « Site IREM2 », « Actualités » (éventuellement) et des « Liens sélectionnés ».

a) Dans des revues publiées par l'IREM de Strasbourg

Les Annales de Didactiques et de Sciences Cognitives

Claire DUPUIS et Suzette ROUSSET-BERT, *De l'influence des représentations disponibles sur la résolution de problèmes élémentaires de probabilité et sur l'acquisition du concept d'indépendance*, 67-87, **6**, 1998.

L'Ouvert, Journal de l'APMEP d'Alsace et de l'IREM de Strasbourg

La consultation en ligne d'articles de l'Ouvert des numéros 95 à 110 est possible sur le site <http://irem.u-strasbg.fr>

Nicole VOGEL, *Nos élèves refont l'histoire des probabilités*, **95**, juin 1999.

Gérard KUNTZ, *Probabilités, réalités et ... pesanteur(s)*, **104**, Novembre 2001.

Michel EMERY, *Quelques phénomènes curieux en probabilités et statistiques*, **104**, Novembre 2001.

Jacques FRANCHI, *Comment s'enroulent les processus aléatoires*, **105**, Avril 2002.

b) Publications en ligne sur le site <http://irem2.u-strasbg.fr>, Rubrique PROBABILITÉS

Nicole VOGEL, *Peut-on imiter le hasard ?*

Bernard KOCH, *Les truites de Pondichéry*.

Christophe KILIAN, *Problème des trois portes*.

Jean-Pierre DAROU, *Adéquation à une loi équirépartie*.

c) Publications dans la revue "Activités mathématiques" de la Mission Laïque Française 9, rue Humblot, 75015 PARIS

Claire DUPUIS, *Situation pour exploiter les probabilités conditionnelles*, **34**, 1998.

Etienne MEYER, *Médiane (en statistique)*, **44**, 2001.

Gérard KUNTZ, *Vache folle : probabilités et réalités*, **46**, 2002.

d) Publications dans la revue *Repères IREM*

GROUPE « Lycée » de l'IREM de Strasbourg, *Clés de contrôle*, **41**, octobre 2000.

Gérard KUNTZ, *Vache folle, probabilités, réalités et ... pesanteur(s)*, **45**, octobre 2001.

Bernard KOCH et le groupe « Statistiques pour le lycée » de l'IREM de Strasbourg, *Les truites de Pondichéry : Un problème d'adéquation au baccalauréat*, **57**, octobre 2004.

- e) Sites mathématiques sélectionnés auxquels participent des animateurs de l'IREM de Strasbourg et comprenant des statistiques et probabilités

Mathématikos, site personnel de Jean-Paul QUELEN

<http://perso.wanadoo.fr/jpq/proba/>

Aromath <http://www.aromath.net>

Ce site est maintenu par un groupe d'enseignants de l'Académie de Strasbourg : Bernard KOCH, Bernard LANGER (webmestre), Gilbert LE CAM, Michèle MATTEI, Jacques OURLIAC, Nicole VOGEL.

- f) Compte rendu de lecture de publications IREM dans des revues

Compte rendu de lecture par F. PLUVINAGE de "*Enseigner les statistiques du CM à la seconde. Pourquoi ? Comment ?*", publication de l'IREM de Lyon.

Ce compte rendu est paru dans le Journal de la Société Française de Statistique, tome 140, n° 1, 1999.

Ce qui donne pour l'ensemble de l'image de la couverture la description suivante

```
// Couverture
#include "colors.inc"
camera {
    location <24, 25, 24>
    direction <1, 1, 1>
    look_at <0, 0, 0>
}
// Macro pour un dé
#macro de(couleur1,couleur2)
    difference
    {
        intersection
        {
            box { <-2.0, -2.0, -2.0>, <2.0, 2, 2.0>
                pigment { color couleur1 }
            }
            sphere { <0, 0, 0>, 3.1
                pigment { color couleur1 }
            }
        }
    }
//
// face 1
sphere { <0, 0.0, -2.0>, 0.25
    pigment { color couleur2 }
}
//
// face 3
sphere { <2, 0.0, 0>, 0.25
    pigment { color couleur2 }
}
sphere { <2, 1.0, 1.0>, 0.25
    pigment { color couleur2 }
}
sphere { <2, -1.0, -1>, 0.25
    pigment { color couleur2 }
}
//
// face 5
sphere { <0, -2.0, 0>, 0.25
    pigment { color couleur2 }
}
sphere { <1, -2.0, 1>, 0.25
    pigment { color couleur2 }
}
sphere { <1, -2.0, -1>, 0.25
    pigment { color couleur2 }
}
sphere { <-1, -2.0, 1>, 0.25
    pigment { color couleur2 }
}
sphere { <-1, -2.0, -1>, 0.25
    pigment { color couleur2 }
}
//
// face 4
sphere { <-2.0, 1, 1>, 0.25
    pigment { color couleur2 }
}
sphere { <-2.0, 1, -1>, 0.25
    pigment { color couleur2 }
}
sphere { <-2.0, -1, 1>, 0.25
    pigment { color couleur2 }
}
sphere { <-2.0, -1, -1>, 0.25
    pigment { color couleur2 }
}
}

//
// face 6
sphere { <1, 0.0, 2.0>, 0.25
    pigment { color couleur2 }
}
sphere { <1, 1, 2>, 0.25
    pigment { color couleur2 }
}
sphere { <1, -1, 2>, 0.25
    pigment { color couleur2 }
}
sphere { <-1, 0.0, 2>, 0.25
    pigment { color couleur2 }
}
sphere { <-1, 1, 2>, 0.25
    pigment { color couleur2 }
}
sphere { <-1, -1, 2>, 0.25
    pigment { color couleur2 }
}
//
// face 2
sphere { <1, 2.0, 1>, 0.25
    pigment { color couleur2 }
}
sphere { <-1, 2.0, -1>, 0.25
    pigment { color couleur2 }
}
}
// Fin de la macro
#end
// Couleur du fond
background {color LightBlue}
// les six dés
object { de(Orange,OrangeRed)
    rotate <45, -130, 35 > }
object { de(Red,Orange)
    rotate <155, -120, 130 >
    translate<6,-3,-1> }
object { de( Yellow,CoolCopper)
    rotate <90, 40, 70 >
    translate <-2,3,5> }
object { de( Goldenrod,CoolCopper)
    rotate <120,-80,70 >
    translate <5,4,-1> }
object { de( Gold,Black)
    rotate <160, -75, 25 >
    translate <4,10,4> }
object { de( Coral,Bronze)
    rotate <60, 75, 55 >
    translate <3,-2,6> }

// Le texte
text { ttf "arial.ttf"
    "Statistique",0.2,0.1
    pigment {color Red}
    scale <2.2,2.5,2.5>
    rotate <20,200,25>
    translate <-16,15,10>
}
text { ttf "arial.ttf"
    "Probabilites",0.2,0.1
    pigment {color White}
    rotate <30,-90,0>
    translate <5,3,0>
    scale <3,3,3>
}

// Eclairage
light_source { <-10, 8, -20> color White }
light_source { <10, 13, 20> color Orange }
light_source { <20, 30, -10> color Yellow }
```

Titre : DE LA STATISTIQUE AUX PROBABILITÉS AU LYCÉE

Auteurs : Bernard KOCH, Emmanuelle ACKER, Claire DUPUIS, Suzette ROUSSET-BERT, Mohamed ATLAGH, Jérôme AUDÉOUD, Jean-Paul QUELEN, Dominique WEIL, membres du groupe de recherche formation "STATISTIQUE" de l'IREM de Strasbourg.

Mots-clés : Statistique – Probabilité - Lycée – Hasard – Aléatoire – Arbre de probabilité – Loi de probabilité – Générateur aléatoire – Loi des grands nombres – Graphique statistique – Simulation – Tableur – Test statistique – Jeux de dés – Loi du Khi-deux – Statistique descriptive.

Date : Octobre 2006.

Nombre de pages : 161 pages A4.

Éditeur : IREM de Strasbourg (S. 193)
<http://irem.u-strasbg.fr>
Pour les brochures :
Courriel : bibirem@math.u-strasbg.fr - Tél. : 01 90 24 01 61.

ISBN : 2-911446-28-3 (dépôt légal : 4^{ème} trimestre 2006).

Public concerné : Professeurs de mathématiques - Élèves de lycée - Grand public.

Résumé : Cette brochure explore quelques thèmes en rapport avec le lien entre statistique et probabilités (loi des grands nombres, adéquation). De nombreuses questions abordées concernent la simulation, ses apports et ses limites (générateurs de nombres au hasard, simulations avec tableur). Elle sera utile aux professeurs de mathématiques désireux d'éveiller leurs élèves à la compréhension des phénomènes aléatoires ainsi qu'aux élèves de lycée curieux de ces questions.

Prix : 14 € (+5 € de frais d'envoi).